Model Robust Regression
Based on Generalized Estimating Equations

by

Seth K. Clark

Dissertation submitted to the faculty of Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Jeffrey B. Birch, Co-Chairman
Oliver Schabenberger, Co-Chairman
George R. Terrell
Keying Ye
Christine M. Anderson-Cook

March 29, 2002
Blacksburg, VA

Keywords: Model Misspecification, Semiparametric, Nonparametric Regression, Local Model

Copyright 2002, Seth K. Clark
Model Robust Regression

Based on Generalized Estimating Equations

by

Seth K. Clark
Jeffrey B. Birch and Oliver Schabenberger, Chairmen
Statistics

Abstract

One form of model robust regression (MRR) predicts mean response as a convex combination of a parametric and a nonparametric prediction. MRR is a semiparametric method by which an incompletely or an incorrectly specified parametric model can be improved through adding an appropriate amount of a nonparametric fit. The combined predictor can have less bias than the parametric model estimate alone and less variance than the nonparametric estimate alone. Additionally, as shown in previous work for uncorrelated data with linear mean function, MRR can converge faster than the nonparametric predictor alone. We extend the MRR technique to the problem of predicting mean response for clustered non-normal data. We combine a nonparametric method based on local estimation with a global, parametric generalized estimating equations (GEE) estimate through a mixing parameter on both the mean scale and the linear predictor scale. As a special case, when data are uncorrelated, this amounts to mixing a local likelihood estimate with predictions from a global generalized linear model. Cross-validation bandwidth and optimal mixing parameter selectors are developed. The global fits and the optimal and data-driven local and mixed fits are studied under no/some/substantial model misspecification via simulation. The methods are then illustrated through application to data from a longitudinal study.
Acknowledgements

Completion of this degree would not have been possible without the continuous support from my advisors Jeffrey Birch and Oliver Schabenberger. I am very fortunate to have had the opportunity to work with them. They invested hundreds of hours to support me in this research, provided invaluable guidance, were patient, and took interest in my well-being. They listened to my talks, read and corrected several drafts, and always made themselves available whenever possible. I thank Dr. Birch especially for bringing my attention to his work—combining parametric and nonparametric regression—I have found it to be a fascinating area. I thank Dr. Schabenberger especially for his expertise with correlated data, the context of my research. Special thanks are in order for my other committee members: George Terrell, Christine Anderson-Cook, and Keying Ye. I appreciate their taking an interest in my research. I thank Dr. Terrell for understanding my change in research interests and his technical expertise—including his help with pesky math questions not related to this work. I thank Key and Christine for motivating comments and sharing experiences with me.

I acknowledge with gratitude the department of statistics for supporting me with advice, assistantships, and education. Special thanks to Michele Marini and Mike Box for assisting me with SAS® and Unix® problems during thousands of hours of computer simulations.

I thank my parents Wilson and Susan Clark and my family for loving support and encouragement to follow my dreams over the years. I thank my friends Jay and Carolyn Magwire, and Ben Smith for convincing me to pursue my graduate degrees. Jay and Carolyn stood by me during some tough times, and always lent an ear when I was frustrated. With gratitude I thank Dr. Abbess Rajia, Dr. Frank Morgan, and Dr. Tekamül Büber for mathematical training during my undergraduate years, friendship, and continued interest to see me through to this end.

I thank my wife Sherry and son John for having the patience and understanding for the many hours and late nights I spent tucked away in my office, in the library, and in the computer lab. Besides being a loving wife, cooking dinner, and running household errands when I was too busy, Sherry proof read each chapter, sometimes at 4 in morning. John, I can now work on giving back to you time that this took from us.

I dedicate this dissertation to the loving memory of my grandmother, Dorothea Streeter Clark (1900–2002). Your love for learning lives in me and you will never be forgotten.
Contents

1 Introduction

2 Global Parametric Estimation

  2.1 Generalized Linear Models
    2.1.1 The Exponential Family and the Three Components of Generalized Linear Models
    2.1.2 Estimation in Generalized Linear Models
    2.1.3 Asymptotic Properties of Estimates and Likelihood Ratios
    2.1.4 Prediction in Generalized Linear Models

  2.2 Generalized Estimating Equations (GEE)
    2.2.1 Quasi-Likelihood Estimation
    2.2.2 Clustered Correlated Data
    2.2.3 The Components of Generalized Estimating Equations
    2.2.4 Parameter Estimation Based on Generalized Estimating Equations
    2.2.5 Asymptotic Properties of Estimates
    2.2.6 Prediction in Generalized Estimating Equations

3 Local Estimation

  3.1 Introduction

  3.2 Locally Weighted Estimation
    3.2.1 Local Estimating Equations (LEE)
    3.2.2 Weight Functions, Bandwidth, and Lack-of-fit Measures

  3.3 Local Polynomial Regression
    3.3.1 The Local Polynomial Model and Estimation
    3.3.2 Asymptotic Bias, Variance, and Optimal Bandwidth for Local Linear Regression

  3.4 Local Polynomial Generalized Linear Models
    3.4.1 Local Polynomial Likelihood
3.4.2 Asymptotic Bias, Variance, and Optimal Bandwidth for Local Linear Generalized Linear Models (LLGLIM) ........................................ 41
3.4.3 Local Linear Generalized Linear Models Estimation ..................... 43
3.5 Local Polynomial Generalized Estimating Equations ......................... 44
  3.5.1 Quasi-likelihood as a Classical Framework for Local Estimation ... 44
  3.5.2 The Local Polynomial GEE (LPGEE) Model .......................... 45
  3.5.3 Asymptotic Bias and Variance and optimal $R$ of the Local Linear GEE (LLGEE) ......................................................... 46
  3.5.4 Local Linear GEE Estimation ............................................ 47

4 Semiparametric Estimation .......................................................... 49
  4.1 Model Robust Regression ....................................................... 51
    4.1.1 PLR, MRR1, and MRR2 .................................................. 51
    4.1.2 Estimation of MRR: $\lambda$ and $h$ ................................... 52
    4.1.3 MRR Convergence Rates ............................................... 53
    4.1.4 Extensions of MRR ..................................................... 54
  4.2 Methods Related to MRR ...................................................... 55
  4.3 Semiparametric Generalized Linear Models .................................. 59
    4.3.1 Generalized Partial Linear Models ................................ 59
    4.3.2 Generalized Single Index Models ................................... 60
    4.3.3 Generalized Additive Models ....................................... 61
  4.4 Semiparametric GEEs .......................................................... 61
    4.4.1 Semiparametric Quasi-Likelihood .................................. 61
    4.4.2 Partial Linear GEE (PLGEE) .......................................... 62

5 Model Robust Generalized Estimating Equations ................................ 65
  5.1 Mixing on Different Scales .................................................. 65
  5.2 Bandwidth Selection .......................................................... 66
    5.2.1 Cross-Validation ....................................................... 66
    5.2.2 Obtaining Deleted Fits ............................................... 69
    5.2.3 Penalized Cross-Validation ......................................... 71
5.2.4 Other Methods .............................................. 74
5.3 Mixing Parameter Estimation ................................. 75
  5.3.1 Minimum Distance ..................................... 76
  5.3.2 Minimum MSE ............................................ 79
5.4 A Relationship Between Selectors ........................... 81
5.5 Simultaneous Estimation of $h$ and $\lambda$ .................... 83

6 Properties of MRGEEI and MRGEEII Fits Under Model Misspecification 89
  6.1 Expectation, Variance and Covariance of the Fits .............. 89
    6.1.1 Deriving $E_P$, $E_{NP}$, $V_P$, $V_{NP}$, and $C$ ............. 89
    6.1.2 Computing $E_P$, $E_{NP}$, $V_P$, $V_{NP}$, and $C$ Under a True Model .... 93
    6.1.3 Estimating $E_P$, $E_{NP}$, $V_P$, $V_{NP}$, and $C$ from Data .......... 96
  6.2 MSE at the Observed Prediction Locations ....................... 97

7 Simulation Studies ............................................. 104
  7.1 Simulation Models .......................................... 104
  7.2 Setup of the Simulations .................................... 108
  7.3 Efficiency of Local GEE Predictions when Using Non-Identity Working Structures ................. 118
  7.4 Validation of the MSE Expressions ......................... 121
    7.4.1 Closeness of $\beta^*$ to $E[\hat{\beta}]$ .................. 122
    7.4.2 Validation of Correlation Parameter Expected Values .......... 124
    7.4.3 Validation of MSE Computations .......................... 125
  7.5 Bandwidth Simulation Study .................................. 130
  7.6 MRGEE Simulation Study .................................... 141
  7.7 Simulation Study of Specific Cases ......................... 160

8 Consistency and Convergence Rates ............................ 171

9 Application, Summary, and Future Research ..................... 175
  9.1 An Application .............................................. 175
  9.2 Summary and Concluding Remarks ............................ 182
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Some Distributions in the Exponential Family</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>Some Examples of Generalized Linear Models</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>The Mean and Variance of GLIM Estimates</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>CD4+ Cell Counts of HIV Infected Men Taken over Time</td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>Method-of-Moment Estimators for Various Correlation Structures</td>
<td>27</td>
</tr>
<tr>
<td>6</td>
<td>The Mean and Variance of GEE Estimates</td>
<td>32</td>
</tr>
<tr>
<td>7</td>
<td>Expectation, Variance and Covariance of MRGEEI and MRGEEII fits</td>
<td>93</td>
</tr>
<tr>
<td>8</td>
<td>Expectation Limits MOM Estimators Applied to GLMM Data</td>
<td>115</td>
</tr>
<tr>
<td>9</td>
<td>Settings of $\sigma$ for No, Low, and High Correlation</td>
<td>115</td>
</tr>
<tr>
<td>10</td>
<td>AvgMSE Comparison of CS and AR(1) to IND Working Structures for LLGEE</td>
<td>120</td>
</tr>
<tr>
<td>11</td>
<td>AvgBias$^2$ and AvgVar Comparison of CS and AR(1) to IND Working Structures for LLGEE</td>
<td>121</td>
</tr>
<tr>
<td>12</td>
<td>Comparison of $E[\hat{\beta}]$ to $\beta^*$</td>
<td>123</td>
</tr>
<tr>
<td>13</td>
<td>Evaluation of First Order Approximations of Expected Value Limits for CS and AR(1) MOM Estimators</td>
<td>124</td>
</tr>
<tr>
<td>14</td>
<td>Validation of AvgMSE Computations for the $\mu$-Scale Poisson Simulation Model</td>
<td>128</td>
</tr>
<tr>
<td>15</td>
<td>Validation of AvgMSE Computations for $\mu$-Scale Binomial Simulation Model</td>
<td>129</td>
</tr>
<tr>
<td>16</td>
<td>Comparison of the Optimal AvgMSE to the Best Possible AvgMSE</td>
<td>135</td>
</tr>
<tr>
<td>17</td>
<td>Bandwidth Simulation Study for the $\mu$-Scale Poisson Simulation Model: Average Estimates</td>
<td>137</td>
</tr>
<tr>
<td>18</td>
<td>Bandwidth Simulation Study for the $\mu$-Scale Poisson Simulation Model: AvgMSE</td>
<td>138</td>
</tr>
<tr>
<td>19</td>
<td>Bandwidth Simulation Study for the $\mu$-Scale Binomial Simulation Model: Average Estimates</td>
<td>139</td>
</tr>
<tr>
<td>20</td>
<td>Bandwidth Simulation Study for the $\mu$-Scale Binomial Simulation Model: AvgMSE</td>
<td>140</td>
</tr>
<tr>
<td>21</td>
<td>Estimated and Optimal Bandwidth and Mixing Parameters for the $\mu$-scale Poisson Simulation Model</td>
<td>152</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gaussian and Epanechnikov Kernel Weight Functions of one Variable</td>
</tr>
<tr>
<td>2</td>
<td>Spherically Symmetric Gaussian and Epanechnikov Kernel Weight Functions of two Variables</td>
</tr>
<tr>
<td>3</td>
<td>MRGEEI Average MSE surface slices for a poisson model with no mean model misspecification</td>
</tr>
<tr>
<td>4</td>
<td>MRGEEI Average MSE surface slices for a poisson model with a high degree of mean model misspecification</td>
</tr>
<tr>
<td>5</td>
<td>MSE Average surface slices for a poisson model with moderate mean misspecification</td>
</tr>
<tr>
<td>6</td>
<td>A 3D representation of Figure 5</td>
</tr>
<tr>
<td>7</td>
<td>$\mu$-Scale Poisson Simulation Model</td>
</tr>
<tr>
<td>8</td>
<td>$\mu$-Scale Binomial Simulation Model</td>
</tr>
<tr>
<td>9</td>
<td>$\eta$-Scale Poisson Simulation Model</td>
</tr>
<tr>
<td>10</td>
<td>$\eta$-Scale Binomial Simulation Model</td>
</tr>
<tr>
<td>11</td>
<td>Correlation Expected Value Limits for $\mu$-Scale Poisson Model</td>
</tr>
<tr>
<td>12</td>
<td>Correlation Expected Value Limits for $\mu$-Scale Binomial Model</td>
</tr>
<tr>
<td>13</td>
<td>Correlation Expected Value Limits for $\eta$-Scale Poisson Model</td>
</tr>
<tr>
<td>14</td>
<td>Correlation Expected Value Limits for $\eta$-Scale Binomial Model</td>
</tr>
<tr>
<td>15</td>
<td>Comparison of PRESS*noV with PRESS**noV Bandwidth Choice with Respect to Theoretical AvgMSE for the $\mu$-Scale Binomial Model</td>
</tr>
<tr>
<td>16</td>
<td>Comparison of CVnoV with PRESS**V Bandwidth Choice with Respect to Theoretical AvgMSE for the $\mu$-Scale Poisson Model</td>
</tr>
<tr>
<td>17</td>
<td>Computed Optimal and Estimated $\lambda$ as a function of Misspecification for the $s = 3$, Corr/Var=Small $\mu$-Scale Poisson Model</td>
</tr>
<tr>
<td>18</td>
<td>Optimal AvgMSE Obtained by Simulation for the $s = 3$, Corr/Var=Small $\mu$-Scale Poisson Model</td>
</tr>
<tr>
<td>19</td>
<td>AvgMSE of the Fits for the $s = 3$, Corr/Var=Small $\mu$-Scale Poisson Model</td>
</tr>
<tr>
<td>20</td>
<td>AvgMSE of the Fits for the $s = 30$, Corr/Var=Small $\mu$-Scale Poisson Model</td>
</tr>
</tbody>
</table>
Computed Optimal and Estimated $\lambda$ as a function of Misspeification for the $s = 3$, Corr/Var=Large $\mu$-Scale Poisson Model .......................... 168

Optimal AvgMSE Obtained by Computation for the $s = 3$, Corr/Var=Large $\mu$-Scale Poisson Model .................................................. 168

AvgMSE of the Fits for the $s = 3$, Corr/Var=Large $\mu$-Scale Poisson Model 169

AvgMSE of the Fits for the $s = 30$, Corr/Var=Large $\mu$-Scale Poisson Model 169

AvgMSE of the Fits for the $n_i = 6$, $s = 3$, Corr/Var=Small $\mu$-Scale Poisson Model ................................................................. 170

AvgMSE of the Fits for the $n_i = 20$, $s = 3$, Corr/Var=Small $\mu$-Scale Poisson Model ................................................................. 170

CD4+ Counts in AIDS Study ................................................................. 176

CD4+ Data with Linear, Quadratic, and Cubic Linear Predictor Parametric Model Estimates and a LLGEE Model Estimate ......................... 180

The Cubic Linear Predictor Parametric Model, LLGEE, and the MRGEE Models Fitted to the CD4 ......................................................... 181

Computing $Var[\bar{y}]$: An Example .................................................. 196

Correlation Structures for the Simulation Models .................................. 210
## Glossary of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIMSE</td>
<td>Asymptotic Integrated Mean Squared Error</td>
<td>75</td>
</tr>
<tr>
<td>AMSE</td>
<td>Asymptotic Mean Squared Error</td>
<td>74</td>
</tr>
<tr>
<td>AR1</td>
<td>Autoregressive(1) correlation structure</td>
<td>24</td>
</tr>
<tr>
<td>AvgMSE</td>
<td>Average Mean Squared Error</td>
<td>37</td>
</tr>
<tr>
<td>CFR</td>
<td>Convergence Failure Rate</td>
<td>119</td>
</tr>
<tr>
<td>CMT</td>
<td>Continuous Mapping Theorem</td>
<td>171</td>
</tr>
<tr>
<td>CS</td>
<td>Compound Symmetric correlation structure</td>
<td>24</td>
</tr>
<tr>
<td>CV</td>
<td>Cross-Validation bandwidth selector</td>
<td>66</td>
</tr>
<tr>
<td>CVnoV</td>
<td>Cross-Validation bandwidth selector without V</td>
<td>131</td>
</tr>
<tr>
<td>CVV</td>
<td>Cross-Validation bandwidth selector with V</td>
<td>131</td>
</tr>
<tr>
<td>DFBETAs</td>
<td>Difference in $\hat{\beta}$ regression diagnostic</td>
<td>71</td>
</tr>
<tr>
<td>DFFITs</td>
<td>Difference in Fits regression diagnostic</td>
<td>71</td>
</tr>
<tr>
<td>EE</td>
<td>Estimating Equations</td>
<td>34</td>
</tr>
<tr>
<td>GEE</td>
<td>Generalized Estimating Equations</td>
<td>22</td>
</tr>
<tr>
<td>GLIM</td>
<td>Generalized Linear Model</td>
<td>5</td>
</tr>
<tr>
<td>GLMM</td>
<td>Generalized Linear Mixed Model</td>
<td>105</td>
</tr>
<tr>
<td>GPLM</td>
<td>Generalized Partial Linear Model</td>
<td>59</td>
</tr>
<tr>
<td>HATLINK</td>
<td>The first MRR procedure that combined “hat” matrices in a convex combination</td>
<td>51</td>
</tr>
<tr>
<td>IMSE</td>
<td>Integrated Mean Squared Error</td>
<td>37</td>
</tr>
<tr>
<td>IND</td>
<td>Independence correlation structure or I</td>
<td>25</td>
</tr>
<tr>
<td>IRLS</td>
<td>Iterated Reweighted Least Squares algorithm</td>
<td>12</td>
</tr>
<tr>
<td>LEE</td>
<td>Local Estimating Equations</td>
<td>34</td>
</tr>
<tr>
<td>LGEE</td>
<td>Abbreviation of LPGEE</td>
<td>47</td>
</tr>
<tr>
<td>LL</td>
<td>Abbreviation of LLGEE</td>
<td>153</td>
</tr>
<tr>
<td>LLGEE</td>
<td>Local Linear Generalized Estimating Equations</td>
<td>46</td>
</tr>
<tr>
<td>LLGLIM</td>
<td>Local Linear Generalized Linear Models</td>
<td>41</td>
</tr>
<tr>
<td>LLR</td>
<td>Local Linear Regression</td>
<td>38</td>
</tr>
<tr>
<td>LPGEE</td>
<td>Local Polynomial Generalized Estimating Equations</td>
<td>45</td>
</tr>
<tr>
<td>LPGLIM</td>
<td>Local Polynomial Generalized Linear Models</td>
<td>41</td>
</tr>
<tr>
<td>LPR</td>
<td>Local Polynomial Regression</td>
<td>38</td>
</tr>
<tr>
<td>MB</td>
<td>Model-Based estimator</td>
<td>29</td>
</tr>
<tr>
<td>mGDybars</td>
<td>Minimum Generalized Distance based on using $y$ in place of the true mean</td>
<td>78</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>mGDys</td>
<td>Minimum Generalized Distance based on using ( y ) in place of the true mean</td>
<td>77</td>
</tr>
<tr>
<td>MISE</td>
<td>Mean Integrated Squared Error</td>
<td>37</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum Likelihood</td>
<td>9</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum Likelihood Estimator</td>
<td>10</td>
</tr>
<tr>
<td>MOM</td>
<td>Method of Moments</td>
<td>26</td>
</tr>
<tr>
<td>MRGEE</td>
<td>Model Robust Generalized Estimating Equations (MRGEEI or MRGEEII or both)</td>
<td>65</td>
</tr>
<tr>
<td>MRGEEI</td>
<td>Model Robust Generalized Estimating Equations I (mean or ( \mu )-scale mixing)</td>
<td>65</td>
</tr>
<tr>
<td>MRGEEII</td>
<td>Model Robust Generalized Estimating Equations II (linear predictor or ( \eta )-scale mixing)</td>
<td>65</td>
</tr>
<tr>
<td>MRI</td>
<td>Abbreviation of MRGEEI</td>
<td>153</td>
</tr>
<tr>
<td>MRII</td>
<td>Abbreviation of MRGEEII</td>
<td>153</td>
</tr>
<tr>
<td>MRQR</td>
<td>Model Robust Quantal Regression</td>
<td>54</td>
</tr>
<tr>
<td>MRR</td>
<td>Model Robust Regression</td>
<td>50</td>
</tr>
<tr>
<td>MRR1</td>
<td>Model Robust Regression 1</td>
<td>51</td>
</tr>
<tr>
<td>MRR2</td>
<td>Model Robust Regression 2</td>
<td>52</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Squared Error</td>
<td>36</td>
</tr>
<tr>
<td>OLS</td>
<td>Ordinary Least Squares</td>
<td>51</td>
</tr>
<tr>
<td>Par</td>
<td>Abbreviation of ParGEE</td>
<td>138</td>
</tr>
<tr>
<td>ParGEE</td>
<td>Parametric Generalized Estimating Equations or GEE</td>
<td>128</td>
</tr>
<tr>
<td>PLGEE</td>
<td>Partial Linear Generalized Estimating Equations</td>
<td>62</td>
</tr>
<tr>
<td>PLR</td>
<td>Partial Linear Regression</td>
<td>51</td>
</tr>
<tr>
<td>PRESS</td>
<td>Prediction Error Sum of Squares</td>
<td>48</td>
</tr>
<tr>
<td>PRESS*</td>
<td>Penalized cross-validation bandwidth selector (penalizes small bandwidths)</td>
<td>72</td>
</tr>
<tr>
<td>PRESS**</td>
<td>Penalized cross-validation bandwidth selector (penalizes large and small bandwidths)</td>
<td>72</td>
</tr>
<tr>
<td>PRESS*noV</td>
<td>PRESS* bandwidth selector without ( V )</td>
<td>131</td>
</tr>
<tr>
<td>PRESS*V</td>
<td>PRESS* bandwidth selector with ( V )</td>
<td>131</td>
</tr>
<tr>
<td>PRESS**noV</td>
<td>PRESS** bandwidth selector without ( V )</td>
<td>131</td>
</tr>
<tr>
<td>PRESS**V</td>
<td>PRESS** bandwidth selector with ( V )</td>
<td>131</td>
</tr>
<tr>
<td>RE</td>
<td>Relative Error</td>
<td>122</td>
</tr>
<tr>
<td>RGU</td>
<td>Rahman, Gokhale, and Ullah (1997) estimator</td>
<td>79</td>
</tr>
<tr>
<td>SMW</td>
<td>Sherman-Morrison-Woodbury theorem</td>
<td>69</td>
</tr>
<tr>
<td>WLS</td>
<td>Weighted Least Squares</td>
<td>38</td>
</tr>
<tr>
<td>WMISE</td>
<td>Weighted Mean Integrated Squared Error</td>
<td>37</td>
</tr>
</tbody>
</table>
### Common Notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbf{Y}_i$</td>
<td>Response vector for cluster $i$</td>
</tr>
<tr>
<td>$Y_{ij}$</td>
<td>$j^{th}$ response for cluster $i$</td>
</tr>
<tr>
<td>$\hat{y}_{i,-i}$</td>
<td>$i^{th}$ cluster fit when the $i^{th}$ cluster is removed</td>
</tr>
<tr>
<td>$\hat{y}_{ij,-i}$</td>
<td>$j^{th}$ fit of the $i^{th}$ cluster when the $i^{th}$ cluster is removed</td>
</tr>
<tr>
<td>$\theta$</td>
<td>Nonparametric (local) parameter vector</td>
</tr>
<tr>
<td>$\hat{\mathbf{y}}_{i,-i}$</td>
<td>Nonparametric (local) parameter vector specific to a prediction location</td>
</tr>
<tr>
<td>$\mathbf{X}_i$</td>
<td>Parametric (global) model matrix</td>
</tr>
<tr>
<td>$\mathbf{X}_{0i}$</td>
<td>Nonparametric (local) model matrix</td>
</tr>
<tr>
<td>$\mathbf{x}^*$</td>
<td>Arbitrary prediction location coordinate vector</td>
</tr>
<tr>
<td>$\mathbf{x}_{ij}^*$</td>
<td>$j^{th}$ prediction location coordinate vector in cluster $i$</td>
</tr>
<tr>
<td>$\mathbf{x}'$</td>
<td>A vector based on $\mathbf{x}^*$ that conforms to a row of $\mathbf{X}_i$</td>
</tr>
<tr>
<td>$\mathbf{x}_{ij}'$</td>
<td>The $j^{th}$ row of $\mathbf{X}_i$</td>
</tr>
<tr>
<td>$\mathbf{x}_{0i}'$</td>
<td>A vector based on $\mathbf{x}^*$ that conforms to a row of $\mathbf{X}_{0i}$</td>
</tr>
<tr>
<td>$\mathbf{x}_{bij}'$</td>
<td>The $j^{th}$ row of $\mathbf{X}_{0i}$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Parametric (global) parameter vector</td>
</tr>
<tr>
<td>$\beta^*$</td>
<td>The parametric (global) parameter vector that minimizes the weighted sum of squares difference between the user’s model and the true mean model</td>
</tr>
<tr>
<td>$\mathbf{R}_i$</td>
<td>Parametric GEE working correlation structure</td>
</tr>
<tr>
<td>$\mathbf{K}_{hi}$</td>
<td>Diagonal matrix of kernel weights for cluster $i$</td>
</tr>
<tr>
<td>$\mathbf{V}_i$</td>
<td>Parametric (global) variance-covariance matrix for cluster $i$</td>
</tr>
<tr>
<td>$\mathbf{V}_{0i}$</td>
<td>Nonparametric (local) variance-covariance matrix for cluster $i$</td>
</tr>
<tr>
<td>$\mathbf{W}_i$</td>
<td>Parametric (global) weight matrix for cluster $i$</td>
</tr>
<tr>
<td>$\mathbf{W}_{0i}$</td>
<td>Nonparametric (local) weight matrix for cluster $i$</td>
</tr>
<tr>
<td>$\mathbf{W}_{0ij}$</td>
<td>Nonparametric (local) weight matrix for cluster $i$, specific to prediction location $\mathbf{x}_{ij}^*$</td>
</tr>
<tr>
<td>$\mathbf{\theta}$</td>
<td>Nonparametric (local) parameter vector specific to a prediction location</td>
</tr>
<tr>
<td>$\mathbf{\theta}^*$</td>
<td>The nonparametric (local) parameter vector that minimizes the weighted sum of squares difference between the local model and the true mean model</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>Mixing parameter</td>
</tr>
<tr>
<td>$\kappa_2$</td>
<td>Scale parameter</td>
</tr>
<tr>
<td>$\phi^2$</td>
<td>Variance function of the mean</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Misspecification or contamination parameter</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>Random effect variance (controls correlation level in the simulation models)</td>
</tr>
</tbody>
</table>

---

$x_{11}$
1 Introduction

Statistical models, in great generality, express outcomes of a random experiment as random fluctuations about a mean structure. Many statistical models accommodate stochastic structure by simultaneously modelling a systematic, fixed component, such as a mean by a mean function and a random component, such as the error about the mean, by a probability distribution. Suppose several instances, $Y_i, i = 1, 2, \ldots, n$, of a response variable $Y$, stacked in a vector $Y$, depend on one or more regressor variables, $x_1, x_2, \ldots, x_k$, whose values are known constants stored in a matrix $X = [ x_1 \ x_2 \ \ldots \ x_k ]$. Suppose $Y_i$ depends on the regressors through a mean function $m(x_{1i}, x_{2i}, \ldots, x_{ki})$. Assume the stochastic variability of the response deviations or errors, $\epsilon = [ \epsilon_i ]_{i=1}^n = Y - E[Y]$, is captured by a probability distribution. This gives rise to a regression model

$$Y = m(x_1, x_2, \ldots, x_k) + \epsilon \text{ where } \epsilon \sim (0, \Sigma).$$

(1)

The errors may be uncorrelated or correlated, depending on the application. The standard regression model assumes uncorrelated errors and hence $\Sigma$, the variance-covariance matrix, is diagonal. In §2.1, where generalized linear models are introduced, this will be the case. In §2.2 however, we will consider an extension of generalized linear models to clustered correlated data using generalized estimating equations. Clustered correlated errors arise in repeated measures and longitudinal studies where measurements are taken repeatedly on the same unit or subject. Such correlations result in a block diagonal structure of the variance-covariance matrix $\Sigma$—each block corresponding to a cluster. An example of a longitudinal study that results in clustered correlated data will be given in §2.2.

The errors $\epsilon_i$ in (1) have the property $E[\epsilon_i] = 0$ which implies that the form of the mean model is assumed to hold for some $m$, a mean model that nature selects. The model $m$ will be called the true model of the mean. In practice the user may pick a model $m^*$,

$$Y = m^*(x_1, x_2, \ldots, x_k) + \epsilon^* \text{ where } \epsilon^* \sim (0, \Sigma^*) \text{ by assumption of the user},$$

(2)

which will be called the user’s model.

If $m^*$ is not the same form as $m$ then $m^*$ is said to be misspecified in functional form. If $m^*$ does not contain the same regressors as $m$ then $m^*$ is said to be misspecified in regressors.
In either case, it may now be true that $E[\epsilon^*_i] \neq 0$ in (2), hence predictions and hypothesis tests with the user’s model may no longer be appropriate. In this work, only misspecification in form will be considered. (Thus by model misspecification we will mean misspecification in form.) It will also be assumed that all regressors included in the model are correct.

Other forms of misspecification of model (1) exist that are related to the error distribution and the manner in which the errors effect the response. For example, the errors may be multiplicative as in the model $Y = m(x_1, x_2, \ldots, x_k)\epsilon$. While it is possible to accommodate this misspecification (see Firth, 1988), it will be assumed that the additive structure of (1) holds. Distributional misspecification in a generalized linear model will occur in the mean, but because the variance depends on the mean, a consequence will be misspecification in variance, whether the variance function is specified correctly or not. For clustered correlated data, the correlation structure may also be misspecified, either because a structure that is too simple is used when the true correlation structure is unknown, or because mean misspecification implies variance and covariance misspecification.

We repeat the assumptions to be made and outline the problem considered in this paper.

**The Assumptions**

1. The model for $Y$ is of the form (1) where the errors are additive. However, the user’s mean model $m^*$ will not be assumed to necessarily have the correct form. That is, there may be model misspecification.

2. All regressors that influence the mean are included in the user’s model $m^*$.

3. – For independent data modeled with a generalized linear model, the proposed error distribution is correct except possibly in mean and variance.
   – For clustered correlated data modeled with generalized estimating equations, the first and second moments may both be incorrect.

The purpose of the following proposed research is to reduce the effect of mean model functional form misspecification in generalized linear models with
and without clustered correlated errors. Our approach is to combine the corresponding global and local regression models by extending previous work in model robust regression (see §4) to the setting of generalized linear models and clustered correlated data. Mays, Birch, and Starnes (2001) have shown that model robust regression predictions based on simple linear regression and local linear regression tend to have smaller prediction variance than the local linear prediction, are asymptotically unbiased, and converge to their true mean at the parametric rate if the parametric model is correct or the nonparametric rate if the parametric model is not correct.

The Problem

1. The user constructs a parametric generalized linear mean model $m^*$, or a mean model $m^*$ for clustered correlated data. First, if $m^*$ is correct ($m^* = m$), then asymptotically there is no prediction bias and the prediction variance can be minimized. Also, the prediction converges quickly to the true mean. Second, if the user’s model is not correct ($m^* \neq m$), then the predictions are biased asymptotically which implies that the prediction does not converge to the true mean at all.

2. The user also considers a nonparametric generalized linear mean model, or a nonparametric mean model $m^*$ for clustered correlated data. The predictions are asymptotically unbiased, regardless of the true mean function $m$. But convergence to the true mean is slower than with parametric prediction. Also, the predictions will tend to have higher prediction variance than the parametric model whether or not the parametric mean model is correct.

3. The user combines the parametric and nonparametric predictions in a convex combination with mixing parameter $\lambda$. The questions that are of interest are as follows.

- Can predictions from a parametric generalized linear model be improved with contributions from a nonparametric fit when the data are uncorrelated?
- Can predictions from a parametric generalized linear model be improved with contributions from a nonparametric fit in the presence of clustered correlated
errors?

- What are the asymptotic properties of the proposed combined parametric and nonparametric predictions for correlated and uncorrelated data as the sample size per cluster or as the number of clusters increases?

More conditions will be introduced for models (1) and (2) for the globally estimated generalized linear models (§2.1) and generalized estimating equations (§2.2). Locally estimated, nonparametric regression procedures are covered in Chapter 3. Mixtures of global and local methods, the basis of this research, appear in Chapter 4. Our model robust methods and estimation techniques are given in Chapter 5. In Chapter 6 the properties of our model robust methods are evaluated under mean model misspecification, and Chapter 7 explores their small sample properties via simulation under differing degrees of misspecification. The consistency and convergence rates of the optimal versions of our methods are determined in Chapter 8. Finally, in Chapter 9 our methods are applied to “real world” data, a summary of the dissertation is given, and topics for future research are discussed.
2 Global Parametric Estimation

The models of this section are estimated from data in a global sense where the estimation of the model parameters does not depend on the location at which prediction is sought. Only one set of parameters is used in the model of the whole dataset. This is to be contrasted with local estimation (§3) where the parameter estimates may be different at each prediction location of interest.

2.1 Generalized Linear Models

2.1.1 The Exponential Family and the Three Components of Generalized Linear Models

A generalized linear model (GLIM), introduced by Nelder and Wedderburn (1972), models independent responses \( Y_i, i = 1, 2, \ldots, n \), with an exponential family distribution as follows

\[
Y_i \sim \text{Expfam}[\vartheta(g^{-1}(x'_i\beta_G)), \psi],
\]

where \( \text{Expfam} \) represents an exponential family member with parameters \( \vartheta \) and \( \psi \); \( \psi \) may be known. The parameter \( \vartheta \) is written as a function of the mean \( g^{-1}(x'_i\beta_G) \), which itself is a function of a linear combination of the regressors. Hence generalized linear models have the following key features (McCullagh and Nelder, 1989).

1. A random component \( Y \) that has a distribution in the exponential family. The density function is given by

\[
f_Y(y) = \exp[(y\vartheta - b(\vartheta))/\psi + c(y, \psi)].
\]

Instances of \( Y, Y_1, Y_2, \ldots, Y_n \), are assumed to be independent. \( \vartheta \) is termed the canonical parameter and is related to \( E[Y] \) through \( b(\cdot) \), namely

\[
\mu = E[Y] = b'(\vartheta).
\]

The variance of \( Y \) is a function of the mean and the scale parameter \( \psi \),

\[
\text{Var}[Y] = \psi \frac{\partial \mu}{\partial \vartheta} = \psi b''(\vartheta).
\]
A critical distinction of generalized linear models from multiple regression and other normal distribution based models is that the variance can depend on the mean. If \( b'(\vartheta) \) is expressed as a function \( a \) of the mean, \( b'(\vartheta) = a(\mu) \), then \( a \) is called the variance function. A few well-known members of the exponential family and expressions for their \( \vartheta, b(\vartheta), \psi, c(y, \psi), E[Y], \) and \( \text{Var}[Y] \) are given in Table 1. Notice that for some distributions the value of \( \psi \) is known.

Table 1: Some Distributions in the Exponential Family

<table>
<thead>
<tr>
<th>Distribution</th>
<th>( \vartheta(\mu) )</th>
<th>( b(\vartheta) )</th>
<th>( \psi )</th>
<th>( c(y, \psi) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>( \mu )</td>
<td>( \frac{1}{2} \vartheta^2 )</td>
<td>( \sigma^2 )</td>
<td>( -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2} y^2 \sigma^{-2} )</td>
</tr>
<tr>
<td>Exponential</td>
<td>( -\mu^{-1} )</td>
<td>( -\log(-\vartheta) )</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Gamma</td>
<td>( -\mu^{-1} )</td>
<td>( -\log(-\vartheta) )</td>
<td>( \alpha^{-1} )</td>
<td>( \alpha \log(\alpha y) - \log(y \Gamma(\alpha)) )</td>
</tr>
<tr>
<td>Poisson</td>
<td>( \log(\mu) )</td>
<td>( e^\vartheta )</td>
<td>1</td>
<td>( -\log(y!) )</td>
</tr>
<tr>
<td>Bernoulli</td>
<td>( \log\left(\frac{\mu}{1-\mu}\right) )</td>
<td>( \log(1+e^\vartheta) )</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Binomial</td>
<td>( \log\left(\frac{\mu}{n-\mu}\right) )</td>
<td>( n \log(1+e^\vartheta) )</td>
<td>1</td>
<td>( \log\left(\frac{n}{y}\right) )</td>
</tr>
<tr>
<td>Negative Binomial</td>
<td>( \log\left(\frac{\mu}{k+\mu}\right) )</td>
<td>( -k \log(1-e^\vartheta) )</td>
<td>1</td>
<td>( \log\left(\frac{y+k-1}{y}\right) )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distribution</th>
<th>( E<a href="%5Cvartheta">Y</a> = \mu )</th>
<th>( \text{Var}<a href="%5Cvartheta">Y</a> )</th>
<th>( \text{Var}<a href="%5Cmu">Y</a> )</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>( \vartheta )</td>
<td>( \sigma^2 )</td>
<td>( \sigma^2 )</td>
<td>( (-\infty, \infty) )</td>
</tr>
<tr>
<td>Exponential</td>
<td>( -\vartheta^{-1} )</td>
<td>( \vartheta^{-2} )</td>
<td>( \mu^2 )</td>
<td>( (0, \infty) )</td>
</tr>
<tr>
<td>Gamma</td>
<td>( -\vartheta^{-1} )</td>
<td>( \alpha^{-1} \vartheta^{-2} )</td>
<td>( \alpha \mu^2 )</td>
<td>( (0, \infty) )</td>
</tr>
<tr>
<td>Poisson</td>
<td>( e^\vartheta )</td>
<td>( e^\vartheta )</td>
<td>( \mu )</td>
<td>( 0, 1, 2, \ldots )</td>
</tr>
<tr>
<td>Bernoulli</td>
<td>( e^\vartheta/(1 + e^\vartheta) )</td>
<td>( e^\vartheta/(1 + e^\vartheta)^2 )</td>
<td>( \mu(1-\mu) )</td>
<td>0, 1</td>
</tr>
<tr>
<td>Binomial</td>
<td>( ne^\vartheta/(1 + e^\vartheta) )</td>
<td>( ne^\vartheta/(1 + e^\vartheta)^2 )</td>
<td>( \mu(1-\mu/n) )</td>
<td>( 0, 1, 2, \ldots, n )</td>
</tr>
<tr>
<td>Negative Binomial</td>
<td>( ke^\vartheta/(1 - e^\vartheta) )</td>
<td>( ke^\vartheta/(n(1 - e^\vartheta)^2 )</td>
<td>( \mu(k + \mu)/k )</td>
<td>( 0, 1, 2, \ldots )</td>
</tr>
</tbody>
</table>

The parameter \( \psi \) is a scale parameter. When it is unknown it must be estimated along with \( \vartheta \).

2. The systematic component of a GLIM is a linear combination of regressor variables, termed the linear predictor \( \eta \),

\[ \eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k. \]

The linear form of the systematic component places the regressors on an additive scale which makes the interpretation of their effects simple. Also, the significance of
each regressor can be tested with a linear hypothesis $H: \beta_i = 0$ versus $K: \beta_i \neq 0$ for $i = 1, 2, \ldots, k$. Although the systematic component is a linear function of the $x$’s, $E[Y]$ need not be linear in the regressors.

3. A link function connects the linear predictor to the mean $E[Y]$. This is done through a monotonic, differentiable function

$$g(\mu) = \eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$ 

The link is a linearizing transformation of the mean—a function that maps the mean onto a scale where regressor effects are linear. One purpose of the link is to allow $\eta$ to range freely while restricting the range of $\mu$. For example, the inverse logit link $\mu = 1/(1 + e^{-\eta})$ maps $(-\infty, \infty)$ onto $(0, 1)$, which is an appropriate range if $\mu$ is a probability. The monotonicity of the link function guarantees that this mapping is one-to-one. Thus we can express the generalized linear model in terms of the inverse link function,

$$E[Y] = m(x_1, \ldots, x_k) = g^{-1}(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k).$$ 

This will be called the true generalized linear mean model.

If the canonical parameter $\vartheta$ itself is set equal to the linear predictor, then $\vartheta(\mu)$ is said to be the canonical link. The canonical link is in many cases a useful link function, and is a reasonable function to try, unless the subject matter suggests otherwise. The canonical link does simplify the estimation method slightly, but there is no need to restrict generalized linear modeling to canonical link functions.

For the generalized linear mean model, we will use the notation $F(x'_i \beta_G)$ for $g^{-1}(x'_i \beta_G)$ and $F(X\beta_G)$ to represent $g^{-1}(x'_i \beta_G)$, $i = 1, 2, \ldots, n$, stacked in a vector. Throughout this work, it will be convenient to denote diagonal matrices, $\text{diag}[c_1, c_2, \ldots, c_n]$, as $\langle c_i \rangle$ or as $\langle c_i \rangle_{i=1,\ldots,n}$ if the values of the subscript are not clear from context. This notation allows the generalized linear model for the entire dataset to be written in the familiar additive form

$$Y = F(X\beta_G) + \epsilon, \text{ where } \epsilon \sim (0, \langle \psi a(\mu_i) \rangle).$$  (3)
Matrices may be represented by their elements as \{c_{ij}\}, or \{c_{ij}\}_{r \times c} if the dimensions of the matrix are not clear. The coordinates of an arbitrary prediction location will be denoted \textbf{x}^\star. If subscripts are added, such as \textbf{x}_i^\star, then the vector will represent the coordinates of an observed prediction location. In the single regressor case, \textbf{x}^\star = x and \textbf{x}_i^\star = x_i so \textbf{x} and \textbf{x}_i can be used in place of the more cumbersome \textbf{x}^\star and \textbf{x}_i^\star. Expression of \textbf{x}^\star or \textbf{x}_i^\star without the * superscript will indicate that the elements of this vector are arranged in the same structure as a row of the model matrix \textbf{X}. (\textbf{x}_i^\prime, as we have already seen, is the \textit{i}^{th} row of \textbf{X}.) For example, if the model matrix \textbf{X} corresponds to the model \( \eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 \) and \textbf{x}^\star = [ 3 \ 5 ]^\prime, then \textbf{x}' = [ 1 \ 3 \ 5 \ 15 \ 9 \ 25 ].

The class of generalized linear models includes many well known statistical models such as multiple regression for normal responses; logistic and probit regression for binary responses, binomial counts, or proportions; Poisson and negative binomial regression; log-linear categorical data analysis models; gamma regression for variance models; and exponential and gamma models for survival time models. These models are summarized in Table 2.

<table>
<thead>
<tr>
<th>E.g.</th>
<th>Response</th>
<th>Regression</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Continuous</td>
<td>Multiple</td>
<td>Normal</td>
</tr>
<tr>
<td>2</td>
<td>Continuous</td>
<td>Gamma</td>
<td>Gamma</td>
</tr>
<tr>
<td>3</td>
<td>Continuous</td>
<td>Exponential</td>
<td>Exponential</td>
</tr>
<tr>
<td>4</td>
<td>Binary</td>
<td>Logistic</td>
<td>Bernoulli/Binomial</td>
</tr>
<tr>
<td>5</td>
<td>Binary</td>
<td>Probit</td>
<td>Bernoulli/Binomial</td>
</tr>
<tr>
<td>6</td>
<td>Counts</td>
<td>Log-Linear</td>
<td>Poisson</td>
</tr>
<tr>
<td>7</td>
<td>Counts</td>
<td>Log-Linear</td>
<td>Negative Binomial</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E.g.</th>
<th>Link ( g(\mu) = \textbf{x}'\beta_G )</th>
<th>Inverse Link ( \mu = F(\textbf{x}'\beta_G) )</th>
<th>Canonical</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \mu )</td>
<td>( \textbf{x}'\beta_G )</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>( 1/\mu )</td>
<td>( 1/\textbf{x}'\beta_G )</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>( 1/\mu )</td>
<td>( 1/\textbf{x}'\beta_G )</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>( \log(\mu/(1 - \mu)) )</td>
<td>( 1/(1 + e^{-\textbf{x}'\beta_G}) )</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>( \Phi^{-1}(\mu) )</td>
<td>( \Phi(\textbf{x}'\beta_G) )</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>( \log(\mu) )</td>
<td>( e^{\textbf{x}'\beta_G} )</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>( \log(\mu) )</td>
<td>( e^{\textbf{x}'\beta_G} )</td>
<td>No</td>
</tr>
</tbody>
</table>
When Nelder and Wedderburn introduced generalized linear models in 1972, they extended Finney’s (1952) application of iteratively reweighted least squares (developed by R. A. Fisher for probit regression) to obtain maximum likelihood estimates. They introduced the term *deviance*, as a measure of model and data agreement. They also considered the change in deviance of a sequential fit of nested models as a generalized analysis of variance, an idea borrowed from Good (1967). McCullagh and Nelder’s first monograph (1983) popularized generalized linear models and their second edition (1989) serves as the standard monograph on generalized linear models today. The literature on generalized linear models and its extensions is voluminous. Generalized linear models have been extended in many ways, such as accommodating random and mixed effects, accommodating correlated data, relaxing distributional assumptions, allowing semiparametric linear predictors, etc. For a review of a few of the extensions, see Smith and Kenward (2000) and Schimek (1997).

### 2.1.2 Estimation in Generalized Linear Models

The parameters in the linear predictor $\eta$ are estimated by the method of *maximum likelihood* (ML). Since the $Y_i$ are assumed to be independent, the joint likelihood is the product of the likelihoods for each $Y_i$. The log likelihood for $\beta_G$, as a function of an arbitrary $\beta$, is then

$$
\ell(\beta|y) = \sum_{i=1}^{n} \left\{ \left[ y_i \vartheta(F(x_i'\beta)) - b(\vartheta(F(x_i'\beta))) \right] / \psi + c(y_i, \psi) \right\},
$$

(4)

where $F(x_i'\beta) = \mu_i$ and $\vartheta$ is written as $\vartheta(\mu_i)$. By the properties of the exponential family (regular domain, twice differentiable $b$) and the fact that the link $g$ is monotonic, the likelihood problem can be solved using elementary calculus. The derivative of the log likelihood is

$$
\frac{\partial \ell(\beta|y)}{\partial \beta} = \sum_{i=1}^{n} \left\{ \frac{f(x_i'\beta)}{\text{var}(Y_i)} (y_i - F(x_i'\beta)) x_i \right\},
$$

(5)

where $f(x_i'\beta) = dF(x_i'\beta)/dx_i'\beta$. Equation (5) is better displayed in matrix notation:

$$
\frac{\partial \ell(\beta|y)}{\partial \beta} = X' \langle f_i \rangle V^{-1} (y - F(X\beta)),
$$

(6)

where $f_i = f(x_i'\beta)$ and $V^{-1} = \left\langle \frac{1}{\text{var}(Y_i)} \right\rangle$. Vector (6) is called the *score function* or score vector and is denoted by $s(\beta)$. The ML estimates of $\beta_G$ are found by solving the score
equation
\[ s_{GLIM}(\hat{\beta}_G) = X'(f_i) V^{-1} (y - F(X\beta)) = 0. \]  
(7)
for \( \hat{\beta}_G \). Notice that the scale parameter \( \psi \) has disappeared from (7). This shows that \( \beta_G \) is estimated independently of the scale parameter.

A Taylor series expansion of \( F(X\beta) \) about \( \beta_G \) yields
\[ F(X\beta) = F(X\beta_G) + \langle f_i \rangle X(\beta - \beta_G) + o(||\beta - \beta_G||) \text{ as } \beta \to \beta_G. \]  
(8)
When the score function is 0, using (8) in (6) implies
\[ \hat{\beta}_G = \beta_G + (X'WX)^{-1}X'W \langle f_i \rangle^{-1} \epsilon + o(||\hat{\beta}_G - \beta_G||) \text{ as } \hat{\beta}_G \to \beta_G, \]  
(9)
where \( \hat{\beta}_G \) has been substituted for \( \beta \) and \( W = \langle f_i \rangle V^{-1} \langle f_i \rangle \). Equations (8) and (9) are results that will be used often.

Agresti (1990) showed that the MLE \( \hat{\beta}_G \) may be obtained using Newton-Raphson, Fisher-scoring, or iterated reweighted least squares. It will be shown how these methods are related to each other in the context of generalized linear models.

**Newton-Raphson**

First consider the Newton-Raphson method, a general purpose numerical method for finding the roots of an equation \( U(\theta) = 0 \). It is derived from a 1\textsuperscript{st} order Taylor series expansion of \( U(\theta) \) or a 2\textsuperscript{nd} order Taylor series expansion of an objective function, \( \ell(\theta) \), about a current estimate. If \( U(\theta) \) is nonlinear in \( \theta \) then Newton-Raphson is an iterative technique. In the maximum likelihood problem, the function \( U \) is the score function \( s \). Consider the second order Taylor series expansion of the log likelihood, \( \ell(\beta) \), centered at \( \beta_G \).
\[ \ell(\beta) = \ell(\beta_G) + \nabla \ell(\beta_G)(\beta - \beta_G) + \frac{1}{2}(\beta - \beta_G)'H_\ell(\beta_G)(\beta - \beta_G) + o(||\beta - \beta_G||^2) \text{ as } \beta \to \beta_G. \]  
(10)
where \( \nabla \ell(\beta_G) \) is the gradient \([\partial \ell/\partial \beta_0 \partial \ell/\partial \beta_1 \cdots \partial \ell/\partial \beta_k]\) of \( \ell \) evaluated at \( \beta_G \) and \( H_\ell(\beta_G) \)
is the Hessian matrix

\[
H_\ell(\beta) = \frac{\partial^2 \ell}{\partial \beta \partial \beta'} = \begin{bmatrix}
\frac{\partial^2 \ell}{\partial \beta_0 \beta_0} & \frac{\partial^2 \ell}{\partial \beta_0 \beta_1} & \cdots & \frac{\partial^2 \ell}{\partial \beta_0 \beta_k} \\
\frac{\partial^2 \ell}{\partial \beta_1 \beta_0} & \frac{\partial^2 \ell}{\partial \beta_1 \beta_1} & \cdots & \frac{\partial^2 \ell}{\partial \beta_1 \beta_k} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 \ell}{\partial \beta_k \beta_0} & \frac{\partial^2 \ell}{\partial \beta_k \beta_1} & \cdots & \frac{\partial^2 \ell}{\partial \beta_k \beta_k}
\end{bmatrix}
\] (11)
evaluated at \(\beta_G\). Letting the first three terms of the series be a second-order approximation of \(\ell(\beta)\), \(\ell\) can be maximized by differentiating and setting the result equal to 0.

\[
\frac{\partial \ell}{\partial \beta} \approx \nabla \ell(\beta_G) + H_\ell(\beta_G)(\beta - \beta_G) = 0
\]
which, when solved for \(\beta\), yields

\[
\hat{\beta}_G \approx \beta_G - H^{-1}_\ell(\beta_G)\nabla \ell(\beta_G).
\] (12)

The Newton-Raphson method uses (12) to update \(\hat{\beta}_G\) through iteration if \(\partial \ell(\beta) / \partial \beta\) is non-linear in \(\beta\). In practice, some initial guess \(\beta_0\), say, is given, then (12) is applied iteratively until some convergence criterion is met, for example when the relative change of \(\hat{\beta}_G\) is less than 0.1%.

**Fisher Scoring**

Recall that the information matrix for \(\beta_G\) is defined as

\[
Var[s(\beta_G)] = E \left[ \left( \frac{\partial \ell}{\partial \beta} \right) \left( \frac{\partial \ell}{\partial \beta} \right)' \right] = -E \left[ \frac{\partial^2 \ell}{\partial \beta \partial \beta'} \right]_{\beta = \beta_G}
\] (13)
which is the negative expected value of the Hessian defined in (11). It can be shown that

\[
-E \left[ \frac{\partial^2 \ell}{\partial \beta \partial \beta'} \right]_{\beta = \beta_G} = -E[H_\ell(\beta_G)] = X'WX
\] (14)
is the information matrix for \(\beta\) in generalized linear models. The method of Fisher scoring replaces the Hessian in (12) with its expected value. Consequently,

\[
\hat{\beta}_G \approx \beta_G + (X'WX)^{-1} \nabla \ell(\beta_G),
\] (15)
is used to obtain estimates. If the link function is canonical, the Hessian is nonstochastic and equals its expected value. This implies that Fisher scoring is the same as Newton-Raphson in this special case.
The Fisher scoring method uses (15) to update $\hat{\beta}_G$ through iteration if $\partial \ell(\beta)/\partial \beta$ is nonlinear in $\beta$. In practice, some initial guess $\beta_0$ is given, then (15) is applied iteratively until some convergence criterion is met.

At convergence of the Fisher scoring algorithm, the final estimate $\hat{\beta}_G$ is the MLE of $\beta_G$ (McCullagh and Nelder, 1989, pp. 41-43). Because the matrix $(X'WX)^{-1}$ is the inverse information matrix for $\beta_G$, it is the variance-covariance matrix of the MLE, $\hat{\beta}_G$. Also $(X'\hat{W}X)^{-1}$, where $\hat{W}$ is $W$ evaluated at the final estimate $\hat{\beta}_G$, is the estimated variance covariance matrix of the MLE, $\hat{\beta}_G$ (Agresti, 1990, pg. 451).

**Iteratively Reweighted Least Squares (IRLS)**

Equation (6) leads to

$$X'WX\beta_G + \nabla \ell(\beta_G) = X'W \left( X\beta_G + \langle f_i \rangle^{-1} \epsilon \right) = X'Wz$$

where $z = X\beta_G + \langle f_i \rangle^{-1} \epsilon$. This result and (15) lead to

$$\hat{\beta}_G \approx (X'WX)^{-1} X'Wz.$$  (16)

This is recognized as a weighted least squares estimate. Note that this result was derived from a second-order Taylor series expansion of $\ell(\beta)$. It may also be derived from a first-order Taylor series expansion of the link as a function of the observations about the estimated mean, which is the approach taken by McCullagh and Nelder (1989, pp. 40-43). Estimation requires iteration if $W$ and $z$ depend on the unknown $\beta_G$, hence the name Iteratively Reweighted Least Squares (IRLS). The working form of IRLS is that of (9), which is the same as (16) when $z$ is expanded.

$$\hat{\beta}_G \approx (X'WX)^{-1} X'W \left( X\beta_G + \langle f_i \rangle^{-1} \epsilon \right) = \beta_G + (X'WX)^{-1} X'W \langle f_i \rangle^{-1} \epsilon$$

The IRLS algorithm for estimating $\beta_G$ for GLIM is as follows.

1. Calculate initial estimates of $\beta$ using ordinary least squares on the transformed response, $F^{-1}(y_i)$, yielding $\hat{\beta}_0 = (X'X)^{-1}X'F^{-1}(y)$. Let $u = 0$. 
2. Compute the weights and derivative components using the current estimate $\hat{\beta}_u$.

Derivative component: $\hat{f}_i = \hat{f}(x'_i\hat{\beta}_u) = dF(\eta)/d\eta|_{\eta=x'_i\hat{\beta}_u}$.

Weight component: $\hat{w}_i = \hat{f}^2(x'_i\hat{\beta}_u)/\text{Var}[Y_i] = \hat{f}^2(x'_i\hat{\beta}_u)/\{\psi a(F(x'_i\hat{\beta}_u))\}$.

3. Update the estimate of $\beta$. Compute

$$\hat{\beta}_{u+1} = \hat{\beta}_u + (X'\hat{W}X)^{-1}X'\hat{W}\langle \hat{f}_i \rangle^{-1}\hat{\epsilon},$$

where $\hat{W} = \langle \hat{w}_i \rangle$ and $\hat{\epsilon} = y - F(X'\hat{\beta}_u)$ is the estimated $\epsilon$.

4. Check a convergence criterion. For example, check if the relative change of the estimate,

$$\frac{(\hat{\beta}_{u+1} - \hat{\beta}_u)}{\hat{\beta}_{u+1}},$$

is less than 0.1%. If the convergence criterion is met, then stop and report $\hat{\beta}_u$ as the MLE estimate of $\beta_G$. Otherwise let $\hat{\beta}_u = \hat{\beta}_{u+1}$, increase $u$ by 1, and return to step 2.

Some comments concerning the IRLS algorithm are necessary.

— If $F^{-1}(y_i)$ does not exist in step 1, use $y_i$ in place of $F^{-1}(y_i)$ or perturb $y_i$ so that $F^{-1}(y_i)$ does exist. For example, for the logit link in logistic regression, if $y_i = 0$ then $\log(\frac{\theta}{1-\theta})$ is undefined, but $\log(\frac{\theta+c}{1-(\theta+c)})$ exists for a small $c > 0$.

— The convergence criterion used in step 4 is similar to The SAS System’s® Proc GENMOD procedure’s criterion.

— The algorithms used in software packages such as The SAS System® may have various modifications to the above. SAS® uses a ridge-stabilized Newton-Raphson algorithm (SAS Institute Inc., 1999) which replaces the updating equation in step 3 with the updating equation (12) in Newton-Raphson.

— If this algorithm has converged at the $u^{th}$ iteration, then $\hat{\beta}_u$ is the MLE for a member of the single parameter exponential family.

— If the exponential family member has two parameters (both $\vartheta$ and $\psi$ are unknown), such as with the negative binomial, then $\psi$ can be estimated from residuals after the estimates of $\beta_G$ have been obtained or Newton-Raphson and Fisher Scoring may be applied to simultaneously estimate all parameters. In the latter case, the vector $\beta$ is augmented with the additional parameter.
2.1.3 Asymptotic Properties of Estimates and Likelihood Ratios

Recall (9) and (8), but now let $\hat{\beta}_G$ represent the $n^{th}$ element of a sequence of random variables that consists of the MLE estimates at convergence, indexed by the sample size used to obtain them.

- The remainder term in (8) has stochastic order $n^{-1/2}$:

$$
F(X\hat{\beta}_G) = F(X\beta_G) + \langle f_i \rangle X(\hat{\beta}_G - \beta_G) + o_p(n^{-1/2}).
$$

(17)

The remainder term in (9) has stochastic order $n^{-1/2}$:

$$
\hat{\beta}_G = \beta_G + (X'WX)^{-1}X'W\langle f_i \rangle^{-1} \epsilon + o_p(n^{-1/2})
$$

(18)

(McCullagh, 1983).

- The MLE $\hat{\beta}_G$ is asymptotically unbiased of order $n^{-1}$:

$$
E[\hat{\beta}_G] = \beta_G + O(n^{-1})
$$

(McCullagh, 1983). This will be expressed as

$$
E_A[\hat{\beta}_G] = \beta_G,
$$

where $E_A$ denotes the asymptotic expectation.

- The variance-covariance matrix of the MLE approaches the Cramér-Rao lower bound, which is the inverse information matrix. This can be expressed as

$$
Var_A[\hat{\beta}_G] = (X'WX)^{-1},
$$

(19)

where $Var_A$ denotes the asymptotic variance. At convergence, the estimated asymptotic variance-covariance is given by

$$
\hat{Var}_A[\hat{\beta}_G] = (X'WX)^{-1}
$$

(20)

The MLEs converge in law (distribution) to a multivariate normal:

\[ \mathcal{L} \left\{ \sqrt{n} \left( \hat{\beta}_G - \beta_G \right) \right\} \to N_p \left( 0, \lim_{n \to \infty} n (X'WX)^{-1} \right), \]

where \( p = k + 1 \) (McCullagh, 1983). Hence we will use the notation

\[ \hat{\beta}_G \sim_n N_p \left( \beta_G, (X'WX)^{-1} \right), \]

where \( \sim_n \) means asymptotically distributed.

The expressions for the mean, variance, and distribution still hold if the exponential family member has two unknown parameters (\( \vartheta \) and \( \psi \)), such as the gamma, negative binomial, or the normal. The MLE vector, the information matrix and hence the asymptotic variance-covariance matrix are augmented appropriately with a row and/or a column corresponding to the additional parameter.

For any likelihood based model, the generalized likelihood ratio test can be employed to test nested hypotheses about the model. The \(-2 \log \text{likelihood ratio} \)

\[ LR = -2 \log \left\{ \frac{L(\hat{\beta}_R|y)}{L(\hat{\beta}_F|y)} \right\} = -2\ell(\hat{\beta}_R|y) - \{ -2\ell(\hat{\beta}_F|y) \}, \]

where \( L(\hat{\beta}_R) \) is the likelihood for the reduced model and \( L(\hat{\beta}_F) \) is the likelihood for the full model. Under certain conditions, \( LR \) has an asymptotic chi-square distribution:

\[ \mathcal{L}\{LR\} \to \chi^2_{f-r}, \]

where \( f \) is the number of parameters in the full model and \( r \) is the number of parameters in the reduced model.

The \(-2 \log \text{likelihood ratio} \) can also be written as

\[ LR = \left( -2\ell(\hat{\beta}_R|y) - \{ -2\ell(\mu = y|y) \} \right) - \left( -2\ell(\hat{\beta}_F|y) - \{ -2\ell(\mu = y|y) \} \right) \]

\[ = D^*(\hat{\beta}_R) - D^*(\hat{\beta}_F). \]  

The components \( D^*(\hat{\beta}_R) \) and \( D^*(\hat{\beta}_F) \) are the \textit{scaled deviances} (scaled by \( \psi \)) of the reduced and the full model, respectively. The \textit{deviance} is

\[ D(\hat{\beta}_G) = -2\psi\ell(\hat{\beta}_G|y) - \{ -2\psi\ell(\mu = y|y) \} = \psi D^*(\hat{\beta}_G) \]  

15
The scaled deviance is a $-2 \log$ likelihood ratio between the proposed model and the saturated model. The saturated model uses the observations $y_i$ to estimate the mean $\mu_i$ directly as $\hat{\mu}_i = y_i$. Clearly then, the scaled deviance and the deviance are lack-of-fit measures for the model. Since the scaled deviance is a $-2 \log$ likelihood ratio, it is asymptotically $\chi^2$ distributed, provided favorable asymptotic conditions hold (See McCullagh and Nelder, 1989, pg. 118 for the binomial data case). An important requirement is that the data are grouped—several observations exist at each $x^*$. 

2.1.4 Prediction in Generalized Linear Models

Estimation of the Mean

Since $F(X\beta) = \mu$, the MLE of the mean is

$$\hat{\mu} = F(X\hat{\beta}_G) \text{ and } \hat{\mu}(x^*) = F(x^*\hat{\beta}_G),$$

for any $x^*$. Note that expansion (8) implies

$$E[F(X\hat{\beta}_G)] = F(X\beta_G).$$

So, $F(X\hat{\beta}_G)$ is also asymptotically unbiased for $\mu$.

Confidence Interval for the Mean

Since $\hat{\beta}_G$ is asymptotically normal,

$$\mathcal{L}\left\{\sqrt{n}(X\hat{\beta}_G - X\beta_G)\right\} \rightarrow N_n\left(0, \lim_{n \to \infty} nX(X'WX)^{-1}X'\right),$$

which implies

$$x^*\hat{\beta}_G \sim N\left(x^*\beta_G, x'(X'WX)^{-1}x\right).$$

So,

$$\frac{x^*\hat{\beta}_G - x'*\beta_G}{\sqrt{x'(X'WX)^{-1}x}} \sim N(0,1).$$

Thus, an asymptotic $(1 - \alpha)100\%$ confidence interval of $x'*\beta_G$ is given by

$$x^*\hat{\beta}_G \pm z_{1-\alpha/2}\sqrt{x'(X'WX)^{-1}x}$$

(27)
In practice, the estimate $\hat{W}$ of $W$ and the more conservative $t$ percentile are used in (27):

$$x'\hat{\beta}_G \pm t_{n-p,1-\alpha/2}\sqrt{x'(X'\hat{W}X)^{-1}x}.$$  

(28)

An approximate asymptotic $(1 - \alpha)100\%$ confidence interval of the mean prediction at $x^*$ is calculated by transforming the limits:

$$F\left(x'\hat{\beta}_G \pm t_{n-p,1-\alpha/2}\sqrt{x'(X'\hat{W}X)^{-1}x}\right).$$  

(29)

This transformation assures that the normal based interval is within the range of the mean function.

We conclude this section with a summary of the asymptotic mean and variance expressions for the estimates $\hat{\beta}_G$, $X\hat{\beta}_G$, $F(X\hat{\beta}_G)$ in Table 3. These will be compared to the corresponding results in other models.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>$E_A[\cdot]$</th>
<th>$Var_A[\cdot]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_G$</td>
<td>$\beta$</td>
<td>$(X'WX)^{-1}$</td>
</tr>
<tr>
<td>$X\hat{\beta}_G$</td>
<td>$X\beta$</td>
<td>$X(X'WX)^{-1}X'$</td>
</tr>
<tr>
<td>$F(X\hat{\beta}_G)$</td>
<td>$F(X\beta)$</td>
<td>$X(X'WX)^{-1}X'\langle f_i \rangle$</td>
</tr>
</tbody>
</table>

2.2 Generalized Estimating Equations (GEE)

2.2.1 Quasi-Likelihood Estimation

The generalized linear model may be misspecified in distribution, correlation structure, link, linear component, or the mean function (link and linear component). These misspecifications can result in a larger or smaller variance than expected under the given model. For example, in logistic regression applied to grouped data where the number of successes, $Y_i$, is modeled with a binomial$(n_i, \pi_i)$, the variance of the $Y_i$ is $\mu_i(1-\mu_i/n)$. If the data are actually Poisson, then the variance is $\mu_i$ which is greater than the assumed model variance. Another example pertains to responses that are positively correlated. Since positive correlation inflates the variance of a summation of responses calculated under the assumption of independence, the
variance of such a sum will be larger than the variance assumed by the generalized linear model. When $\text{Var}[Y]$ exceeds the variance of $Y$ for a specified generalized linear model, we say there is overdispersion, meaning that the data are more variable than allowed by the model. If $\text{Var}[Y]$ is less than the variance for the model then we say there is underdispersion, a condition that is much less likely to occur in practice.

The danger of overdispersion is to underestimate the standard errors of $\hat{\beta}$. We should be less confident of the stability of our estimates if the data are overdispersed. While the best fix is to specify the proper distribution, it is often not possible to do so, especially if little is known about the process generating the data. Wedderburn (1974) proposed a solution to this problem by extending GLIMs to quasi-likelihood. Quasi-likelihood accounts for the overdispersion of a model by multiplying an overdispersion (scale) parameter, $\phi^2$, with the variance of $Y_i$: $\phi^2 \text{Var}[Y_i]$. This changes the GLIM score function to

$$\frac{1}{\phi^2} \cdot X' \langle f_i \rangle V^{-1}(y - F(X\beta)),$$

where $V = \langle \psi a(\mu_i) \rangle$.

In general, the theory of quasi-likelihood assumes only that the forms of the mean and variance functions are known, but the form of the distribution is not known. The variance is assumed to be some function, $\phi^2 a(\mu)$, of the mean, but the function $a(\mu)$ does not have to be the GLIM variance. The general quasi-score for the quasi-likelihood is given by

$$s_{QL}(\beta) = \frac{\partial Q^\ell(\beta|y)}{\partial \beta} = \frac{1}{\phi^2} \cdot D' V^{-1}(y - \mu(\beta)),$$

where $Q^\ell$ is the log quasi-likelihood, $D$ is the derivative matrix $[\partial \mu_i/\partial \beta_j]_{n \times p}$ and $V$ is the scaled (by $\phi^2$) diagonal variance-covariance matrix, $\langle a(\mu_i(\beta)) \rangle$. Note that (30) is a special case of (31) because $D' = X' \langle f_i \rangle$ for a model whose inverse mean function (the link in GLIM) is linear in $\beta$ as in $\mu(\beta) = F(X\beta)$.

Quasi-likelihood estimation is analogous to generalized least squares estimation. In fact, setting (31) to 0 would produce the generalized least squares normal equations if $\mu(\beta)$ is linear in $\beta$. However, the quasi-score’s antiderivative is not a log-likelihood in general.

Wedderburn showed that log-likelihood derivatives have characteristics such as a 0 mean, a variance that is the reciprocal of the variance of $Y_i$, and a negative expected derivative.
that is the reciprocal of the variance of $Y_i$. The asymptotic behavior of maximum likelihood estimates, as illustrated for generalized linear models, is related to these facts. The simple expression,

$$\frac{\partial Q(\mu|y)}{\partial \mu} = \frac{1}{\phi^2} \cdot V^{-1}(y - \mu),$$

for the quasi-likelihood score has the same characteristics and has nice asymptotic properties even though only the forms of the first and second moments have been specified. In the case of regression on one or more covariates, $\mu$ is taken as $\mu(\beta)$. Differentiation of the quasi-score with respect to $\beta$ instead of $\mu$ yields (31), the quasi-score for $\beta$. Note that this result allows for intrinsically nonlinear mean functions of $\beta$.

The quasi-likelihood estimating equations are as follows:

$$D'V^{-1}(y - \mu(\hat{\beta})) = 0,$$

which must be solved for $\hat{\beta}$, while keeping in mind that $D$ and $V$ may depend on $\beta$ in a nonlinear way.

Estimation of $\beta$ is very similar to estimation of GLIMs discussed in §2.1.2. The algorithms are the same except when $\mu(\beta)$ is not of the form $F(X\beta)$. If $\mu(\beta)$ is intrinsically nonlinear in $\beta$, we would use $D'$ in place of $X' \langle f_i \rangle$.

Note that the overdispersion parameter $\phi^2$ dropped out of (32). Therefore the estimation of $\beta$ does not depend on $\phi^2$ and $\phi^2$ cannot be estimated using the quasi-score. Instead, either one of two method-of-moments estimators are often used, Pearson’s chi-square

$$\frac{1}{df_E} \sum_{i=1}^{n} \left( \frac{y_i - F(x_i'\hat{\beta})}{\sqrt{Var[Y_i]}} \right)^2 = \frac{X^2}{df_E},$$

or a scaled deviance statistic

$$\frac{1}{df_E} \left\{ -2\ell(\hat{\beta}|Y) - \{-2\ell(\mu = y|Y)\} \right\} = \frac{D^*(\hat{\beta})}{df_E},$$

where $df_E = n - p$ with $n$ being the number of groups if the data are grouped.

The asymptotic properties of quasi-likelihood estimates are essentially the same as for generalized linear model estimates, provided that the scaled variance-covariance matrix $V$ is diagonal. We only need to be careful that the overdispersion parameter $\phi^2$ is included.
For example, the asymptotic variance-covariance matrix of $\hat{\beta}$ is now $\phi^2 (D'V^{-1}D)^{-1}$, or simply $\phi^2 (X'WX)^{-1}$ instead of $(X'WX)^{-1}$ if $\mu(\beta)$ has the form $F(X\beta)$. Furthermore all prediction results from §2.1.4 are very similar—the only change being the multiplication of the standard errors by $\hat{\phi}$.

Finally, note that quasi-likelihood does not completely fix the overdispersion problem mentioned earlier. In that problem a Poisson count was modeled as a binomial count. Quasi-likelihood attaches the overdispersion parameter to the binomial variance as $\phi^2 \mu_i (1 - \mu_i/n)$ which is not equivalent to the Poisson variance $\mu_i$. All that can be hoped for in this case is that the scale parameter approximately offsets the shrinkage $(1 - \mu_i/n)$. However, quasi-likelihood may still be a viable solution (or patch) if it were unknown that $Y_i$ is Poisson.

### 2.2.2 Clustered Correlated Data

Thus far observations have been assumed to be uncorrelated, but there are many studies that produce dependent data. For example, repeated measures experiments may produce repeated measurements on independent subjects and longitudinal studies generate datasets with several observations collected over time. These data have a clustered correlated structure. The clusters are defined as groups of observations that are correlated, but are uncorrelated with observations in other clusters. Consider an example.

The HIV virus becomes detectable when the body produces antibodies, a process called seroconversion. A human immune cell, called the CD4+ cell, is targeted by the HIV virus. The count of CD4+ cells can be used to assess the body’s immune response to HIV. CD4+ cell counts and several covariates were repeatedly measured on 369 HIV infected males in the Multicenter AIDS Cohort Study (Diggle et al., 1994; Kaslow et al., 1987). A small portion of the data is shown in Table 4.

A vector of CD4+ counts for each man forms a cluster of observations. Since the CD4+ count of a subject at one time point may depend on previous measurements due to autocorrelation over time or depend on the subject’s other measurements because they are made on the same person, the observations in each cluster may be correlated. Different subjects, however, are likely to be independent. Therefore, these data are clustered correlated. Note for some regressors shown in Table 4, their values change within and across clusters (Time
Table 4: CD4+ Cell Counts of HIV Infected Men Taken over Time from the Multicenter AIDS Cohort Study (Diggle et al., 1994; Kaslow et al., 1987)*

<table>
<thead>
<tr>
<th>Subj. ID</th>
<th>Time†</th>
<th>Adj. Age‡</th>
<th>Cigs‡</th>
<th>CD4+ Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>10362</td>
<td>-2.8</td>
<td>2.06</td>
<td>3</td>
<td>1211</td>
</tr>
<tr>
<td>10362</td>
<td>-2.3</td>
<td>2.06</td>
<td>2</td>
<td>1452</td>
</tr>
<tr>
<td>10362</td>
<td>-1.8</td>
<td>2.06</td>
<td>2</td>
<td>703</td>
</tr>
<tr>
<td>10362</td>
<td>0.2</td>
<td>2.06</td>
<td>2</td>
<td>853</td>
</tr>
<tr>
<td>10362</td>
<td>0.7</td>
<td>2.06</td>
<td>3</td>
<td>793</td>
</tr>
<tr>
<td>10132</td>
<td>-1.3</td>
<td>9.33</td>
<td>2</td>
<td>853</td>
</tr>
<tr>
<td>10132</td>
<td>-0.8</td>
<td>9.33</td>
<td>2</td>
<td>1369</td>
</tr>
<tr>
<td>10132</td>
<td>-0.2</td>
<td>9.33</td>
<td>4</td>
<td>1049</td>
</tr>
<tr>
<td>10132</td>
<td>0.2</td>
<td>9.33</td>
<td>3</td>
<td>682</td>
</tr>
</tbody>
</table>

† Time in years relative to the time of occurrence each subject’s seroconversion.
‡‡ Age of the subject relative to a reference age.
‡ The number of packs of cigarettes smoked by the subject per day.

* These data are in the public domain and may be obtained at Peter J. Diggle’s web site: http://www.maths.lancs.ac.uk/~diggle/lda/Datasets/lda.dat

and Cigs), while the values of another covariate remain fixed within each cluster (Adj. Age). The former covariates are called time-varying or observation-specific. The latter covariate is called time-stationary or cluster-specific.

A way to model clustered correlated data is to consider the observed cluster response vectors, \( y_i \), \( i = 1, 2, \ldots, s \), as independent realizations from a multivariate density. There has been much development in the area of multivariate regression for normal responses. Here, brood survival is clearly non-normal. This raises an important question. How do we create distributions for non-normal multivariate data? Formulating such distributions can be difficult and the results are often complicated. Consider a probability model for multivariate binary data,

\[
Pr(Y_i = y_i) \propto \exp \left\{ \sum_{j=1}^{n_i} \theta_{ij} y_{ij} + \sum_{j_1 < j_2} \theta_{ij_1 j_2} y_{ij_1} y_{ij_2} + \cdots + \theta_{i1 \cdots n_i} y_{1i} \cdots y_{ni} \right\}, \tag{35}
\]

where \( y_i = [y_{ij}]_{n_i \times 1} \) (Bishop, Fienberg, and Holland, 1975, pp. 42-45). The 2nd order mo-
ment parameters measure the association between pairs of responses conditional on all other observations within a cluster (Diggle, Liang, and Zeger, 1994, pg. 147). This conditional interpretation of the parameters prevents us from allowing the cluster sizes to vary. This is a problem for observational studies that collect repeated measurements on subjects for different lengths of time and for studies that involve missing data. Also, if the parameters are linked to covariates, the covariate effects will depend on the conditioned observations, making interpretation difficult.

There is in general a lack of off-the-self multivariate distributions for non-normal responses. What should be used for clustered correlated Poisson counts? What about clustered correlated gamma deviates? An alternative to specifying a multivariate model is to formulate a marginal model, for each response, $Y_{ij}$, in the cluster. Marginal models are appropriate for modeling the population-average of all clusters that the data were sampled from. Liang and Zeger (1986), who developed generalized estimating equations (GEEs), assumed the marginal distribution of $Y_{ij}$ was from the exponential family. They accommodated the correlation of the observations in the estimating equations for $\beta_{GEE}$ and provided a semiparametric method for estimation of the standard errors of the parameter estimates. Zeger and Liang (1986) generalized their proposed exponential family marginal likelihood to quasi-likelihood. This had the advantage of freeing the user not only from multivariate distribution specification, but also from marginal distribution specification. Only the forms of the mean and the variance of the marginal distribution are required. The correlation structure is also needed, but it is not required to be correct because the GEE standard errors are robust to correlation structure misspecification.

### 2.2.3 The Components of Generalized Estimating Equations

Since GEEs are based on quasi-likelihood, the important components of a GEE model are the following.

1. The form of the mean function. The mean function is assumed to be of the same form, $F(x'_i\beta_{GEE})$, as the inverse link in generalized linear models. That is, it is assumed that $F$ is a differentiable and monotonic function, and the effects of the covariates are
additive on the $F^{-1}$ scale. The notation for generalized linear models, however, has to be extended to account for the clustered correlated structure of the data. For cluster $i$, the matrix whose columns consist of the fixed covariate measurements will be denoted $X_i, i = 1, 2, \ldots, s$. The $X$ matrix will denote the matrix of stacked $X_i$. Furthermore, $Y$ will represent the stacked response vectors $Y_i, i = 1, 2, \ldots, s$ for each cluster. This notation is illustrated in the mean model

$$E[Y] = E\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{bmatrix} = F\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_s \end{bmatrix}\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} = F(X\beta_{GEE}),$$

where

$$Y_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{bmatrix}$$

and

$$X_i = \begin{bmatrix} 1 & x_{1i1} & x_{2i1} & \cdots & x_{ki1} \\ 1 & x_{1i2} & x_{2i2} & \cdots & x_{ki2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1in_i} & x_{2in_i} & \cdots & x_{kin_i} \end{bmatrix}.$$
\[ \text{define } \phi^2 A_i^{1/2}(\mu) R_i(\rho) A_i^{1/2}(\mu), \]

where \( \rho \) is the vector of correlation parameters. The generalized linear model provides many reasonable variance functions for several response types, so the diagonal matrix of scaled variances of \( Y_i, A_i^{1/2}(\mu) A_i^{1/2}(\mu) = A_i(\mu) \), is often taken to be \( \langle \psi(a(\mu)) \rangle \), where \( \psi(a(\mu)) \) is the variance of \( Y_i \) in a generalized linear model (see \( \S 2.1.1 \)). This is not necessary, however, because quasi-likelihood allows the variance function to take on any functional form of the mean.

The matrix \( R_i \) is called the \textit{working correlation structure} for cluster \( i \) and reflects the user’s assumed correlation pattern between the observations in a cluster. The form of the true correlation structure depends on the stochastic process that generates the data, but this process is often unknown. There are however, reasonable suggestions that can be made for the working correlation matrix in different situations. For example, if the cluster consists of observations measured over time at equally spaced time points, correlation structures from time series may be employed. The most common of these is the \textit{autoregressive}(1) or \( AR(1) \) structure,

\[
R_i^{AR(1)}(\rho) = \begin{bmatrix}
1 & \rho & \rho^2 & \ldots & \rho^{n_i-1} \\
\rho & 1 & \rho & \ldots & \rho^{n_i-2} \\
\rho^2 & \rho & 1 & \ldots & \rho^{n_i-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho^{n_i-1} & \rho^{n_i-2} & \rho^{n_i-3} & \ldots & 1
\end{bmatrix}_{n_i \times n_i},
\]

where \( j \) and \( j' = 1, 2, \ldots, n_i \). This structure is derived from a first-order autoregressive time series model, which regresses the current observation on the previous observation with constant mean and independent errors. It has the property that observations that are farther apart in time are less correlated. Another common correlation structure is the \textit{compound symmetric} (CS) or \textit{exchangeable} structure,

\[
R_i^{CS}(\rho) = \begin{bmatrix}
1 & \rho & \rho & \ldots & \rho \\
\rho & 1 & \rho & \ldots & \rho \\
\rho & \rho & 1 & \ldots & \rho \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho & \rho & \rho & \ldots & 1
\end{bmatrix}_{n_i \times n_i} \cdot \rho J_{n_i \times n_i} + (1 - \rho) I_{n_i},
\]

where \( J_{n_i \times n_i} = [1]_{n_i \times n_i} \). This structure may be appropriate when it is believed that the observations in the cluster are all correlated with each other, but without a par-
ticular ordering. For example, the study of arthritis in horses may involve taking measurements on all legs of each horse. Since there is no reason to give one leg any preference over any other leg, the measurements on each are exchangeable with those of the others. All response pairs from the legs of each horse may be equally correlated. The CS structure also arises naturally in split-plot designs and subsampling designs with independent, nested error terms.

Many other structures can be used. A structure equivalent to the AR(1) for equally spaced data, but also appropriate for unequally spaced data is the exponential structure, $R_{iEX}(\rho) = [\rho^{\lvert u_{ij} - u_{ij}' \rvert}]_{n_i \times n_i}$, where $u_{ij}$ and $u_{ij}'$ are the times when the $j^{th}$ and $j'^{th}$ observations in cluster $i$ are measured. A structure that allows for any pattern of correlations is the unstructured (U) correlation structure, $R_{iU}(\rho) = [\rho_{jj'}]_{n_i \times n_i}$. Other structures include independence (IND) or $I$, the $p^{th}$ order autoregressive or $AR(p)$, the $(p,q)^{th}$ order autoregressive moving average or $ARMA(p,q)$, Huynh-Feldt, Toeplitz, banded, etc. (Schabenberger and Pierce, 2001).

Finally, we issue some warnings. Bahadur (1961), who was the first to provide distribution (35) for multivariate binary data, parameterized it in terms of correlations, showing that if the higher order moments are assumed to be 0, the marginal means place constraints on the correlations. More recently, Gilliland and Schabenberger (2001) have shown for equi-correlated binary data that, depending on the parameterization, the correlations can be severely bounded from above and below even when all the higher moments exist. Also, Prentice (1988) pointed out that the generalized estimating equations ignore these constraints, so the estimation method for $\beta_{GEE}$ may fail. Liang, Zeger, and Qaqish (1992a) and Fitzmaurice, Laird, and Rotnitzky (1993) also noted the constraint problem and suggest that one consider parameterizing $Var[y_{ij}]$ in terms of conditional or marginal odds ratios for the binary response case.

A final point to consider is that a specified GEE model may be completely vacuous with respect to a probability model for the response. There may not exist any multivariate distribution that has the given marginal mean and variance models and the given correlation structure. Nevertheless, the asymptotic results that allow us to do inference on $\beta_{GEE}$ and
2.2.4 Parameter Estimation Based on Generalized Estimating Equations

The quasi-score function for a GEE model is given by

$$s_{GEE}(\beta) = \sum_{i=1}^{s} X_i' \langle f_{ij} \rangle V_i^{-1}(y_i - F(X_i \beta))$$ or more compactly,

$$= X' \langle f_i \rangle V^{-1}(y - F(X \beta)),$$

where $f_i$ is $\langle f_{ij} \rangle_{n_i \times n_i}$, the diagonal matrix of elements $dF(\eta)/d\eta|_{\eta = x'_{ij} \hat{\beta}}$. Note that the only changes made in quasi-score function (31) to account for cluster correlation is the change from a diagonal (scaled) variance-covariance matrix to a block diagonal variance-covariance and the change of the single indexing of the observations to double indexing.

Estimation of $\beta_{GEE}$ is thus based on the generalized estimating equations,

$$X' \langle f_i \rangle V^{-1}(y - F(X \beta)) = 0,$$

which are solved for $\beta$ to provide the parameter estimates $\hat{\beta}_{GEE}$.

Newton-Raphson, Fisher scoring, and iteratively reweighted least squares (§2.1.2) remain appropriate methods for solving (37). However, before an algorithm can be given, estimation of the overdispersion parameter and the covariance matrix must be discussed. The overdispersion parameter cannot be estimated using the generalized estimating equations because it drops out of (37). However, the overdispersion parameter can be estimated with Pearson’s $\chi^2$ method-of-moments (MOM) estimator,

$$\hat{\phi}^2 = \frac{1}{df_E} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left( y_{ij} - F(x'_{ij} \hat{\beta}_{GEE}) \right)^2 \frac{1}{a(F(x'_{ij} \hat{\beta}_{GEE}))} \frac{X^2}{df_E},$$

which is applied to clustered correlated data. Note that $a(F(x'_{ij} \hat{\beta}_{GEE}))$ is the $j^{th}$ diagonal element of $\hat{A}_i$ and $df_E$ are the error degrees of freedom, $\sum n_i - p = N - p$.

Estimation of the correlation parameters can be done using GEE’s themselves. Prentice (1988) and Liang, Zeger, and Qaqish (1992a) specified another set of GEEs for the purpose of estimating the covariance parameters explicitly. However, covariance parameters are considered nuisance parameters in our application; we only want to account for them in the
estimation of the mean. We therefore use the (corrected) method-of-moment estimators of Liang and Zeger (1986). They are simple to calculate and are commonly used\(^1\).

Define the \(i^{th}\) Pearson residual as

\[
p_{ij} = \frac{y_{ij} - F(x_{ij}' \hat{\beta}_{GEE})}{\sqrt{\text{Var}[Y_{ij}]}},
\]

where \(\text{Var}[Y_{ij}] = \hat{\phi}^2 a(F(x_{ij}' \hat{\beta}_{GEE}))\) is the estimated variance of \(Y_{ij}\). Let \(I(j, j' \leq n_{i'})\) indicate if \(j\) and \(j'\) are both \(\leq n_{i'}\) for cluster \(i'\). Then the method-of-moments estimators of \(\rho\) for various correlation structures are given in Table 5 (Liang and Zeger, 1986, with corrections).

Table 5: Method-of-Moment Estimators for Various Correlation Structures

<table>
<thead>
<tr>
<th>(R_i(\rho))</th>
<th>Form</th>
<th>MOM* Estimator of (\rho)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(1)((\rho))</td>
<td>([p_{ij} - p_{i'j'}]) (n_i \times n_i)</td>
<td>(\frac{\sum_{i'=1}^{n_i} \sum_{j=1}^{n_{i'}-1} p_{i'j}p_{i'j'+1}}{N-s-p})</td>
</tr>
<tr>
<td>CS((\rho))</td>
<td>(\rho J_{n_i \times n_i} + (1 - \rho)I_{n_i})</td>
<td>(\frac{\sum_{i'=1}^{n_i} \sum_{j&lt;j'}^{n_{i'}} p_{i'j}p_{i'j'} \sum_{i'=1}^{n_i} n_{i'}(n_{i'}-1)/2-p}{\sum_{i'=1}^{n_i} n_{i'}(n_{i'}-1)/2-p})</td>
</tr>
<tr>
<td>EXP((\rho))</td>
<td>([p_{ij} - p_{i'j'}]) (n_i \times n_i)</td>
<td>(\exp \left{ \frac{\sum_{i'=1}^{n_i} \sum_{j&lt;j'}^{n_{i'}}</td>
</tr>
<tr>
<td>U((\rho))</td>
<td>([p_{ij'}]) (n_i \times n_i)</td>
<td>(\frac{\sum_{i'=1}^{n_i} p_{i'j}p_{i'j}I(j,j' \leq n_{i'})\sum_{i'=1}^{n_i} I(j,j' \leq n_{i'}) - p}{\sum_{i'=1}^{n_i} I(j,j' \leq n_{i'}) - p} )</td>
</tr>
</tbody>
</table>

* These are MOM estimators adjusted by degrees of freedom.

\(^1\) Unfortunately, they may not always provide consistent estimates of the correlation structure as Crowder (1995) has shown. If \(R_i(\rho)\) is not the true correlation matrix then there is no guarantee that a true value of \(\rho\) even exists. See Chaganty (1997) for alternative methods of estimating the correlation structure consistently.
Now all the necessary components for an estimation algorithm are available. The GEE IRLS algorithm for estimating \( \beta_{GEE} \) is as follows.

1. Calculate initial estimates, \( \hat{\beta}_0 \), of \( \beta_{GEE} \), using the IRLS algorithm for estimating \( \beta_G \) for GLIM. Let \( u = 0 \).

2. Estimate the correlation parameter(s) and compute the weight and derivative components using the current estimate \( \hat{\beta}_u \).

   **Derivative component:**
   
   Compute \( \hat{f}_{ij} = \hat{f}(x'_{ij}\hat{\beta}_u) = dF(\eta)/d\eta|_{\eta=x'_{ij}\hat{\beta}_u} \) and construct \( \langle \hat{f}_{ij} \rangle \).

   **Weight component:**
   
   Compute \( \hat{a}_{ij}^{1/2} = \sqrt{a(F(x'_{ij}\hat{\beta}_u))} \),
   
   \( \hat{A}_i^{1/2} = \langle \hat{a}_{ij}^{1/2} \rangle_{n_i \times n_i} \),
   
   \( \hat{V}_i = \hat{\phi}^2 \hat{A}_i^{1/2} R_i(\hat{\rho}) \hat{A}_i^{1/2} \),
   
   and then \( \hat{W}_i = \langle f_{ij} \rangle \hat{V}_i^{-1} \langle f_{ij} \rangle \).

3. Update the estimate of \( \beta_{GEE} \). Compute

   \[
   \hat{\beta}_{u+1} = \hat{\beta}_u + (X'\hat{W}X)^{-1}X'\hat{W}\langle \hat{f} \rangle^{-1} \hat{\epsilon}
   \]

   where \( \hat{W} = \langle \hat{W}_i \rangle \) and \( \hat{\epsilon} = y - F(X'\hat{\beta}_u) \) is the estimated \( \epsilon \).

4. Check a convergence criterion such as the relative change of the estimate,

   \( (\hat{\beta}_{u+1} - \hat{\beta}_u)/\hat{\beta}_{u+1} \). If the convergence criterion is met, stop and report \( \hat{\beta}_{GEE} = \hat{\beta}_u \) as the GEE estimate of \( \beta \). Otherwise let \( \hat{\beta}_u = \hat{\beta}_{u+1} \), increase \( u \) by 1, and return to step 2.

2.2.5 Asymptotic Properties of Estimates

Let \( \hat{\beta}_{GEE} \) represent the \( s^{th} \) element of a sequence of random variables that consist of the GEE estimates at convergence, indexed by the number of clusters, \( s \), used to obtain them.

- Expansions (17) and (18) still hold, but with modifications for clustered correlated data. The remainder term of (17) for clustered correlated data has stochastic order
\[ F(X\hat{\beta}_{GEE}) = F(X\beta_{GEE}) + \langle f_i \rangle X(\hat{\beta}_{GEE} - \beta_{GEE}) + o_p(s^{-1/2}). \] (39)

The remainder term of (18) for clustered correlated data has stochastic order \( s^{-1/2} \):

\[ \hat{\beta}_{GEE} = \beta_{GEE} + (X'WX)^{-1}X'W\langle f_i \rangle^{-1}\epsilon + o_p(s^{-1/2}) \] (40)

- The estimate \( \hat{\beta}_{GEE} \) is asymptotically unbiased of order \( s^{-1} \):

\[ E[\hat{\beta}_{GEE}] = \beta_{GEE} + O(s^{-1}). \]

So, \( E_A[\hat{\beta}_{GEE}] = \beta_{GEE} \).

- Using (40), what is called the asymptotic variance-covariance matrix of the GEE estimate \( \hat{\beta}_{GEE} \) for finite \( s \), is given by

\[ V_{GEE} = \left( \sum_{i=1}^{s} X_i'\hat{W}_iX_i \right)^{-1} \left( \sum_{i=1}^{s} X_i'\hat{W}_i\langle f_{ij} \rangle^{-1}Var[Y_i]\langle f_{ij} \rangle^{-1}\hat{W}_iX_i \right)^{-1} \left( \sum_{i=1}^{s} X_i'\hat{W}_iX_i \right)^{-1} \]

\[ = Var_A[\hat{\beta}_{GEE}], \] (41)

where \( Var[Y_i] \) is the true, but unknown, variance-covariance matrix of cluster \( i \). Notice that if we assume that \( R_i \) has been specified correctly, then (41) collapses into the \textit{model-based} (MB) asymptotic variance,

\[ V_{MB} = \left( \sum_{i=1}^{s} X_i'\hat{W}_iX_i \right)^{-1} = (X'WX)^{-1}, \]

which is similar to the variance-covariance matrix (19) used in GLIM. In practice, because \( R_i \) is not assumed to be correct, the estimated asymptotic variance-covariance

\[ \hat{V}_{GEE} = \left( \sum_{i=1}^{s} X_i'\hat{W}_iX_i \right)^{-1} \left( \sum_{i=1}^{s} X_i'\hat{W}_i\langle f_{ij} \rangle^{-1}\hat{Var}[Y_i]\langle f_{ij} \rangle^{-1}\hat{W}_iX_i \right)^{-1} \left( \sum_{i=1}^{s} X_i'\hat{W}_iX_i \right)^{-1}, \]

\[ \] (42)

called the \textit{sandwich estimator}, is used, where \( \hat{Var}[Y_i] \) is the method-of-moments estimator,

\[ \hat{\epsilon}_i\hat{\epsilon}_i' = (Y_i - F(X_i\hat{\beta}_{GEE})) (Y_i - F(X_i\hat{\beta}_{GEE}))', \]

of \( Var[Y_i] \) (Zeger and Liang, 1986).
• Assuming that \( \hat{\phi}^2 \) and \( \hat{\rho} \) are \( \sqrt{s} \) consistent estimates of \( \phi^2 \) and \( \rho \), respectively, (and that \( \partial \hat{\rho} / \partial \phi \) is bounded in probability,) the estimates \( \hat{\beta}_{GEE} \) converge in law to a multivariate normal:

\[
\mathcal{L}\left\{ \sqrt{s} (\hat{\beta}_{GEE} - \beta) \right\} \to N_p (0, V_G),
\]

where

\[
V_G = \lim_{s \to \infty} s V_{GEE}
\]

(Zeger and Liang, 1986). Thus we can write

\[
\hat{\beta}_{GEE} \overset{A}{\sim} N_p (\beta, V_{GEE})
\]

Zeger and Liang (1986) show that these results hold for any \( \sqrt{s} \) consistent estimators of \( \phi^2 \) and \( \rho \). Since \( \hat{\epsilon}, \hat{\epsilon}', \hat{\phi}^2 \), and \( \hat{\rho} \) are consistent for \( \text{Var}[Y_i], \phi^2 \), and \( \rho \), respectively, \( \hat{\beta}_{GEE} \) is consistent for \( \beta \). Zeger and Liang note that \( s \hat{V}_{GEE} \) will also be consistent for \( V_G \) under the additional assumption that a weighted average of the estimated correlation matrices converges to a fixed matrix. They show that \( \hat{\beta}_{GEE} \) and \( s \hat{V}_{GEE} \) are consistent even when the working matrix, \( R_i \) (which is involved in \( W_i \)) is not correct. In this sense, the estimate of \( V_{GEE} \), and hence the standard errors of \( \hat{\beta}_{GEE} \) are robust to working correlation misspecification. Moreover, Rotnitzky and Jewell (1990) point out that \( \hat{V}_{GEE} \) is also consistent even when the cluster sizes vary, as long as the correlation matrices for small clusters are submatrices of the correlation matrices for larger clusters.

There is, of course, the advantage of more efficient estimates when choosing the correct form for \( R_i \), or at least a form that is closer to the true correlation matrix. Simulation studies and examples by Liang and Zeger (1986), Liang, Zeger, and Qaqish (1992a), and Fitzmaurice, Laird, and Rotnitzky (1993) in general have indicated this. Liang and Zeger (1986) and Liang et al. (1992a) have suggested that when the correlations between pairs of responses are small (say less than 0.4) then the independence-based estimating equations (GEE with diagonal working correlation) produce estimates that are nearly as efficient. More recently, Mancl and Leroux (1996) have shown that efficiency of estimates from the independence-based estimating equations really depends on the covariate variation within and between clusters as well as the size of the correlations. Small correlations do not necessarily imply
that working independence GEE will produce estimates that are nearly as efficient as GEE estimates obtained using non-diagonal working structures.

The instability of the sandwich estimator has been criticized by Prentice (1988), Firth (1992), Drum and McCullagh (1993), among others. The MOM estimator $\hat{\epsilon}_i \hat{\epsilon}_i'$ can be very unstable, so likewise $\hat{V}_{GEE}$ tends to be quite variable. This is the price paid for robustness and illustrates the robustness-efficiency trade-off common to many methods achieving robustness with respect to one particular aspect of model misspecification.

### 2.2.6 Prediction in Generalized Estimating Equations

The results of this section are very similar to those of §2.1.4.

#### Estimation of the Mean

Since $F(X\beta) = \mu$, the customary substitution estimator of $\mu$ is given by

$$\hat{\mu} = F(X\hat{\beta}_{GEE})$$ and $\hat{\mu}(x^*) = F(x'\hat{\beta}_{GEE}),$

(45)

for an arbitrary prediction location $x^*$. Note that expansion (39) implies

$$E[F(X\hat{\beta}_{GEE})] = F(X\beta_{GEE}) + O(s^{-1}).$$

(46)

So $F(X\hat{\beta}_{GEE})$ is also asymptotically unbiased for $\mu$ of order $s^{-1}$.

#### Confidence Interval for the Mean

Since $\hat{\beta}_{GEE}$ is asymptotically normal,

$$L \left\{ \sqrt{s} \left( X\hat{\beta}_{GEE} - X\beta_{GEE} \right) \right\} \rightarrow \mathcal{N}_{\sum n_i} \left( 0, \lim_{s \to \infty} sXV_{GEE}X' \right),$$

which implies

$$x'\hat{\beta}_{GEE} \sim \mathcal{N}(x'\beta_{GEE}, x'V_{GEE}x).$$

So,

$$\frac{x'\hat{\beta}_{GEE} - x'\beta}{\sqrt{x'V_{GEE}x}} \sim \mathcal{N}(0, 1).$$

Thus an asymptotic $(1 - \alpha)100\%$ confidence interval of $x'\beta_{GEE}$ is given by

$$x'\hat{\beta}_{GEE} \pm z_{1-\alpha/2}\sqrt{x'V_{GEE}x}$$

(47)
In practice, the estimate $\hat{V}_{GEE}$ of $V_{GEE}$ and the more conservative $t$ percentile are used in (47):

$$x'\hat{\beta}_{GEE} \pm t_{N-p,1-\alpha/2} \sqrt{x'\hat{V}_{GEE}x}.$$  \hfill (48)

An approximate asymptotic $(1-\alpha)100\%$ confidence interval of the mean prediction at $x^*$ is calculated by transforming the limits as was done in (29):

$$F\left(x'\hat{\beta}_{GEE} \pm t_{N-p,1-\alpha/2} \sqrt{x'\hat{V}_{GEE}x}\right).$$  \hfill (49)

We conclude this section with a summary (Table 6) of the asymptotic mean and variance expressions for the estimates as we did in §2.1. In comparison with Table 3, the most notable change in the asymptotic moments is the use of the sandwich variance, $V_{GEE}$, as opposed to the model-based variance, $(X'WX)^{-1}$.

Table 6: The Mean and Variance of GEE Estimates

<table>
<thead>
<tr>
<th>Statistic</th>
<th>$E_A[\cdot]$</th>
<th>$Var_A[\cdot]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_{GEE}$</td>
<td>$\beta_{GEE}$</td>
<td>$V_{GEE}$</td>
</tr>
<tr>
<td>$X\hat{\beta}_{GEE}$</td>
<td>$X\beta_{GEE}$</td>
<td>$XV_{GEE}X'$</td>
</tr>
<tr>
<td>$F(X\hat{\beta}_{GEE})$</td>
<td>$F(X\beta_{GEE})$</td>
<td>$\langle f_i \rangle XV_{GEE}X' (f_i)$</td>
</tr>
</tbody>
</table>
3 Local Estimation

3.1 Introduction

Generalized linear and generalized estimating equation global models were specified parametrically in §2. They are applied to the dataset in its entirety with equal or model based weights used in the estimation of the parameters. In this section we consider local estimation, which concentrates model fitting at a specific prediction location, by fitting models with emphasis only on nearby observations in a small neighborhood about this prediction location. Additional weights, as functions of the prediction location, are incorporated to define the neighborhood that makes the global model estimation act locally. Local regression estimates nature’s mean model $m$ under the assumption that it is a smooth function. The variance function is the same as in (1) except that $\Sigma$ is allowed to be a smooth function of the prediction location.

Recall the notation used in the previous sections. The vector of coordinates of the prediction location is denoted $\mathbf{x}^\star$, if there are multiple covariates, or $\mathbf{x}^\star = x$ if there is only one covariate. If $\mathbf{x}^\star$ is an observed data point, then it is denoted by $\mathbf{x}_i^\star$ or $\mathbf{x}_{ij}^\star$ to represent the coordinates of the $i^{th}$ or $ij^{th}$ observed point, respectively. The $i^{th}$ row of the parametric model matrix $\mathbf{X}$ for GLIM and the $ij^{th}$ row of the model matrix $\mathbf{X}_i$ for GEE are denoted by $\mathbf{x}_i'$ and $\mathbf{x}_{ij}'$, respectively, so there is a clear distinction between the coordinates of a point and the representation of that point as a row of a model matrix. We will continue to use $\mathbf{x}^\star$ to denote the prediction location, but additional notation will be introduced to denote nonparametric model matrices and their row vectors, nonparametric parameter vectors, and nonparametric variance and weight matrices to make a clear distinction between parametric and nonparametric models.
3.2 Locally Weighted Estimation

3.2.1 Local Estimating Equations (LEE)

Many classical estimation methods estimate a parameter $\theta$ using independently observed data $y_i, i = 1, \ldots, n$ by finding the roots of the estimating equations (EEs),

$$Q(\theta^*; y) = \sum_{i=1}^{n} Q_i(\theta; y_i) = 0$$

(Carroll, Ruppert, and Welsh, 1998). The score equations (7) for generalized linear models, the quasi-score equations (32), and the generalized estimating equations (37) can be written in this form for appropriately defined $Q$, $\theta$, and $n$. For example, if $Q_i(\theta; y_i) = [x_i f_i / \phi^{-2} \sigma^{-2} (F(x'_i \beta))] (y_i - F(x'_i \beta))$ for quasi-likelihood estimation, then $Q(\theta; y) = \frac{1}{\phi^2} X' (f_i) V^{-1} (y - F(X \beta))$, the quasi-score equations, where $\theta = \beta$. Also, the normal equations of ordinary least squares, the score equations of maximum likelihood, and the $\Psi$ function for M estimation can be written in this form. We refer to the EE method of estimation as global estimation.

Local estimation differs from this approach in two important aspects. First, the underlying model is estimated separately at each point $x^*$ where predictions are sought, yielding parameter estimates $\hat{\theta}(x^*)$ specific to $x^*$. Second, for uncorrelated data, neighborhood-defining weights are applied to the terms $Q_i$ in (50) to produce the local estimating equations (LEEs),

$$\sum_{i=1}^{n} w_i(x^*_i, x^*) Q_i(\theta(x^*); y_i) = 0, \quad (51)$$

which are solved for $\theta(x^*)$ to obtain $\hat{\theta}(x^*)$. The uncorrelated nature of the data allow the weights to be placed outside of $Q_i(\theta(x^*); y_i)$.

3.2.2 Weight Functions, Bandwidth, and Lack-of-fit Measures

The weights in (51) define the neighborhood around $x^*$ as a function of $x^*_i$ and $x^*$. They are often a function of the distance of $x^*_i$ from $x^*$ and of a scale parameter $h$, called the bandwidth. Such weight functions, denoted $K_h(||x^*_i - x^*||)$, are called the kernel functions.

Many kernel functions used for density estimation and kernel regression problems involving a single variable can be written in the form $K_h(x_i, x) = K((x_i - x)/h)/h$, where $K$ is
a normalized kernel function. We will assume $K$ is a symmetric density with mean 0, and variance 1. Common $K$ include Gaussian, Epanechnikov, triangle, uniform, among many others. The Epanechnikov kernel is known to be optimal among a wide class of kernels in the context of density estimation (Epanechnikov, 1969) and local polynomial regression (Fan and Gijbels, 1996, pg. 75), but many other kernels are nearly as efficient (see Wand and Jones, 1995, pg. 31). The choice of the kernel function is known to have little effect on the bias and variance of the prediction in a variety of estimation settings (see Wand and Jones, 1995, pg. 13, for example). Hence, when a specific univariate kernel is required, we will use either the Gaussian, $K_G(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$, or the Epanechnikov, $K_E = 0.75(1 - u^2)I(|u| \leq 1)$, at our convenience. These are shown in Figure 1. Notice that the Epanechnikov has bounded support, $u \in (-1, 1)$, and the Gaussian has unbounded support, $u \in (-\infty, \infty)$.

There also many multivariate kernels used for multiple nonparametric regression. Those that are functions of distance, $K_h(||x_i^* - x^*||)$, are called spherically symmetric kernel functions (Wand and Jones, 1995). They can depend on only one bandwidth parameter if the degree of smoothing in each dimension is forced to be the same. In this case, they give the same weight to every point in an sphere around the prediction location. All the univariate kernels mentioned above are special cases. The spherically symmetric Gaussian and Epanechnikov kernels for two variables are shown in Figure 2. Many of the following sections are based on univariate kernels because the mathematics is less cumbersome, but several of the results can be extended to the multiple regressor case (see Ruppert and Wand, 1994, for example). Sections §5 and §6 are not restricted to the univariate case, but §7 is, in order to limit the size and scope of the simulation study.

The bandwidth, $h$, is related to the variance of the kernel weight function, $K_h$. It determines the window width, or the effective size of the neighborhood about $x^*$ of highly weighted observations. The bandwidth can have profound effect on the estimates and their properties because it determines the variability of $\hat{\theta}(x^*)$ over values of $x^*$, i.e., the smoothness of the fitted model. A small bandwidth can result in a highly variable set of estimates over $x^*$, because observations far from $x^*$ receive little weight. A large bandwidth, on the other hand, may produce a set of estimates that are all very similar to a single global model estimate, $\hat{\theta}$, because it assigns nearly equal kernel weight to the observations. The bandwidth may be
chosen by the user, but an objective analysis requires automatic bandwidth selection. This is a difficult problem as evidenced by the volume of literature in simple settings (see Jones, Marron, and Sheather, 1996). Bandwidth selection is not the primary focus of this work, though we explore this issue as the need arises.

Lack-of-fit of the estimate $\hat{m}$ from $m$ is typically measured in one of two ways—pointwise or global. Pointwise lack-of-fit involves measuring the discrepancy of fit at $x^*$. A classical discrepancy measure of a parameter from its estimator is the mean squared error (MSE)

$$MSE(x^*) = E \left[ (\hat{m}(x^*) - m(x^*))^2 \right]$$

$$= (E[\hat{m}(x^*)] - m)^2 + E \left[ (\hat{m}(x^*) - E[\hat{m}(x^*)])^2 \right]$$

$$= bias^2[\hat{m}(x^*)] + var[\hat{m}(x^*)]$$

\[52\]
(Wand and Jones, 1995, pg. 14). The bias and variance decomposition of the MSE shows that minimizing the MSE may shrink both the bias and the variance of the fit, may trade larger bias for variance reduction or may trade larger variance for bias reduction. This phenomenon is commonly called the bias and variance trade-off of MSE.

A natural extension of $MSE(x^*)$ that provides an overall measure of lack-of-fit is the integrated $MSE(x^*)$,

\[
MISE = E \left[ \int (\hat{m}(x^*) - m(x^*))^2 dx^* \right] = \int MSE(x^*) dx^*
\]

\[
= \int bias^2[\hat{m}(x^*)]dx^* + \int var[\hat{m}(x^*)]dx^*,
\]

assuming that integration and expectation can be exchanged. This is called the mean integrated squared error, and as (53) shows, it decomposes into integrated squared bias and integrated variance terms. Often the MISE criterion is evaluated with a weight function, $w(x^*)$, as $WMISE = \int [\hat{m}(x^*) - m(x^*)]^2 w(x^*) dx^*$. The weight function $w(x^*)$ can be used to inflate discrepancies in certain regions of the $x^*$ space. Another measure of overall fit is the average MSE over the observed $x^*_i$,

\[
AvgMSE = \frac{1}{n} \sum (\hat{m}(x^*_i) - m(x^*_i))^2
\]

\[
= \frac{1}{n} \sum bias^2[\hat{m}(x^*_i)] + \frac{1}{n} \sum Var[\hat{m}(x^*_i)].
\]

This is simpler than the MISE, but accounts for discrepancy only at the observed covariate locations.

3.3 Local Polynomial Regression

3.3.1 The Local Polynomial Model and Estimation

The simplest form of local polynomial regression, kernel regression, was developed independently by Nadaraya (1964) and Watson (1964). This procedure fits a degree 0 polynomial, i.e., a constant, at each $x_0$. The fit is a weighted average,

\[
\hat{m}(x) = \sum_{i=1}^{n} \frac{K_h(x_i, x)}{\sum_{i=1}^{n} K_h(x_i, x)} Y_i,
\]

of the responses $Y_i$. 

37
Though this is very simple, Fan and Gijbels (1996, pp. 17-18) consider kernel regression to be outdated technology because it has a number of problems, including bias when the distribution of the $x_i$ in the $x$ space is highly variable and *edge-effect* biases near the boundaries of $x$ space, where less data are available to estimate the trend. A higher-degree local polynomial regression offers a simple and more satisfying solution to these problems.

*Local polynomial regression* (LPR) generalizes the Nadaraya-Watson estimator by replacing the local constant with a local polynomial. This is motivated by the fact that a Taylor series expansion of the mean function $m$ about $x$,

\[
m(x_i) \approx m(x) + m'(x)(x_i - x) + \cdots + \frac{1}{p!} m^{(p)}(x)(x_i - x)^p,
\]

provides a good polynomial approximation of $m$ for $x_i$ close to $x$.

There is an obvious naming of higher order polynomial estimators: a 1st degree polynomial estimator is called *local linear regression* (LLR), a local 2nd degree fit is called *local quadratic regression*, and so on. Fan and Gijbels (1996, pp. 76-80) have shown that odd order local polynomial estimators have less bias than even order local polynomial estimators. Also, they have shown that LLR performs nearly as well as other higher odd order local polynomial estimators. Thus LLR is usually the LPR method of choice for estimating $m(x)$. These results carry over to more complicated settings and hence motivate us to consider local linear methods for our research.

Local polynomial fitting was first used as a time series smoothing technique in the 1930’s. However, Stone (1977) considered local linear regression in the context of nonparametric regression. Cleveland (1979) set the stage for $n^{th}$ degree local polynomial regression by creating *lowess*, a local polynomial smoother with robustness weights. Cleveland’s procedure is a combination of a nearest neighbor method, which defines the local window by the $r$ nearest $x_i$, where $r$ is some pre-specified fraction of the sample size, and a kernel method where the observations in the nearest neighbor window receive kernel weights. (Recently, Assaid and Birch (2000) have proposed robust local linear regression, a kernel only version of lowess with automatic bandwidth selection.)

A local polynomial smoother uses weighted least squares (WLS) to fit a polynomial to the data centered at $x_0$, where the weights are determined by the kernel. The weighted sum
of squares that is minimized is given by

$$WSS(\theta) = \sum_{i=1}^{n} K_h(x_i - x) (Y_i - \theta_0 - \theta_1(x_i - x) - \ldots - \theta_p(x_i - x)^p)^2$$

$$= (Y - \tilde{X}_0 \theta)' W_0 (Y - \tilde{X}_0 \theta), \quad (57)$$

where $W_0$ is the diagonal matrix $\langle K_h(x_i - x) \rangle$ and

$$\tilde{X}_0 = \begin{bmatrix}
1 & x_1 - x & \cdots & (x_1 - x)^p \\
1 & x_2 - x & \cdots & (x_2 - x)^p \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_n - x & \cdots & (x_n - x)^p
\end{bmatrix}. \quad (58)$$

The subscript of $\tilde{X}_0$ indicates that $\tilde{X}_0$ is a model matrix for nonparametric estimation, and the tilde indicates that the regressors in the matrix are centered about the prediction location. The $i^{th}$ row of $\tilde{X}_0$ will be denoted by $\tilde{x}_{0i}'$. The centering is not necessary but is commonly done in the nonparametric regression setting.

This optimization problem leads to LEEs with

$$Q(\theta; y) = \tilde{X}_0' W_0 (y - \tilde{X}_0 \theta) = \sum K_h(x_i - x) \tilde{x}_{0i} (y_i - \tilde{x}_{0i}' \theta) = \sum_{i=1}^{n} w_i(x_i, x) Q_i(\theta; y_i) = 0,$$

which give the WLS estimator,

$$\hat{\theta}(x) = (\tilde{X}_0' W_0 \tilde{X}_0)^{-1} \tilde{X}_0' W_0 Y, \quad (59)$$

of the coefficients of the local polynomial model at $x$. Since the $x_i$ have been centered (about $x$), the estimate of the mean function, $m(x)$, is given by the intercept estimate, $\hat{\theta}_0$. The estimate for $m(x)$ may be written in matrix notation if we define a vector, $S_0$, of indicators $[1 \ 0 \ \cdots \ 0]'$ such that a 1 appears only in the first row. $S_0$ is used to “select” the estimate of $\theta_0$ from the weighted least squares solution. Thus

$$\hat{m}(x) = S_0' (\tilde{X}_0' W_0 \tilde{X}_0)^{-1} \tilde{X}_0' W_0 Y. \quad (60)$$

Since the parameters in the local polynomial fit coincide with the coefficients of the terms in the Taylor series expansion (56), it is also easy to see that

$$d! S_d' (\tilde{X}_0' W_0 \tilde{X}_0)^{-1} \tilde{X}_0' W_0 Y$$
is an estimate of the $d^{th}$ derivative of $m$ at $x$. (See Jones, 1994, for a comparison of this approach with computing $\hat{m}^{(d)}(x)$)

### 3.3.2 Asymptotic Bias, Variance, and Optimal Bandwidth for Local Linear Regression

Suppose that $x_0$ is in the interior of the $x$ space and $m'''$ exists and is continuous. Assume that the arrangement of $x$ values (the design) is fixed. Let $K$ have compact support and let $h = h_n$ be a sequence of bandwidths such that $h \to 0$ and $nh \to \infty$ as $n \to \infty$. Also let $R(K) = \int_{-\infty}^{\infty} K^2(u)du$.

Fan (1992) showed for local linear regression that

\[
\text{Bias}[\hat{m}(x)] = \frac{1}{2} h^2 m''(x) + o(h^2) + O(n^{-1})
\]

\[
\text{Var}[\hat{m}(x)] = \frac{R(K)\sigma^2(x)}{nh} + o((nh)^{-1}),
\]

While these results assume a fixed design, several authors (Fan, 1992, Ruppert and Wand, 1994, Wand and Jones, 1995, Fan and Gijbels, 1996) consider a more general setting of a random design, where $x_i$ are considered realizations of a design density, $f_X$. Fan (1992) showed that incorporating $f_X$ has the advantage of allowing the bandwidth selection (and hence the estimation) to adapt to the distribution of the $X_i$ in the $x$ space. For a random design, the $MSE$ conditioned on $X_i$ for local linear regression is given by

\[
MSE[\hat{m}(x)|X_1, X_2, \ldots, X_n] = \frac{1}{4} h^4 [m''(x)]^2 + \frac{R(K)\sigma^2(x)}{nhf_X(x)} + o_p(h^4 + (nh)^{-1}).
\]

Fan considered the weighted MISE criterion, $WMISE$, for assessing global fit. Ruppert and Wand (1994), among others, used the design density as the weighting function, $w(x)$, which causes the criterion to place more emphasis on regions of the design that have more density (more x’s). The WMISE for local linear regression is simply

\[
WMISE[\hat{m}(x)] = \frac{1}{4} h^4 \int [m''(x)]^2 f_X(x)dx + \frac{R(K)}{nh} \int \sigma^2(x)dx + o_p(h^4 + (nh)^{-1}).
\]

Minimization of $WMISE$ leads to the asymptotically optimal bandwidth for local linear regression,

\[
h_{opt} = \left( \frac{R(K) \int \sigma^2(x)dx}{\int [m''(x)]^2 f_X(x)dx} \right)^{1/5} n^{-1/5}
\]
This involves unknown terms which can be estimated. See Ruppert, Sheather, and Wand (1995) for details.

3.4 Local Polynomial Generalized Linear Models

3.4.1 Local Polynomial Likelihood

The models of this section involve re-expressing estimating equations (50) as local estimating equations (51) with $Q_i$ depending on $\theta(x) = [\theta_0 \theta_1 \cdots \theta_p]$ only through $\eta_0(x_i) = \theta_0 + \theta_1(x_i - x) + \cdots + \theta_p(x_i - x)^p$.\(^2\) We shall consider local polynomial likelihood estimation for generalized linear models (LPGLIM), which maximizes the local log-likelihood

$$
\ell_K(\theta(x)|Y) = \sum_{i=1}^n K_h(x_i - x)\ell(\theta(x)|Y) = \sum_{i=1}^n K_h(x_i - x)\left\{\left[y_i \vartheta(F(\eta_0(x_i))) - b(\vartheta(F(\eta_0(x_i))))\right]/\psi + c(y_i, \psi)\right\}, \tag{66}
$$

where the linear score argument, $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, of $F(\cdot)$ (for multiple regressors) has been replaced with the local linear score, $\eta_0(x_i)$ (for a single regressor). This was first introduced in the context of local linear regression with a uniform kernel by Tibshirani and Hastie (1987). It was extended by Fan, Heckman, and Wand (1995) to local polynomial regression with an arbitrary kernel for generalized linear models and by Fan, Farmen, and Gijbels (1998) to local polynomial regression for arbitrary likelihood-based regression problems.

The LPGLIM model will serve as the basis for a GLIM extension of model robust regression. The properties of this mixed estimate will depend on the properties of each component. The properties of the GLIM prediction were covered in §2.1. Now we consider the properties of the nonparametric estimate.

3.4.2 Asymptotic Bias, Variance, and Optimal Bandwidth for Local Linear Generalized Linear Models (LLGLIM)

Let $h = h_n$ be a sequence of bandwidths such that as $h \to 0$, $nh^3 \to \infty$ as $n \to \infty$. Fan, Heckman, and Wand (1995) showed that the asymptotic bias and variance of the local linear

\(^2\)Note that since $\theta(x)$ is estimated separately at each $x$, the parameters $\theta_0, \ldots, \theta_p$ depend on $x$.\]
predictor, $\eta_0$, are
\begin{align*}
  \text{Bias}_A[\hat{\eta}_0(x)] &= \frac{1}{2} h^2 \eta_0''(x) \quad (67) \\
  \text{Var}_A[\hat{\eta}_0(x)] &= \frac{R(K)\sigma^2(x)}{nhf_X(x)f^2(\eta_0(x))}, \quad (68)
\end{align*}
where $\sigma^2(x) = \psi h(F(\eta_0(x)))$ and $f(\eta_0(x)) = dF(\eta_0(x))/d\eta_0(x)$. These hold for both the interior and boundary design points as did the asymptotic bias and variance expressions for local polynomial regression. Fan et al. (1995) showed that $\hat{\eta}_0(x)$ converges in law to a standard normal:
\begin{equation}
  \mathcal{L} \left[ \frac{\hat{\eta}_0 - \text{Bias}_A[\hat{\eta}_0(x)] - \eta_0(x)}{\sqrt{\text{Var}_A[\hat{\eta}_0(x)]}} \right] \rightarrow N(0, 1), \quad (69)
\end{equation}
or
\begin{equation}
  \hat{\eta}_0(x) - \text{Bias}_A[\hat{\eta}_0(x)] \sim A N(\eta_0(x), \text{Var}_A[\hat{\eta}_0(x)]). \quad (70)
\end{equation}
This gives asymptotic confidence intervals,
\begin{equation*}
  \hat{\eta}_0(x) - \text{Bias}_A[\hat{\eta}_0(x)] \pm Z_{1-\alpha/2} \sqrt{\text{Var}_A[\hat{\eta}_0(x)]},
\end{equation*}
for $\eta_0(x)$. The asymptotic confidence intervals for $m(x)$ are
\begin{equation}
  F \left( \frac{\hat{\eta}_0(x) - \text{Bias}_A[\hat{\eta}_0(x)] \pm Z_{1-\alpha/2} \sqrt{\text{Var}_A[\hat{\eta}_0(x)]}}{\sqrt{\text{Var}_A[\hat{\eta}_0(x)]}} \right), \quad (71)
\end{equation}
which are obtained by transforming the limits on $\eta_0(x)$. In practice, $\text{Bias}_A[\hat{m}(x)]$ and $\text{Var}_A[\hat{m}(x)]$ have to be estimated.

Fan et al. (1995) also gave the asymptotic results for $\hat{m}(x)$. These are similar to (67), (68) and (69) with adjustment by the factor $F^{-1'}(m(x)) = 1/F'(\eta_0(x)) = 1/f(\eta_0(x))$. The asymptotic bias and variance of $\hat{m}(x)$ are
\begin{align*}
  \text{Bias}_A[\hat{m}(x)] &= \frac{1}{2} h^2 \eta_0'' f(\eta_0(x)) \\
  \text{Var}_A[\hat{m}(x)] &= \frac{R(K)\sigma^2(x)}{nhf_X(x)}. \\
\end{align*}
$\hat{m}(x)$ converges in law to a standard normal, which implies
\begin{equation*}
  \hat{m}(x) - \text{Bias}_A[\hat{m}(x)] \sim A N(m(x), \text{Var}_A[\hat{m}(x)]).
\end{equation*}
The WMISE of $\hat{\eta}_0(x)$ is given by
\[
WMISE = \frac{1}{4} h^4 \int [\eta_0''(x)]^2 f_X(x) dx + \frac{R(K)}{nh} \int \frac{\sigma^2(x)}{f(\eta_0(x))} dx,
\]
which when minimized, gives the asymptotically optimal bandwidth,
\[
h_{opt} = \left( \frac{R(K) \int \sigma^2(x) f^{-1}(\eta_0(x)) dx}{\int [\eta_0''(x)]^2 f_X(x) dx} \right)^{1/5} n^{-1/5}.
\]
for local linear generalized linear models.

### 3.4.3 Local Linear Generalized Linear Models Estimation

The local log-likelihood (66) is the conditional local log-likelihood if the $x_i$ are random. Hence maximization of (66) does not depend on the design density $f_X$, assuming that the bandwidth is known a priori. Fan et al. (1995) maximize (66) by applying the Newton-Raphson algorithm after computing a crude estimate of $h$ and assuming this estimate is fixed. As shown in §3.5.1, estimation of $m(x)$ using (66) is equivalent to estimating a GLIM centered at $x$, with weighted variances. Thus our approach will be to apply the IRLS algorithm for GLIM centered at $x$ with fixed $h$. Maximization of (66) to simultaneously estimate $h$ and $\eta_0$ is not a viable option because this will always send $h$ to 0.

Notice that for fixed $x$ and fixed $h$, (66) is the same form as (4) but with constant weights. These weights carry over to the estimating equations, (7), which are a special case of kernel weighted quasi-likelihood estimating equations to be given in the next section.

### Estimates of $h$

Fan et al. (1995) discourage cross-validation bandwidth (Härdle and Marron, 1985) from being used because it has been shown to have large sampling variability. They recommend plugging in estimates of the terms in (72), a method called plug-in in other bandwidth selection settings. How these estimates are obtained is still an open problem. Fan et al. (1995) suggest a somewhat crude plug-in method that involves fitting a higher degree polynomial to obtain a more precise estimate of $\eta_0(x)$, then evaluating the terms in (72) using this estimate. If the design is random, the design density, $f_X$, will also have to be estimated using some density estimation method (which also requires its own bandwidth selection method).
Suffice it to say, bandwidth selection is a difficult problem that may involve very computer intensive methods.

3.5 Local Polynomial Generalized Estimating Equations

3.5.1 Quasi-likelihood as a Classical Framework for Local Estimation

It will be helpful to place the LEEs in a classical framework so that some intuition can be gained about the effect of the localizing weights on the model of the data. Let \( L_K(\theta(x)|Y) = \text{Exp}[\ell_K(\theta(x))] \) be the local likelihood function of \( \theta(x) \) and \( L(\theta(x)|Y) = \text{Exp}[\ell(\theta(x))] \) be the likelihood function of \( \theta(x) \). We shall cast the local log-likelihood as a pseudo log-likelihood of a different model:

\[
L_K(\theta(x)|Y) = \text{Exp}\left\{ \frac{1}{n} \sum_{i=1}^{n} \log L(\theta(x)|Y) K_h(x_i - x) \right\}
\]

While it is not clear if a probability model exists that has this likelihood, it is clear that the kernel weights act on the dispersion of \( Y_i \). Thus (73) corresponds to a quasi-likelihood model with mean \( \mu_0 = F(\eta_0(x)) = F(\theta_0) \) and variance function \( \phi^2 \psi a(\mu_0)/K_h(x_i - x) \). In this perspective, the kernel weights diminish the effect of observations far from \( x \) on the estimates by inflating the variance of the corresponding observations.

Recall that \( Q^\ell \) denotes log quasi-likelihood. Fan et al. (1995) also introduced local polynomial quasi-likelihood estimation, which maximizes the local log quasi-likelihood

\[
Q^\ell_K(\theta(x)|Y) = \sum_{i=1}^{n} K_h(x_i - x) Q^\ell_i(\theta(x)|Y)
\]

over \( \theta \), where \( Q^\ell_i \) is the log quasi-likelihood for observation \( i \) and \( h \) is fixed. If the local mean function is given by \( F(\eta_0(x)) \) and the variance function is \( \phi^2 a(m(x)) \) then (74) leads to the LEEs

\[
Q(\theta; Y) = \frac{\partial Q^\ell_K(\theta(x)|Y)}{\partial \theta} = \sum_{i=1}^{n} K_h(x_i - x) \left[ \tilde{x}_i \phi^{-2} a(F(\tilde{x}_i \theta)) \right] (y_i - F(\tilde{x}_i \theta))
\]
\[
\sum_{i=1}^{n} \left[ \hat{x}_{0i} f_{i} \phi^{-2} a(F(\hat{x}_{0i}')) K_{h}(x_{i} - x) \right] (y_{i} - F(\hat{x}_{0i}'))
\]
\[
= \hat{X}'_{0} (f_{i}) V_{0}^{-1} (y - F(\hat{X}_{0} \theta)) = 0,
\]
(75)

where \( V_{0} = \langle \phi^{2} a(F(\hat{x}_{0}' \theta))/K_{h}(x_{i} - x) \rangle \) and \( \hat{x}_{0i}' \) is the \( i^{th} \) row of \( \hat{X}_{0} \). These LEEs have exactly the same form as the EEs (32) for quasi-likelihood. As long as the bandwidth is fixed, local polynomial quasi-likelihood estimation is quasi-likelihood estimation at \( x_{0} \) with variance function \( \phi^{2} a(F(\hat{x}_{0}' \theta))/K_{h}(x_{i} - x) \). Thus quasi-likelihood will be the framework in which we imbed local estimation. We will see that this framework, when applied to GEEs, is more useful for our purpose than the LEE framework of Carroll, Ruppert, and Welsh (1998) (§3.2.1).

3.5.2 The Local Polynomial GEE (LPGEE) Model

Let \( \eta_{0}(x_{ij}) = \theta_{0} + \theta_{1}(x_{ij} - x) + \cdots + \theta_{p}(x_{ij} - x)^{p} \) be the local linear score. New issues come about in local estimation on clustered data. Should the same smoothing weight be applied to an entire cluster as in the LEE framework of Carroll, Ruppert, and Welsh (1998) or should smoothing weights be applied within the clusters? Where should the kernel weights be placed in the EEs? How should the estimation account for the correlation within clusters on a local scale? We address these questions in this section.

The responses within a cluster should have the same smoothing weight if the GEE covariate is cluster-specific. However, if the covariate is observation-specific, then the weights should be allowed to vary across the observations in each cluster. Since the former weighting scheme is a special case of the latter weighting scheme, we will always use the latter. Hence the LEEs of Carroll et al. (1998) are not always suitable. Lin and Carroll (2000) have recently extended generalized estimating equations to the local estimation setting. They consider two possible kernel weighting schemes of the GEEs:

\[
\sum_{i=1}^{s} \bar{X}'_{0i} (f_{0ij}) V^{-1}_{0i} K_{hi} (Y_{i} - F(\hat{X}_{0i} \theta)) = 0
\]
(76)

and

\[
\sum_{i=1}^{s} \bar{X}'_{0i} (f_{0ij}) K^{1/2}_{hi} V^{-1}_{0i} K^{1/2}_{hi} (Y_{i} - F(\hat{X}_{0i} \theta)) = 0,
\]
(77)
where $K_{hi} = \langle K_h(x_{ij} - x) \rangle$ is the matrix of kernel weights for cluster $i$, $V_{0i}$ is the variance-covariance matrix of $Y$ as a function of the local mean, $f_{0ij} = f(\hat{x}_{0ij}^\theta)$, and $\hat{X}_{0i}$ is the centered nonparametric model $X$ matrix (like (58) in §3.3) for cluster $i$. (76), due to Severini and Staniswalis (1994, see §4.4.1), localizes the estimation in the traditional way by downweighting the residuals $(Y_i - \mu_i)$ that are far from $x$. (77) localizes the estimation by weighting $V_{0i}^{-1}$ symmetrically. Lin and Carroll (2000) proposed (77) because the asymptotic theory for it was much simpler than that of (76). If we imbed these EEs into the quasi-likelihood framework, we can see why. Estimating equations (76) correspond to a quasi-likelihood model that has an \textit{asymmetric} variance-covariance matrix. Thus (76) is the estimating equation of a vacuous model. The quasi-likelihood model perspective of (77) is much more favorable. Taking $K_{hi}^{-1/2}V_{0i}K_{hi}^{-1/2}$ as the variance-covariance matrix, we see that (77) corresponds to a quasi-likelihood model with a \textit{symmetric} variance covariance matrix. Furthermore, since correlation is scale invariant, the symmetric weighting preserves the correlations in each cluster. This is consistent with the clustered correlated structure of the data and is our method of choice for local GEE estimation.

### 3.5.3 Asymptotic Bias and Variance and optimal $R$ of the Local Linear GEE (LLGEE)

Let $h = h_s$ be a sequence of bandwidths such that as $h \to 0$, $sh \to \infty$ as $s \to \infty$. Lin and Carroll (2000) showed that the asymptotic bias and variance of the local linear predictor, $\eta_0$, at $x$ are (with a correction)

$$
Bias_A[\hat{\eta}_0(x)] = \frac{1}{2}h^2\eta''_0(x),
$$

$$
Var_A[\hat{\eta}_0(x)] = \frac{R(K)}{shf^2(\eta_0(x))} \frac{\sum_{k=1}^{n} v_{kk}^2 a(F(\eta_0(x)))f_{X_{ij}}(x)}{\left(\sum_{k=1}^{n} v_{kk}f_{X_{ij}}(x)\right)^2},
$$

where $\forall n_i = n$, $v_{kk}$ is the $k^{th}$ diagonal element of $V_{0i}^{-1}$, and $f_{X_{ij}}$ is the marginal density of $X_{ij}$, assuming a random design ($x_{ij}$ is a realization of $X_{ij}$ which are independent across clusters). A comparison of (78) and (79) with (67) and (68), respectively, illustrates that the LLGEE and the LLGLIM estimates of the mean have the same asymptotic bias, but the LLGEE estimates have a more complicated asymptotic variance. The asymptotic variance
involves the diagonal elements of $V_{oi}^{-1}$ and also $V_{oi}$. It is through $v_{kk}$, $k = 1, \ldots, n$, that the covariance parameters of $V_{oi}$ come into play. Finally note that the asymptotic variance simplifies to (68) if $R = I$. This is because LPGLIM is a special case of LPGEE.

The optimal working correlation matrix in the sense of minimizing the asymptotic variance of the local linear GEE estimate turns out to be $R = I$. This means that locally (in the sense of $h \rightarrow 0$), the correlation structure is not needed and results in a less efficient estimate if used. This result is the opposite of what happens when the correct or nearly correct $R$ is chosen for global GEE estimation.

### 3.5.4 Local Linear GEE Estimation

Lin and Carroll (2000) apply the Fisher scoring method (see §2.1.2) to derive an iterated least squares algorithm to solve EEs (77) and to obtain local polynomial parameter estimates $\hat{\theta}_{LGEE}$. This can be done using steps similar to those in §2.1.2 for GLIM estimates. For fixed bandwidth, the LPGEE IRLS algorithm is simply the IRLS algorithm of the global GEE model (§2.2.4) centered at $x$ with variance-covariance matrix $V^*_0 = K_{hi}^{-1/2} V_{oi} K_{hi}^{-1/2}$ for cluster $i$. This is a consequence of the quasi-likelihood framework for this model.

Since the data are centered, the locally estimated GEE yields an estimate of $\eta_0$ at $x$ that is simply $\hat{\theta}_0$, denoted by $\hat{\theta}_{0GEE}$. Thus, the prediction of $m(x)$ is obtained by $\hat{m}(x) = F(\hat{\theta}_{0GEE})$. Lin and Carroll (2000) show that a robust sandwich estimator can be used to estimate variance of the prediction. Let $\hat{V}_{0GEE} =$

$$
\left( \sum_{i=1}^s \tilde{X}'_{0i} \hat{W}_{0i} \tilde{X}_{0i} \right)^{-1} \left( \sum_{i=1}^s \tilde{X}'_{0i} \hat{W}_{0i} \langle \hat{f}_{0ij} \rangle^{-1} \hat{V}ar[Y_i] \langle \hat{f}_{0ij} \rangle^{-1} \hat{W}_{0i} \tilde{X}_{0i} \right)^{-1} \left( \sum_{i=1}^s \tilde{X}'_{0i} \hat{W}_{0i} \tilde{X}_{0i} \right)^{-1},
$$

(80)

where $\hat{f}_{0ij} = f(\tilde{x}'_{0ij} \hat{\theta}_{LGEE})$, $\hat{W}_{0i} = \langle \hat{f}_{0ij} \rangle \hat{V}_0^{*-1} \langle \hat{f}_{0ij} \rangle$, and $\hat{V}ar[Y_i] = (Y - F(\tilde{X}_0 \hat{\theta}_{LGEE})) (Y - F(\tilde{X}_0 \hat{\theta}_{LGEE}))'$. Then $\hat{V}_{0GEE}$ is a sandwich variance estimate for $\theta$, and $\hat{V}ar[\hat{\eta}_0(x)] = S_0' \hat{V}_{0GEE} S_0$. Notice that this estimator agrees with the estimator based on (41) when LPGEE is placed in the quasi-likelihood framework. The variance of the prediction $\hat{m}(x)$ is approximately $S_0' \hat{V}_{0GEE} S_0 f^2(\hat{\theta}_{0GEE})$, obtained from a Taylor series expansion of $F(\hat{\theta}_{0GEE})$ about $\theta_{0GEE}$.
Estimates of $h$

Lin and Carroll (2000) discuss two methods of estimating the bandwidth. The first is cross-validation involving a prediction error sum of squares (PRESS; Allen, 1974) in the form of a Pearson-like statistic,

$$CV(h) = \sum_{i=1}^{s} \sum_{j=1}^{n_i} \frac{(Y_{ij} - m_{-i}(x_{ij}))^2}{\phi^2 \sigma(m_{-i}(x_{ij}))},$$

where $m_{-i}(x_{ij}) = F(\hat{\eta}_{ij}(x_{ij})) = F(\hat{\theta}_{0GEE})$ is estimated with the $i^{th}$ cluster removed. The $CV$ is minimized by searching over $h$—a very calculation intensive procedure because the local model must be fit $s$ times at each $h$ and $x$. Consequently, Lin and Carroll (2000) suggest using the empirical bias bandwidth selection (EBBS) method of Ruppert (1997). This involves empirical estimation of $MSE(x)$ for $\eta_0$. The empirical variance term is estimated by the sandwich method given above and the empirical bias is estimated by regressing $\eta_0$ on $h$ with the polynomial $\alpha_0 + \alpha_1 h^2 + \alpha_2 h^3 + \cdots + \alpha_k h^{k+1}$ for some pre-specified value of $k$ ($k = 2$ is suggested). The form of this relationship comes from the asymptotic bias expansion (78). The empirical MSE is minimized for each $x$ over $h$. 

48
4 Semiparametric Estimation

The mean model’s flexibility in global estimation, as we have defined it, is only through its parameters, while local estimation allows mean model $m$’s flexibility to be bounded only by smoothness constraints on the derivatives. Global estimation is appropriate if the parameters of the correctly specified model are all that is unknown about $m$ and local estimation is appropriate when the mean function is completely unknown. In this section we consider a situation that lies between these two extremes. We imagine that a model builder specifies a parametric model that he or she believes may only be partly correct. Then hybridization of global (parametric) and local (nonparametric) models is used to formulate semiparametric mean models that draw on the strengths of global and local regression for estimation in this setting.

We consider two classes of semiparametric regression methods, each with somewhat different motivations. The first is the partial linear models. Let the set of covariates for regression be partitioned into two matrices $X$ and $Z$. Suppose that a user is able to specify that covariates in $X$ relate to the mean linearly, but the relationship of the covariates in $Z$ with the mean is unknown. A partial linear model combines a linear model, specified for covariates in $X$, and nonparametric model, for the functional relationship of the covariates in $Z$, with the mean as follows.

$$ Y = X\beta + g(Z) + \epsilon, \quad (81) $$

where $g$ is some unknown function. The user can only specify part of the regression model, but knows which regressors are linearly related to the mean. For example, in an analysis of variance setting, a linear model may emerge as a consequence of a carefully designed experiment. However, if useful axillary information is taken on the experiment as a covariate, ANCOVA will be appropriate only in the case where the covariate is linearly related to the response. A partial linear model would apply if the relationship of the covariate to the adjusted mean (adjusted by the design effects) is unknown. The covariate would be placed in the $Z$ matrix of the partial linear model.

Estimation methods for partial linear models include splines, penalized least squares, penalized quasi-likelihood, and kernel weighting. Estimation with kernel weighting was in-
introduced by Speckman (1988). He used partial residuals to simultaneously estimate the parametric and nonparametric components of the model. His estimates take on the form

\[
\hat{\beta}_p = (X^*X^*)^{-1}X^*Y^*
\]

\[
\hat{g}_p = H_{KER}^*(Y - X\hat{\beta}_p),
\]

where \(H_{KER}^*\) is a kernel smoother matrix \([K_h(z_j - z_i)/\sum_{k=1}^n K_h(z_k - z_i)]_{n \times n}\) such that the kernel fit at \(z_i\), \(\hat{y}_{KER}^*(z_i)\), is the \(i^{th}\) row of \(H_{KER}^*Y\), \(X^* = (I - H_{KER}^*)\), and \(Y^* = (I - H_{KER}^*)Y\).

Then \(E[Y] = X\hat{\beta}_p + \hat{g}_p\) represents the estimated means from the partial linear model.

The splines and penalized methods (see Heckman, 1986, and Green, Jennison, and Seheult, 1985) involve a different concept of local estimation than we have defined here. Rather than localizing a parametric global model with kernel weights, these methods penalize a nonparametric global model for roughness of the fit.

The second class of semiparametric models we will consider are of the model robust regression (MRR) (Mays, Birch, and Starnes, 2001) model types. These involve fitting a global model, say \(X\beta\), and a nonparametric model, \(m\), to the same set of regressors, and combining these fits together. Model robust regression assumes that the model builder has in mind a parametric global model for the data, but is uncertain if it is entirely correct. This is different from the approach of partial linear models since model robust regression does not assume that the user knows what part of the model can be specified correctly. Indeed the user’s entire model may be incorrect. Model robust methods are therefore an approach to a more arbitrary problem of misspecification than partial linear regression.

One may argue that nonparametric regression alone eliminates mean model misspecification and thus MRR methods are not necessary. However, the mean model produced by nonparametric regression is almost completely determined by the data and consequently for small samples it can have relatively large estimated response MSE, due to the relatively large contribution from the variance of the estimated mean response. If the user has a reasonable model in mind, perhaps from subject matter, that information should be used in the model building process to reduce the MSE and thereby produce estimates that are closer to the true mean function more often. MRR allows this possibility while nonparametric regression ignores such model information.
4.1 Model Robust Regression

We now summarize MRR and its extensions.

Einsporn (1987) and Einsporn and Birch (1993a) developed a procedure called HATLINK for combining nonparametric and parametric simple linear regression weights for the purpose of improving prediction over either method alone. They show that a convex combination of the predictions from simple linear regression and kernel regression can be obtained via a convex combination of the weight matrices as follows.

\[ H(\lambda) = \lambda H^{KER} + (1 - \lambda) H^{OLS}, \]

(83)

where \( H^{KER} \) is kernel smoother matrix \( [K_h(x_j - x_i)/\sum_{k=1}^{n} K_h(x_k - x_i)]_{n \times n} \) and \( H^{OLS} \) is the ordinary least squares projection (HAT) matrix. The mixing parameter \( \lambda \) controls the portion of smoothing allowed by the ordinary least squares (OLS) and kernel weights. When \( \lambda = 1 \) the model is completely nonparametric, when \( \lambda = 0 \) the model is purely parametric, and for \( 0 < \lambda < 1 \) the model is semiparametric. Estimates of \( \lambda \) give indication of the degree of model misspecification. Einsporn and Birch (1993b) have shown through examples and simulations that the empirical MSE of the HATLINK can outperform that of either estimate alone.

4.1.1 PLR, MRR1, and MRR2

Mays et al. (2001) extended the HATLINK procedure by replacing kernel regression with local linear regression. They also improved the bandwidth and mixing parameter estimation. They named the procedure model robust regression 1 (MRR1), to reflect that this hybrid estimator provides robustness against model misspecification for a properly chosen \( \lambda \). A MRR1 prediction is simply

\[ \hat{y}^{MRR1}(x) = \lambda \hat{y}^{LLR}(x) + (1 - \lambda) \hat{y}^{OLS}(x), \quad 0 \leq \lambda \leq 1, \]

(84)

where \( \hat{y}^{LLR}(x) \) is the LLR prediction at \( x \) and \( \hat{y}^{OLS}(x) \) is the simple linear regression prediction at \( x \).

Mays and Birch (1998) also considered two other variations of MRR, one that uses the partial linear model of Speckman (1988) with \( X = Z \) called PLR and one called model robust
regression 2 (MRR2) that uses the resmoothing philosophy of Tukey (1977). In MRR2 the
residuals from a parametric fit are smoothed and then mixed back into the fit \( \hat{y}^{OLS} \) as follows.

\[
\hat{y}^{MRR2}(x) = \hat{y}^{OLS}(x) + \lambda h^{LLR}_0(y - \hat{y}^{OLS}), \quad 0 \leq \lambda \leq 1,
\]  

(85)

where \( h^{LLR}_0 = S'_0(\tilde{X}'_0W_0\tilde{X}_0)^{-1}\tilde{X}'_0W_0 \) are the local linear smoothing weights at \( x \) (see §3.3). The parameter \( \lambda \) here determines the proportion of the smoothed OLS residuals added back into the OLS fit. We think of it as the degree to which the parametric model can be improved by using MRR2. Thus, like MRR1, larger \( \lambda \) implies more evidence of model misspecification.

Mays et al. (2001) argue that MRR2 is better than MRR1 on the basis of simulation studies which show that MRR2 can achieve smaller MISE and smaller model degrees of freedom (\( df_{model} \)) than ordinary least squares, local linear regression, MRR1, and PLR for small sample sizes. See Mays (1995) for a detailed simulation study.

4.1.2 Estimation of MRR: \( \lambda \) and \( h \)

Estimation of MRR1 and MRR2 occurs in two stages. First, parametric and nonparametric estimates are obtained by appropriate estimation methods, such as ordinary least squares and local linear regression, respectively. The bandwidth for the nonparametric part is chosen automatically using a cross-validation (PRESS-like) bandwidth selection. Second, the mixing parameter is computed based on these estimates (that is conditional on the estimated bandwidth) either with the asymptotically optimal bandwidth or another cross-validation selector.

Let \( \theta \) be either the bandwidth or the mixing parameter of the model. The form of their proposed cross-validation estimators in Mays et al. (2001) is

\[
PRESS^{**} = \frac{PRESS}{n - tr(H) + (SSE_{max} - SSE_{\theta})/SSE_{max}},
\]

where \( PRESS = \sum_{i=1}^{n}(y_i - \hat{y}_{i,-i})^2 \), \( \hat{y}_{i,-i} \) is the fit at \( x_i \) with \( y_i \) removed, SSE\(_{max} \) is the maximum SSE = \( \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \) over all \( \theta \), and \( H \) is the matrix such that \( \hat{y} = Hy \). For bandwidth selection in local linear regression, Mays et al. (2001) use \( H = H^{LLR} \) and \( \theta = h \), the bandwidth. They also suggested using \( PRESS^{**} \) to estimate the mixing parameter for small samples, where \( H = H^{MRR1} = \lambda H^{LLR} + (1 - \lambda)H^{OLS} \) for MRR1 or \( H = H^{MRR2} = \)
$$H_{OLS} + \lambda H_{LLR}^{OLS}(I - H_{OLS})$$ for MRR2 and $\theta = \lambda$. Note that $H_{LLR}$ for MRR2 may depend on a different bandwidth than $H_{LLR}$ for MRR1 because MRR2 smooths residuals rather than responses. Since the estimate of $\lambda$ is conditional on $h$, other bandwidth selectors, such as the plug-in ideas of Ruppert, Sheather, and Wand (1995) and empirical bias idea of Ruppert (1997), may be applied here to estimate $h$ as well.

Mays et al. (2001) derive the asymptotically optimal mixing parameters, $\lambda_{opt}$, by minimizing the Euclidean distance, $\delta = \sum (\hat{y}^{MRR}(x_i) - m(x_i))^2$, between the vector of true regression function evaluations, $[m(x_i)]$, and the vector of MRR predictions, $[\hat{y}^{MRR}(x_i)]$. This is based on methods of estimation due to Stein (1956) and Burman and Chaudhuri (1992). Let $\langle u_1, u_2 \rangle$ be the dot product between vectors $u_1$ and $u_2$ and let $\|u\|$ be the Euclidean norm of a vector $u$. Then

$$\hat{\lambda}_{opt,MRR1} = \frac{\langle \hat{y}_{LLR}^{OLS} - \hat{y}_{OLS}^{OLS}, y - \hat{y}_{OLS}^{OLS} \rangle}{\|\hat{y}_{LLR}^{OLS} - \hat{y}_{OLS}^{OLS}\|^2}$$

is the estimated asymptotically optimal mixing parameter (Mays et al., 2001). The terms $\hat{y}_{LLR}^{OLS}$ and $\hat{y}_{OLS}^{OLS}$ are the respective fits with observation $i$ deleted. For the MRR2 case the optimal estimated bandwidth is given as

$$\hat{\lambda}_{opt,MRR2} = \frac{\langle h_0^{LLR}(\hat{y} - \hat{y}_{OLS}^{OLS}), y - \hat{y}_{OLS}^{OLS} \rangle}{\|h_0^{LLR}(\hat{y} - \hat{y}_{OLS}^{OLS})\|^2}$$

(Mays et al., 2001). Thus, the estimated optimal mixing parameters are coefficients of the projections of OLS residuals.

### 4.1.3 MRR Convergence Rates

Recent work in model robust regression has focused on the asymptotic convergence rates of the distance $\delta$. Mays et al. (2001), Starnes (1999), and Burman and Chaudhuri (1992) proved that if the parametric mean model is misspecified in form and the asymptotically optimal $\lambda_{opt,MRR1}$ is used, then MRR1 predictions will converge to the mean at the nonparametric rate. However, if the mean model is correct then MRR1 predictions will converge to the mean at the faster parametric rate ($n^{-1/2}$). MRR therefore takes advantage of correct mean model specification and is robust to mean model misspecification, asymptotically. Starnes
(1999) showed that the same rates of convergence for MRR1 apply to MRR2. We will refer to these results as the \textit{golden rule} of model robust regression.

These results do not hold for either MRR1 or MRR2 if PRESS** is used to estimate $\lambda$. However, predictions using PRESS** do perform well, often better than those using $\hat{\lambda}_{\text{optMRR1}}$ or $\hat{\lambda}_{\text{optMRR2}}$ for small sample sizes as simulation studies have shown (Mays et al. 2001).

4.1.4 Extensions of MRR

Other work on model robust regression has been done in the areas of dual models, outlier robust regression, and quantal regression. Dual model regression simultaneously models the mean and variance as a function of regressors. Robinson (1997) and Robinson and Birch (2000) applied the model robust methodology to the mean model, the variance model, and both the mean and variance models. In all cases, model robust techniques were shown to have smaller MISE than the parametric and nonparametric models alone for the specific examples that Robinson (1997) considered.

Assaid (1997) investigated outlier resistant schemes for MRR in the single and multiple regressor cases. He used M-estimation linear regression for the parametric model and robust local linear regression (see Assaid and Birch, 2000) for the nonparametric estimation. Assaid (1997) found similar success as that of MRR1 and MRR2 for robust MRR in this setting. For example, under the conditions of large sample size, extreme model misspecification, and highly (outlier) contaminated model errors, robust MRR had comparable asymptotic MISE with Loess, Cleveland’s (1979) robust scatterplot smoother.

Nottingham (1995) and Nottingham and Birch (2000) applied kernel regression and local linear regression to proportions in the grouped data case (see Nottingham, 1995, chapter 4) of quantal regression, a subclass of generalized linear models for binary responses that includes logistic and probit regression. They mixed these predictions with predictions from global quantal regression in the same manner as in MRR1 and MRR2. Comparisons of asymptotic MISE efficiencies of model robust quantal regression (MRQR, based on MRR1) and local linear regression relative to logistic regression for small sample sizes showed that MRQR can be more MISE efficient than logistic regression when the user’s model is incorrect and more MISE efficient than local linear regression when the user’s model is either correct or
incorrect. Also, small sample simulation results demonstrated that the model robust method can have superior effective dose coverage probabilities as compared to logistic regression and the nonparametrically smoothed estimates alone (Nottingham and Birch, 2000).

We now summarize the benefits of using MRR for both large and small samples and consider the following results as a motivation for applying the MRR methodology to our problem.

**The motivation for using MRR**

- For large samples \((h \rightarrow 0, nh \rightarrow \infty \text{ as } n \rightarrow \infty)\):
  - MRR predictions converge to the true mean regardless of the parametric model functional form specification (correct or not correct).
  - If the form of the parametric mean function is correct, MRR predictions converge to the true mean function at the parametric rate of convergence \((n^{-1/2})\) which is much faster than the nonparametric rate. MRR predictions converge at the nonparametric rate if the parametric mean function form is not correct.

- For small samples:
  - MRR predictions often have smaller bias than parametric predictions and smaller variance than nonparametric predictions.
  - MRR models are more flexible than the nonparametric models alone.

- MRR has been extended to other types of regression with similar results.
  - Dual model regression (Robinson, 1997)
  - Robust regression (Assaid, 1997)
  - Quantal Regression (Nottingham and Birch, 2000)
  - Nonlinear (See Fan and Ullah, 1999 and §4.2)

### 4.2 Methods Related to MRR

Several methods similar to model robust regression have been proposed. Some of the methods come from the econometric literature and a few can be found in statistical journals.
The earliest paper proposing to mix parametric and nonparametric estimates in a convex combination was by Olkin and Spiegelman (1987) in the density estimation setting. They estimated bandwidth first and selected their mixing proportion conditional on the bandwidth. Olkin and Spiegelman (1987) showed that if the parametric density is correct, the convex estimate approaches it at the parametric rate. If it is not correct, the convex estimate approaches the underlying density at the nonparametric rate.

There are several methods that involve fitting parametric and nonparametric models to data for the purpose of testing the parametric model against nonparametric alternatives. Eubank and Spiegelman (1990) considered fitting the model \( y_i = \beta_0 + \beta_1 x_i + f(x_i) + \epsilon_i \) in two ways, one using series methods (which the authors claim perform poorly) and another using spline smoothing of the residuals of the OLS fit. This is akin to Speckman (1988) who focused mostly on kernel smoothing. Eubank and Spiegelman (1990) developed a goodness-of-fit test of \( H_0: f \equiv 0 \) for simple linear regression. A few other methods like these are given by Robinson (1988), Wooldridge (1992), Ullah and Vinod (1993), and Rahman, Gokhale, and Ullah (1997).

Ullah and Vinod (1993) (pg. 86) advocated the model robust methodology by stating that nonparametric regression is useful for “capturing any missing features” of the parametric model. They advised using nonparametric methods on the residuals from a parametric fit—which is what MRR2 does. Ullah and Vinod (1993) (pg. 92) formulated the MRR1 model but used least squares regression to estimate the mixing parameter by estimating the no-intercept regression model \( y_i - x_i' \hat{\beta} = \lambda (\hat{y}_i^{LLR} - x_i' \hat{\beta}) + \epsilon_i \) conditional on a pre-selected bandwidth. The Ullah and Vinod (1993) estimator is similar to \( \hat{\lambda}_{optMRR1} \), except that they placed less emphasis on prediction by not using leave-one-out fits.

Burman and Chaudhuri (1992) developed a hybrid estimator, \( \theta(x_i) = \lambda f(\beta, x_i) + (1 - \lambda) \hat{g}(x_i) \), of the mean function \( m(x_i) \) for \( y_i \), which is similar to MRR1. They placed few restrictions on \( f(\beta, x_i) \) by allowing \( f(\beta, x_i) \) to be nonlinear provided it has certain regularity conditions. The nonparametric estimator, \( \hat{g} \), may be any nonparametric estimator of the mean, provided it can be written as a weighted sum of the observations \( y_i \) such that the weights sum to 1. Kernel, local linear, spline estimators, and their multivariate versions satisfy these requirements. Burman and Chaudhuri (1992) showed that if the parametric
mean model is correct, the hybrid estimator will converge to the true mean at the parametric rate. If the parametric model is not correct, then it will converge to the true mean function at the nonparametric rate. This result was inspired by that of Olkin and Spiegelman (1987).

Burman and Chaudhuri (1992) also considered another situation that is worth mentioning. Rather than assuming the parametric model is correct or not correct, suppose that it is close to the true model. Considering closeness to the true model in an asymptotic sense, we might imagine that the user looks at a plot of the data for every sample size $n$ and improves the parametric model above mean model specifications based on smaller sample sizes. Let the specified models converge to the true mean function at the rate $O_p(\delta_n)$, i.e., the rate that the parametric models’ distance from the true mean converges to 0. Burman and Chaudhuri (1992) show that if $\delta_n$ converges as slow or slower than the nonparametric rate, then the hybrid estimator $\theta(\cdot)$ will converge to the true mean function at the nonparametric rate. Likewise, if the specified parametric model converges to the true model as fast or faster than the parametric rate of convergence, then the hybrid estimator will converge to the true mean at the parametric rate. Finally, and most important, if the specified parametric model converges to the true model at a rate that is between the parametric and the nonparametric rates, then so will the hybrid estimator. These results imply that the hybrid estimator gains (in terms of convergence rate) from the closeness of parametric model to the true mean function (Burman and Chaudhuri, 1992).

Starnes (1999) worked out the same results for MRR1 and MRR2. The hybrid estimator of Burman and Chaudhuri (1992) is equivalent to MRR1 if $\hat{\lambda}_{optMRR1}$ is used as an estimate of the mixing parameter. Starnes (1999) also developed the same convergence rate properties of MRQR, Nottingham’s (1995) model robust version of quantal regression. Starnes (1999) (pg. 107) claims that his asymptotic results can be extended to all generalized linear models.

Fan and Ullah (1999) have extended the results obtained by Burman and Chaudhuri (1992) for a slightly modified version of Burman and Chaudhuri’s hybrid estimator that is based on Ullah and Vinod’s (1993) conditional least squares method of estimating $\lambda$ (see above). Fan and Ullah’s motivation is that their estimator does not require the calculation

---

3This is based on the corrected estimate of Burman and Chaudhuri’s (1992) mixing parameter. Their mistake was pointed out by Starnes (1999) and Rahman et al. (1997).
of leave-one-out fits for the parametric model. They do, however, use leave-one-out fits for the nonparametric part of their $\lambda$ estimator to guard against overfitting, which agrees in principle with Mays et al. (2001) and Burman and Chaudhuri (1992). Fan and Ullah studied a random design and product Nadaraya-Watson kernel regression, an extension of Nadaraya-Watson (1964) kernel regression to the multiple covariate case. Fan and Ullah (1999) show that their hybrid and mixing parameter estimators are asymptotically normal.

Nonparametric smoothing methods typically suffer from a “curse of dimensionality” problem such that fits are poor when several dimensions are involved (Hall, 1989). Interestingly, Fan and Ullah (1999) also showed that the rate of convergence of the hybrid estimator to the true mean function does not depend on the number of covariates (i.e., the number of dimensions) if the parametric model specification is converging to the true mean function faster than the nonparametric rate. Hence, Fan and Ullah’s hybrid estimator is essentially free of the “curse of dimensionality” problem when the specified parametric model is suitably close to the true model. This result is conjectured by Mays et al. (2001) for MRR1 and MRR2 when applied to the multiple covariate case.

A simulation study by Rahman et al. (1997) compared several variations of MRR1-type estimators. Rahman et al. (1997) reviewed existing estimators, including Burman and Chaudhuri (1992), Wooldridge (1992), Ullah and Vinod (1993), and proposed three new estimators, two of which were modifications of Wooldridge’s estimator and one which was a modification of Ullah and Vinod’s estimator. The modifications included incorporating leave-one-out estimation in each component and estimating the mixing parameter with a no-intercept regression as in Ullah and Vinod (1993). Simulation studies in Rahman et al. showed that Burman and Chaudhuri’s (1992) and Ullah and Vinod’s (1993) estimator of the mixing parameter ($\lambda_{\text{optMRR1}}$) had as small or smaller bias and variance overall than the other estimators for the examples considered. Rahman et al. (1997) pointed out that simultaneous estimation of the parametric model and the mixing parameter does not work as well since such estimators fared poorly. Note that this is not the same as simultaneous estimation of the bandwidth and the mixing parameter (see §5.5). Unfortunately, Rahman et al. (1997) did not study estimators MRR1 and MRR2 that use PRESS**.

We have reviewed papers that are the most similar to MRR. Two common themes that
appear when combining parametric and nonparametric estimation are that \( \lambda \) is conditional on the estimate of the nonparametric portion’s bandwidth, and the convergence rates of the methods achieve the golden rule of MRR. There are several other methods that combine parametric and nonparametric components with these themes (see Glad 1998 and Gozalo and Linton 2000). We would like to point out that only a few of these methods emphasize using nonparametric estimation to robustify parametric model specification. That is the aim of model robust regression.

4.3 Semiparametric Generalized Linear Models

To the best of our knowledge, neither MRR nor related methods have been extended to generalized linear models or generalized estimating equations. However, there has been considerable work on models that are partially parametric and partially nonparametric in these areas. Most of the progress has involved extending the partial linear model, which as we pointed out above, does not robustify the entire form of the parametric mean function.

4.3.1 Generalized Partial Linear Models

The partial linear model, \( Y_i = x_i' \beta + g(z_i) + \epsilon_i \) may be extended to generalized linear models by letting the linear predictor, \( \eta_i \), be \( x_i' \beta + \gamma(z_i) \). Thus the generalized partial linear model (GPLM) has mean function \( F(x_i' \beta + \gamma(z_i)) \) and variance function \( \psi(a(\mu)) \).

There has been relatively little work done on GPLM in the specific context of generalized linear models. This is mainly because there are several developments in more general model settings, such as quasi-likelihood and generalized estimating equations of which GLIM is a special case. Interestingly, the first development of GPLM appeared over 16 years ago by Green and Yandell (1985) in the context of spline smoothing. They used penalized (for roughness) maximum likelihood to simultaneously obtain estimates of the linear model portion, \( x_i' \beta \), and the nonparametric model portion, \( \gamma(\cdot) \). A discussion of this approach may be found in Green and Silverman (1994) (pp. 104-114). Another significant contribution by Emond and Self (1997) was based on an efficient score function for semiparametric models defined by Bickel et al. (1993). Emond and Self (1997) worked out an estimation method
for the parametric and nonparametric components of the efficient score equation where the nonparametric part is smoothed with splines.

4.3.2 Generalized Single Index Models

Even with multiple covariates, the generalized linear model has only a single canonical parameter to which the covariates are linked. This is possible because the linear predictor, $x'_i\beta$, is a single index of the covariates. Single index models (SIM) model a linear combination (a linear predictor) of the covariates nonparametrically. The model takes the form $Y = \eta_0(x'_i\beta) + \epsilon$, where $\eta_0$ is an unknown function and the parameter vector $\beta$ is normalized for model identifiability. Single index models were labeled as such by Härdle, Hall, and Ichimura (1993) and are related to projection pursuit regression (Friedman and Stuetzle, 1981). There are four different methods for estimation of SIMs, a popular one being average derivative estimation which is based on kernel smoothing of expected derivatives of the mean function (Schimek, 1997).

In the context of generalized linear models, SIMs are essentially GLIMs with a nonparametric link function (see Weisberg and Welsh, 1994). Unfortunately, the normalization constraint for identifiability of the model only allows estimates of $\beta$ to be known up to the normalization constant. Nevertheless, Carroll et al. (1997) proposed generalized partially linear single index models (GPLSIM) which merge a generalized partial linear model with a single index model. The mean function is of the form $E[Y] = F(x'_i\beta + \eta_0(z'_\alpha))$. As with other partial linear models, the covariates are partitioned into those that are linearly related to $E[Y]$ and those that are related through the nonparametric portion. Note that the mean function contains a link function $F^{-1}$. This function assures that the predictions are properly bounded as in GLIMs. Carroll et al. (1997) find estimates of $\beta$, $\eta_0$, and $\alpha$ by iterating between local linear quasi-likelihood estimation for $\eta_0$ and maximum quasi-likelihood for $\beta$ and $\alpha$.

A primary motivation for single index models is dimension reduction, and to avoid the aforementioned “curse of dimensionality.” Though these methods are useful, they are not so important in the context of model robust regression which is applied to the entire form of the mean function. A single index model is too restricted to fully robustify a parametric
4.3.3 Generalized Additive Models

Generalized additive models (Hastie and Tibshirani, 1986) are another interesting approach to relaxing the mean model assumptions of GLIM. Rather than specifying the mean as a linear combination of the covariates, generalized additive models (GAM) only require that covariates enter the model additively, that is,

\[ E[Y] = F(f_0 + f_1(x_1) + f_2(x_2) + \cdots + f_k(x_k)), \]

where \( f_0 \) is a constant and \( f_j, j = 1, \ldots, k \) are smooth functions of \( x_j, j = 1, \ldots, k \), respectively. GAMs use an estimation algorithm called “backfitting” that iteratively updates each function estimate, \( \hat{f}_j' \), by smoothing the partial residuals from the other \( \hat{f}_j, j \neq j' \). One advantage of this method is that any scatterplot smoother may be used (local polynomial, spline, Lowess, etc.). Hastie and Tibshirani (1990) provide a detailed discussion of GAMs and offer several examples.

GAMs also handle the “curse of dimensionality” problem with success, but these models are often described as being too limited to model arbitrary relationships (due to their additive-only nature, see Ruppert and Wand 1994 and Fan, Heckman, and Wand 1995). Therefore, GAMs would not be as useful for the nonparametric component of MRR as a method like local polynomial regression, which makes weak assumptions about the smoothness of the mean function.

4.4 Semiparametric GEEs

4.4.1 Semiparametric Quasi-Likelihood

A quasi-likelihood formulation of a partial linear model for uncorrelated data is due to Severini and Staniswalis (1994). The model assumes the mean and variance functions

\[
E[Y] = F(x'\beta + \gamma(z')) \\
\text{Var}[Y] = \phi^2 a(E[Y])
\]
hold, where \(a(\cdot)\) is a known variance function and \(F\) is a known inverse link function. Severini and Staniswalis (1994) estimate \(\beta\) and \(\gamma\) using generalized profile likelihood of Severini and Wong (1992), which we will refer to as the profile/kernel method. This method amounts to holding an initial estimate \(\hat{\beta}_0\) of \(\beta\) fixed, while estimating the nonparametric component, \(\gamma\), via a kernel quasi-likelihood regression, then holding the estimate \(\hat{\gamma}\) fixed, while estimating the parametric component via global parametric quasi-likelihood, and finally holding \(\hat{\beta}\) fixed, while estimating \(\gamma\) again. Iteration is not continued beyond this point though we speculate it can be. Severini and Staniswalis (1994) estimate the overdispersion parameter, \(\phi^2\) as

\[
\hat{\phi}^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{(Y_i - F(x'_i\hat{\beta} + \hat{\gamma}(z'_i)))^2}{a(F(x'_i\hat{\beta} + \hat{\gamma}(z'_i))},
\]

which is similar to the MOM Pearson \(\chi^2\) statistic estimator for global parametric quasi-likelihood estimation. \(\hat{\phi}^2\) is updated using the fits from each iteration cycle of the algorithm. Note that their estimate does not correct for the degrees-of-freedom, but simply averages over the Pearson residuals.

Like Robinson (1988), Severini and Staniswalis (1994) study the asymptotic properties of the parametric portion (i.e., \(\hat{\beta}\)) with the motivation being towards inference on the corresponding parameters. They show that the estimates of \(\beta\) are \(\sqrt{n}\)-consistent and asymptotically normal. Severini and Staniswalis (1994) provide an estimator for the variance covariance matrix of \(\hat{\beta}\) that is an estimate of the inverse information matrix for \(\beta\). Furthermore they show that if there is doubt about the variance functional form specification, then a consistent sandwich estimate of the variance can be used.

### 4.4.2 Partial Linear GEE (PLGEE)

Severini and Staniswalis (1994) also considered a generalized estimating equations version of their partial linear model for cluster specific covariates when \(n_i = n, \forall i\). Again, the mean and variance functions were given by \(E[Y] = F(x'\beta + \gamma(z))\) and \(Var[Y] = \phi^2a(E[Y])\), but they allowed for cluster correlation by incorporating a working correlation matrix into the cluster variance-covariance matrix in the usual way, as \(V_i = \phi^2A_i^{1/2}(\mu)R_i(\rho)A_i^{1/2}(\mu)\). They simply applied Severini and Wong’s (1992) profile/kernel methodology to the appropriate estimating equations.
Severini and Staniswalis (1994) obtain initial estimates for $\beta$ and $\gamma$ using their quasi-likelihood profile/kernel algorithm for unclustered data. For estimation at $z = z_{00}$ in the nonparametric part, they hold the estimate of the parametric part, $\hat{\beta}$, fixed. They use

$$\sum_{i=1}^{s} [f_{i1} f_{i2} \cdots f_{in}] V_i^{-1} K_{hi} \left( Y_i - F(X_i \hat{\beta} + [\gamma(z_{ij})]_{n_i \times 1}) \right) = 0, \tag{86}$$

where $K_{hi}$ is a diagonal matrix of kernel weights $K_h(z_i - z_{00})$, to solve for $\gamma$ to obtain the updated estimate $\hat{\gamma}$. This is essentially equivalent to applying the local constant Lin and Carroll (2000) GEEs to a GEE model (see §3.5.2). The linear predictor only has an additional known constant. For the parametric part, they hold the estimate of the nonparametric part, $\hat{\gamma}$ fixed. Then they solve

$$\sum_{i=1}^{s} \mathbf{X}_i' \langle f_{ij} \rangle V_i^{-1} \left( Y_i - F(X_i \beta + [\hat{\gamma}(z_{ij})]_{n_i \times 1}) \right) = 0 \tag{87}$$

for $\beta$ to get an updated estimate of $\beta$. Finally, Severini and Staniswalis (1994) use

$$\hat{\phi}^2 = \frac{1}{ns} \sum_{i=1}^{s} \frac{(Y_{ij} - F(x_{ij}' \hat{\beta} + \hat{\gamma}(z_{ij})))^2}{a(F(x_{ij}' \hat{\beta} + \hat{\gamma}(z_{ij})))},$$

to estimate the scale parameter, $\phi^2$.

Severini and Staniswalis (1994) showed that the parameter estimates $\hat{\beta}$ have an asymptotic normal distribution and gave a sandwich estimator for the variance. Their sandwich estimator is of the same form as $\hat{V}_{\text{GEE}}$ in §2.2.5, where the mean function is replaced with the partial linear estimate.

Lin and Carroll (2001) extend Severini and Staniswalis (1994) PLGEE model, but like Robinson (1988), they also look into the $\sqrt{n}$-consistency of $\hat{\beta}$. Their extension involves using local linear instead of local constant estimation for the nonparametric part and using the wrapped kernel matrix form (77) of the estimating equations, which we favor (see §3.5.2) over Severini and Staniswalis’s (1994) kernel weighting method, (76) and (86). Lin and Carroll (2001) also allow the correlation matrices for the parametric and nonparametric portions to be different. They do this because Lin and Carroll (2000) concluded that the correlation structure should be ignored asymptotically, i.e., $R_i$ should be set to $I$, when computing a local GEE estimate.
Lin and Carroll (2001) show the following. First, the parametric estimates $\hat{\beta}$ are $\sqrt{n}$-consistent only if the correlation structure of the data is ignored in both the nonparametric and parametric profile/kernel estimating equations. Second, even when the true correlation matrix is used for the parametric part, $\hat{\beta}$, is not $\sqrt{n}$-consistent unless undersmoothing is done on the nonparametric part, that is, a smaller than optimal bandwidth for the nonparametric part must be chosen. Third, estimates from the profile/kernel method are not semiparametric efficient, as defined by Bickel et al. (1993).

These negative properties of PLGEE are disturbing. However, they do not apply to the MRR method of estimation because the method of MRR estimates an entirely parametric model and an entirely nonparametric model and then combines the estimates. Also, previous asymptotic results of MRR suggest that MRR applied to GEE would not face such problems.
5 Model Robust Generalized Estimating Equations

We extend the recent work of Nottingham and Birch (2000) and Starnes and Birch (2000) to generalized estimating equations for clustered correlated data by combining parametric (global) GEE estimation and nonparametric (local) GEE estimation. This will also be an extension of similar work done by Fan and Ullah (1999) in nonlinear regression (see §4.2).

In §3, the nonparametric model matrix $\tilde{X}_0$ was centered about the prediction location, while the parametric model matrix in §2 remained uncentered. Though centering the nonparametric model matrix is a common practice, it is not necessary as local estimates may be easily obtained using an uncentered $X_0$ model matrix. Since we prefer to view the nonparametric model in a quasi-likelihood framework (§3.5.1), not centering the model matrix will be convenient in this and the following chapters which strive to integrate the parametric and nonparametric modelling concepts.

5.1 Mixing on Different Scales

There are several possible ways of combining parametric and nonparametric models in the GEE setting. We consider the extension of the MRR1 model to GEE because it is relatively straightforward. There are two essential ways of motivating a particular GLIM or GEE model. One could try to determine the form of the entire mean function, or one could alternatively choose an inverse link function that properly bounds the linear predictor and then try to determine the form of the linear predictor. With this in mind, we feel it is necessary to consider mixing on both the mean scale ($\mu$-scale),

$$\hat{y}^{MRGEEI}(x^*) = (1 - \hat{\lambda})F(x'\hat{\beta}_{GEE}) + \hat{\lambda}F(\hat{\eta}_0),$$

which we call MRGEEI estimation, and the linear predictor scale ($\eta$-scale),

$$\hat{y}^{MRGEEII}(x^*) = F\left((1 - \hat{\lambda})x'\hat{\beta}_{GEE} + \hat{\lambda}\hat{\eta}_0\right),$$

which we call MRGEEII estimation.

The two approaches cater to two views of model building and model interpretation. When one studies the responses, a sense of the mean model functional form can sometimes
be ascertained and a mean-scale model can be built. Or the user may already have in mind a deterministic model developed from previous studies or the subject matter at hand that uses the scale of the data and hence is a mean-scale specified model. Mixing on the mean scale would be appropriate in these contexts as the mixing parameter is interpreted as simply the proportion of adjustment that the given mean model needs due to its misspecification. On the other hand, it is often simpler to think of, and to construct, a model on the linear predictor scale because it is analogous to constructing a multiple regression model for a transformed response. This approach is especially useful for constructing ANOVA and ANCOVA models since effects, especially interactions, can be interpreted in the usual way. Thus, both \( \mu \)-scale and \( \eta \)-scale approaches have advantages and one may be preferred over the other according to one’s needs. It is not clear, however, which of MRGEEI or MRGEEII would perform better. This is an important question that will be addressed.

Two important parameters, the bandwidth (\( h \)) and the mixing parameter (\( \lambda \)), need to be estimated along with the model parameters. The mixing parameter will be estimated conditional on the bandwidth choice. Simultaneous estimation of bandwidth and mixing parameter appears to provide some improvement over conditional estimation (§5.5), but we prefer the former approach because the convex combinations in (88) and (89) contrast the parametric fit with the best possible nonparametric fit.

### 5.2 Bandwidth Selection

#### 5.2.1 Cross-Validation

Finding the optimal bandwidth is a difficult problem in nonparametric regression. Bandwidth estimation for local GEE models in particular is nearly an unexplored topic. As of yet, as discussed in §3.5.4, only Lin and Carroll (2000) have suggested two estimators, one based on Ruppert’s (1997) empirical biased bandwidth and another, more classical estimator, based on cross-validation (CV). This estimator is

\[
CV(h) = \sum_{i=1}^{s} \sum_{j=1}^{n_i} \frac{(Y_{ij} - \hat{y}_{ij, -i})^2}{\hat{\phi}^2 a(\hat{y}_{ij, -i})} = \sum_{i=1}^{s} (Y_i - \hat{y}_{i, -i})' V^{-1} (Y_i - \hat{y}_{i, -i}),
\]  

(90)
where \( \hat{y}_{ij,-i} = F(\hat{\eta}_{ij,-i}) \) is the local GEE estimate at \( x_{ij} \) with the \( i^{th} \) cluster removed. Cross-validation bandwidth selection involves measuring the distance of the deleted fits (leave-one-out) from the data. This approach has been in extensive use. Clark (1975) first used it in the context of kernel smoothing for the purpose of choosing a bandwidth.

Allen (1974) used deleted fits to estimate prediction error sum of squares (PRESS), which assesses the prediction capability of a mean model. Choosing a bandwidth that minimizes PRESS for the purpose of optimizing the predictability of the model is useful, but there is another, more essential reason for using deleted fits. The PRESS objective function is unbiased (up to an additive constant) for an ideal, but unknown, objective function that measures the distance between the fit and the true mean (Härdle, 1990). This unbiasedness is achieved by the leave-one-out approach because the response at a prediction location is uncorrelated with the leave-one-out fit at that location. It turns out that the same holds for the cross-validation objective function for local GEE. However, to achieve unbiasedness, it is necessary that an entire \( \text{cluster} \) be removed. A simple analog of Härdle’s result for the more general CV function in (90) is given by the following theorem.

**Theorem 1** Let \( d(h) = \sum_{i=1}^{s} (m_i - \hat{y}_{LGE}^{LGE})'W_i(m_i - \hat{y}_{LGE}^{LGE}) \) be our ideal objective function and let \( \hat{d}(h) = \sum_{i=1}^{s} (y_i - \hat{y}_{LGE}^{LGE})'W_i(y_i - \hat{y}_{LGE}^{LGE}) \) be the user’s objective function, where \( \hat{y}_{LGE}^{LGE} \) is the vector of nonparametric GEE predictions for cluster \( i \) when the responses for cluster \( i \) are not used in fitting the model and \( W_i \) is a fixed \( n_i \times n_i \) matrix. Then \( \hat{d}(h) \) is unbiased for \( E[d(h)] \) up to an additive constant. That is, \( E[\hat{d}(h)] \) has the same minimum as \( E[d(h)] \).

The proof is found in Appendix A.1.

This theorem implies that

\[
\hat{d}_1(h) = \sum_{i=1}^{s} (y_i - \hat{y}_{LGE}^{LGE})'(y_i - \hat{y}_{LGE}^{LGE})
\]
\[
\hat{d}_2(h) = \sum_{i=1}^{s} (y_i - \hat{y}_{LGE}^{LGE})'(\text{diag}(V))^{-1}(y_i - \hat{y}_{LGE}^{LGE}), \text{ and}
\]
\[
\hat{d}_3(h) = \sum_{i=1}^{s} (y_i - \hat{y}_{LGE}^{LGE})'V^{-1}(y_i - \hat{y}_{LGE}^{LGE}),
\]
are unbiased for

\[
d_1(h) = \sum_{i=1}^{s} (m_i - \hat{y}_{LGE}^{LGE})'(m_i - \hat{y}_{LGE}^{LGE}),
\]

67
\[ d_2(h) = \sum_{i=1}^{s} (m_i - \hat{y}_{i,-i}^{LGEE})' (\text{diag}(V))^{-1} (m_i - \hat{y}_{i,-i}^{LGEE}), \quad \text{and} \]
\[ d_3(h) = \sum_{i=1}^{s} (m_i - \hat{y}_{i,-i}^{LGEE})' V^{-1} (m_i - \hat{y}_{i,-i}^{LGEE}), \]
respectively, up to an additive constant.

These objective functions optimize the fit with different goals in mind. The objective function \( d_1 \) seeks to make the fit as close as possible to the true mean function uniformly throughout the data. The function \( d_2 \), used by Lin and Carroll (2000), has the same objective, but places emphasis on regions where the variability of the response is less. The last function, \( d_3 \), has the same objectives as the first two, but also takes into account the covariation of observations within a cluster. For example, if \( Y_{11} \) and \( Y_{12} \) have a large positive covariance, then it can easily be shown for the \( N = 2 \) case that \( \hat{y}_{11}^{LGEE} \) and \( \hat{y}_{12}^{LGEE} \) will be penalized more when one is above the mean and the other is below the mean but penalized less if both are above the mean.

All three of these objective functions are useful and perhaps the last one corresponds most appropriately with analysis of clustered correlated data. In practice, however, using \( V^{-1} \) creates several problems. First, \( V \) may not be correct because the correlation structure and/or the variance function may not be correct. In this case, the scales on which \( d_2(h) \) or \( d_3(h) \) measure the closeness of the fit to the mean function are artificial. Second, the variance depends on the mean in GEE and GLIM models so even if the variance function is correct, an incorrect mean because of a misspecified model or a bandwidth that is too large will cause the variance function to be evaluated at the wrong means and thereby produce the wrong variances. Third, \( V \) has to be estimated from the data, thus the results of the theorem may not follow through. Estimation of the scale parameter, \( \phi \) is of particular concern. The estimator (38) is based on Pearson residuals and varies directly with the bandwidth. Small bandwidths, for example, have corresponding fits that are close to the data which yield small residuals. These small residuals imply a small \( \hat{\phi} \). Consequently, if \( \hat{\phi} \) is used as a weight in a bandwidth selector, its relationship with bandwidth may significantly distort the objective function. Notice that the simplest estimator, \( d_1 \), circumvents all of these problems.
5.2.2 Obtaining Deleted Fits

The use of deleted fits in cross-validation is a simple way to make the objective function unbiased, and in principle, they are easy to calculate. All one needs to do is remove the $i^{th}$ cluster from the dataset, estimate the model fits for the prediction locations of the removed cluster, and then repeat this process for all $i$. However, this method has the serious drawback of being very computationally demanding. For each evaluation of a bandwidth objective function, $s$ models have to be fitted. Evaluating the CV function over the grid $0.05, 0.1, \ldots, 1$ on a data set with 30 clusters and 10 prediction locations could involve up to $20 \times 30 \times 10 = 6,000$ GEE estimations.

In ordinary least squares, the deleted fits can be obtained from the data in a single pass through the data utilizing the Sherman-Morrison-Woodbury theorem (Myers, 1990). Let the subscript $-k$ of a matrix (or vector) denote that the $k^{th}$ row of the matrix (or vector) has been removed. The Sherman-Morrison-Woodbury (SMW) theorem states that

$$(X'X - x_kx_k')^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_kx_k'(X'X)^{-1}}{1 - x_k'(X'X)^{-1}x_k}. \tag{91}$$

It can be easily shown that $X_{-k}'X_{-k} = X'X - x_kx_k'$ and $X_{-k}'y_{-k} = X'y - x_ky_k$, where $x_k'$ is the $k^{th}$ row of $X$. It follows that the estimate of the model parameters without the $k^{th}$ data point can be calculated as

$$\hat{\beta}_{-k} = (X_{-k}'X_{-k})^{-1}(X_{-k}'y_{-k}) = \left[(X'X)^{-1} + \frac{(X'X)^{-1}x_kx_k'(X'X)^{-1}}{1 - x_k'(X'X)^{-1}x_k}\right] (X'y - x_ky_k),$$

where $\hat{\beta}_{-k}$ denotes that the estimate of $\beta$ is obtained without the $k^{th}$ observation.

This can be generalized to the GEE context where we remove an entire cluster rather than a single observation. Since the details are provided by Preisser and Qaqish (1996), we will only outline the development. First, recognize that the GEE updating equation,

$$\hat{\beta}_{u+1} = \hat{\beta}_u + \left(\sum_{i=1}^s X_i\hat{W}_iX_i\right)^{-1} \left(\sum_{i=1}^s X_i\hat{W}_i\langle f_{ij}\rangle^{-1}(y_i - F(X_i\hat{\beta}_u))\right),$$

from the IRLS algorithm, is similar to the weighted least squares estimator. In fact, the extension of (16) to the GEE context,

$$\hat{\beta}_{u+1} = (X'\hat{W}X)^{-1} X'\hat{W}\hat{z},$$

69
where $\hat{W}$ is a block diagonal weight matrix and $\hat{z} = X\hat{\beta}_u + \left\langle \hat{f}_{ij} \right\rangle^{-1} (y - F(X\hat{\beta}_u))$, looks very similar to a weighted least squares estimator. Second, clearly

$$X'_{-k}\hat{W}_{-k}X_{-k} = \sum_{i=1, i \neq k}^s X'_i\hat{W}_iX_i = X'\hat{W}X - X'_k\hat{W}_kX_k$$

and

$$X'_{-k}\hat{W}_{-k}\hat{z}_{-k} = \sum_{i=1, i \neq k}^s X'_i\hat{W}_i\hat{z}_i = X'\hat{W}\hat{z} - X'_k\hat{W}_k\hat{z}_k,$$

where $X_{-k}$ is $X$ with the rows that correspond to the $k^{th}$ cluster removed, and $\hat{W}_{-k}$ is $\hat{W}$ with the rows and columns that correspond to the $k^{th}$ cluster removed. Thus, the $k^{th}$ cluster deleted estimate of $\beta$, denoted by $\hat{\beta}_{-k}$, may be obtained by iteration on

$$\hat{\beta}_{-k,u+1} = \left( X'_{-k}\hat{W}_{-k}X_{-k} \right)^{-1} X'_{-k}\hat{W}_{-k}\hat{z}_{-k} = \left( X'\hat{W}X - X'_k\hat{W}_kX_k \right)^{-1} \left( X'\hat{W}\hat{z} - X'_k\hat{W}_k\hat{z}_k \right).$$

(92)

A way to speed the calculation of deleted estimates is to avoid evaluating the inverse

$$\left( X'\hat{W}X - X'_k\hat{W}_kX_k \right)^{-1}$$

for each deleted cluster. In ordinary least squares, the SMW theorem allows one to compute only one inverse, $(X'X)^{-1}$ for all the deleted points. Because of the similarities of the equation above to ordinary least squares, a more general form of SMW may be useful. While Nottingham (1995) has provided an extension of the theorem for a diagonal weight matrix, there is another very general result that suits the problem at hand.

It turns out that the SMW result is a special case of the inverse of a Schur complement (Searle, 1982, pg. 261). For any matrices $A$, $B$, $C$, and $D$ that can be placed in a single matrix of partitioned form,

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix},$$

with $A$ nonsingular, the Schur complement of $A$ is the matrix $D - CA^{-1}B$. Its inverse is given by

$$(D - CA^{-1}B)^{-1} = D^{-1} + D^{-1}C \left( A - BD^{-1}C \right)^{-1} BD^{-1}$$

(Searle, 1982, pg. 261). Thus,

$$\left( X'\hat{W}X - X'_k\hat{W}_kX_k \right)^{-1} =$$

$$\left( X'\hat{W}X \right)^{-1} + \left( X'\hat{W}X \right)^{-1}X'_k \left( \hat{W}^{-1}_k - X_k(X'\hat{W}X)^{-1}X'_k \right)^{-1} X_k(X'\hat{W}X)^{-1},$$

(93)
is a generalized SMW result.

Notice that in the special case of uncorrelated data, where $\hat{W}$ is diagonal, $X_k$ is the $k^{th}$ row of $X$ (denoted by $x_k'$), and $\hat{w}_k$ is the $k^{th}$ diagonal element of $\hat{W}$, (93) simplifies to

$$\left(X'\hat{W}X - x_k\hat{w}_kx_k'\right)^{-1} = \left(X'\hat{W}X\right)^{-1} + x_k\left(X'\hat{W}X\right)^{-1}x_k' \left(X'\hat{W}X\right)^{-1},$$

which is equivalent to the generalization of the SMW theorem in Nottingham (1995). In addition, (93) further simplifies to (91) when $\hat{w}_k = 1$.

Finally, (93) allows us to compute (92) using the single inverse $(X'\hat{W}X)^{-1}$ for all cluster deletions. Unfortunately, if $\hat{W}$ is not diagonal, no gain has been made as we also must compute the $n_i \times n_i$ inverse $\left(W_k^{-1} - X_k(X'\hat{W}X)^{-1}X_k'\right)^{-1}$ for each cluster $k$, which is more costly than computing the $p \times p$ matrix $(X'\hat{W}X - X_k\hat{W}_kX_k)^{-1}$ when $n_i > p$. The usefulness of the SMW approach has reached its limit; it is not computationally practical for non-diagonal $\hat{W}$.

Preisser’s and Qaqish’s (1996) motivation was to obtain closed form one-step updates to regression diagnostics such as DFBETAs, DFFITs, and others. A one-step update avoids the multiple iterations that ordinarily would be done to move GEE parameter estimates toward convergence. We have found in practice that the one-step update provides a good approximation of the fully iterated (92). Thus, our approach will be to use only the left-hand-side of (93) in a one-step update,

$$\hat{\beta}_k \approx \left( \sum_{i=1, i \neq k}^s X_i'\hat{W}_iX_i \right)^{-1} \left( \sum_{i=1, i \neq k}^s X_i'\hat{W}_i\hat{z}_i \right).$$

(94)

5.2.3 Penalized Cross-Validation

Using deleted fits in cross-validation is one of several methods to reduce the bias of the data-based distance measure $SSE = \sum_{i=1}^s (y_i - \hat{y}_{i}^{LGE})'(y_i - \hat{y}_{i}^{LGE})$. Several researchers (e.g., Craven and Whaba, 1979; Shibata, 1981; Rice, 1984) applied penalty functions to this distance measure for the purpose of reducing its bias in the polynomial smoothing context. See Rice (1984) for a review and Härdle (1990, pp. 154-155) for a discussion.

Mays (1995) indicates that there is also a need for applying penalties to the cross-validation estimator. The CV method “...has often been observed to still select bandwidths
smaller than desired” (Mays, 1995, pg. 22). The simplest penalized cross-validation objective function,
\[
\text{PRESS}^*(h) = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2}{n - \text{tr}(H^{LPR})},
\]
was developed by Einsporn (1987). The term \( n - \text{tr}(H^{LPR}) \) penalizes cross-validation for small bandwidths and is similar to the penalty function used in generalized cross-validation (Craven and Whaba, 1979). \( H^{LPR} \) is a local polynomial smoother matrix (see §4.1.2). It is constructed in the following manner. First, express the local fit at \( x_k \) as a linear combination of the observations, \( \hat{y}^{LPR}(x_k) = \sum_{i=1}^{n} c_i y_i \). Since the (uncentered) local polynomial estimate at prediction location \( x_k \) is given by \( \hat{\theta}(x_k) = (X_0'W_0kX_0)^{-1}X_0'W_0k y \), the linear combination can be expressed as
\[
\hat{y}^{LPR}(x_k) = \sum_{i=1}^{n} c_i y_i = x_0'(X_0'W_0kX_0)^{-1}X_0'W_0k y,
\]
where \( W_0k = \langle K_h(x_i - x_k) \rangle \) is the matrix of kernel weights and \( x_0' \) is the \( k^{th} \) row of \( X_0 \). Now, stack (95) without \( y \) for each observed \( x_k \) in the matrix
\[
H^{LPR} = \begin{bmatrix}
x_0'(X_0'W_1X_0)^{-1}X_0'W_01 \\
x_0'(X_0'W_2X_0)^{-1}X_0'W_02 \\
\vdots \\
x_0'(X_0'W_nX_0)^{-1}X_0'W_0n
\end{bmatrix}.
\]
This smoother matrix behaves as a “hat” matrix does in OLS, for it has the property that \( \hat{y}^{LPR} = H^{LPR}y \). Also, \( tr(H^{LPR}) \) is thought of as a measure of the degrees of freedom of a local polynomial model (Cleveland, 1979). Thus, the penalty used in PRESS* is analogous to the degrees of freedom of \( SSE \) for local polynomial regression. Note however, since it is neither idempotent nor symmetric, it is not a projection matrix.

Mays and Birch (2002) found in simulation studies that the penalty for PRESS* overcorrected for small bandwidths. They proposed a modification that additionally penalizes for large bandwidths,
\[
\text{PRESS}^{**}(h) = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2}{n - \text{tr}(H^{LPR}) + (n - p)\frac{SSE_{max} - SSE_h}{SSE_{max}}},
\]
Here, \( p \) is the number of parameters in the local polynomial, \( SSE_h = \sum_{i=1}^{n} (y_i - \hat{y}_{i, NP})^2 \) is the SSE of the local polynomial with bandwidth \( h \), and \( SSE_{max} \) is the SSE of the
local polynomial as $h \to \infty$, i.e., the largest possible $SSE$. The term $(n - p)(SSE_{\text{max}} - SSE_h)(SSE_{\text{max}})^{-1}$ penalizes for large bandwidths. Since $0 \leq SSE_h < SSE_{\text{max}}$ (provided there are no replications), we have $1 \geq \frac{SSE_{\text{max}} - SSE_h}{SSE_{\text{max}}} > 0$. For large bandwidths, $SSE_h$ will be large, $\frac{SSE_{\text{max}} - SSE_h}{SSE_{\text{max}}}$ will be small, and the penalty for CV will be large.

To generalize the selectors PRESS* and PRESS** to local GEE bandwidth selection, a suitable “hat” matrix is needed that, at least approximately, allows us to write a fit as a linear combination on the $\eta$-scale. This is possible because the parameter estimates at $x_{ij}$ are obtained from the last iteration of

$$\hat{\theta}(x_{ij}^*) = \left( X'_0 \hat{W}_{0ij} X_0 \right)^{-1} X'_0 \hat{W}_{0ij} \check{z}_{ij},$$

where $\hat{W}_{0ij} = \text{blockdiag} \left( \hat{f}_{0ij}^{1/2} \hat{V}_{0ij}^{-1/2} \hat{f}_{0ij}^{1/2} \right)$ and the subscripts on $\check{z}_{ij}$ indicate that $\check{z}$ also depends on $\hat{\theta}(x_{ij}^*)$. Since $\hat{\eta}_{ij}^{LGE} = x'_{0ij} \hat{\theta}(x_{ij}^*)$, a “hat” matrix can be formed as follows:

$$H^{LGE} = \begin{bmatrix} x'_{011}(X'_0 \hat{W}_{011} X_0)^{-1}X'_0 \hat{W}_{011} \\ x'_{012}(X'_0 \hat{W}_{012} X_0)^{-1}X'_0 \hat{W}_{012} \\ \vdots \\ x'_{0ns}(X'_0 \hat{W}_{0ns} X_0)^{-1}X'_0 \hat{W}_{0ns} \end{bmatrix},$$

(96)

The $tr(H^{LGE})$ does not necessarily measure the degrees of freedom of the LGEE model. As $h \to 0$, the local model eventually interpolates the averages at the prediction locations and the degrees of freedom approaches $N^*$, the number of unique prediction locations. However, when the weight matrix $\hat{W}_0$ is not diagonal, $tr(H^{LGE})$ may be larger than $N^*$ (Appendix A.2). Consequently, we restrict its use to the case where $\hat{V}_0$ is diagonal.

The generalized PRESS* for local GEE with independence working correlation structure is given by

$$\text{PRESS}^*(h) = \frac{\sum_{i=1}^{s} \left( y_i - \hat{y}_{i,-i}^{LGE} \right)' W_i \left( y_i - \hat{y}_{i,-i}^{LGE} \right)}{N^* - tr(H^{LGE})},$$

(97)

where $W_i$ is a positive-definite matrix. The $W_i$ determine the scales on which these penalized distances are measured. The discussion in §5.2.1, regarding the appropriate form of the weight matrix, applies here as well. Two examples are $V_{0i}^{-1}$ and $I_{n_i}$.

Extending PRESS** requires extending $SSE_h$ and $SSE_{\text{max}}$ to the present setting. It seems natural to use $SSE_h = \sum_{i=1}^{s} \left( y_i - \hat{y}_{i}^{LGE} \right)' W_i \left( y_i - \hat{y}_{i}^{LGE} \right)$ and $SSE_{\infty}$ which is
based on a nonparametric fit obtained using equal kernel weights (when \( h \to \infty \), i.e., a “globalized” local model). If \( V_0^{-1} \) is used for \( W_i \), it has to be estimated in practice. A good estimate for a given bandwidth may be obtained from the nonparametric fit. Unfortunately, it is then possible that \( SSE_h > SSE_\infty \). For simplicity and to ensure boundedness, we redefine \( SSE_h \) and \( SSE_\infty \) without \( V_0^{-1} \):

\[
SSE_h = \sum_{i=1}^{s} \left( y_i - \hat{y}_{i, LGEE} \right)' \left( y_i - \hat{y}_{i, LGEE} \right).
\]

Then,

- \( SSE_0 \leq SSE_h < SSE_\infty \), where \( SSE_0 = \sum_{i=1}^{s} (y_i - \bar{y}_i)'(y_i - \bar{y}_i) \),

- \( (SSE_\infty - SSE_h)(SSE_\infty - SSE_0)^{-1} \to 1 \) as \( h \to 0 \), and

- \( (SSE_\infty - SSE_h)(SSE_\infty - SSE_0)^{-1} \to 0 \) as \( h \to \infty \) (Appendix A.2).

The generalized PRESS** for local GEE with independence working correlation structure is given by

\[
\text{PRESS**(h)} = \frac{\sum_{i=1}^{s} \left( y_i - \hat{y}_{i, LGEE} \right)' W_i \left( y_i - \hat{y}_{i, LGEE} \right)}{N^* - tr(H_{LGEE}) + (N^* - p)\frac{SSE_\infty - SSE_h}{SSE_\infty - SSE_0}} \quad (98)
\]

where again, \( W_i \) is a positive-definite matrix that is left up to the user.

Generalizations for PRESS* and PRESS** for local GEEs with other working structures are possible if \( rank(H_{LGEE}) = tr(H_{LGEE}^{-1} H_{LGEE}) \), where \( H_{LGEE}^{-1} \) is the generalized inverse of \( H_{LGEE} \), is used in place of \( tr(H_{LGEE}) \) (see Appendix A.2 for additional comments).

5.2.4 Other Methods

Many other approaches to bandwidth selection are possible. Lin and Carroll (2001) offered an adapted version of Ruppert’s (1997) bandwidth selector, which involves empirical estimation of the MSE (see §3.5.4). Another approach is to minimize the asymptotic MSE and plug-in estimates for the unknown quantities. The asymptotic bias (78) and variance (79) of local linear GEE (§3.5.3), can be used to compute the asymptotic MSE (AMSE) for local linear GEE as a function of bandwidth by adding (79) to the square of (78). Then the AMSE can be integrated over the design space to obtain the asymptotic integrated mean squared error
\[
AIMSE[\eta_0] \approx \frac{1}{4}h^4 \int [\eta_0''(x)]^2 \, dx + \frac{R(K)}{sh} \int \frac{\sum_{k=1}^n v_{kk}^2 \phi^2 a(F(\eta_0(x))) f_{X_{ij}}(x)}{f^2(\eta_0(x)) \left( \sum_{k=1}^n v_{kk} f_{X_{ij}}(x) \right)^2} \, dx.
\]

This expression is of the form \( ah^4 + b \frac{1}{h} \) and has a minimum at \( h = (b/4a)^{1/5} \), or

\[
h_{AIMSE\text{opt}} = \left[ R(K) \int \frac{\sum_{k=1}^n v_{kk}^2 \phi^2 a(F(\eta_0(x))) f_{X_{ij}}(x)}{f^2(\eta_0(x)) \left( \sum_{k=1}^n v_{kk} f_{X_{ij}}(x) \right)^2} \, dx \right]^{1/5} \left( \int [\eta_0''(x)]^2 \, dx \right)^{-1/5} s^{-1/5} \tag{99}
\]

If \( R_i \) is set to \( I_{ni} \), as Lin and Carroll (2001) suggest, then (99) reduces to an \( \eta \)-scale based asymptotically optimal bandwidth selector that is similar to (72), the GLIM \( h_{opt} \).4

A “quick-and-dirty” bandwidth selector can be chosen by estimating the \( \eta \) function with a crude 3\textsuperscript{rd} degree polynomial then calculating and substituting for the quantities in \( h_{AIMSE\text{opt}} \). More sophisticated plug-in or solve-the-equation approaches, along the lines of Ruppert, Sheather, and Wand (1995), may also be possible. The behavior of such approaches could be a topic for future research.

### 5.3 Mixing Parameter Estimation

Estimating the mixing parameter involves finding the best convex combination of parametric fit, \( F(x' \hat{\beta}) \), and nonparametric fit, \( F(x_0' \hat{\theta}) \), using the data. It is as difficult a problem as bandwidth selection. We show several related methods of estimation, all of which revolve around minimization of the MSE. The MSE, which decomposes into the sum of the squared bias and the variance, is the appropriate performance measure because a large bias is a serious issue in a misspecified parametric model and large variance is often associated with a nonparametric model. Minimizing the MSE, therefore, should lead us to the best combination of the parametric and nonparametric components in terms of bias and variance.

4It is not identical to GLIM \( h_{opt} \) because Lin and Carroll (2000) allow the \( X_{ij} \) to be correlated within clusters, where as Fan, Heckman, and Wand (1995) assume the \( X_{ij} \) are independent.
5.3.1 Minimum Distance

Mays et al. (2001) considered minimizing PRESS (cross-validation), PRESS* and PRESS** to estimate the mixing parameter. They found that, while PRESS* and PRESS** can perform well in certain situations, they do not perform well asymptotically in terms of the rate of convergence of the MRR estimate to the true mean. They recommended using a corrected version of Burman and Chaudhuri’s (1992) estimator

$$\hat{\lambda}_{optMRR1} = \frac{\langle \hat{y}_{i}^{LLR} - \hat{y}_{i}^{OLS}, y - \hat{y}^{OLS} \rangle}{\| \hat{y}_{i}^{LLR} - \hat{y}^{OLS} \|^2}$$ (100)

for MRR1, the model robust regression method on which our $\mu$-scale mixing MRGEEI is based (see pg. 53 and §5.1). This estimator was developed by minimizing the squared Euclidean distance, $\delta = \sum (\hat{y}_{MRR1}^{MRR1} - m(x_i))^2$ between the fit and the true mean, which is

$$\frac{\langle \hat{y}_{LLR}^{LLR} - \hat{y}_{OLS}^{OLS}, m - \hat{y}_{OLS}^{OLS} \rangle}{\| \hat{y}_{LLR}^{LLR} - \hat{y}_{OLS}^{OLS} \|^2} = \frac{\langle \hat{y}_{LLR}^{LLR} - \hat{y}_{OLS}^{OLS}, m \rangle - \langle \hat{y}_{LLR}^{LLR} - \hat{y}_{OLS}^{OLS}, \hat{y}_{OLS}^{OLS} \rangle}{\| \hat{y}_{LLR}^{LLR} - \hat{y}_{OLS}^{OLS} \|^2},$$

and requires $\langle \hat{y}_{LLR}^{LLR} - \hat{y}_{OLS}^{OLS}, m \rangle = \sum_{i=1}^{n}(\hat{y}_{LLR}^{LLR}(x_i) - \hat{y}_{OLS}^{OLS}(x_i))m(x_i)$ to be estimated. This is accomplished by substituting $y_i$ for $m(x_i)$. Also, deleted fits are used to “safeguard against the possibility of putting an excessive amount of weight on $\hat{g}[\hat{y}_{LLR}]$ in the hybridization process” (Burman and Chaudhuri, 1992). Note that the deleted fits are only used in this term and not throughout (100).

The same approach can be applied to the current setting. We seek to minimize the squared generalized distance,

$$GD_I(\lambda) = (m - m^{MRGEEI})'W_I(m - m^{MRGEEI}),$$ (101)

i.e., the squared distance between the MRGEEI ($\mu$-scale mixing) fit and the true mean at the prediction locations. Or, for $\eta$-scale mixing, minimize

$$GD_{II}(\lambda) = (\eta - \eta^{MRGEEII})'W_{II}(\eta_i - \hat{\eta}_{i}^{MRGEEII}),$$ (102)

the squared distance between the MRGEEII fit and the true $\eta$-scale predictor at the prediction locations. The weight matrices $W_I$ and $W_{II}$ determine the scales on which distances are measured and are assumed to be positive definite. Recommendations include $V_0^{-1}$ for
\( W_I \) and \( (f_{0ij}) V_0^{-1} (f_{0ij}) \) for \( W_{II} \), when one wishes to measure the distance according to the variability of \( y \). We also suggest that the scale parameter be held fixed due to the Pearson estimator’s relationship with residuals (see §5.2.1). Or, one may use \( I_N \) for both \( W_I \) and \( W_{II} \) when simple estimators are desired.

Assuming \( W_I \) and \( W_{II} \) are fixed, the minimizers of \( GD_I \) and \( GD_{II} \) are given by

\[
\lambda_{optMRGEEI} = \frac{(\hat{y}^{LGE}_I - \hat{y}^{GEE}_I) W_I (m - \hat{y}^{GEE}_I)}{(\hat{y}^{LGE}_I - \hat{y}^{GEE}_I) W_I (\hat{y}^{LGE}_I - \hat{y}^{GEE}_I)} \tag{103}
\]

and

\[
\lambda_{optMRGEEII} = \frac{(\hat{\eta}^{LGE}_I - \hat{\eta}^{GEE}_I) W_{II} (\eta - \hat{\eta}^{GEE}_I)}{(\hat{\eta}^{LGE}_I - \hat{\eta}^{GEE}_I) W_{II} (\hat{\eta}^{LGE}_I - \hat{\eta}^{GEE}_I)} \tag{104}
\]

(Appendix A.3).

Estimators are still needed for \((\hat{y}^{LGE}_I - \hat{y}^{GEE}_I) W_I m\) and \((\hat{\eta}^{LGE}_I - \hat{\eta}^{GEE}_I) W_{II} \eta\). Following Burman and Chaudhuri (1992) and Mays et al. (2001), we substitute an estimate for \( m \) and use deleted fits in place of \((\hat{y}^{LGE}_I - \hat{y}^{GEE}_I)\) in (103). The estimate for \( m \) should not depend on the parametric model, as it is biased under misspecification. Also, the estimate should not depend on the nonparametric estimate or the estimate of \( \lambda \) will surely be biased toward 1 (the nonparametric fit). Several estimators are possible, including the raw data \( y \) the sample means at each prediction location \( \bar{y} \) or, deleted nonparametric fits \( \hat{y}_{i,-i} \).

We will work with the first two of these. The deleted nonparametric fits may still bias the estimate toward 1.

If \( y \) is used to estimate \( m \) in \( \lambda_{optMRGEEI} \) then analogously substitute \( F(y) \) for \( \eta \) in \( \lambda_{optMRGEEII} \) and use deleted \( \eta \)-scale fits in place of \((\hat{\eta}^{LGE}_I - \hat{\eta}^{GEE}_I)\) in (104). Caution must be exercised if the data are not grouped because \( F^{-1}(y) \) may not exist (e.g., with a binary response).

When \( W_I \) and \( W_{II} \) are block diagonal, the minimum generalized distance \( \lambda \) estimators based on \( y \) (mGDys) are

\[
\hat{\lambda}_{optMRGEEI} = \frac{\sum_{i=1}^{s} (\hat{\eta}^{LGE}_{i,-i} - \hat{\eta}^{GEE}_{i,-i}) W_{I,i} (y_{i} - \hat{y}^{GEE}_{i})}{\sum_{i=1}^{s} (\hat{\eta}^{LGE}_{i} - \hat{\eta}^{GEE}_{i}) W_{I,i} (\hat{y}^{LGE}_{i} - \hat{y}^{GEE}_{i})} \tag{105}
\]

and

\[
\hat{\lambda}_{optMRGEEII} = \frac{\sum_{i=1}^{s} (\hat{\eta}^{LGE}_{i} - \hat{\eta}^{GEE}_{i}) W_{II,i} (F^{-1}(y_{i}) - \hat{\eta}^{GEE}_{i})}{\sum_{i=1}^{s} (\hat{\eta}^{LGE}_{i} - \hat{\eta}^{GEE}_{i}) W_{II,i} (\hat{\eta}^{LGE}_{i} - \hat{\eta}^{GEE}_{i})} \tag{106}
\]
If there are multiple observations at each prediction location, then one may want to use \( \bar{y} \) in place of \( m \). The elements of these vectors correspond only to the unique prediction locations:

\[
\bar{y} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_{N^*} \end{bmatrix} \quad \text{and} \quad m = \begin{bmatrix} m(x_1^*) \\ m(x_2^*) \\ \vdots \\ m(x_{N^*}^*) \end{bmatrix}.
\]

Unlike the mGDys estimators, the minimum generalized distance estimators based on \( \bar{y} \) (mGDybars) will not involve the observations directly.

The resulting estimators are

\[
\hat{\lambda}_{\text{optMRGEEI}} = \frac{(\hat{y}_{LGEE} - \hat{y}_{GEE})' W_{\bar{y},I} (\bar{y} - \hat{y}_{GEE})}{(\hat{y}_{LGEE} - \hat{y}_{GEE})' W_{\bar{y},I} (\hat{y}_{LGEE} - \hat{y}_{GEE})} \tag{107}
\]

and

\[
\hat{\lambda}_{\text{optMRGEEII}} = \frac{(\hat{\eta}_{LGEE} - \hat{\eta}_{GEE})' W_{\bar{y},II} (F^{-1}(\bar{y}) - \hat{\eta}_{GEE})}{(\hat{\eta}_{LGEE} - \hat{\eta}_{GEE})' W_{\bar{y},II} (\hat{\eta}_{LGEE} - \hat{\eta}_{GEE})}. \tag{108}
\]

\( W_{\bar{y},I} \) and \( W_{\bar{y},II} \) are assumed to be positive definite.

A recommendation for \( W_{\bar{y},I} \) in this case is the inverse variance-covariance matrix for \( \bar{y} \),

\[
[\text{Var}[\bar{y}]]^{-1} = \left[ \frac{1}{r_{k'j'} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_i'}} I(x_{ij}^* = x_{k}^*) I(x_{ij'}^* = x_{k}^*) \text{Cov}[y_{ij}, y_{ij'}'] \right]_{N^* \times N^*}^{-1},
\]

where \( I(x_{ij}^* = x_{k}^*) \) is the indicator function

\[
I(x_{ij}^* = x_{k}^*) = \begin{cases} 1 & \text{if } x_{ij}^* = x_{k}^* \\ 0 & \text{if } x_{ij}^* \neq x_{k}^* \end{cases}.
\]

And for \( W_{\bar{y},II} \),

\[
[\text{Var}[F^{-1}(\bar{y})]]^{-1} \approx \left\langle f(F^{-1}(\bar{y})) \right\rangle (\text{Var}[\bar{y}])^{-1} \left\langle f(F^{-1}(\bar{y})) \right\rangle,
\]

where the notation \( \langle a \rangle \) means \( \text{diag}(a_1, a_2, \ldots, a_n) \). (These weight matrices are developed in Appendix A.4.)

Interestingly, the results of Theorem 1 also apply to mixing parameter selection for \( \mu \)-scale mixing. All one needs to do is substitute \( \hat{y}_{MGRGEE}^{MRGEE} \) for \( \hat{y}_{LGEE}^{MRGEE} \). The proof (§A.1) is identical and the comments in §5.2.1 apply as well. In the case of \( \eta \)-scale mixing, the theorem only holds asymptotically; it requires the approximation \( E[F^{-1}(\bar{y}_i)] \approx \eta_i \). This leads us to a
third way to estimate \( GD_I \) and \( GD_{II} \) that are extensions of the RGU \( \lambda \) estimator in Rahman et al. (1997). Estimate \( \mathbf{m} \) with \( \mathbf{y} \) in \( GD_I \) and \( \mathbf{\eta} \) with \( \mathbf{F}^{-1}(\mathbf{y}) \) in \( GD_{II} \), which was done to obtain (105) and (106), respectively, but now use deleted fits throughout the estimate, i.e.,

\[
\hat{\lambda}_{\text{optMRGEEI}} = \frac{\sum_{i=1}^{s}(\hat{\mathbf{y}}_{I,i}^{\text{LGEE}} - \hat{\mathbf{y}}_{I,i}^{\text{GEE}})' \mathbf{W}_{I,i}(\mathbf{y}_{i} - \hat{\mathbf{y}}_{I,i}^{\text{GEE}})}{\sum_{i=1}^{s}(\hat{\mathbf{y}}_{I,i}^{\text{LGEE}} - \hat{\mathbf{y}}_{I,i}^{\text{GEE}})' \mathbf{W}_{I,i}(\hat{\mathbf{y}}_{I,i}^{\text{LGEE}} - \hat{\mathbf{y}}_{I,i}^{\text{GEE}})}
\]

and

\[
\hat{\lambda}_{\text{optMRGEEII}} = \frac{\sum_{i=1}^{s}(\hat{\mathbf{n}}_{I,i}^{\text{LGEE}} - \hat{\mathbf{n}}_{I,i}^{\text{GEE}})' \mathbf{W}_{II,i}(\mathbf{F}^{-1}(\mathbf{y}_{i}) - \hat{\mathbf{n}}_{I,i}^{\text{GEE}})}{\sum_{i=1}^{s}(\hat{\mathbf{n}}_{I,i}^{\text{LGEE}} - \hat{\mathbf{n}}_{I,i}^{\text{GEE}})' \mathbf{W}_{II,i}(\hat{\mathbf{n}}_{I,i}^{\text{LGEE}} - \hat{\mathbf{n}}_{I,i}^{\text{GEE}})}.
\]

We should point out, however, that these estimators are not necessarily unbiased. They are derived from the estimated objective functions

\[
\sum_{i=1}^{s}(\mathbf{y}_{i} - \hat{\mathbf{y}}_{I,i}^{\text{MRGEEI}})' \mathbf{W}_{I,i}(\mathbf{y}_{i} - \hat{\mathbf{y}}_{I,i}^{\text{MRGEEI}})
\]

and

\[
\sum_{i=1}^{s}(\mathbf{F}^{-1}(\mathbf{y}_{i}) - \hat{\mathbf{n}}_{I,i}^{\text{MRGEEII}})' \mathbf{W}_{II,i}(\mathbf{F}^{-1}(\mathbf{y}_{i}) - \hat{\mathbf{n}}_{I,i}^{\text{MRGEEII}}),
\]

rather than the expected values of these functions, so they are related to the ideal objective functions in a non-linear fashion. Also, Rahman et al. (1997) found that the RGU estimator did not perform nearly as well as Burman and Chaudhuri’s (1992) corrected \( \lambda \) selector, which is a special case of (105). We recommend using (105) and (106) because they are based on \( \lambda \) selectors that have performed well in practice (see Mays et al. 2001).

### 5.3.2 Minimum MSE

As an alternative to measures of the distance between the model fit and the true mean, consider minimizing the MSE which involves the bias and the variance of the fit. First, a minimum MSE-at-a-point estimator of \( \lambda \) is derived. Since MSE is minimized at each point of interest, the estimate of \( \lambda \) can be expressed as a function of \( \mathbf{x}^* \). Define the following:

\[
E[\hat{\mathbf{y}}^P(\mathbf{x}^*)] = E^P(\mathbf{x}^*)
\]
\[
E[\hat{\mathbf{y}}^{NP}(\mathbf{x}^*)] = E^{NP}(\mathbf{x}^*)
\]
\[
\text{Var}[\hat{\mathbf{y}}^P(\mathbf{x}^*)] = V^P(\mathbf{x}^*)
\]
\[
\text{Var}[\hat{\mathbf{y}}^{NP}(\mathbf{x}^*)] = V^{NP}(\mathbf{x}^*)
\]
\[
\text{Cov}[\hat{\mathbf{y}}^P(\mathbf{x}^*), \hat{\mathbf{y}}^{NP}(\mathbf{x}^*)] = C(\mathbf{x}^*)
\]
where \( \hat{y}^P \) and \( \hat{y}^{NP} \) are the parametric and nonparametric fits at \( \mathbf{x}^* \), respectively. This notation will only allow expression of an estimator for MRGEEI (\( \mu \)-scale mixing). However, if the definition of the notation is changed to expectation, variances and covariances of \( \eta \)-scale (linear predictor) fits and if \( \eta \) is exchanged for \( m \), then all of the following will apply to the MRGEEII case. The arguments of these functions are suppressed for brevity in what follows.

The MSE(\( x^* \)) of \( \hat{y}^{MRGEEI}(x^*) = (1 - \lambda)\hat{y}^{GEE}(x^*) + \lambda\hat{y}^{LGEE}(x^*) \) is

\[
MSE(x^*) = \text{Bias}^2(x^*) + \text{Var}(x^*)
\]

\[
= \lambda^2 \left[ (E^P - E^{NP})^2 + V^P + V^{NP} - 2C \right]
- 2\lambda \left[ (E^P - E^{NP})(E^P - m) + V^P - C \right] + (E^P - m) + V^P. \tag{109}
\]

A unique minimum of \( \lambda \) is obtained at

\[
\lambda(x^*) = \frac{(E^P - E^{NP})(E^P - m) + V^P - C}{(E^P - E^{NP})^2 + V^P + V^{NP} - 2C} \tag{110}
\]

(Appendix A.5). This provides a value of \( \lambda \) for each point \( x^* \) of interest, which allows MRGEEI and MRGEEII to adjust to different degrees of misspecification throughout the \( x^* \) space. Regions where the user’s model performs well can be identified by plotting \( \lambda(x^*) \) versus \( x^* \) (assuming there are less than three regressors). This feature is attractive but comes at a price. This selector may introduce additional variance into the fit because of its flexibility. We recommend that it be used for exploratory purposes. An example of its use will be illustrated in §9.

Though (110) is concise, the estimators of \( E^P, E^{NP}, V^P, V^{NP} \) and \( C \) are rather detailed and will not be given here (see §6.1.3). \( \lambda(x^*) \) can be computed by substituting these estimators and an estimate for \( m(x^*) \). As with most bandwidth selection methods, estimation of the mean function often requires, at some level, the mean function itself. A popular solution is to provide another, less sophisticated estimate that is based on a model with more assumptions. For example, one could temporarily assume that the mean function is a 3\(^{rd} \) degree polynomial, fit it to the data, and use the fit as a crude estimate of \( m \). Or one could try the previous approaches—using the raw data \( \mathbf{y} \) or using \( \bar{\mathbf{y}} \), the averages at the prediction locations, if they are available. Most importantly, the estimate of \( m \) should not depend on the global model or the nonparametric model so that \( \lambda \) is not biased toward 0 or 1.
The variability of $\hat{\lambda}(x^*)$ can be reduced by averaging $\text{MSE}(x^*)$ across the data locations. This will simplify $\hat{\lambda}(x^*)$ to a single mixing parameter estimate for the whole model. If averaging is done at the design points, the *minimum average* $\lambda$ selector is

$$\lambda_{\text{AvgMSE}_{\text{opt}}} = \frac{\sum_{i=1}^s \sum_{j=1}^n (E_P(x_{ij}^*) - E_{NP}(x_{ij}^*)) (E_P(x_{ij}^*) - m(x_{ij}^*)) + V_P(x_{ij}^*) - C(x_{ij}^*)}{\sum_{i=1}^s \sum_{j=1}^n (E_P(x_{ij}^*) - E_{NP}(x_{ij}^*))^2 + V_P(x_{ij}^*) + V_{NP}(x_{ij}^*) - 2C(x_{ij}^*)}$$

(111)

(Appendix A.5). If integration is done, the *minimum IMSE* $\lambda$ selector is

$$\lambda_{\text{IMSE}_{\text{opt}}} = \int \left[ (E_P(x^*) - E_{NP}(x^*)) (E_P(x^*) - m(x^*)) + V_P(x^*) - C(x^*) \right] dx^* \div \int \left[ (E_P(x^*) - E_{NP}(x^*))^2 + V_P(x^*) + V_{NP}(x^*) - 2C(x^*) \right] dx^*$$

(112)

(Appendix A.5), which for obvious reasons, may be difficult to compute.

### 5.4 A Relationship Between Selectors

The mixing parameter selectors mentioned above and the distance-based bandwidth selectors (PRESS, PRESS*, and PRESS**) are related to MSE. First notice that $\text{MSE}(x^*)$ and $\sum \text{MSE}(x_{k}^*)$,

$$\text{MSE}(x^*) = E \left[ (\hat{y}(x^*) - \mu(x^*))^2 \right] \text{ and}$$

$$\sum \text{MSE}(x_{k}^*) = E \left[ \sum(\hat{y}(x^*) - \mu(x^*))^2 \right] = E \left[ (\hat{y} - m)'(\hat{y} - m) \right],$$

are themselves expected squared distance measures. An ideal objective function for bandwidth selection, such as

$$(m - \hat{y})'(m - \hat{y}),$$

(113)

is an estimate of $\sum \text{MSE}(x_{k}^*)$. If deleted fits are used instead of $\hat{y}$ and $y$ is used to estimate $m$, as in the PRESS function, then $E[\text{PRESS}]$ is merely the same constant shift of $E[(m - \hat{y}_{-i})'(m - \hat{y}_{-i})]$ for all bandwidths, which is approximately the $\sum \text{MSE}(x_{k}^*)$.

If a weight matrix $W$ is used in the distance measure then the same comparisons can be made if a weighted MSE,

$$W \text{MSE} = E \left[ (m - \hat{y})'W(m - \hat{y}) \right]$$

were the criterion. The scales on which to optimize the fit are mainly up to the user. The fits that are being assessed by the criterion already account for the dispersion seen or believed to be in the data.
The same connections and comments hold for the distance based $\lambda$ selectors. In particular though, there is an interesting relationship between (111) and the optimal $\lambda^*$ selector of Mays et al. (2001).

Notice the following about the denominator of (110):

\[
(E^P - E^{NP})^2 + V^P + V^{NP} - 2C
\]

\[
= (E[\hat{y}^P - \hat{y}^{NP}])^2 + Var[(\hat{y}^P - \hat{y}^{NP})] \\
= E[(\hat{y}^P - \hat{y}^{NP})^2] \\
= E[\|\hat{y}^P(\mathbf{x}^*) - \hat{y}^{NP}(\mathbf{x}^*)\|^2].
\]

Notice the following about the numerator:

\[
(E^P - E^{NP})(E^P - m) + V^P - C
\]

\[
= E[\hat{y}^P - \hat{y}^{NP}]E[\hat{y}^P - m] + V^P - C \\
= E[(\hat{y}^P - \hat{y}^{NP})(\hat{y}^P - m)] - Cov[\hat{y}^P - \hat{y}^{NP}, \hat{y}^P - m] + V^P - C \\
= E[(\hat{y}^P - \hat{y}^{NP})(\hat{y}^P - m)] \\
= E \left[ \langle \hat{y}^P(\mathbf{x}^*) - \hat{y}^{NP}(\mathbf{x}^*), \hat{y}^P(\mathbf{x}^*) - m(\mathbf{x}^*) \rangle \right].
\]

Hence,

\[
\lambda(\mathbf{x}^*) = \frac{E \left[ \langle \hat{y}^P(\mathbf{x}^*) - \hat{y}^{NP}(\mathbf{x}^*), \hat{y}^P(\mathbf{x}^*) - m(\mathbf{x}^*) \rangle \right]}{E \left[ \|\hat{y}^P(\mathbf{x}^*) - \hat{y}^{NP}(\mathbf{x}^*)\|^2 \right]}
\]

and likewise

\[
\lambda_{\text{AvgMSE opt}} = \frac{E \left[ \frac{1}{N} \sum_{i=1}^{s} \sum_{j=1}^{n_i} (\hat{y}^P(\mathbf{x}_{ij}^*) - \hat{y}^{NP}(\mathbf{x}_{ij}^*)) (\hat{y}^P(\mathbf{x}_{ij}^*) - m(\mathbf{x}_{ij}^*)) \right]}{E \left[ \frac{1}{N} \sum_{i=1}^{s} \sum_{j=1}^{n_i} (\hat{y}^P(\mathbf{x}_{ij}^*) - \hat{y}^{NP}(\mathbf{x}_{ij}^*))^2 \right]} \\
= \frac{E \left[ \langle \hat{y}^P - \hat{y}^{NP}, \hat{y}^P - m \rangle \right]}{E \left[ \|\hat{y}^P - \hat{y}^{NP}\|^2 \right]}.
\]

Thus, $\lambda_{\text{AvgMSE opt}}$ is the same as $\lambda^*$, except for the expectation in the numerator and the denominator. This connection is trivial because the method of Stein (1956) seeks to minimize $\sum_{i=1}^{s} \sum_{j=1}^{n_i} (\hat{y}^M_{RGEI}(\mathbf{x}_{ij}^*) - m(\mathbf{x}_{ij}^*))^2$, while the minimum AvgMSE method seeks to minimize $\frac{1}{N} \sum_{i=1}^{s} \sum_{j=1}^{n_i} E[(\hat{y}^M_{RGEI}(\mathbf{x}_{ij}^*) - m(\mathbf{x}_{ij}^*))^2]$. Similar statements can be made about the connection of $\lambda(\mathbf{x}^*)$ to

\[
\lambda^*(\mathbf{x}^*) = \frac{(y^P(\mathbf{x}^*) - \hat{y}^{NP}(\mathbf{x}^*)) (\hat{y}^P(\mathbf{x}^*) - m(\mathbf{x}^*))}{(\hat{y}^P(\mathbf{x}^*) - \hat{y}^{NP}(\mathbf{x}^*))^2},
\]

82
a special case of $\lambda^*$ not considered by Mays et al. (2001). It is easy to see that an estimator of this quantity could be quite variable. It also cautions us that

$$\lambda(x^*) = \frac{E\left[(\hat{y}^P(x^*) - \hat{y}^{NP}(x^*)) \left(\hat{y}^P(x^*) - m(x^*)\right)\right]}{E\left[(\hat{y}^P(x^*) - \hat{y}^{NP}(x^*))^2\right]},$$

which is (110), may have large variation unless stable estimates of the expectation can be obtained.

### 5.5 Simultaneous Estimation of $h$ and $\lambda$

All estimators proposed in this section assume that the bandwidth has been chosen at the time of $\lambda$-estimation. This is how model robust regression has always been done. In principle, we think of the mixing parameter as a measure that contrasts the performance of the parametric fit against the nonparametric fit. Consequently, the bandwidth has to be known or estimated to make this comparison. It is interesting, however, to wonder if simultaneous estimation of $h$ and $\lambda$ is possible and worthwhile.

Consider the example in Figures 3–5. These show slices through a MRGEEI average MSE surface over $h$ and $\lambda$ obtained by simulation. The model used is a $\mu$-scale contaminated Poisson generalized linear mixed model, called the $\mu$-Poisson model (see §7.1), with $s = 3$, $n_i = 10$ for each cluster, a low degree of correlation/variability ($\sigma = 0.12$), no misspecification ($\gamma = 0$, Figure 3), moderate misspecification ($\gamma = 0.5$, Figure 5) and a high degree of misspecification ($\gamma = 1$, Figure 4). The parametric model used is a GEE with compound symmetric correlation working structure and mean and variance functions identical to those of the $\mu$-Poisson model under no misspecification. The nonparametric model is a local linear GEE with an identity working structure and a variance function that is also identical to that of the $\mu$-Poisson model. Note that for large bandwidths ($h$), the local model fits poorly. This reflects the fact that the $\mu$-Poisson model has an overall quadratic trend on the $\eta$-scale, while the local linear GEE is linear on $\eta$-scale for large bandwidths.

In Figures 4 and 5, as $\lambda$ increases, the optimal $h$ also increases, illustrating that $h$ and $\lambda$ are directly related in this example. Notice for moderate misspecification (Figure 5) that a unique, simultaneously optimal $(h, \lambda)_{opt}$ exists at $h = 0.08$ and $\lambda = 0.5$ and that this is not
the same as \( h = 0.1 \) and \( \lambda = 0.6 \), the conditionally optimal \( \lambda_{opt|\lambda_{opt}} \). However, the simulated MSEs at \((h, \lambda)_{opt}\) and \(\lambda_{opt|h_{opt}}\) are 8.65 and 8.72, respectively, which are quite close. Notice especially that both the conditional and the simultaneously optimal values do not occur at a boundary, meaning that \textit{MRGEEI can improve the average MSE of the fit over both the parametric and the nonparametric fit based on the average MSE optimal} \( h \). This result parallels that by Mays (1995, pg. 137).
Figure 3: MRGEEI Average MSE surface slices for a poisson model with no mean model misspecification

Figure 4: MRGEEI Average MSE surface slices for a poisson model with a high degree of mean model misspecification
Figure 5: MRGEEI Average MSE surface slices for a poisson model with moderate mean misspecification

Figure 6: A 3D representation of Figure 5
Figure 4 suggests to use the nonparametric fit entirely, if one can find the optimal bandwidth, and Figure 3 suggests to use the only the parametric fit. It seems reasonable from these examples, that the simultaneously optimal $\lambda$ will be 1 when there is a high degree of misspecification, i.e., always use the nonparametric model. And under no misspecification, the simultaneously optimal $\lambda$ should be 0, i.e., always use the parametric model. Why use MRGEE then? Because we assume that the user does not know whether the model is misspecified. The estimate of $\lambda$ can provide information towards that end. There are large MSE penalties if one’s bandwidth selector performs poorly, so choosing a good $\lambda$ with the knowledge of the parametric model can lessen the costs of using the nonparametric fit alone. Furthermore, for slight or moderate departures from the mean model, where the parametric model is still useful, the MRGEE fit can provide MSE improvement over both the parametric and the nonparametric fits (at least in these examples).

Although they are related, $\lambda$ and $h$ are not interchangeable. One might imagine that if $h$ were near 0, so that the nonparametric fit interpolated the data, that $\lambda$ could play the same role as $h$. But Figure 5 illustrates that no matter which $\lambda$ is chosen, the optimal average MSE of the nonparametric fit at $h = 0.1$ can not be achieved. In other examples (not shown), where the “globalized” local model ($h \to \infty$) was equivalent to the parametric model, no setting of $\lambda$ at bandwidths near 0 were optimal. Thus, we conclude that $h$ and $\lambda$ are jointly identifiable parameters for the mean function. Certainly though, if bandwidth is chosen poorly, adjusting the mixing parameter can allay the problem to some extent. This relationship is not entirely symmetric though—for $\lambda$ close to 0, the value of the bandwidth will have little impact on the MRGEE fit. While the role of the bandwidth and mixing parameters are somewhat intertwined, the mixing parameter allows information about the parametric model to enter into the fitting process while the nonparametric method alone does not.

Simultaneous estimation of $(h, \lambda)$ would seem to be computationally daunting because a two-dimensional parameter space is to be searched. But obtaining the objective function in the $\lambda$ direction is easy: only convex combinations have to be computed. In fact, each slice is often a simple function of $\lambda$ (e.g., concave up quadratic). Simultaneous estimation can be done as follows. Search over a grid in the $h$ direction. For each $h$, compute the optimal $\lambda$
and evaluate the objective surface (e.g., cross-validation) at that \((h, \lambda_{opt})\). Then choose the \(h\) that gives the smallest value of the objective function along the path \((h, \lambda_{opt})\), \(h \in (0, \infty)\). If the grid on \(h\) is fine enough, this point must be \((h, \lambda)_{opt}\), the simultaneous optimum (if it exists). Metaphorically, this search over \(h\) is akin to tracing out a river at the bottom of a valley. The lowest point along the river gives the lowest point in the valley.

Finally, it seems reasonable to suspect that the dependence shown between \(h\) and \(\lambda\) in these examples would occur quite often in practice. This suggests that simultaneous estimation may be useful as some of this association may be recoverable in the estimate. And we have seen that it is easy to compute the simultaneous estimate, so the method looks promising. However, in this current research, we will choose \(\lambda\) conditionally on \(h\) to allow comparisons with previous research. Further exploration of simultaneous estimation could be a topic of future work.
6 Properties of MRGEEI and MRGEEII Fits Under Model Misspecification

6.1 Expectation, Variance and Covariance of the Fits

The MRGEE estimators will be studied theoretically with MSE, as this penalizes for both the (potentially) large bias of the misspecified parametric model and the (potentially) large variance of the nonparametric model. The MSE bias and variance decomposition requires knowledge of the expectation and variance of the fits, and since MRGEE is a convex combination of two fits, their covariance is also needed. These are developed here using the asymptotics of Taylor series expansions and matrix expectation and variance operators.

These results are shown to be generalizations of the work by Nottingham (1995) in the special case of quantal regression with uncorrelated data. The MSE expressions are used in §7 to find the asymptotically optimal bandwidth and mixing parameters for specific simulation models, and to provide the machinery required to compute estimates for (110), the optimal MSE mixing parameter (see §6.1.3). Furthermore, these expressions can be used to explore the feasibility or practicality of simultaneous estimation of \( h \) and \( \lambda \). For example, to determine how much improvement in MSE can be made by simultaneous rather than conditional estimation for a particular model, one could use the equations here to plot the entire MSE surface (see §5.5).

6.1.1 Deriving \( E^P, E^{NP}, V^P, V^{NP}, \text{ and } C \)

The expressions are first developed for the \( \eta \)-scale (with MRGEEII in mind), then with further Taylor series expansions, the \( \mu \)-scale versions are given (with MRGEEI in mind). The results are summarized in Table 7. Recall the notation \( E^P, E^{NP}, V^P, V^{NP}, \text{ and } C \) for the expectation, variance, and covariance of the parametric and nonparametric fits, respectively (§5.3.2). For the purpose of the following development, we utilize the \( \eta \)-scale definitions of these notations, specifically:

\[
E[\eta^P(x^*)] = E^P_\eta(x^*),
\]
\[
E[\eta^{NP}(x^*)] = E^{NP}_\eta(x^*),
\]
\[
\begin{align*}
\text{Var}[\hat{\eta}^P(x^*)] &= V_{\eta}^P(x^*), \\
\text{Var}[\hat{\eta}^{NP}(x^*)] &= V_{\eta}^{NP}(x^*), \quad \text{and} \\
\text{Cov}[\hat{\eta}^P(x^*), \hat{\eta}^{NP}(x^*)] &= C_{\eta}(x^*).
\end{align*}
\]

Again, the arguments of these functions are dropped for brevity.

These moments are calculated under the assumption that nature’s mean model is \(m(x^*)\), which may or may not be equal to the user’s mean function, \(m^*(x^*) = F(x^\prime \beta)\). Hence, these results apply both when there is mean model misspecification and when the mean model is correct. We use the quasi-likelihood framework for local estimation to derive asymptotic bias and variance expressions rather than the more popular asymptotic development which Lin and Carroll (2000) have followed. These results apply to the multiple covariate case and they do not depend on limiting conditions such as \(h \to 0\) and \(sh \to \infty\). Hence, we conjecture that they are more accurate for small samples than (78) and (79) of Lin and Carroll (2000).

The \(\eta\)-scale nonparametric fit, \(\hat{\eta}_{LGE}(x^*)\), may be obtained as \(x_0^\prime \hat{\theta}\) if the nonparametric model matrix is not centered or as \(\hat{\theta}_0 = S_0^\prime \hat{\theta}\) if the model matrix is centered (cf. 58). The work in this section is done with the former approach in mind, but all the results apply for the latter if \(x_0\) is replaced with \(S_0\) (see pg. 39) in expressions \(E_{\eta}^{NP}, V_{\eta}^{NP}\), and the right side of \(C_{\eta}\) where \(x_0\) appears.

Often one employs a first order Taylor series about \(x^\prime \beta\) or \(\beta\) to find expectations or variances of the parameters, but since \(E[Y_{ij}]\) may not be equal to \(F(x_{ij}^\prime \beta)\), \(F(x_{ij}^\prime \hat{\beta})\) may not converge to \(m(x^*)\). There is no longer necessarily a “true” parameter \(\beta\). However, when the parametric model is wrong, it can still adapt its shape toward \(m\) to a certain extent. Consequently, it is reasonable to assume that for some \(\beta^*\), \(F(x_{ij}^\prime \hat{\beta})\) will converge to or be close to \(F(x_{ij}^\prime \beta^*)\) in the sense that \(E[\hat{\beta}] \approx \beta^*\). This way, Taylor series approximations about \(\beta^*\), and \(x^\prime \beta^*\) will be reasonably accurate.

For the nonparametric model, similar comments can be made. Though more flexible than the parametric model, the local model \(F(x_0^\prime \hat{\theta})\) still has some degree of misspecification at a given bandwidth if the user’s mean model is not correct. Consequently, the local model also has no “true” parameter \(\theta\) and we assume there is some \(\theta^*\) such that \(E[\hat{\theta}] \approx \theta^*\). This \(\theta^*\) will be a function of \(x^*\), the prediction location.
Consider the updating equation
\[
\hat{\beta}_{u+1} = \hat{\beta}_u + \left( \sum_{i=1}^{s} X_i' \tilde{W}_i X_i \right)^{-1} \sum_{i=1}^{s} X_i' \tilde{W}_i \left( \hat{f}_{ij} \right)^{-1} (y_i - F(X_i \hat{\beta}_u)).
\] (114)

As in McCullagh and Nelder (1989), set \( \hat{\beta}_u \) equal to \( \beta^* \) and set \( \hat{\beta}_{u+1} \) equal to the converged estimate \( \hat{\beta} \). Then the asymptotic expectation is given by
\[
E_{\eta}^P(\mathbf{x}^*) \approx \mathbf{x}' \beta^* + \mathbf{x}' \left( \sum_{i=1}^{s} X_i' \mathbf{W}_i X_i \right)^{-1} \sum_{i=1}^{s} X_i' \mathbf{W}_i \left( f_{ij} \right)^{-1} (m_i - F(X_i \beta^*)).
\] (115)

For the variance expression, a sandwich variance calculation is utilized, while keeping in mind that the mean is \( m(\mathbf{x}^*) \). \( V_{\eta}^P(\mathbf{x}^*) \approx \)
\[
\mathbf{x}' \left( \sum_{i=1}^{s} X_i' \mathbf{W}_i X_i \right)^{-1} \left( \sum_{i=1}^{s} X_i' \mathbf{W}_i \left( f_{ij} \right)^{-1} \text{Var}[Y_i] \left( f_{ij} \right)^{-1} \mathbf{W}_i X_i \right) \left( \sum_{i=1}^{s} X_i' \mathbf{W}_i X_i \right)^{-1} \mathbf{x},
\] (116)
where \( \text{Var}[Y_i] = E[(y_i - m_i)(y_i - m_i)'] \).

The expectation and variance for the nonparametric model are the same as for the parametric model except one needs to replace \( \beta^* \) with \( \theta^* \), \( \mathbf{W}_i \) with \( \mathbf{W}_{0i} \), and \( \mathbf{X}_i \) with \( \mathbf{X}_{0i} \), to indicate that the nonparametric model matrix may be different from the parametric model matrix. Also, the weight matrix \( \mathbf{W}_{0i} = \left( f_{0ij} \right) \mathbf{K}_{hi}^{1/2} \mathbf{V}_{0i}^{-1} \mathbf{K}_{hi}^{1/2} \left( f_{0ij} \right) \) involves the kernel weights centered at \( \mathbf{x}^* \) and other terms involving \( \theta^* \) that depend on \( \mathbf{x}^* \). The results for the nonparametric model are
\[
E_{\eta}^{NP}(\mathbf{x}^*) \approx \mathbf{x}' \theta^* + \mathbf{x}' \left( \sum_{i=1}^{s} X_{0i}' \mathbf{W}_{0i} X_{0i} \right)^{-1} \sum_{i=1}^{s} X_{0i}' \mathbf{W}_{0i} \left( f_{0ij} \right)^{-1} (m_i - F(X_{0i} \theta^*)),
\] (117)
and \( V_{\eta}^{NP}(\mathbf{x}^*) \approx \)
\[
\mathbf{x}' \left( \sum_{i=1}^{s} X_{0i}' \mathbf{W}_{0i} X_{0i} \right)^{-1} \left( \sum_{i=1}^{s} X_{0i}' \mathbf{W}_{0i} \left( f_{0ij} \right)^{-1} \text{Var}[Y_i] \left( f_{0ij} \right)^{-1} \mathbf{W}_{0i} X_{0i} \right) \left( \sum_{i=1}^{s} X_{0i}' \mathbf{W}_{0i} X_{0i} \right)^{-1} \mathbf{x}.
\] (118)

An estimate of the covariance term requires more work, but quite naturally is a cross between the variance of the parametric and the nonparametric fit. It is given by \( C_{\eta}(\mathbf{x}^*) = \text{Cov}[\hat{\theta}_{GEE}(\mathbf{x}^*), \hat{\eta}_{LGE}(\mathbf{x}^*)] = \)
\[
\mathbf{x}' \left( \sum_{i=1}^{s} X_i' \mathbf{W}_i X_i \right)^{-1} \left( \sum_{i=1}^{s} X_i' \mathbf{W}_i \left( f_{ij} \right)^{-1} \text{Var}[Y_i] \left( f_{ij} \right)^{-1} \mathbf{W}_i X_i \right) \left( \sum_{i=1}^{s} X_i' \mathbf{W}_i X_i \right)^{-1} \mathbf{x}.
\] (119)
The derivation of the covariance is found in Appendix B.1.

The results for $\mu$-scale fits follow easily from the expressions above. The first-order Taylor series approximations,

$$F(x'\hat{\beta}) \approx F(x'\beta^\star) + f(x'\beta^\star)(x'\hat{\beta} - x'\beta^\star)$$

and

$$F(x_0'\hat{\theta}) \approx F(x_0'\theta^\star) + f(x_0'\theta^\star)(x_0'\hat{\theta} - x_0'\theta^\star),$$

yield

$$E[F(x'\hat{\beta})] \approx F(x'\beta^\star) + f(x'\beta^\star)(E[x'\hat{\beta}] - x'\beta^\star),$$

$$E[F(x_0'\hat{\theta})] \approx F(x_0'\theta^\star) + f(x_0'\theta^\star)(E[x_0'\hat{\theta}] - x_0'\theta^\star),$$

$$Var[F(x'\hat{\beta})] \approx f^2(x'\beta^\star)Var[x'\hat{\beta}],$$

$$Var[F(x_0'\hat{\theta})] \approx f^2(x_0'\theta^\star)Var[x_0'\hat{\theta}],$$

and

$$Cov[F(x'\hat{\beta}), F(x_0'\hat{\theta})] \approx f(x'\beta^\star)Cov[x'\hat{\beta}, x_0'\hat{\theta}]f(x_0'\theta^\star).$$

With the following notation, the results can be written compactly:

$$\left(\sum_{i=1}^{s} X_i'W_iX_i\right)^{-1} = B,$$

$$\left(\sum_{i=1}^{s} X_{0i}'W_{0i}X_{0i}\right)^{-1} = B_0,$$

$$X_i'W_i\langle f_{ij}\rangle^{-1} = C_i,$$

$$X_{0i}'W_{0i}\langle f_{0ij}\rangle^{-1} = C_{0i},$$

$$m_i - F(X_i\beta^\star) = \epsilon_i^P,$$

$$m_i - F(X_{0i}\theta^\star) = \epsilon_i^{NP},$$

and

$$y_i - m_i = \epsilon_i.$$

The results are shown in Table 7.

In Table 7 notice the symmetry between the parametric and nonparametric expressions and their similarity to the sandwich estimators (42) and (80). This is a consequence of studying both methods in the quasi-likelihood framework (§3.5.1).
Table 7: Expectation, Variance and Covariance of MRGEEI and MRGEEII fits

<table>
<thead>
<tr>
<th>Scale</th>
<th>Type</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>$EP(x^*)$</td>
<td>$F(x'\beta^<em>) + f(x'\beta^</em>)x'B\sum_{i=1}^{s}C_i\epsilon_i^P$</td>
</tr>
<tr>
<td>(MRGEEI)</td>
<td>$ENP(x^*)$</td>
<td>$F(x_0'\theta^<em>) + f(x_0'\theta^</em>)x_0'B_0\sum_{i=1}^{s}C_0i\epsilon_i^{NP}$</td>
</tr>
<tr>
<td></td>
<td>$V^P(x^*)$</td>
<td>$f(x'\beta^<em>)x'B(\sum_{i=1}^{s}C_iE[\epsilon_i\epsilon_i']C_i')Bxf(x'\beta^</em>)$</td>
</tr>
<tr>
<td></td>
<td>$V^{NP}(x^*)$</td>
<td>$f(x_0'\theta^<em>)x_0'B_0(\sum_{i=1}^{s}C_0iE[\epsilon_i\epsilon_i']C_0i_0)B_0x_0f(x_0'\theta^</em>)$</td>
</tr>
<tr>
<td></td>
<td>$C(x^*)$</td>
<td>$f(x'\beta^<em>)x'B(\sum_{i=1}^{s}C_iE[\epsilon_i\epsilon_i']C_i')B_0x_0f(x_0'\theta^</em>)$</td>
</tr>
</tbody>
</table>

| $\eta$  | $E_P(\eta^*)$ | $x'\beta^* + x'B\sum_{i=1}^{s}C_i\epsilon_i^P$                          |
| (MRGEEII) | $E^{NP}(\eta^*)$ | $x_0'\theta^* + x_0'B_0\sum_{i=1}^{s}C_0i\epsilon_i^{NP}$            |
|         | $V_P(\eta^*)$ | $x'B(\sum_{i=1}^{s}C_iE[\epsilon_i\epsilon_i']C_i')Bx$                 |
|         | $V^{NP}(\eta^*)$ | $x_0'B_0(\sum_{i=1}^{s}C_0iE[\epsilon_i\epsilon_i']C_0i_0)B_0x_0$ |
|         | $C(\eta^*)$   | $x'B(\sum_{i=1}^{s}C_iE[\epsilon_i\epsilon_i']C_i')B_0x_0$            |

6.1.2 Computing $E^P$, $E^{NP}$, $V^P$, $V^{NP}$, and $C$ Under a True Model

Given the first and second moments of a true model, the asymptotic $\text{MSE}(x^*)$ can be computed for a given user’s model estimate and a given nonparametric model estimate using expressions displayed in Table 7. Likewise, from these results the $\text{MSE}(x^*)$ can be computed for MRGEEI and MRGEEII estimates on either the $\eta$-scale or $\mu$-scale, as will be shown in §6.2.

The quantities in Table 7 require $\beta^*$ and $\theta^*$ as reasonable initial approximations of $E[\hat{\beta}]$ and $E[\hat{\theta}]$, respectively, to center the Taylor series expansions. A reasonable approach to obtain these values for the parametric model would be to minimize the sum of squares

$$\sum_{i=1}^{s}(m_i - F(X_i\beta))^tV_i^{-1}(m_i - F(X_i\beta))$$

over $\beta$ at the prediction locations, and for the nonparametric model, minimize

$$\sum_{i=1}^{s}(m_i - F(X_0i\theta))^tK_{hi}^{1/2}V_{0i}^{-1}K_{hi}^{1/2}(m_i - F(X_0i\theta))$$
over $\theta$ at the prediction locations, centering the kernel weights at each $x^*$ of interest. Such
minimizations lead to quasi-score estimating equations

$$\sum_{i=1}^{s} X'_i \langle f_{ij} \rangle V_i^{-1} (m_i - F(X_i \beta^*)) = 0, \quad (120)$$

and

$$\sum_{i=1}^{s} X'_0i \langle f_{0ij} \rangle K_{hi}^{1/2} V_{0i}^{-1} K_{hi}^{1/2} (m_i - F(X_0 \theta^*)) = 0, \quad (121)$$

which are solved iteratively.

The $\beta^*$ and $\theta^*$ parameters are not the only unknown quantities in the expressions given
in Table 7. Within the weight matrices, $V_i$ and $V_{0i}$ there is (are) correlation parameter(s)
whose values are typically estimated from the data. When the true model is known, however,
we would like to be able to evaluate $E^P$, $E^{NP}$, $V^P$, $V^{NP}$, and $C$ solely from the fixed
properties of that model. After all, $E^P$, $E^{NP}$, $V^P$, $V^{NP}$, and $C$ are not random variables.
Due to difficulties in specifying a GEE model (see §7), it is not as simple as replacing the
correlation parameters with values from a true model. The true model is not commonly
expressed with an explicit correlation structure. It can be derived, but often the appropriate
correlation structure does not match the one specified in the user’s GEE model and may not
even have the same parameters.

Our approach will be to replace the correlation parameters in $V_i$ and $V_{0i}$ with the
expected value of their respective MOM estimators. This is done for two common correlation
structures, AR1 and CS (see §2.2.3) whose MOM estimators are given in Table 5. The
details are found in Appendix B.2. For simplicity, the cluster sizes $n_i$ are assumed to be
equal to $n$ and the expressions are given using the notation for the parametric model (the
nonparametric results follow immediately.) The CS and AR1 estimators depend on the scale
parameter estimate, so the expectation of

$$\hat{\phi}^2 = \frac{1}{sn - p} \sum_{i=1}^{s} \sum_{j=1}^{n} \frac{(Y_{ij} - F(x'_{ij} \hat{\beta}))^2}{a^*(F(x'_{ij} \hat{\beta}))},$$

where $a^*(\cdot)$ is the user’s variance function, is found first. It is given by

$$E[\hat{\phi}^2] \approx \frac{1}{sn - p} \sum_{i=1}^{s} \sum_{j=1}^{n} \frac{\text{Var}[Y_{ij}] + \left( m(x^*_{ij}) - F(x'_{ij} \beta^*) \right)^2}{a^* \left( F(x'_{ij} \beta^*) \right)} \overset{\text{define}}{=} \tilde{\phi}^2, \quad (122)$$
where the variance \( \text{Var}[Y_{ij}] \) is determined from the given model. Notice that if the user’s mean model is correct \( (m(x^*) = F(x^'\beta)) \) and the user’s variance function is correct, then

\[
E[\hat{\phi}^2] \approx sn(sn-p)^{-1} \approx 1 \text{ for large } s, \text{ i.e., on average, no adjustment of the variance function by the scale parameter is needed. The compound symmetry (CS) estimator,}
\]

\[
\hat{\rho}_{CS} = \left(\frac{1}{2} sn(n-1) - p\right)^{-1} \sum_{i=1}^{s} \sum_{j<k}^{n} \frac{(Y_{ij} - F(x_{ij}'\hat{\beta}))(Y_{ik} - F(x_{ik}'\hat{\beta}))}{\sqrt{\hat{\phi}^2 a^*(F(x_{ij}'\beta))} \sqrt{\hat{\phi}^2 a^*(F(x_{ik}'\beta))}},
\]

has approximate expectation \( E[\hat{\rho}_{CS}] \approx \)

\[
\left(\frac{1}{2} sn(n-1) - p\right)^{-1} \sum_{i=1}^{s} \sum_{j<k}^{n} \text{Cov}[Y_{ij}, Y_{ik}] + \left(m(x^*_{ij}) - F(x^*_{ij}\beta^*)\right) \left(m(x^*_{ik}) - F(x^*_{ik}\beta^*)\right).
\]

(123)

The approximate expectation of the autoregressive(1) (AR1) estimator,

\[
\hat{\rho}_{AR1} = \frac{1}{(n-1)s-p} \sum_{i=1}^{s} \sum_{j=1}^{n-1} \frac{(Y_{ij} - F(x_{ij}'\hat{\beta}))(Y_{ij+1} - F(x_{ij+1}'\hat{\beta}))}{\sqrt{\hat{\phi}^2 a^*(F(x_{ij}'\beta))} \sqrt{\hat{\phi}^2 a^*(F(x_{ij+1}'\beta))}},
\]

is determined in a similar manner. It is given by \( E[\hat{\rho}_{AR1}] \approx \)

\[
\frac{s}{(n-1)s-p} \sum_{j=1}^{n-1} \text{Cov}[Y_{ij}, Y_{ij+j+1}] + \left(m(x^*_{ij}) - F(x^*_{ij}\beta^*)\right) \left(m(x^*_{ij+j+1}) - F(x^*_{ij+j+1}\beta^*)\right).
\]

(124)

If the user’s mean model is correct and the variance function is correct, then the terms in these expectations reduce to true model correlations, \( \text{corr}[Y_{ij}, Y_{ik}] \). Additionally, if the correlation structure is correct, then these approximate expectations reduce to approximate averages over the true correlation parameters. For example, if the true model is compound symmetric, then (123) reduces to (approximately) the average of the true model’s correlations over all combinations \( Y_{ij} \) and \( Y_{ik} \), \( j \neq k \). On the other hand, when the mean model is misspecified, there are two consequences. First, the variance functions \( a^* \), even if correct, do not provide the correct variances because their arguments are not necessarily correct. Thus, the terms in (123) and (124) are no longer correlations. Secondly, there is a contribution that depends on which side of the true mean the user’s optimal mean prediction \( F(x^*_{ij}\beta^*) \) falls. If, for both points \( x^*_{ij} \) and \( x^*_{ik} \), \( j \neq k \), the user’s mean model prediction is above the
true mean, then the contribution to the covariance is positive. If the user’s predictions are on opposite sides of the true mean, the contribution to the covariance is negative.

Using these estimates, the equations in Table 7 may be computed from a true model as follows.

1. Solve (120) to provide an initial estimate of \( \beta^\star \) by iterating on the updating equation,

\[
\beta_{u+1}^\star = \beta_u^\star + \left( \sum_{i=1}^s X_i' W_i X_i \right)^{-1} \sum_{i=1}^s X_i' W_i \langle f_{ij} \rangle^{-1} (m_i - F(X_i \beta_u^\star)),
\]  

(125)

where \( W_i = \langle f_{ij} \rangle V^{-1} \langle f_{ij} \rangle \) and \( V \) is diagonal.

2. Compute approximate expected values of correlation estimators for the AR1 or CS structures using properties of the true model (mean, variance, covariance of observations). Use these estimates to compute the whole \( V \) matrix.

3. Update \( \beta^\star \) using (125) with the whole \( V \) matrix.

4. Repeat steps 2 and 3 until a convergence criterion, such as the relative change in the estimates, is met (see pg. 28).

5. Repeat steps 1–4 for the nonparametric model by replacing \( \beta^\star \) with \( \theta^\star \) and \( X \) with \( X_0 \), the nonparametric model matrix. Also, replace the respective correlation structure, correlation parameters, scale parameter, and user’s variance function in \( V \), replace \( f_{ij} \) with \( f_{0ij} \), replace \( W_i \) with \( W_{0i} = \langle f_{0ij} \rangle K_{hi}^{1/2} V_0^{-1} K_{hi}^{1/2} \langle f_{0ij} \rangle \), and solve (121) instead of (120). The resulting \( \theta^\star \) and \( E[\hat{\rho}_{CS}] \) or \( E[\hat{\rho}_{AR1}] \) for the nonparametric model are functions of \( x^\star \) and as such, these steps must be repeated at each point \( x^\star \) of interest, with appropriate adjustments to the kernel weights.

6. Compute \( E[\epsilon_i \epsilon_i'] \), the variance-covariance matrix of \( y \), based on the true model. Compute \( \epsilon_i^P \) and \( \epsilon_i^{NP} \). Finally, use the computed values obtained in all previous steps to compute the expressions in Table 7.

### 6.1.3 Estimating \( E^P, E^{NP}, V^P, V^{NP}, \) and \( C \) from Data

The quantities \( E^P, E^{NP}, V^P, V^{NP}, \) and \( C \) in the minimum MSE estimator (110) in §5.3.2 can be estimated from the data by using the scale, correlation, \( \hat{\theta} \), and \( \hat{\beta} \) estimates from the
data and applying the sandwich variance approach to approximate \( E[\epsilon_i\epsilon'_i] = Var[Y_i] \) in the expressions in Table 7. By the sandwich variance approach, we imply to estimate \( E[\epsilon_i\epsilon'_i] \) with \( \epsilon_i\epsilon'_i = (y_i - m_i)(y_i - m_i)' \). As an estimate of \( m \), we use the averages \( \bar{y} \) obtained at the unique observed \( x^* \) locations. Unfortunately, \( \lambda(x^*) \) can then only be computed at these points. Other potential substitutions for \( m \) are discussed on page 80.

The sandwich variance estimator for GEE has been shown to have considerable variability in practice (Prentice, 1988). This variability problem is even more serious for estimates based on \( \hat{\epsilon}_i\hat{\epsilon}'_i = (y_i - \bar{y}_i)(y_i - \bar{y}_i)' \), because the \( \bar{y}_i \) are not as smooth as the model-based predictions. An alternative suggestion is to use a nonparametric, model-based estimate of \( Var[Y_i] \), i.e., \( (\sum_{i=1}^s X'_0 W_0 X_0)^{-1} \), where the means needed for the MOM correlation parameter estimators are estimated from a nonparametric leave-one-out fit. Our experience has been that this latter technique makes the average minimum MSE \( \lambda \) selectors based on (111) very competitive, but not uniformly better than the minimum distance estimators, (105)–(108).

### 6.2 MSE at the Observed Prediction Locations

Bias and variance properties of the parametric, nonparametric, MRGEEI, and MRGEEII fits at the observed data locations serve as a useful summary of model performance. These terms can indicate effects of model misspecification across all the points. This section parallels §7.1 of Nottingham (1995), who developed bias, variance, and MSE properties in the case of quantal regression with uncorrelated data. His results are shown to be special cases of the \( \mu \)-scale results presented here. Several of these expressions follow directly from those in Table 7.

Mean model misspecification can occur though the link function, the linear predictor, or both. For example, the user’s parametric mean model, \( F(x'_j\beta) \), may be misspecified in \( \eta = x'_j\beta \), may have an incorrect function \( F \), or be incorrect in link and linear predictor. Regardless of the form of misspecification, the parametric model can be expressed through \( \eta \) misspecification only. Since \( F \) has an inverse, \( m(x'_{ij}) \equiv F(\eta(x'_{ij})) \) implies that \( \eta(x^*_{ij}) = F^{-1}(m(x^*_{ij})) \) and \( m(x^*_{ij}) = F(F^{-1}(m(x^*_{ij}))) \), as long as \( m \) and \( F \) have the same range. Consequently, without loss of generality we assume that the inverse link function \( F \) is always
correct. This is especially important for the MRGEEII case, because the mixing of parametric and nonparametric fits occurs on the $\eta$-scale.

If nature’s mean model agrees with the user’s model, then $\hat{\eta}^P = x_{ij}^T \hat{\beta}$ and $\hat{y}^P = F(x_{ij}^T \hat{\beta})$ are asymptotically unbiased at observed location $x_{ij}^*$. Otherwise there is a bias of

$$\text{Bias}^P[x_{ij}^T \hat{\beta}] \approx E_\eta(x_{ij}^*) - \eta(x_{ij}^*) = x_{ij}^T \beta^* - \eta(x_{ij}^*) + x_{ij}^T B \sum_{i=1}^s C_i \epsilon_i^P$$

$$= x_{ij}^T \beta^* - \eta(x_{ij}^*) + x_{ij}^T \left( \sum_{i=1}^s X_i^T W_i X_i \right)^{-1} \sum_{i=1}^s X_i^T W_i \langle f_{ij} \rangle^{-1} (m_i - F(X_i \beta^*)),$$

on the $\eta$-scale and a bias of

$$\text{Bias}^P[F(x_{ij}^T \hat{\beta})] \approx E^P(x_{ij}^*) - m(x_{ij}^*) = F(x_{ij}^T \beta^*) - m(x_{ij}) + x_{ij}^T f(x_{ij}^T \beta^*) B \sum_{i=1}^s C_i \epsilon_i^P$$

$$= F(x_{ij}^T \beta^*) - m(x_{ij}) + x_{ij}^T f(x_{ij}^T \beta^*) \left( \sum_{i=1}^s X_i^T W_i X_i \right)^{-1} \sum_{i=1}^s X_i^T W_i \langle f_{ij} \rangle^{-1} (m_i - F(X_i \beta^*))$$

on the $\mu$-scale. For $\mu$-scale prediction at the data locations, the bias can be re-expressed as

$$\text{Bias}^P[X \hat{\beta}^*] \approx \left( \langle f_{ij} \rangle X \left( \sum_{i=1}^s X_i^T W_i X_i \right)^{-1} X^T W \langle f_{ij} \rangle^{-1} - I_N \right) (m - F(X \beta^*))$$

$$= \left( \text{Var}_{MB}[F(X \beta)] V^{-1} - I_N \right) (m - F(X \beta^*)) \tag{126}$$

This shows that the bias incurred from a misspecified model is approximately a linear transformation of the difference between the true mean and the optimal user’s model fit. The transformation matrix is a measure of how much the product of the model-based variance matrix (see pg. 29) of the fits and the inverse model-based variance matrix of the responses differs from the identity matrix. A corresponding result can be found in Nottingham (1995, pg. 52) for the special case of uncorrelated data.

The variance of the $\eta$-scale fit for all data locations is

$$\left\{ \text{Var}^P(x_{ij}^*) \right\}_{N \times 1} = \text{vecdiag} \left( X B \left( \sum_{i=1}^s C_i \text{Var}[Y_i | C_i'] \right) B X' \right)$$

$$= \left\{ x_{ij} \left( \sum_{i=1}^s X_i^T W_i X_i \right)^{-1} \left( \sum_{i=1}^s X_i^T W_i \langle f_{ij} \rangle^{-1} \text{Var}[Y_i | \langle f_{ij} \rangle^{-1} W_i X_i \left( \sum_{i=1}^s X_i^T W_i X_i \right)^{-1} x_{ij} \right) \right\}_{N \times 1}$$

and the variance for the $\mu$-scale fit is

$$\left\{ \text{Var}^P(x_{ij}^*) \right\}_{N \times 1} = \langle f_{ij}^2 \rangle \left\{ \text{Var}^P(x_{ij}^*) \right\}_{N \times 1}, \tag{128}$$

98
recalling the notation \( \{b_{k\ell}\}_{r \times c} \), which denotes a \( r \times c \) dimensioned matrix by its \( k\ell^{th} \) element.

Note that \( \text{Var}[Y_i] \) can not be expressed as \( \phi^2 A_i^{1/2} R_i A_i^{1/2} \) because \( R_i \) may be incorrect. However, for uncorrelated data in the GLIM context, \( \text{Var}[Y_i] = A_i(m_i) \), and (128) reduces to

\[
\langle f_{ij} \rangle X \left( \sum_{i=1}^{s} X'_i W_i X_i \right)^{-1} \sum_{i=1}^{s} X'_i W_i \left( \frac{A(m(X_i))}{A(F(X_i \beta))} \right) X_i \left( \sum_{i=1}^{s} X'_i W_i X_i \right)^{-1} X' \langle f_{ij} \rangle
\]  

(129)

(see Appendix B.3 for details). This is analogous to Nottingham’s (1995, pg 52) variance expression and equal to it in the special case of quantal regression. Equation (129) further reduces to

\[
\langle f_{ij} \rangle X B X' \langle f_{ij} \rangle = \langle f_{ij} \rangle X \left( \sum_{i=1}^{s} X'_i W_i X_i \right)^{-1} X' \langle f_{ij} \rangle
\]  

(130)

if there is no mean model misspecification. This last term is the asymptotic variance-covariance matrix of the fits in a GLIM model (Table 3).

The bias expression for the nonparametric model at a single data location is the same as for the parametric model after replacing \( \beta^{*} \) with \( \theta^{*} \), \( W_i \) with \( W_{0i} \), and \( X \) with the nonparametric model matrix \( X_{0i} \), provided \( h \) is fixed. This duality is a consequence of embedding local models in a quasi-likelihood framework; the results will not be given. However, the bias and variance at all data locations can not be expressed in the same manner as (126) and (128), respectively, because the kernel weights and \( \theta^{*} \) may change with each prediction location. At the risk of being redundant, expressions for the bias and variance are shown in order to emphasize what quantities must change across the data points.

Let the local parameter estimate, \( \hat{\theta}(x_{ij}^{*}) \), be denoted by \( \hat{\theta}_{ij} \). Denote as \( \theta_{ij}^{*} = \theta^{*}(x_{ij}^{*}) \) the expansion locus of the Taylor series in the nonparametric model. Let the weight matrix \( W_{0k} \) for the \( k^{th} \) cluster, which contains the kernel weights centered at \( x_{ij}^{*} \) and other terms that depend on \( \theta_{ij}^{*} \), be denoted \( W_{0k,ij} \). Also, let \( \ell = 1, \ldots, n_i \) index the observations within a cluster. Then, the bias at the prediction locations can be formed by stacking the biases at the observed prediction locations. For the \( \eta \)-scale, the nonparametric bias is

\[
\text{Bias} \left[ \hat{\eta}^{NP} \right] \approx \left\{ E_{\eta}^{NP}(x_{ij}^{*}) \right\}_{N \times 1} - \eta \\
= \left\{ x_{0ij}^{'} \theta_{ij}^{*} \right\}_{N \times 1} - \eta \\
+ \left\{ x_{0ij}^{'} \left( \sum_{k=1}^{s} X_{0k} W_{0k,ij} X_{0k} \right)^{-1} \left( \sum_{k=1}^{s} X_{0k}' W_{0k,ij} f(x_{0k,ij}^{*}) \right)^{-1} (m_k - F(X_{0k} \theta_{ij}^{*})) \right\}_{N \times 1} \\
\]  

(131)

99
For the \( \mu \)-scale, the bias is

\[
\text{Bias} \left[ \hat{y}^{NP} \right] \approx \left\{ E^{NP}(x_{ij}^*) \right\}_{N \times 1} - m
\]

\[
= \left\{ F(x_{0ij}^* \theta_{ij}^*) \right\}_{N \times 1} - m
\]

\[
+ \left\{ f(x'_{0ij} \theta_{ij}^*) x_{0ij} \left( \sum_{k=1}^{s} X'_{0k} W_{0k,ij} X_{0k} \right)^{-1} \left( \sum_{k=1}^{s} X'_{0k} W_{0k,ij} f(x'_{0k,ij}) \right)^{-1} (m_k - F(X_{0k} \theta_{ij}^*)) \right\}_{N \times 1}
\]

which can not be expressed in the same form as (126).

The variance at all prediction locations for the nonparametric fits is also found by stacking. On the \( \eta \)-scale,

\[
\left\{ \text{Var} \left[ \hat{y}^{NP}(x_{ij}^*) \right] \right\}_{N \times 1} \approx \left\{ V^{NP}(x_{ij}^*) \right\}_{N \times 1}
\]

\[
= \left\{ x_{0ij} \left( \sum_{k=1}^{s} X'_{0k} W_{0k,ij} X_{0k} \right)^{-1} \left( \sum_{k=1}^{s} X'_{0k} W_{0k,ij} f(x'_{0k,ij}) \right)^{-1} \text{Var}[Y_k]
\]

\[
\times \left\{ f(x'_{0k,ij} \theta_{ij}^*) \right\}^{-1} W_{0k,ij} X_{0k} \left( \sum_{k=1}^{s} X'_{0k} W_{0k,ij} X_{0k} \right)^{-1} x_{0ij} \right\}_{N \times 1},
\]

and on the \( \mu \)-scale,

\[
\left\{ \text{Var} \left[ \hat{y}^{NP}(x_{ij}^*) \right] \right\}_{N \times 1} \approx \left\{ V^{NP}(x_{ij}^*) \right\}_{N \times 1}
\]

\[
= \left\{ f(x'_{0ij} \theta_{ij}^*) x_{0ij} \left( \sum_{k=1}^{s} X'_{0k} W_{0k,ij} X_{0k} \right)^{-1} \left( \sum_{k=1}^{s} X'_{0k} W_{0k,ij} f(x'_{0k,ij}) \right)^{-1} \text{Var}[Y_k]
\]

\[
\times \left\{ f(x'_{0k,ij} \theta_{ij}^*) \right\}^{-1} W_{0k,ij} X_{0k} \left( \sum_{k=1}^{s} X'_{0k} W_{0k,ij} X_{0k} \right)^{-1} x_{0ij} \right\}_{N \times 1}.
\]

These variance expressions can not be put into the same forms as (127) and (128) because every component, except the nonparametric model matrix, changes with the prediction point \( x_{ij}^* \).

Nottingham approached the local estimation problem by applying kernel and local linear regression directly to the sample proportions. His nonparametric estimator can be expressed as follows: Let \( \mathbf{p} \) be the vector of sample proportions at all data locations in a quantal regression. Let \( \mathbf{H}^{NP} \) be a local polynomial nonparametric smoother matrix. Then,

\[
\hat{\mathbf{p}}^{NP} = \mathbf{H}^{NP} \mathbf{p}
\]

100
This shows that Nottingham’s nonparametric fits are equivalent to GEE nonparametric fits in the special case \( F(\eta) = \eta \) (identity link), and \( V = \sigma^2 I \). Hence, his nonparametric bias and variance results are also special cases of (132) and (134), respectively.

The MSE for MRGEEI at \( x^* \) is given by (109) and developed in Appendix A.5 by minimizing over \( \lambda \). This will be re-examined to show how the MSE of MRGEEI and MRGEEII are related to the bias and variance of their respective parametric and nonparametric fits.

Rewrite \( m \) and \( \eta \) as convex combinations:

\[
\eta = (1 - \lambda) \eta + \lambda \eta, \\
m = (1 - \lambda) m + \lambda m,
\]

and write \( m \) as \( F(F^{-1}(m)) \). Then it can easily be shown that

\[
\text{Bias}[\hat{y}_{\text{MRGEEI}}] = (1 - \lambda) \text{Bias}[\hat{y}^P] + \lambda \text{Bias}[\hat{y}^\text{NP}],
\]

\[
\text{Bias}[\hat{\eta}_{\text{MRGEEII}}] = (1 - \lambda) \text{Bias}[\hat{\eta}^P] + \lambda \text{Bias}[\hat{\eta}^\text{NP}],
\]

and

\[
\text{Bias}[\hat{y}_{\text{MRGEEII}}] \approx \langle f(F^{-1}(m(x^*_{ij}))) \rangle \left\{ (1 - \lambda) \text{Bias}[\hat{\eta}^P] + \lambda \text{Bias}[\hat{\eta}^\text{NP}] \right\}.
\]

Consequently, the biases of the MRGEEI fits on the \( \mu \)-scale and the MRGEEII fits on the \( \eta \)-scale are convex combinations of the respective biases of the parametric and nonparametric components.

Let

\[
\mathbf{v}_{\text{MRGEEI}} = \left\{ \text{Var}[\hat{y}_{\text{MRGEEI}}(x^*_{ij})] \right\}_{N \times 1},
\]

\[
\mathbf{v}_{\text{MRGEEII}} = \left\{ \text{Var}[\hat{y}_{\text{MRGEEII}}(x^*_{ij})] \right\}_{N \times 1},
\]

101
\[ \begin{align*}
\mathbf{v}^P &= \{V^P(x^*_ij)\}_{N \times 1}, \\
\mathbf{v}^{NP} &= \{V^{NP}(x^*_ij)\}_{N \times 1}, \\
c &= \{C(x^*_ij)\}_{N \times 1}, \\
\mathbf{v}^{\text{MRGEEII}}_\eta &= \{\text{Var}[\hat{\eta}^{\text{MRGEEII}}(x^*_ij)]\}_{N \times 1}, \\
\mathbf{v}^P_\eta &= \{V^P(\mathbf{x}^*_ij)\}_{N \times 1}, \\
\mathbf{v}^{NP}_\eta &= \{V^{NP}(\mathbf{x}^*_ij)\}_{N \times 1}, \\
c_\eta &= \{C_\eta(x^*_ij)\}_{N \times 1}.
\end{align*} \]

Then,
\[ \mathbf{v}^{\text{MRGEEI}} = (1 - \lambda)^2 \mathbf{v}^P + \lambda^2 \mathbf{v}^{NP} + 2\lambda(1 - \lambda)\mathbf{c}, \]
\[ \mathbf{v}^{\text{MRGEEII}}_\eta = (1 - \lambda)^2 \mathbf{v}^P_\eta + \lambda^2 \mathbf{v}^{NP}_\eta + 2\lambda(1 - \lambda)\mathbf{c}_\eta, \]

and
\[ \mathbf{v}^{\text{MRGEEII}} \approx \left( f^2(F^{-1}(m(x^*_ij))) \right) \mathbf{v}^{\text{MRGEEII}}_\eta. \]

These quantities assume that \( \lambda \) and \( h \) are fixed.

The variance expressions for \( \mathbf{v}^{\text{MRGEEI}} \) and \( \mathbf{v}^{\text{MRGEEII}}_\eta \) reduce to the parametric variance of the fits (Table 6) when \( \lambda = 0 \) or to the nonparametric variance of the fits when \( \lambda = 1 \). \( \mathbf{v}^{\text{MRGEEII}} \) reduces in the same way if there is no misspecification when \( \lambda = 0 \) or if \( m(x^*_ij) = F(x'_0\theta) \) when \( \lambda = 1 \). Note however, that when there is mean model misspecification, the Taylor series expansion locus \( F^{-1}(m) \) is not the same as either of the loci \( x'\beta^* \) or \( x'_0\theta^* \) used to obtain the \( \mu \)-scale expressions in Table 7. For instance, this implies that the computed LLGEE MSE on the \( \mu \)-scale will not be exactly equal to the computed MRGEEII MSE on the \( \mu \)-scale when \( \lambda = 1 \). If a convex combination \( (1 - \lambda)x'\beta^* + \lambda x'_0\theta^* \) of the expansion loci is used in place of \( F^{-1}(m) \), then the \( \mu \)-scale AvgMSE of MRGEEII is based on

\[ \text{Bias}[\hat{y}^{\text{MRGEEII}}] \approx \left\{ F((1 - \lambda)x'_{ij}\beta^* + \lambda x'_{0ij}\theta^*_{ij}) \right\}_{N \times 1} - \mathbf{m} \]

\[ + \left\{ f( (1 - \lambda)x'_{ij}\beta^* + \lambda x'_{0ij}\theta^*_{ij}) \left[ (1 - \lambda)(E^P_{\eta}(x^*_ij) - x'_{ij}\beta^*) + \lambda(E^{NP}_{\eta}(x^*_ij) - x'_{0ij}\theta^*_{ij}) \right] \right\}_{N \times 1} \]

and
\[ \mathbf{v}^{\text{MRGEEII}} \approx \left( f^2((1 - \lambda)x'_{ij}\beta^* + \lambda x'_{0ij}\theta^*_{ij}) \right) \mathbf{v}^{\text{MRGEEII}}_\eta. \]
When $\lambda = 1$ these results give exactly the same bias and variance as that of LLGEE (132 and 134). We have also found that this approach works better in practice, over the full domain of $\lambda$. The MRGEEII estimator $(1 - \lambda)x'\hat{\beta} + \lambda x'_0\hat{\theta}$ tends to be closer to the Taylor series locus $(1 - \lambda)x'\beta^\star + \lambda x'_0\theta^\star$ than to $F^{-1}(m(x^\star))$, because it is of the same form.
7 Simulation Studies

To better understand, compare, and observe their practical performance, the methods developed in this work will be evaluated with simulation studies. This is done by generation of data from an assumed model and application of the estimation methods to the generated data. Statistical properties such as expectation, variance, and distribution can be approximated through simple statistics of the simulated outcomes.

Many assumptions and approximations have been made to obtain the expectation, variance, covariance, and MSE properties of the MRGEE estimators in §6. An extensive simulation study will show how well these approximations work, and will provide us with some guidelines as to when to use them. More importantly, we want to illustrate the improvement in MSE of the MRGEE methods over both the parametric and nonparametric methods and demonstrate how it carries through to the practical implementation of the methods. The simulation studies will also help identify situations in which the MRGEE methods perform well and when they do not.

7.1 Simulation Models

Data from a specified GLIM model can be easily generated from the model’s corresponding exponential family member distribution, however, generating data for GEE models is much more difficult. Zeger (1988) and Lindsey (1993) have suggested that it may not be possible to generate data for some marginal models at all. On first thought, this appears to be due to the generality of the GEE model; since only the first two moments are specified, it is not clear if a probability mechanism with such moments exists. Nevertheless, GEE models are useful because they provide consistent estimates of the parameters whether the correlation structure is specified correctly or not.

On second thought, there are other issues involved. Because the variance is assumed to be a function of the mean, it is hard to imagine a non-normal model that could have correlation structures as simple as CS or AR(1). Since the means depend on the regressors, so does the variance. One would expect the correlation to also depend on the regressors, unless the effect of the regressors on the variance and covariance can cancel in the correlation
calculation. Another issue is the lack of probability models for cluster correlated data, especially for categorical data. Formulating multivariate distributions in the simplest cases, like for correlated binary data, have proven to be quite cumbersome (see Bahadur, 1961). This problem provides motivation for using GEE since only the first two moments have to be specified, but it also makes it difficult to construct a data generating mechanism.

Moment matching is one technique that has been used in the literature (Gotway and Wolfinger, 2001). Data are generated from a well known multivariate distribution in such a way that, except for the correlation structure, the first and second moments match with the moments of the specified GEE model. For example, data for a “Poisson” GEE model that has an identity variance function can be generated from a multivariate normal distribution by simply setting the mean parameters equal to the variance parameters. However, such a method does not seem practically meaningful. The generated data are still correlated normal deviates, not correlated Poisson counts.

Our approach is to provide a plausible generating mechanism for correlated data that has moments similar to a generalized linear model. For this we use a generalized linear mixed model (GLMM) with a latent random effect in the linear predictor. The random effect varies from cluster to cluster, thus inducing marginal correlation through a latent variable process.

To continue the $\mu$-scale and $\eta$-scale development that we have followed, the functional form of the mean is contaminated in two ways: by adding a portion of a contamination function $G$ on the $\mu$-scale to formulate a $\mu$-scale contaminated simulation model or by adding a portion of a contamination function $g$ to the linear predictor to formulate an $\eta$-scale contaminated simulation model. The contamination functions do not contaminate the mean simultaneously; they are either both identically 0 (no contamination) or one is identically 0. It is believed that MRGEEI will perform best for $\mu$-scale contaminated models, while MRGEEII will be superior in the case of $\eta$-scale contamination.
General GLMM Simulation Model

\[ E[Y_{ij}|e_i] = F(x'_i\beta + g(x^*_i) + e_i) + G(x^*_i), \]
\[ e_i \overset{iid}{\sim} N(0, \sigma^2), \]
\[ Y_{ij}|e_i \overset{iid}{\sim} \text{Expfam}(E[Y_{ij}|e_i], a(E[Y_{ij}|e_i])), \]

\( \beta_S \) is the parametric model \( \beta \) (to be given later), and \( \text{Expfam} \) is an exponential family member distribution.

This probability model is similar to a split-plot linear model where the whole-plot error induces a compound symmetry correlation structure for observations within the same whole plot. Variability in \( Y_{ij} \) is introduced through both the exponential family distribution and the random effect. However, unlike the split-plot model, the correlation structure induced by a GLMM model is not necessarily compound symmetric, since the variances of the observations depend on the mean. Data can easily be created from this model by generating \( e_i \), computing \( E[Y_{ij}|e_i] \), and then generating \( Y_{ij}|e_i \) from an exponential family member distribution.

Properties of \( Y_{ij} \)

The unconditional properties of \( Y_{ij} \) can be determined via numerical methods or exactly in specific cases. Alternatively, a first order Taylor series approximation of \( F(x'_i\beta + g(x^*_i) + e_i) \) about \( E[e_i] = 0 \) can be used to obtain first order properties for arbitrary \( F \). Let \( g_i = \{g(x^*_i)\}_{n_i \times 1}, e_i = 1_{n_i}e_i, \) and \( G_i = \{G(x^*_i)\}_{n_i \times 1} \). Then,

\[ E[Y_i] \approx F(X_i\beta + g_i) + G_i, \]

and,

\[ \text{Var}[Y_i] \approx A(F(X_i\beta + g_i) + G_i) + \sigma^2 \left< f(x'_i\beta + g(x^*_i)) \right>, \]

where \( A(\cdot) = \langle a(\cdot) \rangle \). The derivation is shown in Appendix C.1.

For example, a Poisson simulation model with log link and Gaussian random effect assumes
\[ E[Y_{ij} | e_i] = e^{\eta_{ij} + g_{ij} + e_i} + G_{ij}, \]
\[ e_i \overset{iid}{\sim} N(0, \sigma^2), \]
\[ Y_{ij} | e_i \overset{iid}{\sim} \text{Poisson } (E[Y_{ij} | e_i]), \]

where \( \eta_{ij} = x'_{ij} \beta_S, \) \( g_{ij} = g(x'_{ij}), \) and \( G_{ij} = G(x'_{ij}). \) The exact expectation and variance for this model are obtained as

\[ E[Y_i] = e^{\sigma^2/2} e^{\eta_i + g_i} + G_i, \]

and

\[ \text{Var}[Y_i] = \left(e^{\sigma^2/2} e^{\eta_i + g_i} + G_i \right) + e^{\sigma^2} \left(e^{\sigma^2} - 1 \right) \left(e^{\eta_i + g_i + \eta_k + g_k} \right)_{n_i \times n_i}, \]

as shown in Appendix C.2. These are similar in form to (138) and (139). At \( \sigma^2 = 0.55^2, \) the largest random effect variance that will be used in the simulations involving this model, the exact mean is \( E[Y_i] = 1.1633 e^{\eta_i + g_i} + G_i, \) which is close to the approximate mean \( E[Y_i] \approx e^{\eta_i + g_i} + G_i. \) The exact variance, \( \text{Var}[Y_i] = \langle 1.1633 e^{\eta_i + g_i} + G_{ij} \rangle + 0.4780 \{ e^{\eta_i + g_i + \eta_k + g_k} \}_{n_i \times n_i}, \) is also close to the approximate variance,

\[ \text{Var}[Y_i] \approx \langle e^{\eta_i + g_i} + G_{ij} \rangle + 0.3025 \{ e^{\eta_i + g_i + \eta_k + g_k} \}_{n_i \times n_i}. \]

This comparison gives us some confidence that the approximate moments will be reasonable. We will use these exact moments for the Poisson model because they are as easy to calculate as the approximate moments.

The variance expressions show that the correlation structure for an arbitrary model is not of any of the common forms that are typically used in GEE estimation. Only if the regressors are cluster-specific, so that the cross products of the derivatives \( f(x'_{ij} \beta) \) in the approximate variance expression are constant within clusters, does \( \text{Var}[Y_i] \) appear to have an approximate compound symmetric correlation structure. However, the correlation changes from cluster to cluster, which is not a common assumption of GEE. We are not overly concerned with the form of the correlation structure in the simulations, as GEE’s permit such misspecification. However, the correlation matrix should be preserved across the clusters, which can be accomplished by using the same design points in every cluster.

The overall magnitude of the correlation can be controlled by the random effect variance \( \sigma^2. \) Large values of \( \sigma^2 \) produce large correlations within the cluster. In §7.2 we show how \( \sigma \) relates to \( E[\hat{\rho}_{CS}] \) and \( E[\hat{\rho}_{AR1}], \) which illustrates \( \sigma \)'s relationship with the MOM estimators.
7.2 Setup of the Simulations

The simulations are restricted to two model types—Poisson and binomial, which are commonly used to model count data. To reduce the complexity of the simulation study, only a single regressor \((x)\) is considered. The design is fixed and uniform over \([0, 1]\). A parameter \(\gamma \in [0, 1]\) controls the degree of mean model misspecification. The impact of \(\gamma\) on the variance-covariance matrix of the responses is lessened by constructing contamination functions that have 0 functional averages over \([0, 1]\). The random effect is Gaussian distributed.

First, the \(\mu\)-scale contaminated models are presented \((g \equiv 0)\).

### \(\mu\)-Scale Poisson Model

\[
E[Y_{ij}|e_i] = e^{3+2x_{ij}-2x_{ij}^2+e_i} + \gamma 5 \sin(4\pi x_{ij}),
\]

\(e_i \sim iid N(0, \sigma^2)\),

\(Y_{ij}|e_i \sim Poisson (E[Y_{ij}|e_i])\),

which has \(E[Y_{ij}] = e^{\sigma^2/2}e^{3+2x_{ij}-2x_{ij}^2} + \gamma 5 \sin(4\pi x_{ij})\) and,

\[
Var[Y_i] = \langle E[Y_{ij}] \rangle + e^{\sigma^2} \left( e^{\sigma^2} - 1 \right) \left\{ e^{6+2x_{ij}-2x_{ij}^2+2x_{ij}-2x_{ij}^2} \right\}_{n_i \times n_i}.
\] (140)

The conditional mean involves the inverse of the log link function (canonical) with a sinusoidal contamination on the \(\mu\) scale. Plots of \(E[Y_{ij}|e_i = 0] \approx E[Y_{ij}]\) various \(\gamma\) are shown in Figure 7. Notice that the quadratic behavior of the linear predictor is present also on the mean scale. Also, the random effect can have a profound effect on the conditional mean, though this cannot be observed in Figure 7.

### \(\mu\)-Scale Binomial Model

\[
E[Y_{ij}|e_i] = 50 \left( 1 + e^{-(x_{ij}-0.5)/(0.2+e_i)} \right)^{-1} + \gamma (50)(-1.2)(x_{ij} - 0.5) \left( 1 - \left( \frac{x_{ij} - 0.5^2}{0.5} \right) \right)^2,
\]

\(e_i \sim iid N(0, \sigma^2)\),

\(Y_{ij}|e_i \sim Binomial (50, E[Y_{ij}|e_i]/50)\),

which has \(E[Y_{ij}] \approx E[Y_{ij}|e_i = 0]\) and

\[
Var[Y_i] \approx \langle E[Y_{ij}|e_i = 0]E[Y_{ij}|e_i = 0]/50 \rangle.
\]
\[ + \sigma^2 \left\{ \frac{50^2 e^{-(x_{ij}+x_{ik}-1)/0.2}}{(1 + e^{-(x_{ij}+x_{kj}-0.5)/0.2})^2(1 + e^{-(x_{ij}+x_{ik}-0.5)/0.2})^2} \right\}_{n_i \times n_i} \]  

(141)

for small \( \sigma^2 \). The binomial counts are out of 50. This bound is larger than we had hoped to use—Nottingham (1995) had counts out of 10, but 50 was the smallest bound on the counts that would allow the GEE models to converge often (>95% of the time). Convergence failure is due to the large variability in the counts produced by even moderate values of \( \sigma \). A large bound on the count reduces the variability of the proportions to a practical level.

The conditional mean under no misspecification is based on the logistic CDF or inverse logit (canonical). The contamination function,

\[ \gamma(50)(-1.2)(x_{ij} - 0.5) \left( 1 - \left( \frac{x_{ij} - 0.5^2}{0.5} \right) \right)^2, \]

is known as a **bisquare** function in M-estimation. Use of this function in \( E[Y_{ij}|e_i] \) was found to produce mean function profiles nearly identical to the mixtures of logistic CDFs used by Nottingham (1995). His models were not directly transferable to our context because the level of misspecification \( (\gamma) \) effected the cluster correlations too strongly. Several plots of \( E[Y_{ij}|e_i = 0]/50 \approx E[Y_{ij}]/50 \) are shown in Figure 8. These show the mean proportion rather than the mean count out of 50.

The \( \eta \)-scale contaminated models are the following.

**\( \eta \)-Scale Poisson Model**

\[ E[Y_{ij}|e_i] = e^{(1.5)x_{ij}+2+\gamma(0.4)\sin(4\pi x_{ij}+0.5)+e_i}, \]

\[ e_i \sim iid \quad N(0, \sigma^2), \]

\[ Y_{ij}|e_i \sim Poisson (E[Y_{ij}|e_i]), \]  

(142)

which has \( E[Y_{ij}] = e^{\sigma^2/2}E[Y_{ij}|e_i = 0] \) and,

\[ Var[Y_i] = e^{\sigma^2/2} \langle E[Y_{ij}|e_i = 0] \rangle + e^{\sigma^2} \left( e^{\sigma^2} - 1 \right) \{ E[Y_{ij}|e_i = 0|E[Y_{ik}|e_i = 0] \}_{n_i \times n_i} \]

\[ = \langle \mu_i \rangle + \left( e^{\sigma^2} - 1 \right) \mu_i \mu_i', \]  

(143)

where \( \mu_i = \{ E[Y_{ij}] \}_{n_i \times 1} \). The variance-covariance matrix can be written as a function of the mean because the \( \eta \)-scale contamination function is part of the variance calculation. Note that the \( \mu \)-scale variance-covariance matrices can not be written as functions of the mean
because the contamination function $G$ drops out of the variance calculation in the 2$^{nd}$ term. The effect of the sinusoidal function on $E[Y_{ij}|e_i] \approx E[Y_{ij}]$ can be seen in Figure 9 to be no longer additive. The differences are more extreme between the curves at different $\gamma$ on the right side of the graph.

$\eta$-Scale Binomial Model

$$E[Y_{ij}|e_i] = 50 \left(1 + e^{-(5x_{ij} - 2.5 + \gamma(0.6)\sin(4\pi x_{ij} + 0.5)+e_i)}\right)^{-1},$$

$$e_i \overset{iid}{\sim} N(0, \sigma^2),$$

$$Y_{ij}|e_i \overset{ind}{\sim} \text{Binomial}\ (50, E[Y_{ij}|e_i]/50),$$

which has, for small $\sigma^2$, $E[Y_{ij}] \approx E[Y_{ij}|e_i = 0]$ and

$$Var[Y_i] \approx \langle E[Y_{ij}|e_i = 0](1 - E[Y_{ij}|e_i = 0]/50) \rangle + \sigma^2 \{E[Y_{ij}|e_i = 0](1 - E[Y_{ij}|e_i = 0]/50)E[Y_{ik}|e_i = 0](1 - E[Y_{ik}|e_i = 0]/50)\}_{n_i \times n_i}$$

$$\approx \langle \mu_i \rangle \langle 1_{n_i} - \mu_i/50 \rangle + \sigma^2 \langle \mu_i \rangle \langle 1_{n_i} - \mu_i/50 \rangle 1_{n_i} \langle 1_{n_i} - \mu_i/50 \rangle,$$

which is the approximate variance-covariance matrix as a function of the mean. As with the $\mu$-scale binomial model (under no misspecification), the conditional mean has the form of a logistic CDF. The mean proportions for several values of $\gamma$ are displayed in Figure 10. Unlike the $\eta$-scale Poisson model, sinusoidal contamination function here appears to act additively on the $\mu$-scale.

For the majority of the simulations, the misspecification levels are $\gamma = 0$, 0.25, 0.5, and 1. A finer scale is used for smaller $\gamma$ because it is believed that this is where the MRGEE methods are most effective—at a slight level of misspecification. This is in fact demonstrated in the results. We set the user’s parametric mean model to be the conditional mean $E[Y_{ij}|e_i = 0]$, which is only approximately equal to the unconditional mean. This has a very important implication: the parametric model used in $\gamma = 0$ cases is not the true model, it is the least misspecified model. Thus, $\gamma = 0$ only implies that the mean model is correct to a first order approximation about $E[e_i] = 0$. When $\sigma^2 = 0$, i.e., when the data are uncorrelated, the user’s parametric model is completely correct. At higher levels of correlation, the first order approximation becomes poorer implying that the meaning of $\gamma = 0$ changes somewhat with
the level of correlation. Hence, from this point forward, “no misspecification” will mean no misspecification of a first order approximation of the unconditional mean. We work with the assumption that the user is naïve about the true form of the mean model and uses a common mean model from a generalized linear model.

Since we are primarily interested in small sample properties, the number of observations per cluster is kept small. However, there must be enough design points to capture the features illustrated in Figures 7–10. Unless specified otherwise, ten uniformly spaced design points, \( x = 0, \frac{1}{9}, \frac{2}{9}, \ldots, 1 \), are used for all clusters. The number of clusters is set at two levels. A “large” sample size of \( s = 30 \) clusters, which corresponds to the minimum sample size rule-of-thumb for asymptotic distribution theory (§2.2.5) to be reasonably accurate, and a “small” sample size \( \frac{1}{10} \) as large, \( s = 3 \). Three clusters is less than what should be used in practice (since \( 3 = s < n = 10 \)), but we want to illustrate the effect of “very small” samples. These settings apply to the vast majority of the simulations, but variations follow as specific effects become of interest.
Figure 7: $\mu$-Scale Poisson Simulation Model for $\gamma = 0, 0.1, \ldots, 1$

Figure 8: $\mu$-Scale Binomial Simulation Model for $\gamma = 0, 0.1, \ldots, 1$
Figure 9: $\eta$-Scale Poisson Simulation Model for $\gamma = 0, 0.1, \ldots, 1$

Figure 10: $\eta$-Scale Binomial Simulation Model for $\gamma = 0, 0.1, \ldots, 1$
An unfortunate consequence of using a latent process model for the simulations is that the marginal variability of the responses increases with the magnitude of the correlations as \(\sigma\) increases. This implies that the effect of the level of correlation is confounded with the effect of the magnitude of the variance. Nevertheless, it is still interesting to determine how the correlation/variability (corr/var) effects our new methods.

An overall measure of the level of correlation would be useful, because the correlations vary within the true correlation matrix. Rather than computing some summary metric of the correlation matrix, we choose to study the limiting expected values of the CS and AR(1) MOM correlation estimators. A large limit implies the model correlations are overall “large” and likewise for small limits\(^5\). This way, we are assured that our large corr/var setting will not only produce data with “large” correlations, but also produce a large MOM correlation estimate. The size of the MOM correlation estimate is an issue because small estimates have little impact on the converged GEE fits.

The expected values for \(\hat{\rho}_{CS}\) and \(\hat{\rho}_{AR1}\) are approximated in Appendix B.2 and their limits as \(s \to \infty\) are shown in Table 8. The limits are computed in Appendix C.3, based on the assumption that the clusters are iid, which holds for these simulation models. They are shown for the four simulation models in Figures 11–14. The limits for the Poisson models under no misspecification are the most accurate. The limits for the binomial models under no misspecification rely on a first order approximation of the variance, (139). All other limits rely on the same approximation and a first order Taylor series approximation of the MOM estimators. Vertical reference lines on the plots show the settings of \(\sigma\) that we choose for low (0.1 < \(E[\hat{\rho}_{CS}]\), \(E[\hat{\rho}_{AR1}] < 0.4\)) and high correlation (0.6 < \(E[\hat{\rho}_{CS}]\), \(E[\hat{\rho}_{AR1}] < 0.9\)). These values are displayed in Table 9. Appendix C.4 shows the correlation matrices for the \(\sigma > 0\) settings at no misspecification (\(\gamma = 0\)). At other degrees of misspecification (0 < \(\gamma \leq 1\)), the correlation matrices are nearly the same as the respective correlation structures under no misspecification for all simulation models.

Structures for the Poisson models appear to be almost compound symmetric. The binomial models have correlation matrices with the following properties. As distance between the

\(^5\)“Large” is placed in quotes because we do not imply that every correlation in the correlation matrix is large, but rather that the correlations are large on average.
prediction locations increases, correlation decreases (similar to AR(1)). Pairs of responses at prediction locations near the midpoint of the $x$ space have the largest correlation, which decreases towards pairs near the edges of the $x$ space.

Table 8: Expectation Limits MOM Estimators Applied to GLMM Data

<table>
<thead>
<tr>
<th>Model</th>
<th>Approximate Expectation Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\phi}^2$</td>
<td>$\frac{1}{n} \sum_{j=1}^{n} \frac{a(F(x'<em>{ij}\beta^*))+\sigma^2 f^2(x'</em>{ij}\beta^<em>)+(m(x'<em>i)-F(x'</em>{ij}\beta^</em>))^2}{a^<em>(F(x'_{ij}\beta^</em>))} = \frac{\gamma_\phi}{n}$</td>
</tr>
<tr>
<td>$\hat{\rho}_{CS}$</td>
<td>$\frac{1}{2\gamma_\phi(n-1)} \sum_{j&lt;k}^{n} \frac{a^2 f(x'<em>{ij}\beta^*) f(x'</em>{ik}\beta^<em>) + (m(x'<em>i)-F(x'</em>{ij}\beta^</em>)) (m(x'<em>k)-F(x'</em>{ik}\beta^<em>))}{a^</em>(F(x'<em>{ij}\beta^<em>)) a^</em>(F(x'</em>{ik}\beta^<em>))} \sqrt{a^</em>(F(x'<em>{ij}\beta^<em>)) a^</em>(F(x'</em>{ik}\beta^*))}$</td>
</tr>
<tr>
<td>$\rho_{AR1}$</td>
<td>$\frac{n}{\gamma_\phi(n-1)} \sum_{j=1}^{n-1} \frac{a^2 f(x'<em>{ij}\beta^*) f(x'</em>{i,j+1}\beta^<em>) + (m(x'<em>i)-F(x'</em>{ij}\beta^</em>)) (m(x'<em>{i,j+1})-F(x'</em>{i,j+1}\beta^<em>))}{a^</em>(F(x'<em>{ij}\beta^<em>)) a^</em>(F(x'</em>{i,j+1}\beta^<em>))} \sqrt{a^</em>(F(x'<em>{ij}\beta^<em>)) a^</em>(F(x'</em>{i,j+1}\beta^*))}$</td>
</tr>
</tbody>
</table>

Table 9: Settings of $\sigma$ for No, Low, and High Correlation

<table>
<thead>
<tr>
<th>Model</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>None</td>
</tr>
<tr>
<td>$\mu$-Scale Poisson</td>
<td>0</td>
</tr>
<tr>
<td>$\mu$-Scale Binomial</td>
<td>0</td>
</tr>
<tr>
<td>$\eta$-Scale Poisson</td>
<td>0</td>
</tr>
<tr>
<td>$\eta$-Scale Binomial</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure 11: Correlation Expected Value Limits for $\mu$-Scale Poisson at $\gamma = 0, 0.2, \ldots, 1$

Figure 12: Correlation Expected Value Limits for $\mu$-Scale Binomial at $\gamma = 0, 0.2, \ldots, 1$
Figure 13: Correlation Expected Value Limits for $\eta$-Scale Poisson at $\gamma = 0, 0.2, \ldots, 1$

Figure 14: Correlation Expected Value Limits for $\eta$-Scale Binomial at $\gamma = 0, 0.2, \ldots, 1$
7.3 Efficiency of Local GEE Predictions when Using Non-Identity Working Structures

Lin and Carroll (2000) showed for local linear GEE estimation that (79), the asymptotic variance, is minimized when the identity working structure is used \((R = I)\). This holds even if the true correlation structure is known. They claim that for large \(s\) and small \(h\) such that \(sh\) is large, there is nothing to be gained by accounting for the correlations within the clusters. This counters the parametric GEE case, where it has been shown that the correlation structure, especially the correct one, improves efficiency of the fits over using the independence working structure. Lin and Carroll’s (2000) result is valid, but we wonder if it holds for small samples where asymptotic expressions are not often highly accurate.

A simulation study was carried out on the \(\eta\) scale models to address this issue. The local linear GEE estimates were based on the IND (independence), CS, and AR(1) structures and were summarized in terms of AvgMSE in Table 10. The squared bias and variance components of the AvgMSE decomposition are shown in Table 11. The variance functions used for these simulation models are of the correct form (see §7.2), but the true mean functions are highly misspecified \((\gamma = 1)\) from their parametric forms. The optimal bandwidths were found using the equations developed in §6 for each model, corr/var level (for \(\sigma > 0\)), \(s\), and working structure. Since they differed only slightly across the correlation structures, only the optimal bandwidths for the IND structure are given in Table 10. The AvgMSEs for the Binomial models are much smaller than those of the Poisson models. We scale them by \(10^5\) here and in every table that follows in the remainder of this chapter. These simulations are based on 1,000 runs.

We make the following observations based on these results.

1. Of the IND, CS, and AR(1) structures, IND is least similar to the true correlation structures given in Appendix C.4. Yet, the Monte Carlo average variance (AvgVar) is minimized for the IND (Independence) working structure in nearly all cases (Table 11). The exception is the \(\eta\)-Scale Binomial, \(s = 3\), high corr/var case, which sees a 4% increase in efficiency for CS over the IND structure. This case indicates that in small samples, it may be possible to improve efficiency in a local linear GEE by utilizing the
correlation structure.

2. The variance efficiency appears to depend on how close the working structure is to the true structure. In Appendix C.4, it was shown that the Poisson model’s true correlation structure is nearly CS and the binomial model’s true structure is closer to a CS than an AR(1) structure. This is in agreement with Table 11 which shows that CS is more efficient than AR(1).

3. The correlation structure does not effect the asymptotic bias (78) for local linear GEE, but it appears in Table 11 that the Monte Carlo average squared bias (AvgBias$^2$) is minimized for the CS structure in most cases. That is, utilizing the correlation structure in LLGEE can improve the bias for small samples, at least in these instances.

4. Like the average variance, the structure that improves the average bias the most is the CS structure, which is closest to the true correlation structure among IND, CS, and AR(1).

5. The AvgMSE shown in Table 10 additively combines the average squared bias, of which CS performs the best, and the average variance, of which IND performs the best. As a consequence, neither structure uniformly outperforms the other. In fact, the AvgMSE efficiency of the CS structure over the IND structure is 1 or almost 1 in all cases. However, the CS structure does appear to lead to better AvgMSE performance in most of the binomial model cases, while the IND structure is superior in the Poisson model cases.

6. There is a significant problem with the convergence failure rate (CFR), or the percentage of simulation runs (out of 1,000) that fail to converge, for the methods that use non-identity working structures. Failure rates as high as 23.8% are observed here for the small sample binomial with high corr/var. The convergence failure rates were found to be even worse for IMSE computations, where CFRs as high as 40.3% occurred. The convergence of the IRLS algorithm (pp. 12–13) at non-design points is difficult when correlation structures are involved. Modifications to improve the IRLS
algorithm included “1/2 stepping,” used in nonlinear regression estimation, and “subiteration steps,” where the estimate of the correlation structure is held constant for a number of iterations. These techniques did not significantly improve the convergence failure rates.

In summary, the optimal working structure result of Lin and Carroll (2000) works quite well in the cases studied here even though the sample sizes are small. However, the correlation structure does appear to play a role in the bias of the fits which implies that the MSE can be improved for some cases by using the correct correlation structure in small samples. However, for these simulation models overall, the IND structure is AvgMSE competitive with non-identity structures and is nearly immune to convergence failure, which is a significant problem for CS and AR(1). Also, using IND implies fast LLGEE estimation which is critical for local estimation. For these reasons, we use LLGEE based on the IND structure for the remainder of the simulations in this chapter.

Table 10: AvgMSE Comparison of CS and AR(1) Working Structures to the Independence Working Structure for Local Linear GEE Estimation of the $\eta$-scale Models. $\gamma = 1$ for all cases.

<table>
<thead>
<tr>
<th>Model</th>
<th>Corr/Var</th>
<th>$s$</th>
<th>$h_{opt}$</th>
<th>AvgMSE†</th>
<th>MSE Efficiency</th>
<th>CFR* (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AvgMSE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>IND</td>
<td>CS AR(1)</td>
<td>IND CS AR(1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CS</td>
<td>AR(1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>IND</td>
<td>CS AR(1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CS</td>
<td>AR(1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>IND</td>
<td>CS AR(1)</td>
<td></td>
</tr>
<tr>
<td>Poisson</td>
<td>Low</td>
<td>3</td>
<td>0.0553</td>
<td>8.59</td>
<td>8.61 11.06</td>
<td>1.00 0.78</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>3</td>
<td>0.0545</td>
<td>72.06</td>
<td>72.15 78.54</td>
<td>1.00 0.92</td>
</tr>
<tr>
<td>Poisson</td>
<td>Low</td>
<td>30</td>
<td>0.0385</td>
<td>0.96</td>
<td>0.96 1.22</td>
<td>1.00 0.79</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>30</td>
<td>0.0383</td>
<td>7.43</td>
<td>7.44 7.79</td>
<td>1.00 0.95</td>
</tr>
<tr>
<td>Binomial</td>
<td>Low</td>
<td>3</td>
<td>0.0550</td>
<td>14.93</td>
<td>14.78 21.49</td>
<td>1.01 0.69</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>3</td>
<td>0.0539</td>
<td>65.43</td>
<td>62.81 66.61</td>
<td>1.04 0.98</td>
</tr>
<tr>
<td>Binomial</td>
<td>Low</td>
<td>30</td>
<td>0.0383</td>
<td>1.74</td>
<td>1.73 2.31</td>
<td>1.00 0.75</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>30</td>
<td>0.0381</td>
<td>10.58</td>
<td>10.53 10.95</td>
<td>1.00 0.97</td>
</tr>
</tbody>
</table>

† The AvgMSE for the binomial cases are multiplied by 10^5.

* CFR = Convergence Failure Rate is the percentage of runs that did not converge.
Table 11: AvgBias² and AvgVar Comparison of CS and AR(1) Working Structures to the Independence Working Structure for Local Linear GEE Estimation of the $\eta$-scale Models. $\gamma = 1$ for all cases.

<table>
<thead>
<tr>
<th>Model</th>
<th>Corr/Var</th>
<th>$s$</th>
<th>$h_{opt}$</th>
<th>AvgBias²†</th>
<th>AvgVar†</th>
<th>Var Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>IND CS AR(1)</td>
<td>IND CS AR(1)</td>
<td>CS AR(1)</td>
</tr>
<tr>
<td>Poisson</td>
<td>Low</td>
<td>3</td>
<td>0.0553</td>
<td>0.67 0.51 0.50</td>
<td>7.92 8.10 10.57</td>
<td>0.98 0.75</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>3</td>
<td>0.0545</td>
<td>0.83 0.52 0.72</td>
<td>71.23 71.63 77.82</td>
<td>0.99 0.92</td>
</tr>
<tr>
<td>Poisson</td>
<td>Low</td>
<td>30</td>
<td>0.0385</td>
<td>0.01 0.01 0.15</td>
<td>0.94 0.95 1.07</td>
<td>0.99 0.88</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>30</td>
<td>0.0383</td>
<td>0.02 0.01 0.23</td>
<td>7.41 7.43 7.57</td>
<td>1.00 0.98</td>
</tr>
<tr>
<td>Binomial</td>
<td>Low</td>
<td>3</td>
<td>0.0550</td>
<td>1.40 1.08 1.09</td>
<td>13.53 13.70 20.40</td>
<td>0.99 0.66</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>3</td>
<td>0.0539</td>
<td>5.88 5.74 5.88</td>
<td>59.55 57.08 60.73</td>
<td>1.04 0.98</td>
</tr>
<tr>
<td>Binomial</td>
<td>Low</td>
<td>30</td>
<td>0.0383</td>
<td>0.11 0.08 0.40</td>
<td>1.63 1.66 1.92</td>
<td>0.98 0.85</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>30</td>
<td>0.0381</td>
<td>4.39 4.30 4.47</td>
<td>6.20 6.24 6.48</td>
<td>0.99 0.96</td>
</tr>
</tbody>
</table>

† The AvgBias² and AvgVar for the binomial cases are multiplied by 10⁵.

7.4 Validation of the MSE Expressions

Simulation studies are performed in an effort to support our claims in Chapter 6 that the MSE computations are “reasonably accurate.” This is done in part by validating the “reasonable” accuracy of the first order approximations to the MOM estimators expected values, and by showing that the Taylor series expansion locus, $\beta^\star$, is close to $E[\hat{\beta}]$. Extensive validation for the $\mu$-Scale simulation models is illustrated by comparisons of calculated AvgMSE to simulated AvgMSE and calculated optimal bandwidths to optimal bandwidths. When we mention that AvgMSE is calculated or computed we mean that it has been obtained by the expressions given in Chapter 6. Unlike the AvgMSEs obtained by simulation, which we may call simulated AvgMSEs, the computed AvgMSEs depend only on the properties of the models involved and are not random. The same terminology will be used to indicate how $h_{opt}$ and $\lambda_{opt}$ are obtained.
7.4.1 Closeness of $\beta^\star$ to $E[\hat{\beta}]$

Ideally the expansion center of the Taylor series in (114) would be close to $E[\hat{\beta}]$ so that the average deviation of the approximation from the true value would be small. This comparison is performed for the parametric GEE model using the CS structure on the four simulation models under high degrees of misspecification ($\gamma = 1$). The high misspecification level was chosen because under no misspecification, the true simulation mean model and the user’s parametric mean model are the same, and as such, $\beta^\star$ is the parameter vector for the true model.

Data are generated from the four models 1,000 times for small ($s = 3$) and large ($s = 30$) sample sizes and small and large levels of corr/var. The parameter estimates, $\hat{\beta}$, are averaged over these runs to determine a Monte Carlo estimate of $E[\hat{\beta}]$, denoted $\hat{E}[\hat{\beta}]$. This estimate is compared to the computed value $\beta^\star$ (see §6.1.2) by the vector relative error (RE) measure

$$RE = \frac{||\beta^\star - \hat{E}[\hat{\beta}]||^{1/2}}{||E[\hat{\beta}]||^{1/2}}.$$ 

The following comments are made.

1. The results in Table 12 show that for most cases $E[\hat{\beta}]$ is close to $\beta^\star$, even for small samples. The greatest discrepancies occur for the $\mu$-Binomial, $s = 30$, corr/var=Large case (11.4%) and the $\eta$-Binomial, $s = 30$, corr/var=Large case (10.5%). The respective vectors for these cases are approximately

$$\hat{E}[\hat{\beta}] = [-1.54 \ 3.09]' \quad \beta^\star = [-1.72 \ 3.44]'$$

and

$$\hat{E}[\hat{\beta}] = [-2.16 \ 4.22]' \quad \beta^\star = [-2.38 \ 4.66]' .$$

These differences are not excessive. The inaccuracy could possibly be due to the approximation of the MOM expected values via a Taylor series, which becomes less accurate as corr/var ($\sigma^2$) increases. For the most part, this argument seems to be supported by the other results in the table.
Table 12: Comparison of $E[\hat{\beta}]$ to $\beta^*$. For all cases $\gamma=1$.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Model</th>
<th>$s$</th>
<th>Corr/Var</th>
<th>RE† (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>Poisson</td>
<td>3</td>
<td>Small</td>
<td>1.0</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Poisson</td>
<td>30</td>
<td>Small</td>
<td>0.3</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Binomial</td>
<td>3</td>
<td>Small</td>
<td>0.5</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Binomial</td>
<td>30</td>
<td>Small</td>
<td>1.1</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Poisson</td>
<td>3</td>
<td>Small</td>
<td>0.4</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Poisson</td>
<td>30</td>
<td>Small</td>
<td>0.1</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Binomial</td>
<td>3</td>
<td>Small</td>
<td>0.5</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Binomial</td>
<td>30</td>
<td>Small</td>
<td>1.2</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Poisson</td>
<td>3</td>
<td>Large</td>
<td>1.0</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Poisson</td>
<td>30</td>
<td>Large</td>
<td>0.1</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Binomial</td>
<td>3</td>
<td>Large</td>
<td>8.2</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Binomial</td>
<td>30</td>
<td>Large</td>
<td>11.4</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Poisson</td>
<td>3</td>
<td>Large</td>
<td>2.2</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Poisson</td>
<td>30</td>
<td>Large</td>
<td>0.1</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Binomial</td>
<td>3</td>
<td>Large</td>
<td>6.9</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Binomial</td>
<td>30</td>
<td>Large</td>
<td>10.5</td>
</tr>
</tbody>
</table>

† RE is the vector relative error of $\beta^*$ from $E[\hat{\beta}]$.

2. The RE is larger for larger corr/var except for the $\mu$-Poisson, $s = 30$ case, in which the RE drops slightly. Again, this increase in RE is probably due to degradation of accuracy of the MOM Estimators Expected values with increases in corr/var.

3. For the Poisson models, larger samples have smaller RE. Unexpectedly, the opposite effect occurs for binomial models. This may be due to the use of approximate moments for the binomial simulation model. The moments used for the Poisson cases are exact.

In conclusion, the Taylor series locus $\beta^*$ seems to be reasonable in most cases. Unfortunately, the $\mu$- and $\eta$-scale binomial, high corr/var cases may lead to less accurate MSE computations than the other cases.
7.4.2 Validation of Correlation Parameter Expected Values

The correlation expected values are somewhat crude first order approximations of the true expectations. Inaccuracies in these expectations lead to inaccurate weights in $V$ and $V_0$ which can effect $\beta^*$, $\theta^*$, and the bias and variance of the parametric and nonparametric models when computed by the expressions in Chapter 6. A simulation study is carried out to compare the expected value limits for the correlation parameter estimators to Monte Carlo expected value limits.

The true limits are approximated by averaging over the MOM estimates based on fitting parametric GEE’s to 30 sets of $s = 1,000$ simulated clusters. Given the variability ($< 0.0004$) of the estimates, 30 runs was determined to be sufficient in order to find $E[\hat{\rho}_{CS}]$ and $E[\hat{\rho}_{CS}]$ to within 2 decimal places. Thus the Monte Carlo estimates, denoted $\hat{E}[\hat{\rho}_{CS}]$ and $\hat{E}[\hat{\rho}_{CS}]$, should be close to the limiting expected values. For all combinations, $\gamma = 1$.

Table 13: Comparison of First Order Approximations of Expected Value MOM CS and AR(1) Limits to Expected Values Obtained Through Simulation. For all cases $\gamma = 1$.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Model</th>
<th>Corr/Var</th>
<th>Sim. Comp.</th>
<th>Sim. Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>Poisson</td>
<td>Small</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Poisson</td>
<td>Large</td>
<td>0.78</td>
<td>0.78</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Binomial</td>
<td>Small</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Binomial</td>
<td>Large</td>
<td>0.54</td>
<td>0.57</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Poisson</td>
<td>Small</td>
<td>0.04</td>
<td>0.09</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Poisson</td>
<td>Large</td>
<td>0.51</td>
<td>0.66</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Binomial</td>
<td>Small</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Binomial</td>
<td>Large</td>
<td>0.62</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Table 13 shows that the first order limits are certainly reasonable. In a majority of cases, they agree within 0.01. The greatest disagreements occur in the large corr/var cases, which is to be expected. These approximations are based on Taylor series expansions about $E[e_i] = 0$ that degrade as $\text{Var}[e_i] = \sigma^2$ increases. The approximation also appears to break down for the $\eta$-Poisson, large corr/var case, but for unknown reasons. Inaccuracies that occur when using the expressions from Chapter 6 may be caused by these breakdowns.
7.4.3 Validation of MSE Computations

Full-scale MSE computations are carried out for 48 cases of the \( \mu \)-scale simulation models using the algorithm given in Chapter 6. The LLGEE AvgMSEs are computed at the AvgMSE optimal bandwidths \( h_{opt} \), which are obtained in two stages. First, the computed AvgMSE objective function is minimized over a fine grid of bandwidths. This gives an initial approximation of \( h_{opt} \). Then the search is refined by a golden section search algorithm in a neighborhood surrounding the initial \( h_{opt} \). If the AvgMSE object function is smooth enough or if the grid is fine enough, the obtained \( h_{opt} \) is the minimum AvgMSE optimal bandwidth.

The AvgMSEs and optimal bandwidths are also found by simulation. For each case, 1,000 datasets are generated. For each dataset, the sum of squared differences of the fits from the true means are computed for each bandwidth over a fine grid. By averaging across these curves, a Monte Carlo estimate of the AvgMSE objective function for \( h \) is obtained. Minimization over this grid produces an initial \( h_{opt} \), of which several improvements are made by finer and finer grid searches.

The identity variance function was used for the Poisson models and the function \( \mu(1 - \mu/n) \) was used for the binomial models in the parametric GEE and LLGEE models. We refer to these functions as the user’s variance functions since these are commonly chosen when modelling Poisson or binomial count data. The true variance functions are not of this form because of the latent variable mechanism used to generate correlated observations (see pg. 108). The variance function is correct, however, when \( \sigma^2 = 0 \), i.e., in the no corr/var (no correlation, no random effect) cases. Also, recall that the user’s parametric model is the conditional mean \( E[Y_{ij}|e_i = 0] \), which is only approximately equal to the unconditional mean. Even if the variance functions are correct and the unconditional mean is used, the user’s model is still misspecified in correlation structure. Thus, \( \gamma = 0 \) only implies that the mean model is correct to a first order approximation about \( E[e_i] = 0 \) and this is what we will mean by “no misspecification”.

All the following simulation results in this chapter are based on these (incorrect) variance functions and the approximately correct parametric mean functions. The results using the true variance functions were looked at in hundreds of cases but there were no noteworthy
differences in the comparisons that follow. We wanted to judge the impact of our methods based on the variance functions that are being used in practice because we assume the user does not know the true model. If he or she did, these simulated data sets would be analyzed with a GLMM not a GEE.

The results are given in Table 14 for the $\mu$-scale Poisson model and in Table 15 for the $\mu$-scale binomial model. The “Str” column of these tables gives the working correlation structure of the user’s parametric model. The working correlation structure of the user’s local model is always IND. Where the true model’s corr/var is “None” (no correlation, no random effect), the appropriate structure, IND, was used. These cases correspond to GLIM models.

The following comments apply.

1. The computed and simulated optimal bandwidths are, for the most part, within 2–3 decimal places of each other. The largest disagreements occur for the $\mu$-Scale binomial model (Table 15). In the $\gamma = 0$ cases, the objective functions are quite flat near $h_{opt}$ because the globalized local model ($h \to \infty$) is equivalent to the parametric model. Consequently, slight shifts in the rate of change between the simulated and calculated objective functions have a profound impact on the optimal bandwidth. There are considerable differences in the optimal bandwidth for $\mu$-Scale binomial, large corr/var cases. This is probably due to a breakdown in the Taylor series approximation of the true model variance, as mentioned previously, or the less favorable Taylor series locus for the MSE expressions that has 8–11% RE (Table 12).

2. The AvgMSEs are also in close agreement throughout most of these tables. Significant disagreements occur at the same places that the computed $h_{opt}$ fails to closely match the $h_{opt}$ obtained by simulation for the same reasons. The computed AvgMSEs tend to be biased toward smaller values. This is expected since the computations do not account for the additional variability in estimating $V$.

3. The LLGEE method theoretically beats the parametric GEE in terms of AvgMSE when there is large misspecification in every case. For the $\mu$-scale Poisson at moderate degrees of misspecification, there appears to be a cross-over point in misspecification at
which the LLGEE starts to perform better than the parametric GEE. However, LLGEE outperforms the parametric model in most cases of the $\mu$-scale binomial model, even where there is no misspecification. The correlation structure is not nearly as correct as it is in the $\mu$-scale Poisson cases. Though GEEs are supposed to be more efficient with the use of a non-identity correlation structure even when it is incorrect, the improvement in efficiency here may be too small to be detected, given the Monte Carlo variability.

4. The larger sample size implies smaller AvgMSEs and smaller optimal bandwidths. As the number of clusters increases, the averages at the unique prediction locations are closer to the true means. Smaller bandwidths bring the LLGEE estimate closer to interpolation of these averages and hence closer to the true means.

5. Larger sample sizes also appear to imply better agreement between the simulated and calculated values for the $\mu$-Scale Poisson model (Table 14). This arrangement does not occur in the $\mu$-Scale Binomial model (Table 15). This counter-intuitive effect most likely originates with the calculation of $\beta^\star$. Table 12 shows that $\beta^\star$ is a poorer approximation of $E[\hat{\beta}]$ as $s$ increases in the binomial cases. See the comments following Table 12.

6. The best agreement between the computed and simulated values occurs for the no correlation cases, which are fitted with GLIM models. The computed values in these cases do not rely on the first order approximations of the MOM estimator’s expected values.

In conclusion, the computed AvgMSEs are seen to perform well in practice in approximating the actual AvgMSEs (obtained by simulation) for most of the cases considered. They will be used when needed in the remainder of this chapter for obtaining the optimal bandwidths and optimal LLGEE AvgMSEs.
Table 14: Comparison of AvgMSE Computations (Comp.) for the $\mu$-Scale Poisson Simulation Model to Those Values Obtained by Simulation (Sim.).

<table>
<thead>
<tr>
<th>Row</th>
<th>Str*</th>
<th>Corr/Var</th>
<th>$s$</th>
<th>$\gamma$</th>
<th>$h_{opt}$</th>
<th>AvgMSE</th>
<th>ParGEE†</th>
<th>LLGEE‡</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sim. Comp.</td>
<td></td>
<td>Sim. Comp.</td>
<td>Sim. Comp.</td>
</tr>
<tr>
<td>1</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0</td>
<td>0.130</td>
<td>0.131</td>
<td>2.50</td>
<td>2.57</td>
</tr>
<tr>
<td>2</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.25</td>
<td>0.122</td>
<td>0.122</td>
<td>3.16</td>
<td>3.23</td>
</tr>
<tr>
<td>3</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.5</td>
<td>0.101</td>
<td>0.101</td>
<td>5.15</td>
<td>5.21</td>
</tr>
<tr>
<td>4</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>1</td>
<td>0.072</td>
<td>0.072</td>
<td>13.09</td>
<td>13.13</td>
</tr>
<tr>
<td>5</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0</td>
<td>0.079</td>
<td>0.079</td>
<td>0.26</td>
<td>0.26</td>
</tr>
<tr>
<td>6</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.25</td>
<td>0.069</td>
<td>0.069</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>7</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.5</td>
<td>0.056</td>
<td>0.056</td>
<td>2.90</td>
<td>2.90</td>
</tr>
<tr>
<td>8</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>1</td>
<td>0.044</td>
<td>0.044</td>
<td>10.81</td>
<td>10.82</td>
</tr>
<tr>
<td>9</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0</td>
<td>0.133</td>
<td>0.135</td>
<td>6.33</td>
<td>6.38</td>
</tr>
<tr>
<td>10</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.25</td>
<td>0.124</td>
<td>0.126</td>
<td>6.97</td>
<td>7.04</td>
</tr>
<tr>
<td>11</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.5</td>
<td>0.103</td>
<td>0.104</td>
<td>8.98</td>
<td>9.02</td>
</tr>
<tr>
<td>12</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>1</td>
<td>0.072</td>
<td>0.072</td>
<td>16.88</td>
<td>16.95</td>
</tr>
<tr>
<td>13</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0</td>
<td>0.080</td>
<td>0.079</td>
<td>0.64</td>
<td>0.64</td>
</tr>
<tr>
<td>14</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.25</td>
<td>0.070</td>
<td>0.069</td>
<td>1.29</td>
<td>1.30</td>
</tr>
<tr>
<td>15</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.5</td>
<td>0.056</td>
<td>0.056</td>
<td>3.28</td>
<td>3.28</td>
</tr>
<tr>
<td>16</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>1</td>
<td>0.045</td>
<td>0.045</td>
<td>11.20</td>
<td>11.20</td>
</tr>
<tr>
<td>17</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0</td>
<td>0.174</td>
<td>0.180</td>
<td>55.60</td>
<td>55.35</td>
</tr>
<tr>
<td>18</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.25</td>
<td>0.167</td>
<td>0.175</td>
<td>57.22</td>
<td>56.01</td>
</tr>
<tr>
<td>19</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.5</td>
<td>0.144</td>
<td>0.155</td>
<td>58.25</td>
<td>57.99</td>
</tr>
<tr>
<td>20</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>1</td>
<td>0.081</td>
<td>0.083</td>
<td>66.95</td>
<td>65.92</td>
</tr>
<tr>
<td>21</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0</td>
<td>0.088</td>
<td>0.086</td>
<td>5.63</td>
<td>5.54</td>
</tr>
<tr>
<td>22</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.25</td>
<td>0.077</td>
<td>0.074</td>
<td>6.29</td>
<td>6.20</td>
</tr>
<tr>
<td>23</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.5</td>
<td>0.060</td>
<td>0.059</td>
<td>8.29</td>
<td>8.18</td>
</tr>
<tr>
<td>24</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>1</td>
<td>0.046</td>
<td>0.046</td>
<td>16.24</td>
<td>16.10</td>
</tr>
</tbody>
</table>

* The Str column indicates which working structure the parametric model uses. The local model always uses IND here.
† The ParGEE column gives the AvgMSE for the parametric GEE model fits.
‡ The LLGEE column gives the AvgMSE for the local linear GEE model fits at their respective optimal bandwidths.
Table 15: Comparison of AvgMSE Computations (Comp.) for the $\mu$-Scale Binomial Simulation Model to Those Values Obtained by Simulation (Sim.). The AvgMSEs have been scaled by $10^5$.

<table>
<thead>
<tr>
<th>Row</th>
<th>Str*</th>
<th>Corr/Var</th>
<th>$s$</th>
<th>$\gamma$</th>
<th>$h_{opt}$</th>
<th>AvgMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sim. Comp.</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0</td>
<td>1.569 1.740</td>
<td>2.06 2.04</td>
</tr>
<tr>
<td>2</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.25</td>
<td>0.181 0.184</td>
<td>4.82 4.74</td>
</tr>
<tr>
<td>3</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.5</td>
<td>0.112 0.112</td>
<td>13.44 13.31</td>
</tr>
<tr>
<td>4</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>1</td>
<td>0.080 0.080</td>
<td>51.41 51.21</td>
</tr>
<tr>
<td>5</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0</td>
<td>1.119 1.180</td>
<td>0.21 0.20</td>
</tr>
<tr>
<td>6</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.25</td>
<td>0.089 0.089</td>
<td>2.85 2.84</td>
</tr>
<tr>
<td>7</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.5</td>
<td>0.066 0.065</td>
<td>11.37 11.36</td>
</tr>
<tr>
<td>8</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>1</td>
<td>0.050 0.049</td>
<td>49.20 49.18</td>
</tr>
<tr>
<td>9</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0</td>
<td>1.680 1.550</td>
<td>9.53 8.85</td>
</tr>
<tr>
<td>10</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.25</td>
<td>0.200 0.197</td>
<td>11.67 11.50</td>
</tr>
<tr>
<td>11</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.5</td>
<td>0.122 0.117</td>
<td>19.39 19.77</td>
</tr>
<tr>
<td>12</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>1</td>
<td>0.083 0.081</td>
<td>56.41 57.01</td>
</tr>
<tr>
<td>13</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0</td>
<td>1.240 1.410</td>
<td>0.96 0.89</td>
</tr>
<tr>
<td>14</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.25</td>
<td>0.101 0.090</td>
<td>3.36 3.53</td>
</tr>
<tr>
<td>15</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.5</td>
<td>0.073 0.066</td>
<td>11.70 12.01</td>
</tr>
<tr>
<td>16</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>1</td>
<td>0.055 0.049</td>
<td>49.28 49.76</td>
</tr>
<tr>
<td>17</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0</td>
<td>3.940 1.490</td>
<td>69.35 73.61</td>
</tr>
<tr>
<td>18</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.25</td>
<td>0.217 0.246</td>
<td>74.60 91.75</td>
</tr>
<tr>
<td>19</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.5</td>
<td>0.180 0.188</td>
<td>81.42 107.56</td>
</tr>
<tr>
<td>20</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>1</td>
<td>0.123 0.097</td>
<td>106.71 136.89</td>
</tr>
<tr>
<td>21</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0</td>
<td>3.300 1.010</td>
<td>10.87 7.53</td>
</tr>
<tr>
<td>22</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.25</td>
<td>0.159 0.101</td>
<td>12.30 11.91</td>
</tr>
<tr>
<td>23</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.5</td>
<td>0.129 0.069</td>
<td>19.65 21.26</td>
</tr>
<tr>
<td>24</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>1</td>
<td>0.092 0.051</td>
<td>55.01 58.08</td>
</tr>
</tbody>
</table>

* The Str column indicates which working structure the parametric model uses. The local model always uses IND here.

† The ParGEE column gives the AvgMSE for the parametric GEE model fits.

‡ The LLGEE column gives the AvgMSE for the local linear GEE model fits at their respective optimal bandwidth.
7.5 Bandwidth Simulation Study

The bandwidth selectors CV, PRESS* and PRESS** were originally developed for linear models of uncorrelated data. It is not entirely clear how their respective extensions to the GEE context will perform on the data structures considered here. It is of special interest as to how these bandwidth selectors perform on small samples. Previous research in MRR methods has shown that PRESS** performs “best” when estimating nonparametric models for uncorrelated normal data. We inquire here whether this result carries over to the generalized PRESS** in the current context.

The weight matrix $W$ used in these bandwidth selectors is up for debate. Setting $W = V^{-1}$ would be a natural choice for data whose variance depends on the mean\textsuperscript{6}. However, there are several problems with this option, as discussed on page 68. In summary, the estimate of $V^{-1}$ may be incorrect because the variance function is incorrect, or because it is evaluated based on a misspecified mean model or when the bandwidth is too large. Also, the Pearson $\chi^2$ MOM estimate of $\phi^2$ will depend on the bandwidth directly which often prevents the CV objective function from increasing when the bandwidth is too large. To eliminate this problem, we choose not to allow $\phi^2$ to be estimated and fix it at 1. This reduces the LLGEE method to a local linear GLIM model but does not change the fits since the scale parameter is not involved in the estimating equations for $\theta$. The question remains whether using $V^{-1}$ for $W$ in the bandwidth selectors is the best choice.

A simulation study is done for the purpose of determining the following: which of CV, PRESS* and PRESS** perform the best, whether or not PRESS** performs well in small samples (as its predecessors have done), and whether or not the variance function (without the scale parameter) should be used in these bandwidth selectors. The study is based on 500 runs rather than 1,000 because the computations required increase dramatically from the previous simulation studies. For each generated dataset, the local model is fit for each bandwidth on a grid, $0, 0.05, \ldots, 1$, for each prediction location. The bandwidth with the minimum objective (CV, PRESS, or PRESS**) is used to define a small neighborhood for a golden section search. The golden section search narrows the bandwidth choice to within 3

---

\textsuperscript{6}Since IND is used in the local model, $V^{-1}$ is diagonal.
decimal places of the optimal.

The average bandwidths for 24 cases of the $\mu$-Scale Poisson model are given in Table 17, the AvgMSEs are given in Table 18. The average bandwidths for 24 cases of the $\mu$-Scale Binomial model are in Table 19, with the AvgMSEs following in Table 20. These are the same cases used in the simulation study that validate the MSE expressions (7.4.3). The average bandwidth tables (17–19) include the simulated optimal bandwidths from Tables 14–15 rather than the computed values. The AvgMSE tables (18–20) include the parametric AvgMSE’s and optimal LLGEE AvgMSEs as bases of comparison.

Denote CV selection without $V$ as CVnoV, and with $V$ as CVV. Likewise, let PRESS*noV and PRESS**noV denote PRESS* and PRESS** without $V$. Let PRESS*V and PRESS**V represent PRESS* and PRESS** with $V$. The following comments apply to Tables 17–20.

1. Overall, the average bandwidth chosen by CVnoV is closest to the optimal bandwidth for both the Poisson models (Table 17) and the Binomial models (Table 19). For the Poisson cases, the general ordering of the average bandwidths with regards to their closeness to the optimal is (closest to farthest) CVnoV, CVV, PRESS**V, PRESS**noV, PRESS*noV, and PRESS*V. This ordering is quite consistent across the cases and indicates that the simplest bandwidth selector performs the best.

2. For the binomial cases (Table 19), the best ordering is not as clear. Roughly, the overall ordering is (closest to farthest) CVnoV, CVV, PRESS**noV, PRESS**V, PRESS*V, and PRESS*noV. However, it is worth considering the ordering in a few special cases. For the small sample size ($s = 3$) under no corr/var (GLIM context) and small corr/var, the PRESS** selectors tend to be closer to the optimal on average than the CVnoV and CVV. For the larger sample size ($s = 30$), largest misspecification ($\gamma = 1$), or the largest corr/var, CVnoV and CVV are closer to optimal than PRESS**noV and PRESS**V. Under no misspecification, PRESS* appears to be the closest to the optimal. This is a consequence of the following. The $\mu$-scale binomial simulation model differs from the $\mu$-scale Poisson simulation model in that it limits to the parametric model as $h \rightarrow \infty$ under no misspecification. Thus the best bandwidth under no misspecification will be provided by the selector that chooses the largest bandwidth.
PRESS*’s penalty on small bandwidths is strong and forces the selector to choose large bandwidths at all levels of misspecification. Thus, PRESS* has superior performance under no misspecification by coincidence.

3. More important to the assessment of these selectors than average bandwidth is the AvgMSE, which accounts for the bias and variance of the fits based on the chosen bandwidths. In general the overall ordering of the selectors in terms of AvgMSE for both the Poisson models (Table 18) and the binomial models (Table 20) is roughly (best to worst) CVnoV, PRESS**V, CVV, PRESS**noV, PRESS*noV and PRESS*V. For the Poisson simulation models, PRESS**V seems to perform best for the small sample size, while CVnoV performs best for the larger sample size. This agrees with previous research in MRR (see §4.1). Note though, that when CVnoV does better than PRESS**V, it is considerably better, whereas when PRESS** performs best, the improvement over CVnoV tends to be slight. For example, in row 16, PRESS**noV is only 32% as efficient as CVnoV; in row 17, CVnoV is 98% as efficient as PRESS**noV.

4. The ordering of the selectors given in comment 3 is not very reliable for the binomial models (Table 20), but CVnoV has an overall advantage over PRESS**V (67% of the cases). PRESS* performs well for the no misspecification cases but this is merely a coincidence as comment 3 points out. PRESS**noV tends to do better than CVnoV for the smaller sample size, though less consistently. Also, we see that the improvement of CVnoV over PRESS**noV is more substantial than the improvement of PRESS**noV over CVnoV (e.g., row 16), though this relationship is not as consistent as the relationship between CVnoV and PRESS**noV described in comment 3.

Notice that the average bandwidth being closer to the optimal does not necessarily imply the AvgMSE is smaller. For example, in the \( \mu \)-scale binomial average bandwidth table (19), the small corr/var, \( s = 3, \gamma = 0.25 \), PRESS*noV and PRESS**noV cases (row 10), the average bandwidth for PRESS**noV (0.22) is the closest to the optimal (0.2) and PRESS*noV (0.99) is the farthest from the optimal. Yet, in the corresponding AvgMSE table (20) the opposite relationship holds! There are two reasons. First, a large difference in average bandwidths does not necessarily imply that
the AvgMSEs will differ by the same degree. Since the $\mu$-scale binomial local linear model is equivalent in mean to the parametric model at large bandwidths, the theoretical AvgMSE curve is nearly flat for a wide range of bandwidths (see Figure 15) when there is little misspecification. Second, the asymmetry of the theoretical AvgMSE curve about the optimal bandwidth observed in Figure 15 implies that any selector choosing bandwidths in this region will tend to incur heavier AvgMSE penalties to the left of the optimal bandwidth than to the right. On average PRESS**noV is close to the optimal but chooses about 1/3 of its bandwidths where the theoretical AvgMSEs are higher than those at large bandwidths. PRESS*noV almost always chooses large bandwidths beyond the boundary 1.01, and as such its bandwidth choices are often set to 1.01. It comes out ahead because it consistently selects large bandwidths which have theoretical AvgMSEs that are close to the optimal, while PRESS**noV selects a
Figure 16: Comparison of CVnoV with PRESS**V Bandwidth Choice Estimated Densities with Respect to Theoretical AvgMSE for the $\mu$-Scale Poisson Model. There is no correlation, no misspecification and $s = 3$.

The no corr/var, $s = 3$, $\gamma = 0$ case in row 1 of the $\mu$-scale Poisson tables (17–18) is another interesting situation. The CVnoV selector is slightly closer to the optimal bandwidth on average, yet PRESS**V has a significantly smaller (> 2 SEs) AvgMSE. Figure 16 shows the estimated densities for CVnoV and PRESS**V $h$ choice. While the CVnoV is closer to optimal $h$ on average, the left side of the MSE curve increases more rapidly than the right, so PRESS**V comes out ahead in terms of AvgMSE. Note that the CVnoV bandwidth choice has more variation. This is not the case when CVnoV performs best. A more in-depth study of the variability of bandwidth selectors could be a topic for future research.

6. Cases 17–20, 21, 23–24 in Table 18 and 5, 13, 21, and 23 in Table 20 have AvgMSEs that
Table 16: Comparison of the Optimal AvgMSE Based on Optimizing a Fixed $h$ for all Datasets to the Best Possible AvgMSE based on Optimizing $h$ for Each Dataset. These cases correspond to rows 17–24 of the $\mu$-scale Poisson Table (18).

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>$s$</th>
<th>Fixed $h$</th>
<th>Best</th>
<th>Optimal</th>
<th>CVnoV</th>
<th>Possible</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>54.18</td>
<td>53.47</td>
<td>47.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>3</td>
<td>55.43</td>
<td>53.61</td>
<td>47.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>3</td>
<td>55.90</td>
<td>53.54</td>
<td>47.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>59.37</td>
<td>56.19</td>
<td>50.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>30</td>
<td>5.77</td>
<td>5.73</td>
<td>4.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>30</td>
<td>5.89</td>
<td>5.89</td>
<td>4.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>30</td>
<td>6.09</td>
<td>6.05</td>
<td>4.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>30</td>
<td>6.33</td>
<td>6.10</td>
<td>5.66</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

are smaller than the optimal. Though this could be due to Monte Carlo variability\(^7\), it is most likely due to fact that the optimal AvgMSE is based on a single fixed bandwidth, whereas the bandwidth selectors are allowed to choose bandwidths on a case-by-case basis, i.e., one for each dataset. The true lower bound of the AvgMSEs is determined by optimizing the bandwidth separately for each dataset observed. This is examined for cases 17–24 in the $\mu$-scale Poisson table (18). The results in Table 16 show that AvgMSE using the best fixed $h$ can not perform nearly as well as optimizing $h$ for each dataset.

7. All the bandwidth selection methods other than PRESS\(^*\) perform better than the parametric model at high degrees of misspecification. Thus, the LLGEE method performs best in both theory and in practice for these cases.

8. In general, the average bandwidth chosen by CV tends to be closer to the optimal for the larger sample size compared to the smaller sample size.

9. CVnoV performs better than CVV in the majority of the cases. For PRESS\(^**\), PRESS\(^**\)V performs best in the majority of the cases.

\(^7\)For example, the most outstanding case is in the $\mu$-scale Poisson table (18), row 20, where the AvgMSE for PRESS\(^**\)V is below the optimal by 3.59, which is only 1.1 standard errors
In conclusion, the best bandwidth selector overall is CVnoV. PRESS** is competitive in small samples and when there is no correlation. This agrees with previous research. Also, the ordering of PRESS* and PRESS** agrees with previous research, which concluded that PRESS* chooses bandwidths that are too large. Whether the variance function (without the scale parameter) should be used in the bandwidth selectors depends on which bandwidth selector is used. The variance function helps PRESS** in small samples, but tends to hinder CV. We need to choose one bandwidth selector for the simulation study of the MRGEE methods to keep the comparisons simple. Our choice is CVnoV.
Table 17: Average Bandwidths Estimated by CV, PRESS* and PRESS** with \( W = I \) (NoV) and \( W = V^{-1} \) (DiagV) for the \( \mu \)-Scale Poisson Simulation Model

<table>
<thead>
<tr>
<th>Row</th>
<th>Str†</th>
<th>Corr/Var</th>
<th>s</th>
<th>( \gamma )</th>
<th>( h_{opt} )††</th>
<th>CV NoV</th>
<th>CV DiagV</th>
<th>PRESS* NoV</th>
<th>PRESS* DiagV</th>
<th>PRESS** NoV</th>
<th>PRESS** DiagV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0</td>
<td>0.13</td>
<td>0.14</td>
<td>0.14</td>
<td>0.27</td>
<td>0.28</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>2</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.25</td>
<td>0.12</td>
<td>0.14</td>
<td>0.14</td>
<td>0.27</td>
<td>0.28</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>3</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.5</td>
<td>0.10</td>
<td>0.13</td>
<td>0.12</td>
<td>0.27</td>
<td>0.28</td>
<td>0.15</td>
<td>0.14</td>
</tr>
<tr>
<td>4</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>1</td>
<td>0.07</td>
<td>0.09</td>
<td>0.09</td>
<td>0.26</td>
<td>0.27</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>5</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0</td>
<td>0.08</td>
<td>0.08</td>
<td>0.06</td>
<td>0.24</td>
<td>0.24</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>6</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.25</td>
<td>0.07</td>
<td>0.07</td>
<td>0.06</td>
<td>0.24</td>
<td>0.24</td>
<td>0.15</td>
<td>0.14</td>
</tr>
<tr>
<td>7</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.5</td>
<td>0.06</td>
<td>0.06</td>
<td>0.05</td>
<td>0.24</td>
<td>0.24</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>8</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>1</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.22</td>
<td>0.21</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>9</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0</td>
<td>0.13</td>
<td>0.15</td>
<td>0.14</td>
<td>0.31</td>
<td>0.33</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>10</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.25</td>
<td>0.12</td>
<td>0.14</td>
<td>0.13</td>
<td>0.31</td>
<td>0.33</td>
<td>0.17</td>
<td>0.16</td>
</tr>
<tr>
<td>11</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.5</td>
<td>0.10</td>
<td>0.13</td>
<td>0.12</td>
<td>0.31</td>
<td>0.33</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>12</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>1</td>
<td>0.07</td>
<td>0.09</td>
<td>0.09</td>
<td>0.31</td>
<td>0.33</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>13</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0</td>
<td>0.08</td>
<td>0.08</td>
<td>0.06</td>
<td>0.27</td>
<td>0.27</td>
<td>0.16</td>
<td>0.15</td>
</tr>
<tr>
<td>14</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.25</td>
<td>0.07</td>
<td>0.07</td>
<td>0.06</td>
<td>0.27</td>
<td>0.27</td>
<td>0.16</td>
<td>0.15</td>
</tr>
<tr>
<td>15</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.5</td>
<td>0.06</td>
<td>0.06</td>
<td>0.05</td>
<td>0.27</td>
<td>0.27</td>
<td>0.14</td>
<td>0.13</td>
</tr>
<tr>
<td>16</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>1</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>0.27</td>
<td>0.27</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>17</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0</td>
<td>0.17</td>
<td>0.19</td>
<td>0.12</td>
<td>0.52</td>
<td>0.62</td>
<td>0.19</td>
<td>0.18</td>
</tr>
<tr>
<td>18</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.25</td>
<td>0.17</td>
<td>0.18</td>
<td>0.12</td>
<td>0.52</td>
<td>0.62</td>
<td>0.18</td>
<td>0.17</td>
</tr>
<tr>
<td>19</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.5</td>
<td>0.14</td>
<td>0.17</td>
<td>0.11</td>
<td>0.52</td>
<td>0.62</td>
<td>0.17</td>
<td>0.16</td>
</tr>
<tr>
<td>20</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>1</td>
<td>0.08</td>
<td>0.13</td>
<td>0.09</td>
<td>0.53</td>
<td>0.64</td>
<td>0.13</td>
<td>0.12</td>
</tr>
<tr>
<td>21</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0</td>
<td>0.09</td>
<td>0.09</td>
<td>0.01</td>
<td>0.49</td>
<td>0.63</td>
<td>0.18</td>
<td>0.17</td>
</tr>
<tr>
<td>22</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.25</td>
<td>0.08</td>
<td>0.08</td>
<td>0.01</td>
<td>0.50</td>
<td>0.63</td>
<td>0.18</td>
<td>0.17</td>
</tr>
<tr>
<td>23</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.5</td>
<td>0.06</td>
<td>0.06</td>
<td>0.01</td>
<td>0.50</td>
<td>0.64</td>
<td>0.17</td>
<td>0.16</td>
</tr>
<tr>
<td>24</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>1</td>
<td>0.05</td>
<td>0.05</td>
<td>0.03</td>
<td>0.51</td>
<td>0.66</td>
<td>0.13</td>
<td>0.12</td>
</tr>
</tbody>
</table>

† The Str column indicates which working structure the parametric model uses. The local model always uses the IND structure.

†† \( h_{opt} \) is the optimal bandwidth obtained by simulation.
Table 18: AvgMSEs of Fits Obtained by CV, PRESS* and PRESS** Bandwidth Estimation with $W = I$ (NoV) and $W = V^{-1}$ (DiagV) for the $\mu$-Scale Poisson Simulation Model. The bold values highlight the best AvgMSE within each case.

<table>
<thead>
<tr>
<th>Row</th>
<th>Str†</th>
<th>Corr/Var</th>
<th>s</th>
<th>$\gamma$</th>
<th>Par</th>
<th>Optimal GEE†</th>
<th>CV NoV</th>
<th>DiagV</th>
<th>PRESS* NoV</th>
<th>DiagV</th>
<th>PRESS** NoV</th>
<th>DiagV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0</td>
<td>2.63</td>
<td>3.61</td>
<td>5.00</td>
<td>5.06</td>
<td>7.16</td>
<td>7.39</td>
<td>4.57</td>
<td>4.55</td>
</tr>
<tr>
<td>2</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.25</td>
<td>3.31</td>
<td>3.92</td>
<td>5.32</td>
<td>5.38</td>
<td>7.80</td>
<td>8.03</td>
<td>4.95</td>
<td>4.95</td>
</tr>
<tr>
<td>3</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.5</td>
<td>5.29</td>
<td>4.67</td>
<td>6.08</td>
<td>6.06</td>
<td>9.55</td>
<td>9.75</td>
<td>5.85</td>
<td>5.82</td>
</tr>
<tr>
<td>4</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>1</td>
<td>13.22</td>
<td>6.14</td>
<td>7.70</td>
<td>7.64</td>
<td>16.11</td>
<td>16.17</td>
<td>7.99</td>
<td>7.93</td>
</tr>
<tr>
<td>5</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0</td>
<td>0.27</td>
<td>0.54</td>
<td>0.62</td>
<td>0.65</td>
<td>3.88</td>
<td>3.85</td>
<td>1.36</td>
<td>1.28</td>
</tr>
<tr>
<td>6</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.25</td>
<td>0.93</td>
<td>0.62</td>
<td>0.71</td>
<td>0.73</td>
<td>4.45</td>
<td>4.41</td>
<td>1.63</td>
<td>1.53</td>
</tr>
<tr>
<td>7</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.5</td>
<td>2.91</td>
<td>0.74</td>
<td>0.83</td>
<td>0.82</td>
<td>6.15</td>
<td>6.07</td>
<td>2.21</td>
<td>2.09</td>
</tr>
<tr>
<td>8</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>1</td>
<td>10.83</td>
<td>0.85</td>
<td>0.89</td>
<td>0.88</td>
<td>12.13</td>
<td>11.46</td>
<td>3.06</td>
<td>3.00</td>
</tr>
<tr>
<td>9</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0</td>
<td>6.47</td>
<td>7.35</td>
<td>8.94</td>
<td>9.08</td>
<td>12.02</td>
<td>12.40</td>
<td>8.42</td>
<td>8.35</td>
</tr>
<tr>
<td>10</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.25</td>
<td>7.14</td>
<td>7.67</td>
<td>9.21</td>
<td>9.34</td>
<td>12.68</td>
<td>13.06</td>
<td>8.79</td>
<td>8.74</td>
</tr>
<tr>
<td>11</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.5</td>
<td>9.09</td>
<td>8.46</td>
<td>9.98</td>
<td>9.99</td>
<td>14.50</td>
<td>14.87</td>
<td>9.69</td>
<td>9.60</td>
</tr>
<tr>
<td>12</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>1</td>
<td>17.03</td>
<td>9.96</td>
<td>11.66</td>
<td>11.52</td>
<td>21.74</td>
<td>22.02</td>
<td>12.01</td>
<td>11.87</td>
</tr>
<tr>
<td>13</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0</td>
<td>0.62</td>
<td>0.91</td>
<td>0.98</td>
<td>1.05</td>
<td>5.37</td>
<td>5.35</td>
<td>1.90</td>
<td>1.76</td>
</tr>
<tr>
<td>14</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.25</td>
<td>1.28</td>
<td>0.99</td>
<td>1.07</td>
<td>1.11</td>
<td>5.99</td>
<td>5.97</td>
<td>2.19</td>
<td>2.04</td>
</tr>
<tr>
<td>15</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.5</td>
<td>3.26</td>
<td>1.12</td>
<td>1.18</td>
<td>1.18</td>
<td>7.85</td>
<td>7.80</td>
<td>2.86</td>
<td>2.67</td>
</tr>
<tr>
<td>16</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>1</td>
<td>11.18</td>
<td>1.23</td>
<td>1.24</td>
<td>1.24</td>
<td>15.23</td>
<td>14.98</td>
<td>3.93</td>
<td>3.82</td>
</tr>
<tr>
<td>17</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0</td>
<td>53.12</td>
<td>54.18</td>
<td>53.47</td>
<td>55.97</td>
<td>63.00</td>
<td>64.77</td>
<td>52.60</td>
<td>52.56</td>
</tr>
<tr>
<td>18</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.25</td>
<td>54.63</td>
<td>55.43</td>
<td>53.61</td>
<td>56.22</td>
<td>63.48</td>
<td>65.55</td>
<td>53.04</td>
<td>53.04</td>
</tr>
<tr>
<td>19</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.5</td>
<td>55.16</td>
<td>55.90</td>
<td>53.54</td>
<td>55.97</td>
<td>64.35</td>
<td>66.41</td>
<td>52.93</td>
<td>52.88</td>
</tr>
<tr>
<td>20</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>1</td>
<td>63.31</td>
<td>59.37</td>
<td>56.19</td>
<td>57.05</td>
<td>72.58</td>
<td>74.68</td>
<td>55.95</td>
<td>55.79</td>
</tr>
<tr>
<td>21</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0</td>
<td>5.44</td>
<td>5.77</td>
<td>5.73</td>
<td>6.14</td>
<td>18.83</td>
<td>21.54</td>
<td>7.15</td>
<td>6.82</td>
</tr>
<tr>
<td>22</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.25</td>
<td>6.15</td>
<td>5.89</td>
<td>5.89</td>
<td>6.23</td>
<td>19.51</td>
<td>22.20</td>
<td>7.57</td>
<td>7.20</td>
</tr>
<tr>
<td>23</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.5</td>
<td>8.16</td>
<td>6.09</td>
<td>6.05</td>
<td>6.29</td>
<td>21.65</td>
<td>24.37</td>
<td>8.59</td>
<td>8.11</td>
</tr>
<tr>
<td>24</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>1</td>
<td>16.03</td>
<td>6.33</td>
<td>6.10</td>
<td>6.20</td>
<td>29.64</td>
<td>32.43</td>
<td>10.80</td>
<td>10.30</td>
</tr>
</tbody>
</table>

† The Str column indicates which working structure the parametric model uses. The local model always uses the IND structure.
‡ The Par column gives the AvgMSE of the parametric model, obtained by simulation.
†† The LLGEE column gives local linear GEE AvgMSE at the the optimal bandwidth, both of which are obtained by simulation.
Table 19: Average Bandwidths Estimated by CV, PRESS* and PRESS** with $W = I$ (NoV) and $W = V^{-1}$ (DiagV) for the $\mu$-Scale Binomial Simulation Model

<table>
<thead>
<tr>
<th>Row</th>
<th>Str†</th>
<th>Corr/Var</th>
<th>$s$</th>
<th>$\gamma$</th>
<th>$h_{opt}$††</th>
<th>CV</th>
<th>PRESS*</th>
<th>PRESS**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NoV</td>
<td>DiagV</td>
<td>NoV</td>
</tr>
<tr>
<td>1</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0</td>
<td>1.57</td>
<td>0.64</td>
<td>0.65</td>
<td>0.99</td>
</tr>
<tr>
<td>2</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.25</td>
<td>0.18</td>
<td>0.46</td>
<td>0.50</td>
<td>0.97</td>
</tr>
<tr>
<td>3</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.5</td>
<td>0.11</td>
<td>0.19</td>
<td>0.23</td>
<td>0.78</td>
</tr>
<tr>
<td>4</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>1</td>
<td>0.08</td>
<td>0.09</td>
<td>0.09</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NoV</td>
<td>DiagV</td>
<td>NoV</td>
</tr>
<tr>
<td>5</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0</td>
<td>1.12</td>
<td>0.69</td>
<td>0.62</td>
<td>1.01</td>
</tr>
<tr>
<td>6</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.25</td>
<td>0.09</td>
<td>0.09</td>
<td>0.11</td>
<td>1.01</td>
</tr>
<tr>
<td>7</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.5</td>
<td>0.07</td>
<td>0.07</td>
<td>0.06</td>
<td>0.86</td>
</tr>
<tr>
<td>8</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>1</td>
<td>0.05</td>
<td>0.05</td>
<td>0.04</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NoV</td>
<td>DiagV</td>
<td>NoV</td>
</tr>
<tr>
<td>9</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0</td>
<td>1.68</td>
<td>0.62</td>
<td>0.62</td>
<td>1.00</td>
</tr>
<tr>
<td>10</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.25</td>
<td>0.20</td>
<td>0.43</td>
<td>0.47</td>
<td>0.99</td>
</tr>
<tr>
<td>11</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.5</td>
<td>0.12</td>
<td>0.18</td>
<td>0.22</td>
<td>0.88</td>
</tr>
<tr>
<td>12</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>1</td>
<td>0.08</td>
<td>0.09</td>
<td>0.10</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NoV</td>
<td>DiagV</td>
<td>NoV</td>
</tr>
<tr>
<td>13</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0</td>
<td>1.24</td>
<td>0.69</td>
<td>0.61</td>
<td>1.01</td>
</tr>
<tr>
<td>14</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.25</td>
<td>0.10</td>
<td>0.09</td>
<td>0.10</td>
<td>1.01</td>
</tr>
<tr>
<td>15</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.5</td>
<td>0.07</td>
<td>0.07</td>
<td>0.05</td>
<td>1.00</td>
</tr>
<tr>
<td>16</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>1</td>
<td>0.06</td>
<td>0.05</td>
<td>0.04</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NoV</td>
<td>DiagV</td>
<td>NoV</td>
</tr>
<tr>
<td>17</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0</td>
<td>3.94</td>
<td>0.49</td>
<td>0.50</td>
<td>1.01</td>
</tr>
<tr>
<td>18</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.25</td>
<td>0.22</td>
<td>0.32</td>
<td>0.36</td>
<td>1.00</td>
</tr>
<tr>
<td>19</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.5</td>
<td>0.18</td>
<td>0.19</td>
<td>0.21</td>
<td>0.98</td>
</tr>
<tr>
<td>20</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>1</td>
<td>0.12</td>
<td>0.12</td>
<td>0.13</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NoV</td>
<td>DiagV</td>
<td>NoV</td>
</tr>
<tr>
<td>21</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0</td>
<td>3.30</td>
<td>0.62</td>
<td>0.46</td>
<td>1.01</td>
</tr>
<tr>
<td>22</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.25</td>
<td>0.16</td>
<td>0.09</td>
<td>0.11</td>
<td>1.01</td>
</tr>
<tr>
<td>23</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.5</td>
<td>0.13</td>
<td>0.07</td>
<td>0.01</td>
<td>1.01</td>
</tr>
<tr>
<td>24</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>1</td>
<td>0.09</td>
<td>0.05</td>
<td>0.01</td>
<td>0.96</td>
</tr>
</tbody>
</table>

† The Str column indicates which working structure the parametric model uses.
The local model always uses the IND structure.

†† $h_{opt}$ is the optimal bandwidth obtained by simulation.
Table 20: AvgMSEs of Fits Obtained by CV, PRESS* and PRESS** Bandwidth Estimation with $W = I$ (NoV) and $W = V^{-1}$ (DiagV) for the $\mu$-Scale Binomial Simulation Model. The bold values highlight the best AvgMSE within each case. The AvgMSEs have been scaled by $10^5$.

<table>
<thead>
<tr>
<th>Row</th>
<th>Str†</th>
<th>Corr/Var</th>
<th>s</th>
<th>$\gamma$</th>
<th>Par GEE‡</th>
<th>Optimal LLGEE††</th>
<th>CV NoV</th>
<th>DiagV</th>
<th>PRESS* NoV</th>
<th>DiagV</th>
<th>PRESS** NoV</th>
<th>DiagV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0</td>
<td>2.10</td>
<td>2.06</td>
<td>3.33</td>
<td>3.06</td>
<td>2.12</td>
<td>2.13</td>
<td>4.05</td>
<td>3.76</td>
</tr>
<tr>
<td>2</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.25</td>
<td>4.89</td>
<td>3.78</td>
<td>5.07</td>
<td>4.93</td>
<td>4.78</td>
<td>4.78</td>
<td>5.04</td>
<td>4.93</td>
</tr>
<tr>
<td>3</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.5</td>
<td>13.51</td>
<td>5.21</td>
<td>7.14</td>
<td>7.43</td>
<td>11.41</td>
<td>11.36</td>
<td>6.20</td>
<td>6.28</td>
</tr>
<tr>
<td>4</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>1</td>
<td>51.50</td>
<td>7.45</td>
<td>9.04</td>
<td>9.44</td>
<td>17.21</td>
<td>18.42</td>
<td>9.20</td>
<td>9.54</td>
</tr>
<tr>
<td>5</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0</td>
<td>0.20</td>
<td>0.21</td>
<td>0.31</td>
<td>0.31</td>
<td>0.20</td>
<td>0.20</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>6</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.25</td>
<td>2.85</td>
<td>0.60</td>
<td>0.72</td>
<td>0.90</td>
<td>2.79</td>
<td>2.79</td>
<td>0.74</td>
<td>0.75</td>
</tr>
<tr>
<td>7</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.5</td>
<td>11.38</td>
<td>0.84</td>
<td>0.94</td>
<td>0.97</td>
<td>9.99</td>
<td>9.42</td>
<td>1.65</td>
<td>1.72</td>
</tr>
<tr>
<td>8</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>1</td>
<td>49.22</td>
<td>1.13</td>
<td>1.19</td>
<td>1.19</td>
<td>10.81</td>
<td>11.91</td>
<td>3.62</td>
<td>3.94</td>
</tr>
<tr>
<td>9</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0</td>
<td>9.76</td>
<td>8.49</td>
<td>9.98</td>
<td>9.75</td>
<td>8.81</td>
<td>8.82</td>
<td>11.03</td>
<td>10.81</td>
</tr>
<tr>
<td>10</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.25</td>
<td>12.02</td>
<td>9.64</td>
<td>11.37</td>
<td>11.30</td>
<td>11.09</td>
<td>11.11</td>
<td>11.47</td>
<td>11.34</td>
</tr>
<tr>
<td>11</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.5</td>
<td>19.69</td>
<td>10.92</td>
<td>13.03</td>
<td>13.42</td>
<td>17.85</td>
<td>17.85</td>
<td>11.91</td>
<td>11.89</td>
</tr>
<tr>
<td>12</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>1</td>
<td>56.78</td>
<td>12.90</td>
<td>14.82</td>
<td>15.27</td>
<td>27.44</td>
<td>27.34</td>
<td>14.88</td>
<td>15.17</td>
</tr>
<tr>
<td>13</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0</td>
<td>0.94</td>
<td>0.90</td>
<td>0.99</td>
<td>0.96</td>
<td>0.88</td>
<td>0.88</td>
<td>1.25</td>
<td>1.25</td>
</tr>
<tr>
<td>14</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.25</td>
<td>3.34</td>
<td>1.13</td>
<td>1.23</td>
<td>1.47</td>
<td>3.22</td>
<td>3.22</td>
<td>1.18</td>
<td>1.18</td>
</tr>
<tr>
<td>15</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.5</td>
<td>11.67</td>
<td>1.34</td>
<td>1.42</td>
<td>1.58</td>
<td>11.37</td>
<td>11.29</td>
<td>1.98</td>
<td>2.02</td>
</tr>
<tr>
<td>16</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>1</td>
<td>49.28</td>
<td>1.64</td>
<td>1.72</td>
<td>1.83</td>
<td>13.93</td>
<td>14.64</td>
<td>4.18</td>
<td>4.40</td>
</tr>
<tr>
<td>17</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0</td>
<td>68.83</td>
<td>65.90</td>
<td>67.87</td>
<td>67.80</td>
<td>66.76</td>
<td>66.87</td>
<td>69.88</td>
<td>69.83</td>
</tr>
<tr>
<td>18</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.25</td>
<td>74.26</td>
<td>62.47</td>
<td>64.27</td>
<td>65.43</td>
<td>64.61</td>
<td>65.66</td>
<td>66.45</td>
<td>66.37</td>
</tr>
<tr>
<td>19</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.5</td>
<td>80.81</td>
<td>61.30</td>
<td>63.33</td>
<td>65.37</td>
<td>69.39</td>
<td>71.02</td>
<td>63.82</td>
<td>63.75</td>
</tr>
<tr>
<td>20</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>1</td>
<td>112.68</td>
<td>58.71</td>
<td>60.22</td>
<td>68.87</td>
<td>86.05</td>
<td>90.88</td>
<td>64.76</td>
<td>64.63</td>
</tr>
<tr>
<td>21</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0</td>
<td>10.62</td>
<td>10.28</td>
<td>10.38</td>
<td>10.41</td>
<td>10.22</td>
<td>10.22</td>
<td>10.66</td>
<td>10.67</td>
</tr>
<tr>
<td>22</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.25</td>
<td>12.06</td>
<td>8.64</td>
<td>9.55</td>
<td>11.13</td>
<td>10.82</td>
<td>10.82</td>
<td>8.72</td>
<td>8.72</td>
</tr>
<tr>
<td>23</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.5</td>
<td>19.31</td>
<td>8.00</td>
<td>9.52</td>
<td>11.40</td>
<td>17.77</td>
<td>17.77</td>
<td>7.85</td>
<td>7.85</td>
</tr>
<tr>
<td>24</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>1</td>
<td>54.66</td>
<td>7.70</td>
<td>9.95</td>
<td>11.30</td>
<td>51.75</td>
<td>50.03</td>
<td>8.27</td>
<td>8.39</td>
</tr>
</tbody>
</table>

† The Str column indicates which working structure the parametric model uses. The local model always uses the IND structure.
‡ The Par column gives the AvgMSE of the parametric model, obtained by simulation.
†† The LLGEE column gives local linear GEE AvgMSE at the the optimal bandwidth, both of which are obtained by simulation.
7.6 MRGEE Simulation Study

Finally, we investigate the proposed methods, MRGEEI and MRGEEII. They are studied on both the $\mu$-scale and the $\eta$-scale with the given simulation models. Some major aspects of interest are the following:

- the ability of the MRGEE methods to detect mean model misspecification,
- the closeness of MRGEE mixing parameter selection to the optimal $\lambda$,
- the difference in performance of MRGEEI and MRGEEII when there is $\mu$-scale or $\eta$-scale contamination, and
- the AvgMSE performance of the MRGEE methods compared to LLGEE and the user’s parametric model—with optimal bandwidths and mixing parameters as well as data-driven bandwidth and mixing parameters.

Also of interest is how these aspects depend on factors such as the degree of misspecification ($\gamma$), sample size ($s$), the degree of correlation/variability (corr/var), the cluster size ($n_i$), and the working correlation structure ($R$) used in the parametric model. Because the number of combinations is potentially large, this simulation study is not conducted as a full factorial.

As with the bandwidth study, data-based estimation of the fits are kept to 500 runs because of computational demand. The AvgMSEs for the user’s parametric model and the optimal LLGEE, optimal MRGEEI and optimal MRGEEII are based on 1,000 runs since computations using fixed $h$ and fixed $\lambda$ are less involved. For each generated dataset, all models are estimated and their fits are summarized by average squared error loss. The average loss of the fits over all the runs yields the Monte Carlo AvgMSE.

The estimated bandwidths are obtained using the CVnoV selector. The optimal bandwidths are taken from those given in the bandwidth simulation study (§7.5) for the $\mu$-scale Poisson and binomial models, which were found by simulation. The remaining optimal bandwidths for the $\eta$-scale Poisson and binomial models are found by computation using the expressions given in Chapter 6. The computed AvgMSE is minimized over the grid $h = 0.02, 0.05, 0.1, \ldots, 1$ followed by a golden section search to determine $h_{opt}$ to within four
decimal places. The optimal LLGEE AvgMSE is obtained by simulation, where the optimal bandwidth is fixed for all generated datasets.

The estimated mixing parameters are found by the mGDys method. Estimator (105) is used for MRGEEI and estimator (106) is used for MRGEEII. Since the weight matrix in the bandwidth selector is $I_{ni}$, the same weight matrix is used for $\lambda$ estimation, i.e., $W_{I,i} = I_{ni}$ and $W_{II,i} = I_{ni}$. Thus, the MRGEEI $\lambda$ estimator used in this simulation study is the same as that suggested by Mays et al. (2001) for uncorrelated data. The optimal $\lambda$ is obtained by minimizing the computed AvgMSE (111) and the MSE expressions from Chapter 6. The optimization is conditional on $h_{opt}$. For MRGEEI, the optimal $\lambda$ is found by optimization of the computed $\mu$-scale AvgMSE. Optimization of $\lambda$ for MRGEEII occurs over the computed $\eta$-scale AvgMSE. Optimal AvgMSEs for MRGEEI and MRGEEII, both on the $\mu$-scale, are approximated by simulation (1,000 runs).

The user’s parametric mean models are given by $E[Y_{ij} | e_i = 0]$, the approximate unconditional mean, for the simulation models (pages 108–110) when $\gamma = 0$. The user’s variance functions are those commonly used in practice, the identity function for the Poisson models and $\mu(1 - \mu/50)$ for the binomial models. The working correlation structure for the parametric model is either CS or IND; CS when the generated data are correlated, IND otherwise. The AR(1) structure was also used for every case involving correlated data that follows, but the difference in the results from using the CS structure was only slight and had no impact with respect to the comparisons made. The AvgMSEs for models using the AR(1) structure tended to be slightly larger than those using CS, a reflection of the fact that CS is more appropriate for these data. The local model working structure is always IND.

The user’s $\mu$-scale binomial, $\eta$-scale Poisson, and $\eta$-scale binomial parametric models have linear predictors that are 1st degree polynomials. Consequently, the mean model of LLGEE is equivalent to the user’s parametric mean model for these simulation models when $h \rightarrow \infty$. We refer to this property as localization of the parametric model. When there is no misspecification ($\gamma = 0$), the optimal bandwidth for LLGEE in these cases will tend to be large or unbounded and the AvgMSE objective function for bandwidth will be flat over a wide range of $h$. For simulated correlated data, the parametric model and the LLGEE

---

8That is, the minimum generalized distance using $y$ in place of $m$ or $F^{-1}(y)$ in place of $\eta$. 

142
models differ only in correlation structure.

The results are shown in Tables 21–28. The η-scale cases follow the μ-scale cases and binomial cases follow the Poisson cases. For each simulation model, there are two tables, the first gives the estimated and optimal bandwidth and mixing parameters and the second shows the data-driven (termed “Estimated”) and optimal AvgMSEs. The following comments apply.

1. Bandwidth:

The η-scale average data-driven bandwidths do not agree as well with the optimal bandwidths as do the μ-scale average data-driven bandwidths. This may be due to \( h_{opt} \) being computed rather than obtained by simulation. The disagreements are the greatest for the low misspecification (0 \( \leq \gamma \leq 0.25 \)) and small sample size cases of the μ-scale binomial, η-scale Poisson, and η-scale binomial models (Tables 23–28). The AvgMSE objective functions for \( h \) tend to be flat because the LLGEE mean model is a localization of the user’s parametric mean model. Consequently, there are wide ranges of bandwidths which yield nearly identical AvgMSEs. Case 18 of the η-scale Poisson model (Table 23), where corr/var=large, \( s = 3 \), and \( \gamma = 0.25 \), is particularly interesting because the misspecification was not detectable by the LLGEE model. Simulation verified that the computed AvgMSE for the objective function is a good approximation, yet there appears to be no AvgMSE advantage for a small bandwidth until the misspecification level reaches 0.34.

2. AvgMSE of the LLGEE:

The optimal LLGEE performs as well or better than the parametric model under no misspecification. Both the optimal and the data-based LLGEE perform better than the parametric under no misspecification and high correlation for every simulation model, except case 21 of μ-Poisson (Table 22). It seems that the use of the CS structure does not give the parametric model an advantage over the simpler GLIM model under high degrees of correlation. For example, in case 9 of Table 26, the AvgMSE for the LLGEE, MRGEEI and MRGEEII (which are essentially equivalent to the LLGEE model) is better than the parametric AvgMSE by 1.04, which is a difference of 2
SEs. Simulation of the parametric model using the CS structure and then the IND structure at 5,000 runs for case 17 of the $\mu$-scale binomial (Table 26) has shown that the parametric model with the IND structure (the same model as LLGEE except for the flexibility provided by the bandwidth) leads to better performance over using the CS structure by 3.9 SEs! Our explanation for this phenomenon is that the parametric model violates the assumption of Liang and Zeger (1986) that the mean is correct, since the parametric model mean function $E[Y_{ij}|e_i = 0]$ is only approximately equal to the unconditional mean. This is a bit surprising though, as the improvement in efficiency by using CS was expected to be substantial and because it was shown that the approximate mean for the Poisson models is quite close to the true unconditional mean. Apparently, when the estimating equations are biased, the ostensible gain in efficiency by using CS over IND may be lost. Employing the user’s incorrect variance functions could also be a problem. Furthermore, the advantage of the LLGEE model with the IND working structure under misspecification may be the consequence of results in Lin and Carroll (2000) taking hold at small $h$ (see §7.3).

3. Optimal $\lambda$:  

a. The optimal $\lambda$ increases with the degree of model misspecification, illustrating that the mixing parameter can measure mean model misspecification. There are a number of exceptions for the $\eta$-scale models: cases 2, 9, 13, 17, and 21 for the Poisson (Table 23), and 9, 17, and 21 for the binomial (Table 27). But they often increase for $\gamma > 0$ when the $\gamma = 0$ case has a large bandwidth (say $h > 1$). This occurs because the local linear mean model is then a localization of the user’s parametric model in which case the parametric and the LLGEE are very competitive. This can be seen when comparing the AvgMSEs for the parametric and the LLGEE, as they tend to both fit the data well. The objective functions for $\lambda$ are rather flat, so slight changes in the AvgMSE objective function cause drastic changes in the $\lambda$ estimate. In other words, $\lambda$ is unstable because any value gives nearly an equivalent fit. Another reason is that in some instances, when corr/var is large, the local model fairs better than parametric. For example, compare cases
17 and 18 on the $\eta$-scale Poisson (Table 23). Both of these have large optimal bandwidths, implying that LLGEE is equivalent to the parametric model in mean (not in working correlation structure), but under no misspecification, LLGEE performs better than the parametric (Table 24) and under slight misspecification the parametric performs better than LLGEE. The optimal $\lambda$ therefore changes from 1 to 0 with an increase in misspecification for these particular cases. The equivalence of the parametric and LLGEE mean models suggests that their role reversal here is due mostly to Monte Carlo variability (also see comment 2).

b. The $\mu$-scale binomial cases (Table 25) have $\lambda_{opt} = 1$ for many cases because LLGEE performs very well for this model (Table 26). See comment 2 for further discussion.

c. The misspecification is detected at smaller levels as the number of clusters increases. For example, at $s = 3$, case 2 of the $\mu$-scale Poisson (Table 21), $\lambda_{opt}$ for MRGEEI and MRGEEII are 0.24 and 0.18, respectively, but at $s = 30$, case 6, the $\lambda_{opt}$ are 0.73 and 0.67, implying a more dramatic $\lambda_{opt}$ increase with increase of misspecification. This effect is examined in more detail in §7.7.

d. The MRGEEII $\lambda_{opt}$ tends to be smaller than the MRGEEI $\lambda_{opt}$ because the parametric model actually tends to do a little better in terms of $\eta$-scale AvgMSE than $\mu$-scale AvgMSE. This is examined in §7.7 for cases 17–20 of the Table 21, which exhibit the largest difference between MRGEEI and MRGEEII $\lambda_{opt}$.

4. Average Estimated $\lambda$:

   a. The average estimated $\lambda$ for MRGEEI and MRGEEII tend to be larger than optimal at low degrees of misspecification and slightly smaller than optimal at larger degrees of misspecification. The problem is especially apparent in the no misspecification ($\gamma = 0$) cases of the $\mu$-scale Poisson model (Table 21) and the no misspecification, no corr/var cases of the other models (Tables 23, 25, and 27). This is a negative feature because in most of these cases, the efficiency of MRGEEI and MRGEEII over the parametric model is less than 90% and several cases have efficiencies less than 80%.
b. The estimated $\lambda$ are able to detect increasing degrees of misspecification, but tend not to change as quickly as $\lambda_{opt}$ when the sample size is small. This suggests that the estimated $\lambda$ are less sensitive to misspecification than the optimal $\lambda$ at small $s$.

c. The average estimated $\lambda$ tends to increase with increase in corr/var for the $\mu$-scale simulation models but there is some decrease in average estimated $\lambda$ for the $\eta$-scale models. This latter relationship is weak, however, as increases in most $\gamma = 0$ cases are seen in the results.

d. The MRGEEII $\lambda$ estimates tend to be farther from optimal than those of MRGEEI for low degrees of misspecification. By far, the worst cases are those pointed out in comment 4, part a. Cases 9–24 in Table 21 are studied over a finer scale in §7.7. For the small and large corr/var cases of the binomial models, the MRGEEI and MRGEEII $\lambda$ average estimators do not differ as much and the MRGEEII estimators tend to correspond better to the optimal MRGEEII $\lambda$. The agreement between MRGEEI and MRGEEII for the binomial models is a bit surprising, as the $\eta$-scale is unbounded while the $\mu$-scale is bounded.

e. The comparison between optimal and average $\lambda$ is made more difficult by each using a different criteria for optimization. Like the optimal $h$, the optimal $\lambda$ is the best mixing parameter when the mixing parameter is held constant across the datasets. The estimated $\lambda$ is allowed to be chosen differently for each dataset and is conditional on the estimated bandwidth which is chosen differently for each dataset. A comparison of the estimated $\lambda$ with the best possible $\lambda$ for each dataset is left to future research.

5. Comparison of Optimal AvgMSEs for the Parametric, LLGEE, MRGEEI, and MRGEEII:

a. Overall, the optimal MRGEEI or MRGEEII performs as well as or better than both the parametric and LLGEE in every combination except cases 19–24 of the $\eta$-scale binomial model (Table 28) where the methods perform only slightly worse than LLGEE (well within Monte Carlo variability) under high misspecification and high corr/var. This exemplifies the potential of the method. The mixing
approach can improve AvgMSE over both the parametric and the nonparametric approaches.

b. The optimal performance of MRGEEI is superior to MRGEEII in most cases, even for the $\eta$-scale models, which are additively contaminated on the $\eta$-scale. However, MRGEEII performs competitively with MRGEEI for the $\eta$-scale Poisson model (Table 24).

c. The optimal MRGEEI often has the same performance as the optimal LLGEE, especially in the $\eta$-Poisson and binomial tables (24, 26, and 28), where the performance of LLGEE and MRGEEI is the same as or better than the parametric model. This is because the local model as a localization of the parametric model can act as either the parametric model, as $h \to \infty$, or as a nonparametric model for small $h$. Its fits can range from interpolation of the means to the estimated parametric model, with an IND working structure. For example, in no corr/var cases 1 and 5 of the $\eta$-Poisson and binomial models, all optimal methods have the same performance. In these cases they represent the same model. For other cases, the LLGEE and MRGEEI can outperform the parametric, even under no misspecification. See comment 2 for explanations why LLGEE performs so well.

d. The improvement in optimal performance of MRGEEI and MRGEEII over the parametric and LLGEE tends to be greatest when misspecification is small to moderate ($\gamma = 0.25$; see especially the $\mu$-scale Poisson cases, Table 22). This indicates that the optimal MRGEE methods tend to be most efficient for low to moderate degrees of misspecification.

6. Comparison of AvgMSEs for the Parametric and the estimated (data-driven) LLGEE, MRGEEI, and MRGEEII:

a. Overall, the performance of the data-driven MRGEEI and MRGEEII falls in-between the parametric and LLGEE when there is no misspecification (see comment 4, part a). They perform about as well as LLGEE under high degrees of misspecification, and in several cases outperform both the parametric and the nonparametric at low levels of misspecification.
b. For the small and no corr/var cases, as misspecification increases from $\gamma = 0$ to 1, there is transition in best performance from the parametric model to the data-driven MRGEEI or MRGEEII and then to the data-driven LLGEE. This phenomenon is termed the cross-over effect: the MRGEE methods perform better than both the parametric and LLGEE over a range of misspecification where the AvgMSE of the parametric crosses above the LLGEE AvgMSE. The advantage of this cross-over effect for MRGEE is not always apparent in these tables because of the coarse scale for $\gamma$. For example, in cases 9–12 of the $\mu$-scale Poisson model (Table 22), there is no point at which the MRGEE methods outperform either the parametric or the nonparametric model, but in §7.7, it is shown that MRGEEI outperforms both (cross-over effect) between $\gamma = 0.6$ and 0.8 (Figure 19).

c. The cross-over effect tends to occur at smaller degrees of misspecification as sample size increases. This holds for both Poisson and binomial models, but is more noticeable in the Poisson models.

d. The high corr/var cases are particularly interesting because the MRGEE methods often perform best when there is no misspecification, and perform the same or about as well as LLGEE under misspecification. Exceptions include cases 17–20 ($s = 3$) of the $\eta$-scale Poisson (Table 24) where the MRGEE methods always perform best, cases 17–20 ($s = 3$) of the $\mu$-scale binomial (Table 26)—where there is a cross-over effect—and cases 17–20 ($s = 3$) of the $\eta$-scale binomial (Table 28) where LLGEE always performs best. Also see comment 2.

e. The data-driven performances of MRGEEI and MRGEEII are similar, but MRGEEI tends to have a smaller AvgMSE more often, which corresponds with the optimal results. For the $\mu$-Poisson model, MRGEEII tends to perform best when $s = 30$ and the mean model is misspecified (Table 22). For the other simulation models, in which LLGEE is a localization of the user’s parametric model, the opposite tends to hold; MRGEEII performs better than MRGEEI when there is no mean model misspecification. This appears to be a consequence of the better performance of the local model in these cases (see comment 2) and MRGEEII’s
tendency to bias \( \lambda \) toward the the local fit at low degrees of misspecification (see comment 4, part a).

7. **Summary of the Effect of Correlation/Variability**

a. The optimal and data-driven bandwidth estimates increase as corr/var increases. This matches our intuition that moderately variable correlated data require more smoothing than less variable uncorrelated data because they carry less information about the mean function.

b. The AvgMSE performance of the LLGEE method with IND working structure improves as corr/var increases—an effect explained by misspecification of the mean model in comment 2.

c. The optimal and average data-driven \( \lambda \) tend to increase slightly as corr/var increases for the \( \mu \)-scale models but the opposite tends to hold for the \( \eta \)-scale models.

d. The data-driven MRGEEI and MRGEEII tend to improve, with respect to the parametric model and LLGEE, as corr/var increases. For the Poisson models, MRGEEI and MRGEEII beat both the parametric and LLGEE at the highest level of corr/var. Thus, *increases in corr/var tend to lead to better data-driven MRGEE performance*.

8. The \( \mu \)-scale binomial no corr/var cases (1–8, Table 26) correspond to models that are similar to those used by Nottingham (1995). It is difficult to compare his results to ours because he did not consider more than 7 design points, his binomial counts are out of 20 or less, he used only a single cluster, his local methods use ordinary kernel regression and local linear regression rather than local GLIM estimation, and he used the PRESS* bandwidth selector rather than CVnoV. Nevertheless, Table 8.5.2.1 of Nottingham (1995, pg. 137), where \( n_i = 7 \) and \( s = 1 \), illustrates that a cross-over effect occurs at \( \gamma = 0.3 \). The cross-over effect in our cases occurred at \( \gamma = 0.25 \) or less.
9. **The Effects of the User’s Naïve Mean Model at $\gamma = 0$**

The user’s mean model is naïvely taken as the mean model from a generalized linear model, given by the conditional mean $E[Y_{ij}|e_i = 0]$, and only equivalent to the true unconditional mean at $\gamma = 0$ to a first order Taylor series approximation. These simulation studies reveal that there are significant consequences of mean model misspecification when the degree of misspecification from the true model is the smallest at $\gamma = 0$. The user’s model at $\gamma = 0$ becomes a poorer representation of the unconditional mean as the degree of corr/var increases. Two important consequences are summarized here.

a. As seen in the bandwidth study and discussed in comment 2 in this section, using the IND working structure rather than the CS structure in the user’s parametric model when the data are correlated leads to better fits. As a consequence, LLGEE with an IND working structure performs better than the parametric model in many cases, even at $\gamma = 0$. Using the IND structure leads to the best performance over the CS structure at the highest levels of correlation because the user’s model at $\gamma = 0$ is the most misspecified when corr/var is large.

b. The optimal and data-driven mixing parameters at $\gamma = 0$ tend to increase, and the corresponding AvgMSE efficiency of the user’s parametric model to the other models tends to decrease, as corr/var increases.

The level of impact of using a naïve mean model in this study was surprising.

10. **Finally, we wish to emphasize that one should interpret these results with some caution because the Monte Carlo AvgMSEs are subject to variation. There are several instances where the Monte Carlo AvgMSEs are too close with respect to the variability to claim a statistically significant difference. The computational demands of these methods enabled us to evaluate estimates for only 500 datasets for each of the data-driven LLGEE and MRGEE models, and for 1,000 datasets for each of the parametric and optimal models. Computing time is also the reason why we use AvgMSE rather than IMSE to summarize the fits.**
In conclusion, the optimal MRGEE methods perform as well or better than the parametric and LLGEE methods. The data-driven MRGEE methods tend to beat both methods at small degrees of misspecification. At times, the performance is not as good as parametric under no misspecification and at small sample size, but the MRGEE methods tend to become more sensitive to misspecification as sample size increases. MRGEEI performs better than MRGEEII in terms of theoretical $\mu$-scale AvgMSE (on which all our methods have been compared), even with $\eta$-scale contaminated models, but the data-driven MRGEEI and MRGEEII are more competitive.
Table 21: Estimated and Optimal Bandwidth and Mixing Parameters for the $\mu$-scale Poisson Simulation Model. $n = 10$ for all cases.

<table>
<thead>
<tr>
<th>Row</th>
<th>Str</th>
<th>Corr/Var</th>
<th>$s$</th>
<th>$\gamma$</th>
<th>$h_{opt}$</th>
<th>$h$</th>
<th>$\lambda_{opt}$</th>
<th>$\lambda$</th>
<th>$\lambda_{opt}$</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0</td>
<td>0.13</td>
<td>0.14</td>
<td>0.10</td>
<td>0.36</td>
<td>0.06</td>
<td>0.42</td>
</tr>
<tr>
<td>2</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.25</td>
<td>0.12</td>
<td>0.14</td>
<td>0.24</td>
<td>0.40</td>
<td>0.18</td>
<td>0.44</td>
</tr>
<tr>
<td>3</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.5</td>
<td>0.10</td>
<td>0.13</td>
<td>0.62</td>
<td>0.52</td>
<td>0.52</td>
<td>0.53</td>
</tr>
<tr>
<td>4</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>1</td>
<td>0.07</td>
<td>0.09</td>
<td>0.98</td>
<td>0.80</td>
<td>0.92</td>
<td>0.78</td>
</tr>
<tr>
<td>5</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0</td>
<td>0.08</td>
<td>0.08</td>
<td>0.02</td>
<td>0.17</td>
<td>0.01</td>
<td>0.36</td>
</tr>
<tr>
<td>6</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.25</td>
<td>0.07</td>
<td>0.07</td>
<td>0.73</td>
<td>0.59</td>
<td>0.67</td>
<td>0.74</td>
</tr>
<tr>
<td>7</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.5</td>
<td>0.06</td>
<td>0.06</td>
<td>1.00</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>8</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>1</td>
<td>0.04</td>
<td>0.04</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>9</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0</td>
<td>0.13</td>
<td>0.15</td>
<td>0.17</td>
<td>0.42</td>
<td>0.06</td>
<td>0.45</td>
</tr>
<tr>
<td>10</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.25</td>
<td>0.12</td>
<td>0.14</td>
<td>0.30</td>
<td>0.46</td>
<td>0.18</td>
<td>0.48</td>
</tr>
<tr>
<td>11</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.5</td>
<td>0.10</td>
<td>0.13</td>
<td>0.65</td>
<td>0.57</td>
<td>0.52</td>
<td>0.56</td>
</tr>
<tr>
<td>12</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>1</td>
<td>0.07</td>
<td>0.09</td>
<td>0.98</td>
<td>0.83</td>
<td>0.91</td>
<td>0.79</td>
</tr>
<tr>
<td>13</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0</td>
<td>0.08</td>
<td>0.08</td>
<td>0.03</td>
<td>0.17</td>
<td>0.01</td>
<td>0.51</td>
</tr>
<tr>
<td>14</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.25</td>
<td>0.07</td>
<td>0.07</td>
<td>0.73</td>
<td>0.59</td>
<td>0.67</td>
<td>0.83</td>
</tr>
<tr>
<td>15</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.5</td>
<td>0.06</td>
<td>0.06</td>
<td>1.00</td>
<td>0.97</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td>16</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>1</td>
<td>0.04</td>
<td>0.04</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>17</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0</td>
<td>0.17</td>
<td>0.19</td>
<td>0.93</td>
<td>0.75</td>
<td>0.10</td>
<td>0.66</td>
</tr>
<tr>
<td>18</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.25</td>
<td>0.17</td>
<td>0.18</td>
<td>0.99</td>
<td>0.76</td>
<td>0.14</td>
<td>0.68</td>
</tr>
<tr>
<td>19</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.5</td>
<td>0.14</td>
<td>0.17</td>
<td>1.00</td>
<td>0.80</td>
<td>0.38</td>
<td>0.71</td>
</tr>
<tr>
<td>20</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>1</td>
<td>0.08</td>
<td>0.13</td>
<td>1.00</td>
<td>0.91</td>
<td>0.89</td>
<td>0.86</td>
</tr>
<tr>
<td>21</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0</td>
<td>0.09</td>
<td>0.09</td>
<td>0.22</td>
<td>0.26</td>
<td>0.02</td>
<td>0.99</td>
</tr>
<tr>
<td>22</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.25</td>
<td>0.08</td>
<td>0.08</td>
<td>0.79</td>
<td>0.66</td>
<td>0.65</td>
<td>1.00</td>
</tr>
<tr>
<td>23</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.5</td>
<td>0.06</td>
<td>0.06</td>
<td>1.00</td>
<td>0.97</td>
<td>0.98</td>
<td>1.00</td>
</tr>
<tr>
<td>24</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>1</td>
<td>0.05</td>
<td>0.05</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

* The Str column indicates which working structure the parametric model uses. The local model always uses the IND structure.

‡ $h$ is estimated using the CVnoV selector. $h_{opt}$ is the optimal bandwidth obtained by simulation.

† $\lambda$ is estimated by mGDys (105) with $W_{I,i} = I_n$, and $\lambda_{opt}$ is obtained by optimizing the computed AvgMSE on the $\mu$-scale conditional on the optimal $h$ obtained by simulation.

†† $\lambda$ is estimated by mGDys (106) with $W_{II,i} = I_n$, and $\lambda_{opt}$ is obtained by optimizing the computed AvgMSE on the $\eta$-scale conditional on the optimal $h$ obtained by simulation.

152
Table 22: Estimated and Optimal AvgMSEs for LLGEE, MRGEEI, MRGEEII, and the Parametric Model for the \( \mu \)-scale Poisson Simulation Model. The bold values in columns 6–9 indicate the best AvgMSE among the optimal LLGEE, MRGEEI, MRGEEII, and the parametric model. The bold values in columns 10–13 indicate the best AvgMSEs among the fully estimated LLGEE, MRGEEI, MRGEEII, and parametric models. \( n = 10 \) for all cases.

<table>
<thead>
<tr>
<th>Row</th>
<th>Str(^*)</th>
<th>Corr/Var</th>
<th>( s )</th>
<th>( \gamma )</th>
<th>Optimal(^\dagger)</th>
<th>Estimated(^\ddagger)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Par†</td>
<td>LL</td>
</tr>
<tr>
<td>1</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0</td>
<td>2.50</td>
<td>3.61</td>
</tr>
<tr>
<td>2</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.25</td>
<td>3.16</td>
<td>3.92</td>
</tr>
<tr>
<td>3</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.5</td>
<td>5.15</td>
<td>4.67</td>
</tr>
<tr>
<td>4</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>1</td>
<td>13.09</td>
<td>6.14</td>
</tr>
<tr>
<td>5</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0</td>
<td>0.26</td>
<td>0.54</td>
</tr>
<tr>
<td>6</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.25</td>
<td>0.92</td>
<td>0.62</td>
</tr>
<tr>
<td>7</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.5</td>
<td>2.90</td>
<td>0.74</td>
</tr>
<tr>
<td>8</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>1</td>
<td>10.81</td>
<td>0.85</td>
</tr>
<tr>
<td>9</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0</td>
<td>6.33</td>
<td>7.35</td>
</tr>
<tr>
<td>10</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.25</td>
<td>6.97</td>
<td>7.67</td>
</tr>
<tr>
<td>11</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.5</td>
<td>8.98</td>
<td>8.46</td>
</tr>
<tr>
<td>12</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>1</td>
<td>16.88</td>
<td>9.96</td>
</tr>
<tr>
<td>13</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0</td>
<td>0.64</td>
<td>0.91</td>
</tr>
<tr>
<td>14</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.25</td>
<td>1.29</td>
<td>0.99</td>
</tr>
<tr>
<td>15</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.5</td>
<td>3.28</td>
<td>1.12</td>
</tr>
<tr>
<td>16</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>1</td>
<td>11.20</td>
<td>1.23</td>
</tr>
<tr>
<td>17</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0</td>
<td>55.60</td>
<td>54.18</td>
</tr>
<tr>
<td>18</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.25</td>
<td>57.22</td>
<td>55.43</td>
</tr>
<tr>
<td>19</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.5</td>
<td>58.25</td>
<td>55.90</td>
</tr>
<tr>
<td>20</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>1</td>
<td>66.95</td>
<td>59.37</td>
</tr>
<tr>
<td>21</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0</td>
<td>5.63</td>
<td>5.77</td>
</tr>
<tr>
<td>22</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.25</td>
<td>6.29</td>
<td>5.89</td>
</tr>
<tr>
<td>23</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.5</td>
<td>8.29</td>
<td>6.09</td>
</tr>
<tr>
<td>24</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>1</td>
<td>16.24</td>
<td>6.33</td>
</tr>
</tbody>
</table>

* The Str column indicates which working structure the parametric model uses. The local model always uses the IND structure.

† The Par columns are the same and give the AvgMSE of the parametric model, obtained by simulation.

‡ The optimal LL column gives the AvgMSE of LLGEE at the optimal \( h \), both obtained by simulation. The optimal MRI and MRII columns give the computed AvgMSE of the MRGEEI and MRGEEII, respectively, at the optimal \( h \) (by sim.) and optimal \( \lambda \) (by comp.).

†† The estimated LL, MRI, MRII columns give the AvgMSE of the LLGEE, MRGEEI, and MRGEEII estimates, respectively. The LLGEE is based on CVnoV \( h \) selection, MRGEEI and MRGEEII are based on mGDys \( \lambda \) selection, and all the AvgMSEs are presented on the \( \mu \)-scale.
Table 23: Estimated and Optimal Bandwidth and Mixing Parameters for the $\eta$-scale Poisson Simulation Model. $n = 10$ for all cases.

<table>
<thead>
<tr>
<th>Row</th>
<th>Str*</th>
<th>Corr/Var</th>
<th>$s$</th>
<th>$\gamma$</th>
<th>LLGEE‡ $h_{opt}$</th>
<th>MRGEEI† $\lambda_{opt}$</th>
<th>MRGEEII†† $\lambda_{opt}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0</td>
<td>16.50</td>
<td>0.00 0.54</td>
<td>0.00 0.44</td>
</tr>
<tr>
<td>2</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.25</td>
<td>1.02</td>
<td>1.00 0.57</td>
<td>1.00 0.47</td>
</tr>
<tr>
<td>3</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.5</td>
<td>0.08</td>
<td>0.85 0.73</td>
<td>0.73 0.57</td>
</tr>
<tr>
<td>4</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>1</td>
<td>0.06</td>
<td>0.99 0.93</td>
<td>0.96 0.89</td>
</tr>
<tr>
<td>5</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0</td>
<td>8.00</td>
<td>0.00 0.38</td>
<td>0.00 0.35</td>
</tr>
<tr>
<td>6</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.25</td>
<td>0.06</td>
<td>0.97 0.95</td>
<td>0.90 0.83</td>
</tr>
<tr>
<td>7</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.5</td>
<td>0.05</td>
<td>1.00 1.00</td>
<td>0.99 0.99</td>
</tr>
<tr>
<td>8</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>1</td>
<td>0.04</td>
<td>1.00 1.00</td>
<td>1.00 1.00</td>
</tr>
<tr>
<td>9</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0</td>
<td>14.60</td>
<td>1.00 0.56</td>
<td>1.00 0.47</td>
</tr>
<tr>
<td>10</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.25</td>
<td>1.14</td>
<td>0.55 0.61</td>
<td>0.59 0.46</td>
</tr>
<tr>
<td>11</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.5</td>
<td>0.08</td>
<td>0.84 0.71</td>
<td>0.73 0.62</td>
</tr>
<tr>
<td>12</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>1</td>
<td>0.06</td>
<td>0.98 0.93</td>
<td>0.97 0.91</td>
</tr>
<tr>
<td>13</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0</td>
<td>9.80</td>
<td>0.99 0.42</td>
<td>0.99 0.46</td>
</tr>
<tr>
<td>14</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.25</td>
<td>0.06</td>
<td>0.97 0.94</td>
<td>0.90 0.83</td>
</tr>
<tr>
<td>15</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.5</td>
<td>0.05</td>
<td>1.00 1.00</td>
<td>0.99 0.99</td>
</tr>
<tr>
<td>16</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>1</td>
<td>0.04</td>
<td>1.00 1.00</td>
<td>1.00 1.00</td>
</tr>
<tr>
<td>17</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0</td>
<td>13.00</td>
<td>1.00 0.59</td>
<td>1.00 0.52</td>
</tr>
<tr>
<td>18</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.25</td>
<td>9.10</td>
<td>0.00 0.59</td>
<td>0.00 0.51</td>
</tr>
<tr>
<td>19</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.5</td>
<td>0.07</td>
<td>0.62 0.60</td>
<td>0.61 0.55</td>
</tr>
<tr>
<td>20</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>1</td>
<td>0.05</td>
<td>0.86 0.78</td>
<td>0.99 0.85</td>
</tr>
<tr>
<td>21</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0</td>
<td>8.75</td>
<td>1.00 0.64</td>
<td>0.95 0.51</td>
</tr>
<tr>
<td>22</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.25</td>
<td>0.06</td>
<td>0.90 0.88</td>
<td>0.41 0.38</td>
</tr>
<tr>
<td>23</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.5</td>
<td>0.04</td>
<td>0.97 0.97</td>
<td>0.83 0.80</td>
</tr>
<tr>
<td>24</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>1</td>
<td>0.04</td>
<td>0.99 0.99</td>
<td>1.00 1.00</td>
</tr>
</tbody>
</table>

* The Str column indicates which working structure the parametric model uses. The local model always uses the IND structure.
† $h$ is estimated using the CVnoV selector. $h_{opt}$ is the optimal bandwidth obtained by computation.
‡ $\lambda$ is estimated by mGDys (105) with $W_{I,i} = I_n$, and $\lambda_{opt}$ is obtained by optimizing the computed AvgMSE on the $\mu$-scale conditional on the optimal $h$ obtained by computation.
†† $\lambda$ is estimated by mGDys (106) with $W_{II,i} = I_n$, and $\lambda_{opt}$ and $\lambda_{opt}$ is obtained by optimizing the computed AvgMSE on the $\eta$-scale conditional on the optimal $h$ obtained by computation.
Table 24: Estimated and Optimal AvgMSEs for LLGEE, MRGEEI, MRGEEII, and the Parametric Model for the $\eta$-scale Poisson Simulation Model. The bold values in columns 6–9 indicate the best AvgMSE among the optimal LLGEE, MRGEEI, MRGEEII, and the parametric model. The bold values in columns 10–13 indicate the best AvgMSEs among the fully estimated LLGEE, MRGEEI, MRGEEII, and parametric models. $n = 10$ for all cases.

<table>
<thead>
<tr>
<th>Row</th>
<th>Str*</th>
<th>Corr/Var</th>
<th>s</th>
<th>$\gamma$</th>
<th>Par†</th>
<th>Optimal‡</th>
<th>Estimated¶</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LL</td>
<td>MRI</td>
</tr>
<tr>
<td>1</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0</td>
<td>1.30</td>
<td>1.30</td>
<td>1.30</td>
</tr>
<tr>
<td>2</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.25</td>
<td>2.95</td>
<td>2.94</td>
<td>2.94</td>
</tr>
<tr>
<td>3</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.5</td>
<td>7.73</td>
<td>3.96</td>
<td>3.83</td>
</tr>
<tr>
<td>4</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>1</td>
<td>27.22</td>
<td>5.08</td>
<td>5.07</td>
</tr>
<tr>
<td>5</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>6</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.25</td>
<td>1.78</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td>7</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.5</td>
<td>6.62</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>8</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>1</td>
<td>26.10</td>
<td>0.59</td>
<td>0.59</td>
</tr>
<tr>
<td>9</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0</td>
<td>4.25</td>
<td>4.23</td>
<td>4.23</td>
</tr>
<tr>
<td>10</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.25</td>
<td>6.08</td>
<td>6.09</td>
<td>6.08</td>
</tr>
<tr>
<td>11</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.5</td>
<td>11.02</td>
<td>7.25</td>
<td>7.08</td>
</tr>
<tr>
<td>12</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>1</td>
<td>31.07</td>
<td>8.57</td>
<td>8.57</td>
</tr>
<tr>
<td>13</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0</td>
<td>0.43</td>
<td>0.43</td>
<td>0.43</td>
</tr>
<tr>
<td>14</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.25</td>
<td>2.09</td>
<td>0.76</td>
<td>0.76</td>
</tr>
<tr>
<td>15</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.5</td>
<td>7.08</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>16</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>1</td>
<td>27.02</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>17</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0</td>
<td>62.13</td>
<td>61.62</td>
<td>61.62</td>
</tr>
<tr>
<td>18</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.25</td>
<td>61.33</td>
<td>63.50</td>
<td>61.33</td>
</tr>
<tr>
<td>19</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.5</td>
<td>67.57</td>
<td>65.87</td>
<td>64.36</td>
</tr>
<tr>
<td>20</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>1</td>
<td>96.53</td>
<td>71.99</td>
<td>71.41</td>
</tr>
<tr>
<td>21</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0</td>
<td>6.23</td>
<td>6.12</td>
<td>6.12</td>
</tr>
<tr>
<td>22</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.25</td>
<td>8.12</td>
<td>6.59</td>
<td>6.55</td>
</tr>
<tr>
<td>23</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.5</td>
<td>14.53</td>
<td>6.74</td>
<td>6.71</td>
</tr>
<tr>
<td>24</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>1</td>
<td>41.20</td>
<td>7.43</td>
<td>7.42</td>
</tr>
</tbody>
</table>

* The Str column indicates which working structure the parametric model uses. The local model always uses the IND structure.

† The Par columns are the same and give the AvgMSE of the parametric model, obtained by simulation.

‡ The optimal LL column gives the AvgMSE of LLGEE at the optimal $h$, both obtained by computation. The optimal MRI and MRII columns give the computed AvgMSE of the MRGEEI and MRGEEII, respectively, at the optimal $h$ (by comp.) and optimal $\lambda$ (by comp.).

¶ The estimated LL, MRI, MRII columns give the AvgMSE of the LLGEE, MRGEEI, and MRGEEII estimates, respectively. The LLGEE is based on CVnoV $h$ selection, MRGEEI and MRGEEII are based on mGDys $\lambda$ selection, and all the AvgMSEs are presented on the $\mu$-scale.
Table 25: Estimated and Optimal Bandwidth and Mixing Parameters for the $\mu$-scale Binomial Simulation Model. $n = 10$ for all cases.

<table>
<thead>
<tr>
<th>Row</th>
<th>Str*</th>
<th>Corr/Var</th>
<th>$s$</th>
<th>$\gamma$</th>
<th>$h_{opt}$</th>
<th>$h$</th>
<th>$\lambda_{opt}$</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0</td>
<td>1.57</td>
<td>0.64</td>
<td>0.00</td>
<td>0.46</td>
</tr>
<tr>
<td>2</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.25</td>
<td>0.18</td>
<td>0.46</td>
<td>0.83</td>
<td>0.61</td>
</tr>
<tr>
<td>3</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.5</td>
<td>0.11</td>
<td>0.19</td>
<td>1.00</td>
<td>0.87</td>
</tr>
<tr>
<td>4</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>1</td>
<td>0.08</td>
<td>0.09</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>5</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0</td>
<td>1.12</td>
<td>0.69</td>
<td>0.00</td>
<td>0.40</td>
</tr>
<tr>
<td>6</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.25</td>
<td>0.09</td>
<td>0.09</td>
<td>1.00</td>
<td>0.98</td>
</tr>
<tr>
<td>7</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.5</td>
<td>0.07</td>
<td>0.07</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>8</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>1</td>
<td>0.05</td>
<td>0.05</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>9</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0</td>
<td>1.68</td>
<td>0.62</td>
<td>0.99</td>
<td>0.54</td>
</tr>
<tr>
<td>10</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.25</td>
<td>0.20</td>
<td>0.43</td>
<td>1.00</td>
<td>0.70</td>
</tr>
<tr>
<td>11</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.5</td>
<td>0.12</td>
<td>0.18</td>
<td>1.00</td>
<td>0.90</td>
</tr>
<tr>
<td>12</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>1</td>
<td>0.08</td>
<td>0.09</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>13</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0</td>
<td>1.24</td>
<td>0.69</td>
<td>0.99</td>
<td>0.75</td>
</tr>
<tr>
<td>14</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.25</td>
<td>0.10</td>
<td>0.09</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>15</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.5</td>
<td>0.07</td>
<td>0.07</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>16</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>1</td>
<td>0.05</td>
<td>0.05</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>17</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0</td>
<td>3.94</td>
<td>0.49</td>
<td>1.00</td>
<td>0.64</td>
</tr>
<tr>
<td>18</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.25</td>
<td>0.22</td>
<td>0.32</td>
<td>1.00</td>
<td>0.80</td>
</tr>
<tr>
<td>19</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.5</td>
<td>0.18</td>
<td>0.19</td>
<td>1.00</td>
<td>0.94</td>
</tr>
<tr>
<td>20</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>1</td>
<td>0.12</td>
<td>0.12</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>21</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0</td>
<td>3.30</td>
<td>0.62</td>
<td>1.00</td>
<td>0.64</td>
</tr>
<tr>
<td>22</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.25</td>
<td>0.16</td>
<td>0.09</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>23</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.5</td>
<td>0.13</td>
<td>0.07</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>24</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>1</td>
<td>0.09</td>
<td>0.05</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

* The Str column indicates which working structure the parametric model uses. The local model always uses the IND structure.

† $h$ is estimated using the CVnoV selector. $h_{opt}$ is the optimal bandwidth obtained by simulation.

‡ $\lambda$ is estimated by mGDys (105) with $W_{I,i} = I_{ni}$ and $\lambda_{opt}$ is obtained by optimizing the computed AvgMSE on the $\mu$-scale conditional on the optimal $h$ obtained by simulation.

†† $\lambda$ is estimated by mGDys (106) with $W_{II,i} = I_{ni}$ and $\lambda_{opt}$ is obtained by optimizing the computed AvgMSE on the $\eta$-scale conditional on the optimal $h$ obtained by simulation.
Table 26: Estimated and Optimal AvgMSEs for LLGEE, MRGEEI, MRGEEII, and the Parametric Model for the $\mu$-scale Binomial Simulation Model. The bold values in columns 6–9 indicate the best AvgMSE among the optimal LLGEE, MRGEEI, MRGEEII, and the parametric model. The bold values in columns 10–13 indicate the best AvgMSEs among the fully estimated LLGEE, MRGEEI, MRGEEII, and parametric models. The AvgMSEs have been scaled by $10^5$. $n = 10$ for all cases.

<table>
<thead>
<tr>
<th>Row</th>
<th>Str*</th>
<th>Corr/Var</th>
<th>s</th>
<th>$\gamma$</th>
<th>Par†</th>
<th>Optimal†</th>
<th>Estimated‖</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LL</td>
<td>MRI</td>
<td>MRII</td>
</tr>
<tr>
<td>1</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0</td>
<td>2.06</td>
<td>2.06</td>
<td>2.06</td>
</tr>
<tr>
<td>2</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.25</td>
<td>4.82</td>
<td>3.78</td>
<td>3.74</td>
</tr>
<tr>
<td>3</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.5</td>
<td>13.44</td>
<td>5.21</td>
<td>5.21</td>
</tr>
<tr>
<td>4</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>1</td>
<td>51.41</td>
<td>7.44</td>
<td>7.44</td>
</tr>
<tr>
<td>5</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0</td>
<td>0.21</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>6</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.25</td>
<td>2.85</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>7</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.5</td>
<td>11.37</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>8</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>1</td>
<td>49.20</td>
<td>1.13</td>
<td>1.13</td>
</tr>
<tr>
<td>9</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0</td>
<td>9.53</td>
<td>8.49</td>
<td>8.49</td>
</tr>
<tr>
<td>10</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.25</td>
<td>11.67</td>
<td>9.64</td>
<td>9.64</td>
</tr>
<tr>
<td>11</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.5</td>
<td>19.39</td>
<td>10.92</td>
<td>10.92</td>
</tr>
<tr>
<td>12</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>1</td>
<td>56.41</td>
<td>12.90</td>
<td>12.90</td>
</tr>
<tr>
<td>13</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0</td>
<td>0.96</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>14</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.25</td>
<td>3.36</td>
<td>1.12</td>
<td>1.12</td>
</tr>
<tr>
<td>15</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.5</td>
<td>11.70</td>
<td>1.34</td>
<td>1.34</td>
</tr>
<tr>
<td>16</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>1</td>
<td>49.28</td>
<td>1.64</td>
<td>1.64</td>
</tr>
<tr>
<td>17</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0</td>
<td>69.35</td>
<td>65.90</td>
<td>65.90</td>
</tr>
<tr>
<td>18</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.25</td>
<td>74.60</td>
<td>62.47</td>
<td>62.47</td>
</tr>
<tr>
<td>19</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.5</td>
<td>81.42</td>
<td>61.30</td>
<td>61.30</td>
</tr>
<tr>
<td>20</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>1</td>
<td>106.71</td>
<td>58.71</td>
<td>58.71</td>
</tr>
<tr>
<td>21</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0</td>
<td>10.87</td>
<td>10.28</td>
<td>10.28</td>
</tr>
<tr>
<td>22</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.25</td>
<td>12.30</td>
<td>8.64</td>
<td>8.64</td>
</tr>
<tr>
<td>23</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.5</td>
<td>19.65</td>
<td>8.00</td>
<td>8.00</td>
</tr>
<tr>
<td>24</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>1</td>
<td>55.01</td>
<td>7.70</td>
<td>7.70</td>
</tr>
</tbody>
</table>

* The Str column indicates which working structure the parametric model uses. The local model always uses the IND structure.
† The Par columns are the same and give the AvgMSE of the parametric model, obtained by simulation.
‡ The optimal LL column gives the AvgMSE of LLGEE at the optimal $h$, both obtained by computation. The optimal MRI and MRII columns give the computed AvgMSE of the MRGEEI and MRGEEII, respectively, at the optimal $h$ (by sim.) and optimal $\lambda$ (by comp.).
‖ The estimated LL, MRI, MRII columns give the AvgMSE of the LLGEE, MRGEEI, and MRGEEII estimates, respectively. The LLGEE is based on CVnoV $h$ selection, MRGEEI and MRGEEII are based on mGDys $\lambda$ selection, and all the AvgMSEs are presented on the $\mu$-scale.
Table 27: Estimated and Optimal Bandwidth and Mixing Parameters for the \( \eta \)-scale Binomial Simulation Model. \( n = 10 \) for all cases.

<table>
<thead>
<tr>
<th>Row</th>
<th>Str*</th>
<th>Corr/Var</th>
<th>s</th>
<th>( \gamma )</th>
<th>LLGEE‡</th>
<th>MRGEEI‡</th>
<th>MRGEEII††</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0</td>
<td>1.80</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.25</td>
<td>0.12</td>
<td>0.62</td>
<td>0.30</td>
</tr>
<tr>
<td>3</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.5</td>
<td>0.08</td>
<td>0.97</td>
<td>0.68</td>
</tr>
<tr>
<td>4</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>1</td>
<td>0.06</td>
<td>1.00</td>
<td>0.93</td>
</tr>
<tr>
<td>5</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0</td>
<td>1.20</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.25</td>
<td>0.06</td>
<td>1.00</td>
<td>0.87</td>
</tr>
<tr>
<td>7</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.5</td>
<td>0.05</td>
<td>1.00</td>
<td>0.98</td>
</tr>
<tr>
<td>8</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>1</td>
<td>0.04</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>9</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0</td>
<td>2.80</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>10</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.25</td>
<td>0.12</td>
<td>0.68</td>
<td>0.38</td>
</tr>
<tr>
<td>11</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.5</td>
<td>0.08</td>
<td>0.97</td>
<td>0.67</td>
</tr>
<tr>
<td>12</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>1</td>
<td>0.06</td>
<td>1.00</td>
<td>0.92</td>
</tr>
<tr>
<td>13</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0</td>
<td>1.50</td>
<td>0.99</td>
<td>0.96</td>
</tr>
<tr>
<td>14</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.25</td>
<td>0.06</td>
<td>1.00</td>
<td>0.86</td>
</tr>
<tr>
<td>15</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.5</td>
<td>0.05</td>
<td>1.00</td>
<td>0.97</td>
</tr>
<tr>
<td>16</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>1</td>
<td>0.04</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>17</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0</td>
<td>2.00</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>18</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.25</td>
<td>0.10</td>
<td>0.68</td>
<td>0.51</td>
</tr>
<tr>
<td>19</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.5</td>
<td>0.07</td>
<td>0.72</td>
<td>0.52</td>
</tr>
<tr>
<td>20</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>1</td>
<td>0.05</td>
<td>0.88</td>
<td>0.75</td>
</tr>
<tr>
<td>21</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0</td>
<td>0.90</td>
<td>0.99</td>
<td>0.97</td>
</tr>
<tr>
<td>22</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.25</td>
<td>0.06</td>
<td>0.97</td>
<td>0.80</td>
</tr>
<tr>
<td>23</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.5</td>
<td>0.05</td>
<td>0.98</td>
<td>0.92</td>
</tr>
<tr>
<td>24</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>1</td>
<td>0.04</td>
<td>0.99</td>
<td>0.97</td>
</tr>
</tbody>
</table>

* The Str column indicates which working structure the parametric model uses. The local model always uses the IND structure.

‡ \( h \) is estimated using the CVnoV selector. \( h_{opt} \) is the optimal bandwidth obtained by computation.

† \( \lambda \) is estimated by mGDys (105) with \( W_{I,i} = I_n \) and \( \lambda_{opt} \) is obtained by optimizing the computed AvgMSE on the \( \mu \)-scale conditional on the optimal \( h \) obtained by computation.

†† \( \lambda \) is estimated by mGDys (106) with \( W_{II,i} = I_n \) and \( \lambda_{opt} \) and \( \lambda_{opt} \) is obtained by optimizing the computed AvgMSE on the \( \eta \)-scale conditional on the optimal \( h \) obtained by computation.
Table 28: Estimated and Optimal AvgMSEs for LLGEE, MRGEEI, MRGEEII, and the Parametric Model for the $\eta$-scale Binomial Simulation Model. The bold values in columns 6–9 indicate the best AvgMSE among the optimal LLGEE, MRGEEI, MRGEEII, and the parametric model. The bold values in columns 10–13 indicate the best AvgMSEs among the fully estimated LLGEE, MRGEEI, MRGEEII, and parametric models. The AvgMSEs have been scaled by $10^5$. $n = 10$ for all cases.

<table>
<thead>
<tr>
<th>Row</th>
<th>Str†</th>
<th>Corr/Var</th>
<th>$s$</th>
<th>$\gamma$</th>
<th>Par‡</th>
<th>Optimal†</th>
<th>Estimated†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LL</td>
<td>MRI</td>
</tr>
<tr>
<td>1</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0</td>
<td>2.06</td>
<td>2.06</td>
<td>2.06</td>
</tr>
<tr>
<td>2</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.25</td>
<td>5.37</td>
<td>4.81</td>
<td>4.50</td>
</tr>
<tr>
<td>3</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>0.5</td>
<td>15.20</td>
<td>6.64</td>
<td>6.63</td>
</tr>
<tr>
<td>4</td>
<td>IND</td>
<td>None</td>
<td>3</td>
<td>1</td>
<td>53.43</td>
<td>8.70</td>
<td>8.70</td>
</tr>
<tr>
<td>5</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0</td>
<td>0.21</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>6</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.25</td>
<td>3.53</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>7</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>0.5</td>
<td>13.38</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>8</td>
<td>IND</td>
<td>None</td>
<td>30</td>
<td>1</td>
<td>51.59</td>
<td>1.05</td>
<td>1.05</td>
</tr>
<tr>
<td>9</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0</td>
<td>9.53</td>
<td>8.49</td>
<td>8.49</td>
</tr>
<tr>
<td>10</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.25</td>
<td>12.53</td>
<td>11.25</td>
<td>10.99</td>
</tr>
<tr>
<td>11</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>0.5</td>
<td>22.09</td>
<td>13.01</td>
<td>13.01</td>
</tr>
<tr>
<td>12</td>
<td>CS</td>
<td>Small</td>
<td>3</td>
<td>1</td>
<td>59.96</td>
<td>14.84</td>
<td>14.84</td>
</tr>
<tr>
<td>13</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0</td>
<td>0.96</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>14</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.25</td>
<td>4.37</td>
<td>1.53</td>
<td>1.53</td>
</tr>
<tr>
<td>15</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>0.5</td>
<td>14.32</td>
<td>1.69</td>
<td>1.69</td>
</tr>
<tr>
<td>16</td>
<td>CS</td>
<td>Small</td>
<td>30</td>
<td>1</td>
<td>52.66</td>
<td>1.74</td>
<td>1.74</td>
</tr>
<tr>
<td>17</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0</td>
<td>61.91</td>
<td>58.44</td>
<td>58.43</td>
</tr>
<tr>
<td>18</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.25</td>
<td>65.36</td>
<td>62.19</td>
<td>61.51</td>
</tr>
<tr>
<td>19</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>0.5</td>
<td>74.32</td>
<td>63.45</td>
<td>63.80</td>
</tr>
<tr>
<td>20</td>
<td>CS</td>
<td>Large</td>
<td>3</td>
<td>1</td>
<td>110.72</td>
<td>64.21</td>
<td>65.33</td>
</tr>
<tr>
<td>21</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0</td>
<td>9.34</td>
<td>8.79</td>
<td>8.79</td>
</tr>
<tr>
<td>22</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.25</td>
<td>13.51</td>
<td>9.90</td>
<td>9.92</td>
</tr>
<tr>
<td>23</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>0.5</td>
<td>24.30</td>
<td>10.27</td>
<td>10.32</td>
</tr>
<tr>
<td>24</td>
<td>CS</td>
<td>Large</td>
<td>30</td>
<td>1</td>
<td>63.94</td>
<td>10.58</td>
<td>10.65</td>
</tr>
</tbody>
</table>

* The Str column indicates which working structure the parametric model uses. The local model always uses the IND structure.
† The Par columns are the same and give the AvgMSE of the parametric model, obtained by simulation.
‡ The optimal LL column gives the AvgMSE of LLGEE at the optimal $h$, both obtained by computation. The optimal MRI and MRII columns give the computed AvgMSE of the MRGEEI and MRGEEII, respectively, at the optimal $h$ (by comp.) and optimal $\lambda$ (by comp.).
†† The estimated LL, MRI, MRII columns give the AvgMSE of the LLGEE, MRGEEI, and MRGEEII estimates, respectively. The LLGEE is based on CVnoV $h$ selection, MRGEEI and MRGEEII are based on mG Dys $\lambda$ selection, and all the AvgMSEs are presented on the $\mu$-scale.
7.7 Simulation Study of Specific Cases

The effect of misspecification on the choice of bandwidth and mixing parameter and the AvgMSE performance of the parametric, LLGEE, MRGEEI and MRGEEII methods are now studied over fine degrees of misspecification or $\gamma$, where $\gamma = 0, 0.05, 0.1, \ldots, 1$, for cases 9–24 of the $\mu$-scale Poisson model (Tables 21–22). The cross-over effect is of particular interest since this is where the MRGEE methods outperform both the user’s parametric and the LLGEE models. Also of particular interest are the large differences between the optimal mixing parameters for MRGEEI and MRGEEII in cases 17–20 of the $\mu$-scale Poisson (Table 21).

The following $\mu$-scale Poisson cases are studied over fine degrees of misspecification:

1. $s = 3, n_i = 10$, small corr/var
2. $s = 30, n_i = 10$, small corr/var
3. $s = 3, n_i = 10$, large corr/var
4. $s = 30, n_i = 10$, large corr/var
5. $s = 3, n_i = 6$, small corr/var
6. $s = 3, n_i = 20$, small corr/var

Cases 1–4 above correspond to cases 9–24 of the $\mu$-scale Poisson model (Tables 21–22), cases 5 and 6 are cluster size variations of case 1. The cluster size in case 5 is considered “small” while the cluster size in case 6 is considered “large”. The estimates, AvgMSEs, and optimal values are obtained as described in §7.6. The following comments apply.

1. **Bandwidth:**

In case 1, the average estimated bandwidth (CVnoV selector) is larger than the optimal bandwidth for the small sample size ($s = 3$, Table 29), but nearly the same for the large sample size ($s = 30$, Table 30). Both the estimated and the optimal bandwidths decrease steadily with increase in misspecification.
2. Optimal and Data-Driven Mixing Parameters

a. The estimated and optimal $\lambda$ increase steadily as misspecification increases for both the $s = 3$, small corr/var (Table 29 and Figure 17) and $s = 30$, small corr/var (Table 30) cases, implying that they both react properly to changes in the misspecification.

b. The average estimated $\lambda$ for MRGEEI is much closer to the optimal when the sample size is large (Table 30) rather than small (Table 29). Thus, increases in sample size improve the estimate of $\lambda$ for MRGEEI. MRGEEII $\lambda$ estimates are biased toward the local model, even more so for the larger sample size.

c. The optimal $\lambda$ for MRGEEI and MRGEEII differ in Figure 17 for the $s = 3$, small corr/var case and in Figure 21, where $s = 3$ and corr/var is large. The difference in the optimal $\lambda$ for the latter case is by far the most extreme out of all cases considered in the simulation studies of this chapter. Careful examination of the AvgMSE objective functions for $\lambda_{opt}$ on the $\mu$-scale for MRGEEI and the $\eta$-scale for MRGEEII using both simulation and computation show that these values are correct. Where $\gamma = 0$, the largest difference in $\lambda_{opt}$ for MRGEEI and MRGEEII occurs as the mixing parameters are optimized at nearly opposite ends of their range. This is not alarming for three reasons. First, the objective functions are relatively flat; the large change in $\lambda_{opt}$ does not imply a large change in AvgMSE. Second, the parametric and LLGEE have the opposite ordering of performance when studying AvgMSE on the $\mu$-scale than the AvgMSE on the $\eta$-scale. This seems unreasonable, yet in fact, the relationship between differences on the $\eta$-scale can be reversed on the $\mu$-scale for the log link function. Third, the MRGEEI optimal $\lambda$ are most effected by the mean model only being correct to a first order approximation of the true mean. Notice that when corr/var increases from small in Figure 17 to large in Figure 21, the optimal $\lambda$ increase. This reflects the fact that the user’s naïve mean model at $\gamma = 0$ is a poorer approximation of the true mean function as corr/var increases.

d. The average estimated $\lambda$ for MRGEEI and MRGEEII agree more “than they
should.” The problem is most drastic in the $s = 3$, large corr/var case (Figure 21). The MRGEEII average $\lambda$ estimate appears to be biased toward LLGEE-only fits. However, the data-driven MRGEEII fits (Figure 23) are better than the fixed optimal MRGEEII fits (Figure 22) in terms of $\mu$-scale AvgMSE, an indication that MRGEEII behaves more like MRGEEI than perhaps it should.

3. Optimal and Data-driven AvgMSE

a. The cross-over effect is best illustrated in Figure 18 for $s = 3$ and small corr/var. As the AvgMSE of the parametric model crosses above the AvgMSE of LLGEE, the optimal MRGEEI and MRGEEII have AvgMSEs below both. Theoretically, this illustrates that MRGEE does as well or better than the best of the parametric model and the nonparametric model no matter which outperforms the other. The practical performance of the data-driven methods is less pleasing. For example, in the $s = 3$, small corr/var case (Table 29 and Figure 19), the MRGEE performance is in-between that of the parametric and LLGEE at low levels of misspecification. A larger sample size is required for the MRGEE procedures, especially MRGEEI, to be competitive with the parametric model at small or no misspecification. In terms of optimal performance in the $s = 30$ case (Table 30), the cross-over effect extends over a wider range of misspecification. This shows that there is more potential for better performance of MRGEEI in the smaller sample size than in the larger sample size, but in practice the data-based MRGEEI requires much more misspecification at $s = 3$ than at $s = 30$ for improved performance over the parametric and LLGEE.

b. Notice in the $s = 3$, large corr/var case, that the local model always performs better over the parametric, both optimally (Figure 22) and data-based (Figure 23), so there is no cross-over. The better performance of the data-driven MRGEE methods over LLGEE in Figure 23 is probably due to Monte Carlo variation, which is fairly large in this case. The cross-over effect is somewhat apparent when the sample size is increased (Figure 24).
4. Data-Driven AvgMSE Performance of the Small and Large Cluster Sizes

a. The reduction in cluster size from \( n_i = 10 \) to \( n_i = 6 \) in the \( s = 3 \), small corr/var case (Figure 19 vs. Figure 25) forces the cross-over effect to occur at even higher levels of misspecification. This seems to be only a consequence of a smaller number of observations. Also, the LLGEE has less efficiency over the parametric for the \( n_i = 6 \) case.

b. Doubling the cluster size from \( n_i = 10 \) to \( n_i = 20 \) in the \( s = 3 \), small corr/var case (Figure 19 vs. Figure 26) forces the cross-over effect to occur at a lower level of misspecification. This seems to be a consequence of sample size, but notice that the cross-over occurs at misspecification levels that are mid-way between the \( s = 3, n_i = 10 \) case (Figure 19) and the \( s = 30, n_i = 10 \) case (Figure 20). Yet, while the \( s = 30, n_i = 10 \) case has a 10-fold increase in the number of responses over the \( s = 3, n_i = 10 \) case, the \( s = 3, n_i = 20 \) case only has twice as many responses. This suggests that a denser design grid may give LLGEE and MRGEE methods more sensitivity to misspecification than a comparable increase in the number of clusters. In addition, notice that the MRGEE methods are more efficient, with respect to the parametric model for \( n_i = 20 \) than for \( n_i = 10 \).

For most cases, the cross-over effect was demonstrated to depend on the sample size \((s)\) and the cluster size \((n_i)\). The MRGEE methods tend to outperform both the parametric and LLGEE at small degrees of misspecification for a large number of responses (large \(s\) or large \(n_i\)), but they tend to outperform the parametric and LLGEE at larger degrees of misspecification for a smaller number of responses (small \(s\) or small \(n_i\)). At the smallest levels of misspecification, the MRGEEI and especially the MRGEEII mGDys \(\lambda\) selectors do not perform nearly as well as the parametric model. The corr/var has a profound effect on the variability of the estimates for all methods. Also, the MRGEEI and MRGEEII optimal estimators can differ substantially with only modest differences in their \(\mu\)-scale or \(\eta\)-scale AvgMSEs.
Table 29: Misspecification Study of the $\mu$-Scale Poisson, $s = 3$, $n_i = 10$, Small Corr/Var, CS Parametric Model Working Structure Case. The bold values in columns 9–12 indicate the best AvgMSE among the parametric model and the optimal LLGEE, MRGEEI, MRGEEII models. The bold values in columns 13–16 indicate the best AvgMSE among the fully estimated parametric, LLGEE, MRGEEI, and MRGEEII models.

<table>
<thead>
<tr>
<th>Row</th>
<th>$\gamma$</th>
<th>$\lambda_{opt}$</th>
<th>$h_{opt}$</th>
<th>LL$^a$</th>
<th>MRI$^b$</th>
<th>MRII$^c$</th>
<th>Par$^d$</th>
<th>Optimal$^e$</th>
<th>Estimated$^f$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LL</td>
<td>MRI</td>
</tr>
<tr>
<td>1</td>
<td>0.00</td>
<td>0.17</td>
<td>0.145</td>
<td>0.06</td>
<td>0.45</td>
<td>6.33</td>
<td>7.35</td>
<td>6.30</td>
<td>6.31</td>
</tr>
<tr>
<td>2</td>
<td>0.05</td>
<td>0.18</td>
<td>0.149</td>
<td>0.07</td>
<td>0.45</td>
<td>6.33</td>
<td>7.36</td>
<td>6.31</td>
<td>6.31</td>
</tr>
<tr>
<td>3</td>
<td>0.10</td>
<td>0.19</td>
<td>0.148</td>
<td>0.08</td>
<td>0.46</td>
<td>6.41</td>
<td>7.40</td>
<td>6.38</td>
<td>6.38</td>
</tr>
<tr>
<td>4</td>
<td>0.15</td>
<td>0.22</td>
<td>0.147</td>
<td>0.10</td>
<td>0.46</td>
<td>6.55</td>
<td>7.47</td>
<td>6.51</td>
<td>6.51</td>
</tr>
<tr>
<td>5</td>
<td>0.20</td>
<td>0.25</td>
<td>0.146</td>
<td>0.14</td>
<td>0.47</td>
<td>6.74</td>
<td>7.57</td>
<td>6.68</td>
<td>6.68</td>
</tr>
<tr>
<td>6</td>
<td>0.25</td>
<td>0.30</td>
<td>0.144</td>
<td>0.18</td>
<td>0.48</td>
<td>6.97</td>
<td>7.67</td>
<td>6.87</td>
<td>6.88</td>
</tr>
<tr>
<td>7</td>
<td>0.30</td>
<td>0.36</td>
<td>0.142</td>
<td>0.23</td>
<td>0.50</td>
<td>7.26</td>
<td>7.80</td>
<td>7.11</td>
<td>7.12</td>
</tr>
<tr>
<td>8</td>
<td>0.35</td>
<td>0.43</td>
<td>0.140</td>
<td>0.30</td>
<td>0.51</td>
<td>7.60</td>
<td>7.95</td>
<td>7.37</td>
<td>7.38</td>
</tr>
<tr>
<td>9</td>
<td>0.40</td>
<td>0.50</td>
<td>0.143</td>
<td>0.37</td>
<td>0.53</td>
<td>8.00</td>
<td>8.11</td>
<td>7.64</td>
<td>7.65</td>
</tr>
<tr>
<td>10</td>
<td>0.45</td>
<td>0.58</td>
<td>0.134</td>
<td>0.45</td>
<td>0.54</td>
<td>8.48</td>
<td>8.30</td>
<td>7.93</td>
<td>7.95</td>
</tr>
<tr>
<td>11</td>
<td>0.50</td>
<td>0.66</td>
<td>0.130</td>
<td>0.52</td>
<td>0.56</td>
<td>8.98</td>
<td>8.46</td>
<td>8.18</td>
<td>8.20</td>
</tr>
<tr>
<td>12</td>
<td>0.55</td>
<td>0.72</td>
<td>0.127</td>
<td>0.59</td>
<td>0.59</td>
<td>9.54</td>
<td>8.64</td>
<td>8.43</td>
<td>8.46</td>
</tr>
<tr>
<td>13</td>
<td>0.60</td>
<td>0.78</td>
<td>0.123</td>
<td>0.65</td>
<td>0.61</td>
<td>10.14</td>
<td>8.80</td>
<td>8.65</td>
<td>8.68</td>
</tr>
<tr>
<td>14</td>
<td>0.65</td>
<td>0.83</td>
<td>0.120</td>
<td>0.70</td>
<td>0.63</td>
<td>10.79</td>
<td>8.96</td>
<td>8.85</td>
<td>8.89</td>
</tr>
<tr>
<td>15</td>
<td>0.70</td>
<td>0.87</td>
<td>0.116</td>
<td>0.75</td>
<td>0.66</td>
<td>11.50</td>
<td>9.11</td>
<td>9.03</td>
<td>9.07</td>
</tr>
<tr>
<td>16</td>
<td>0.75</td>
<td>0.90</td>
<td>0.112</td>
<td>0.79</td>
<td>0.69</td>
<td>12.27</td>
<td>9.27</td>
<td>9.22</td>
<td>9.26</td>
</tr>
<tr>
<td>17</td>
<td>0.80</td>
<td>0.92</td>
<td>0.108</td>
<td>0.82</td>
<td>0.71</td>
<td>13.10</td>
<td>9.43</td>
<td>9.40</td>
<td>9.44</td>
</tr>
<tr>
<td>18</td>
<td>0.85</td>
<td>0.94</td>
<td>0.104</td>
<td>0.85</td>
<td>0.73</td>
<td>13.97</td>
<td>9.57</td>
<td>9.55</td>
<td>9.60</td>
</tr>
<tr>
<td>19</td>
<td>0.90</td>
<td>0.96</td>
<td>0.099</td>
<td>0.88</td>
<td>0.75</td>
<td>14.87</td>
<td>9.69</td>
<td>9.67</td>
<td>9.71</td>
</tr>
<tr>
<td>20</td>
<td>0.95</td>
<td>0.98</td>
<td>0.096</td>
<td>0.90</td>
<td>0.77</td>
<td>15.86</td>
<td>9.83</td>
<td>9.83</td>
<td>9.87</td>
</tr>
<tr>
<td>21</td>
<td>1.00</td>
<td>0.99</td>
<td>0.093</td>
<td>0.91</td>
<td>0.79</td>
<td>16.88</td>
<td>9.96</td>
<td>9.96</td>
<td>9.99</td>
</tr>
</tbody>
</table>

$^a$ These LL columns give the estimated $h$, obtained using the CVnoV selector and $h_{opt}$, the AvgMSE minimal $h$, obtained by computation.

$^b$ These MRI columns give the estimated $\lambda$ using mGDys (105) with $W_{I,i} = I_{n_i}$ and the AvgMSE optimal $\lambda$ obtained by computation, both of which are for the MRGEEI model.

$^c$ These MRII columns give the estimated $\lambda$ using mGDys (106) with $W_{I,i} = I_{n_i}$ and the AvgMSE optimal $\lambda$ obtained by computation, both of which are for the MRGEEII model.

$^d$ The Par columns are the same and give the AvgMSE of the parametric model, obtained by simulation.

$^e$ The optimal LL column gives the AvgMSE of LLGEE at the optimal $h$. The optimal MRI and MRII columns give the AvgMSE of the MRGEEI and MRGEEII estimates, respectively, at the optimal $h$ and respective optimal $\lambda$. All of these AvgMSEs are found by simulation.

$^f$ The estimated Par, LL, MRI, MRII columns give the AvgMSE of the parametric, LLGEE, MRGEEI, and MRGEEII estimates, respectively. The LLGEE is based on CVnoV $h$ selection, MRGEEI and MRGEEII are based on mGDys $\lambda$ selection, and all the AvgMSEs are presented on the $\mu$-scale.
Table 30: Misspecification Study of the \( \mu \)-Scale Poisson, \( s = 30, n_i = 10 \), Small Corr/Var, CS Parametric Model Working Structure Case. The bold values in columns 9–12 indicate the best AvgMSE among the parametric model and the optimal LLGEE, MRGEEI, MRGEEII models. The bold values in columns 13–16 indicate the best AvgMSE among the fully estimated parametric, LLGEE, MRGEEI, and MRGEEII models.

<table>
<thead>
<tr>
<th>Row</th>
<th>( \gamma )</th>
<th>( h_{opt} )</th>
<th>LL ( \lambda )</th>
<th>MRI ( \lambda )</th>
<th>MRII ( \lambda )</th>
<th>Par ( \lambda )</th>
<th>Optimal ( \lambda )</th>
<th>Estimated ( \lambda )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00</td>
<td>0.079</td>
<td>0.03</td>
<td>0.01</td>
<td>0.51</td>
<td>0.64</td>
<td>0.91</td>
<td>0.64</td>
</tr>
<tr>
<td>2</td>
<td>0.05</td>
<td>0.079</td>
<td>0.08</td>
<td>0.05</td>
<td>0.53</td>
<td>0.66</td>
<td>0.91</td>
<td>0.66</td>
</tr>
<tr>
<td>3</td>
<td>0.10</td>
<td>0.077</td>
<td>0.22</td>
<td>0.18</td>
<td>0.59</td>
<td>0.74</td>
<td>0.92</td>
<td>0.72</td>
</tr>
<tr>
<td>4</td>
<td>0.15</td>
<td>0.075</td>
<td>0.40</td>
<td>0.35</td>
<td>0.67</td>
<td>0.87</td>
<td>0.94</td>
<td>0.81</td>
</tr>
<tr>
<td>5</td>
<td>0.20</td>
<td>0.072</td>
<td>0.58</td>
<td>0.52</td>
<td>0.75</td>
<td>1.06</td>
<td>0.97</td>
<td>0.89</td>
</tr>
<tr>
<td>6</td>
<td>0.25</td>
<td>0.069</td>
<td>0.73</td>
<td>0.67</td>
<td>0.83</td>
<td>1.29</td>
<td>0.99</td>
<td>0.95</td>
</tr>
<tr>
<td>7</td>
<td>0.30</td>
<td>0.066</td>
<td>0.84</td>
<td>0.78</td>
<td>0.89</td>
<td>1.59</td>
<td>1.02</td>
<td>1.00</td>
</tr>
<tr>
<td>8</td>
<td>0.35</td>
<td>0.063</td>
<td>0.91</td>
<td>0.86</td>
<td>0.94</td>
<td>1.93</td>
<td>1.05</td>
<td>1.04</td>
</tr>
<tr>
<td>9</td>
<td>0.40</td>
<td>0.061</td>
<td>0.95</td>
<td>0.91</td>
<td>0.97</td>
<td>2.33</td>
<td>1.08</td>
<td>1.07</td>
</tr>
<tr>
<td>10</td>
<td>0.45</td>
<td>0.058</td>
<td>0.98</td>
<td>0.95</td>
<td>0.98</td>
<td>2.77</td>
<td>1.10</td>
<td>1.10</td>
</tr>
<tr>
<td>11</td>
<td>0.50</td>
<td>0.056</td>
<td>1.00</td>
<td>0.97</td>
<td>0.99</td>
<td>3.28</td>
<td>1.12</td>
<td>1.12</td>
</tr>
<tr>
<td>12</td>
<td>0.55</td>
<td>0.054</td>
<td>1.00</td>
<td>0.98</td>
<td>1.00</td>
<td>3.83</td>
<td>1.14</td>
<td>1.14</td>
</tr>
<tr>
<td>13</td>
<td>0.60</td>
<td>0.053</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
<td>4.44</td>
<td>1.15</td>
<td>1.15</td>
</tr>
<tr>
<td>14</td>
<td>0.65</td>
<td>0.051</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>5.10</td>
<td>1.17</td>
<td>1.17</td>
</tr>
<tr>
<td>15</td>
<td>0.70</td>
<td>0.050</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>5.81</td>
<td>1.18</td>
<td>1.18</td>
</tr>
<tr>
<td>16</td>
<td>0.75</td>
<td>0.049</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>6.57</td>
<td>1.19</td>
<td>1.19</td>
</tr>
<tr>
<td>17</td>
<td>0.80</td>
<td>0.048</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>7.39</td>
<td>1.20</td>
<td>1.20</td>
</tr>
<tr>
<td>18</td>
<td>0.85</td>
<td>0.047</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>8.27</td>
<td>1.21</td>
<td>1.21</td>
</tr>
<tr>
<td>19</td>
<td>0.90</td>
<td>0.046</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>9.19</td>
<td>1.22</td>
<td>1.22</td>
</tr>
<tr>
<td>20</td>
<td>0.95</td>
<td>0.045</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>10.17</td>
<td>1.23</td>
<td>1.23</td>
</tr>
<tr>
<td>21</td>
<td>1.00</td>
<td>0.045</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>11.20</td>
<td>1.23</td>
<td>1.23</td>
</tr>
</tbody>
</table>

a  These LL columns give the estimated \( h \), obtained using the CVnoV selector and \( h_{opt} \), the AvgMSE minimal \( h \), obtained by computation.

b  These MRI columns give the estimated \( \lambda \) using mGDys (105) with \( W_{I,i} = I_{n_i} \) and the AvgMSE optimal \( \lambda \) obtained by computation, both of which are for the MRGEEI model.

c  These MRII columns give the estimated \( \lambda \) using mGDys (106) with \( W_{II,i} = I_{n_i} \) and the AvgMSE optimal \( \lambda \) obtained by computation, both of which are for the MRGEEII model.

d  The Par columns are the same and give the AvgMSE of the parametric model, obtained by simulation.

e  The optimal LL column gives the AvgMSE of LLGEE at the optimal \( h \). The optimal MRI and MRII columns give the AvgMSE of the MRGEEI and MRGEEII estimates, respectively, at the optimal \( h \) and respective optimal \( \lambda \). All of these AvgMSEs are found by simulation.

f  The estimated Par, LL, MRI, MRII columns give the AvgMSE of the parametric, LLGEE, MRGEEI, and MRGEEII estimates, respectively. The LLGEE is based on CVnoV \( h \) selection, MRGEEI and MRGEEII are based on mGDys \( \lambda \) selection, and all the AvgMSEs are presented on the \( \mu \)-scale.
Figure 17: Computed Optimal and Estimated $\lambda$ as a function of Misspecification for the $\mu$-Scale Poisson, $s = 3$, $n_i = 10$, Small Corr/Var, CS Parametric Model Working Structure Case.

Figure 18: Optimal AvgMSE Obtained by Simulation for the $\mu$-Scale Poisson, $s = 3$, $n_i = 10$, Small Corr/Var, CS Parametric Model Working Structure Case.
Figure 19: AvgMSE of the Fits for the $\mu$-Scale Poisson, $s = 3$, $n_i = 10$, Small Corr/Var, CS Parametric Model Working Structure Case.

Figure 20: AvgMSE of the Fits for the $\mu$-Scale Poisson, $s = 30$, $n_i = 10$, Small Corr/Var, CS Parametric Model Working Structure Case.
Figure 21: Computed Optimal and Estimated $\lambda$ as a function of Misspecification for the $\mu$-Scale Poisson, $s = 3$, $n_i = 10$, Large Corr/Var, CS Parametric Model Working Structure Case.

![Graph showing computed optimal and estimated $\lambda$ as a function of misspecification](image)

Figure 22: Optimal AvgMSE Obtained by Computation for the $\mu$-Scale Poisson, $s = 3$, $n_i = 10$, Large Corr/Var, CS Parametric Model Working Structure Case.

![Graph showing optimal avgMSE as a function of misspecification](image)
Figure 23: AvgMSE of the Fits for the $\mu$-Scale Poisson, $s = 3$, $n_i = 10$, Large Corr/Var, CS Parametric Model Working Structure Case.

Figure 24: AvgMSE of the Fits for the $\mu$-Scale Poisson, $s = 30$, $n_i = 10$, Large Corr/Var, CS Parametric Model Working Structure Case.
Figure 25: AvgMSE of the Fits for the $\mu$-Scale Poisson, $s = 3$, $n_i = 6$, Small Corr/Var, CS Parametric Model Working Structure Case.

Figure 26: AvgMSE of the Fits for the $\mu$-Scale Poisson, $s = 3$, $n_i = 20$, Small Corr/Var, CS Parametric Model Working Structure Case.
8 Consistency and Convergence Rates

Consistency is an important property that should be established for any new estimation method. An estimator is consistent if it becomes close to the parameter it estimates with high probability as the sample size tends to infinity. We now discuss the conditions under which the parametric, local GEE, and the optimal MRGEEI and MRGEEII estimators are consistent.

A fit is consistent for the true mean at \( x^* \) if

\[
\hat{y}(x^*) \xrightarrow{p} E[Y|x^*],
\]

where \( \xrightarrow{p} \) means convergence in probability. This property is known as weak consistency.

Consistency of the Parametric Model

Provided that the MOM estimators for \( \phi^2 \) and \( \rho \) can be shown to be \( \sqrt{s} \)-consistent and that the mean model is correct, consistency of the parametric model fits follows from Theorem 2 of Liang and Zeger (1986), which gives the limiting distribution of \( \hat{\beta}_{GEE} \). \( \sqrt{s} \)-consistency is defined as convergence in probability to a parameter at the rate \( s^{-1/2} \), e.g., if \( \hat{\phi}^2 - \phi^2 = O_p(s^{-1/2}) \) and \( \hat{\rho} - \rho = O_p(s^{-1/2}) \) then \( \hat{\phi}^2 \) and \( \hat{\rho} \) are \( \sqrt{s} \)-consistent. In Appendix D.1 it is shown that this holds for \( \hat{\phi}^2 \), \( \hat{\rho}_{CS} \), and \( \hat{\rho}_{AR1} \), the MOM estimators given in §2.2.4 and used throughout this work. Thus, the parametric GEE parameter estimates are consistent, if the mean model is correct. Theorem 2 of Liang and Zeger (1986) and a corollary of the continuous mapping theorem (CMT) (Serfling, 1980, pg. 26) imply

\[
\|\hat{\beta}_{GEE} - \beta_{GEE}\| = O_p(s^{-1/2}),
\]

i.e., the parameter estimates are \( \sqrt{s} \)-consistent and converge at the rate \( s^{1/2} \). Theorem 14.6-2 (multivariate delta method) of Bishop, Fienberg, and Holland (1975) implies that

\[
\hat{y}^P - \eta = x' \hat{\beta} - F^{-1}(E[Y|x^*]) = O_p(s^{-1/2}) \quad \text{and} \quad \hat{y}^P - m = F(x' \hat{\beta}) - E[Y|x^*] = O_p(s^{-1/2})
\]

at all locations \( x^* \). Thus, the parametric model fits are \( \sqrt{s} \)-consistent for the true \( \eta \) on the \( \eta \)-scale and for the true mean on the \( \mu \)-scale.

Some caution must be exercised when interpreting MOM correlation estimates when the working structure disagrees with the true correlation structure. Crowder (1995) found a
case in which the correlation AR(1) MOM estimator does not converge to a value that is within the bounds for correlation. Gilliland and Schabenberger (2001) have shown in the equi-correlated binary data case that correlations can be severely bounded from above under certain conditions, illustrating that there are cases where the range of correlations that the MOM estimators can converge to may be very limited. Thus, while consistency may hold, it does not necessarily imply that the parameter the MOM estimator converges to has any meaning with respect to the true correlation structure. However, for consistency of the fits, this issue is not a concern as interpretability of the MOM estimates is not a condition of Liang and Zeger’s (1986) theorem.

### Consistency of the Local Model

Consistency of the parametric model requires correct specification of the mean model. If the model is not correct, then the fits may not converge to the true mean. The local model has the same difficulty for fixed $h$, if one views it in the quasi-likelihood framework (see §3.5.1). But, if $h$ is allowed to shrink with increasing sample size, then the local model can either: 1) interpolate the sample means at the unique $x^*$, if the design points remain fixed as $s \to \infty$, or 2) have increasing information about the form of the mean model by filling-in the $x^*$ covariate space with clusters, if additional assumptions are applied.

The sample means are clearly consistent for the true means in (1), so consider (2). The design needs to be modified so that filling-in is possible. If the clusters are observation-specific, then the cluster size may be increased. But this is troublesome because GEE’s apply to finite cluster sizes, and small ones at that. Infinite cluster sizes should be studied with time series methods, so it is preferred to keep the cluster size fixed. Filling-in may be achieved by simply allowing the covariates to be random. This is the approach of Lin and Carroll (2000).

When $s \to \infty$, $h \to 0$ such that $sh \to \infty$, and $X_{ij}$ are random, the terms in the asymptotic bias and variance for the local polynomial GEE are proportional to $h$ and $(sh)^{-1}$, respectively (Lin and Carroll, 2000)$^9$, which limit to 0. Consequently, the local polynomial GEE $\eta$-scale estimate, $\eta^{NP}(x)$ is consistent for $F^{-1}(m(x))$ and by the CMT, the fit, $\hat{y}^{NP}$ is

---

$^9$The asymptotic bias and variance for local linear GEE were given by (78) and (79), respectively.
consistent for \( m(x) \), whether the local model is correct or not.

**Convergence Rates and Consistency of the Optimal MRGEEI and MRGEEII**

For the simple case where the distance optimal \( \lambda \) selectors, (103) and (104), are used, the convergence rate of MRGEEII can be established if the following conditions hold:

1. \[ \| X\hat{\beta} - X\beta^* \| = O_p(\pi) \] for some non-stochastic sequence \( \pi \), where \( \beta^* \) is the value of \( \beta \) that gives the minimum sum of squares difference between \( F^{-1}(m(x_{ij}^*)) = \eta(x_{ij}^*) \) and \( x'\beta \), assuming it exists. This is similar to a previous definition of \( \beta^* \) given in §6.1.1.

2. \[ \| \{ x'_{bij} \hat{\theta}_{ij} \}_{N \times 1} - F^{-1}(m(x^*)) \| = O_p(\gamma) \]

and \( x^* = \{ x_{ij}^* \}_{N \times 1} \), a vector of stacked cluster coordinate vectors.

3. \[ \lim_{s \to \infty} \pi_s \gamma_s^{-1} = 0. \]

Assumption 1 gives the rate of convergence of the parametric fits to the true \( \eta \). By Theorem 2 of Liang and Zeger (1986) and the multivariate delta method, \( \pi_s = s^{-1/2} \) under no misspecification. When there is misspecification, assumption 1 requires the parametric estimate to be bounded in probability at some rate \( \pi_s \). Assumption 3 requires this rate to be faster than the rate for the local model fits, which is given by \( \gamma_s \) in assumption 2. We do not yet know what the nonparametric convergence rate is, but assuming that it is slower than the parametric rate is reasonable, as this has repeatedly been shown in simpler settings.

**Theorem 2**

Under assumptions 1–3 above,

\[ \| (1 - \lambda_{optMRGEEII}) X\hat{\beta} + \lambda_{optMRGEEII} \{ x'_{bij} \hat{\theta}_{ij} \}_{N \times 1} \| = \begin{cases} O_p(\gamma) & \text{if } F(x'\beta) \text{ is misspecified} \\ O_p(s^{-1/2}) & \text{if } F(x'\beta) \text{ is correct} \end{cases}, \]

a result that holds whether the data are correlated or not. The proof closely follows the proof of Theorem 3.1 in Burman and Chaudhuri (1992), except one has to replace their \( \lambda^* \) with \( 1 - \lambda_{optMRGEEII} \) and \( 1 - \lambda^* \) with \( \lambda_{optMRGEEII} \), and use assumption 1 above in place of their stronger assumption A1. The consistency of the optimal MRGEEII estimator follows immediately from this result. A similar result can be obtained for MRGEEI under assumptions 1–3 with \( X\hat{\beta} \) and \( \{ x'_{bij} \hat{\theta}_{ij} \}_{N \times 1} \) replaced by \( \hat{y}^P \) and \( \hat{y}^{NP} \), respectively, and
$F^{-1}(m(x^*_ij))$ replaced by $m(x^*_ij)$. Thus,

$$||(1 - \lambda_{optMRGEE})\tilde{y}^P + \lambda_{optMRGEE}\tilde{y}^NP|| = \begin{cases} O_p(\gamma_s) & \text{if } F(x'\beta) \text{ is misspecified} \\ O_p(s^{-1/2}) & \text{if } F(x'\beta) \text{ is correct} \end{cases}. $$

The rate $\pi_s = s^{-1/2}$ for the $\mu$-scale model under no misspecification is guaranteed by application of the multivariate delta method to Theorem 2 of Liang and Zeger (1986).

Note that the MRGEE $\lambda$ estimators in Theorem 2 are distance optimal not AvgMSE optimal. (The optimal $\lambda$ presented in the MRGEE simulation study (§7.6) are all AvgMSE optimal.) Interestingly, the distance optimal $\lambda$ gives a smaller AvgMSE than the AvgMSE optimal selector because it is allowed to choose $\lambda$ differently for each dataset. A similar phenomenon is discussed in §7.5 (comment 6) for the bandwidth.

The consistency and convergence rates of the data-driven $\lambda$ selectors are even more interesting, but they are not currently available. The proof of Theorem 2 applies to both dependent and independent data, but a proof involving the data-driven $\lambda$ selectors seems to require inequalities for correlated random variables not yet known to us.
9 Application, Summary, and Future Research

9.1 An Application

We finalize the work presented with an application of some of the methods developed here to a “real world” problem. The simulation studies have shown that the methods hold both theoretical potential and have good practical performance, but the simulations restrict features of the estimation setting to simple cases. For example, most cases assumed that the clusters are of size $n_i = 10$, for all $i$, and that the design points are uniformly spaced and the same for all clusters. However, many longitudinal studies often involve unequal cluster sizes, outliers, and missing data. Application of our methods under less than “ideal” circumstances helps provide another measure of their performance, even if only for a single situation.

We turn our attention to an application presented in §2.2.2. Recall that the CD4+ immune cell counts of 369 men have been measured repeatedly over time, along with other covariates. A small portion of the data are shown in Table 4 and the counts for 80 randomly chosen subjects are shown in Figure 27. Three subjects in the figure are highlighted. Notice that the clusters are of different sizes.

It is of interest to model the mean count over time. Past experience has indicated that CD4+ cell counts may decline rather sharply around the point of seroconversion (Diggle, Liang, and Zeger, 1994). It is of particular interest if a drop in the average CD4+ cell count can be detected by our models and which of the parametric or LLGEE models is recommended by MRGEEI and MRGEEII. Since each subject was repeatedly measured, the observations within a subject are likely to be correlated. We use the CS structure to model these correlations, realizing that this structure is too simple to capture both correlation from an autoregressive process over time, a latent variable process of the subject effect, and the influence of the changing variance on the correlations. However, as a working correlation structure, it is not assumed to be necessarily correct, an advantage of using GEE. The counts are essentially unbounded, which suggests the use of a Poisson-type model. There is much more variation than a Poisson distribution would allow and outliers abound, but because the GEE model only requires specification of the first two moments, the identity variance

---

10 The data reduction was done primarily for speed in estimation and secondarily for clarity in Figure 27.
function can be modified to allow for overdispersion. The counts are also large enough for a normal-based analysis to be reasonable.

Parametric, LLGEE, MRGEEI and MRGEEII models were applied to the CD4 count data shown in Figure 27. The counts do decrease (roughly) after seroconversion, but how they decrease is difficult to see given the variation present. Past information about the mean CD4+ count indicates that the parametric models need to decrease around the seroconversion point. Let $x_{ij}$ be the time in years until seroconversion. Three parametric models of the form

\[
F(x_{ij}' \beta) = e^{\eta(x_{ij})}
\]

\[
\text{Var}[Y_{ij}] = \phi^2 e^{\eta(x_{ij})}
\]

\[
\text{Corr}[Y_{ij}, Y_{ik}] = \rho, \ j \neq k,
\]
with log link, identity variance function, and CS working correlation structure, are tried. In particular, a linear model, \( \eta = \beta_0 + \beta_1 x_{ij} \), a quadratic model, \( \eta = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 \), and a cubic model, \( \eta = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \beta_3 x_{ij}^3 \), each with increasing flexibility on the \( \eta \)-scale, are fit to the CD4+ data. The LLGEE model also uses the identity variance function, but uses an IND working correlation structure. The estimate of the bandwidth is obtained by the CVnoV selection method (see \S7.5).

The parametric model estimates are given in Table 31.

Table 31: Parameter Estimates for the Parametric Linear, Quadratic, and Cubic GEE Models Applied to the CD4 Data.

<table>
<thead>
<tr>
<th>Model</th>
<th>Linear ( \eta )</th>
<th>Quadratic ( \eta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F(\mathbf{x}'<em>{ij} \hat{\beta}) = e^{6.69-0.12 x</em>{ij}} )</td>
<td>( F(\mathbf{x}'<em>{ij} \hat{\beta}) = e^{6.69-0.12 x</em>{ij}-0.003 x_{ij}^2} )</td>
<td></td>
</tr>
<tr>
<td>( \overline{\text{Var}}[Y_{ij}] = 180.05 e^{6.69-0.12 x_{ij}} )</td>
<td>( \overline{\text{Var}}[Y_{ij}] = 181.25 e^{6.69-0.12 x_{ij}-0.003 x_{ij}^2} )</td>
<td></td>
</tr>
<tr>
<td>( \overline{\text{Corr}}[Y_{ij}, Y_{ik}] = 0.4606 )</td>
<td>( \overline{\text{Corr}}[Y_{ij}, Y_{ik}] = 0.4620 )</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cubic ( \eta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F(\mathbf{x}'<em>{ij} \hat{\beta}) = e^{6.73-0.17 x</em>{ij}-0.02 x_{ij}^2+0.009 x_{ij}^3} )</td>
</tr>
<tr>
<td>( \overline{\text{Var}}[Y_{ij}] = 175.01 e^{6.73-0.17 x_{ij}-0.02 x_{ij}^2+0.009 x_{ij}^3} )</td>
</tr>
<tr>
<td>( \overline{\text{Corr}}[Y_{ij}, Y_{ik}] = 0.4759 )</td>
</tr>
</tbody>
</table>

The correlation parameter estimate changes little between the models. The linear and quadratic models are nearly the same. Asymptotic Wald chi-square tests on the parameters show that all terms are significant \((P < 0.03)\) in the both the cubic and the linear models. The quadratic term in the quadratic model was not significant \((P = 0.67)\). Notice that \( \hat{\phi}^2 \) is the smallest for the cubic model. Thus, a standard GEE analysis suggests that the cubic model is most appropriate. The three parametric model fits are contrasted with the LLGEE fit in Figure 28. The CVnoV bandwidth choice for LLGEE is 0.6, which is rather small, but the resulting model fits the data seemingly well.

MRGEE fits to the data are shown in Figure 29. The MRGEEI \( \lambda \) estimates (Table 32) are near 1 for the linear and quadratic models, indicating that the local model is favored over the parametric models. However, the cubic model is more competitive with the local model so a smaller \( \lambda \) is chosen. The MRGEEII \( \lambda \) estimates are smaller than the MRGEEI estimates and indicate that the parametric models are favored on the \( \eta \)-scale.
Table 32: MRGEEI and MRGEEII \( \lambda \) Estimates Corresponding to Various Parametric Estimates for the CD4 Data. Fits are summarized by model degrees of freedom, Pearson \( \chi^2/df \) (38) and deleted Pearson \( \chi^2/df \) (145) statistics. All models use the log link and the identity variance function. The parametric models use the CS structure, and the local model uses the IND structure. The LLGEE bandwidth is \( h = 0.6 \), as chosen by the CVnoV selector.

<table>
<thead>
<tr>
<th>Par</th>
<th>( \eta )</th>
<th>Model</th>
<th>( \lambda )</th>
<th>( \lambda )</th>
<th>( df_{model} )</th>
<th>Par</th>
<th>MRI</th>
<th>MII</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>0.88</td>
<td>0.33</td>
<td>2</td>
<td>6.4</td>
<td>3.7</td>
<td>7.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quadratic</td>
<td>0.96</td>
<td>0.44</td>
<td>3</td>
<td>6.9</td>
<td>4.8</td>
<td>7.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cubic</td>
<td>0.73</td>
<td>0.16</td>
<td>4</td>
<td>6.2</td>
<td>4.5</td>
<td>7.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The local model \( df \) displayed in Table 32 are obtained using \( tr(H^{LGEE}) \) as described in §5.2.3. The MRGEEI and MRGEEII model \( df \) are calculated as convex combinations of the parametric and local model \( df \), using the respective estimated mixing parameters. The \( df_{model} \) are interpreted loosely as one more than the degree of a polynomial linear predictor that achieves the same flexibility\(^\text{11}\) as the evaluated model. The local model has the largest \( df_{model} \) because it is the most flexible. The MRGEEI models have nearly the same \( df_{model} \) as LLGEE because their \( \lambda \) estimates are near 1. On the other hand, the MRGEEII models have \( df_{model} \) closer to the parametric models because their \( \lambda \) estimates are closer to 0.

The models are also summarized by Pearson’s \( \chi^2/df \) goodness-of-fit and a cluster-deleted version of this statistic. Since there is no likelihood associated with the GEE model, other measures such as the deviance and AIC are not applicable. The deleted Pearson \( \chi^2/df \) is given by

\[
\frac{1}{df_E} \sum_{i=1}^{S} \sum_{j=1}^{n_i} \frac{(y_{ij} - \hat{y}_{ij,-i})^2}{a(\hat{y}_{ij,-i})},
\]

where \( a(\hat{y}_{ij,-i}) \) is the variance function evaluated at the deleted fits and \( df_E \) is the error degrees of freedom, \( \sum n_i - df_{model} = N - df_{model} \). This statistic is related to PRESS* (97), where \( W_i = V^{-1} \) and the denominator, \( N^* - tr(H^{LGEE}) \) has been replaced with \( df_E \) (see

\(^{11}\)In this case, this is precisely the interpretation of the parametric model \( df \)
§5.2.3). The deleted Pearson $\chi^2/df$ statistic measures the prediction capability of the model by summarizing the leave-one-out fits. Table 32 shows that LLGEE has the smallest Pearson $\chi^2/df$ goodness-of-fit and that the cubic model is the best of the parametric models. The MRGEEII models do not fair as well as MRGEEI, which is partly a consequence of presenting the results on the $\mu$-scale rather than the $\eta$-scale. Similar conclusions are reached for the deleted Pearson $\chi^2/df$ comparison of the methods, but notice that the quadratic model performs worse in terms of prediction, relative to the other models, than it does in terms of goodness-of-fit. Also the performance of the MRGEE fits is nearly unaffected by the form of the parametric model, illustrating the methods’ adaptive feature.

The MRGEEI and MRGEEII fits are shown with the cubic parametric and LLGEE models in Figure 29. Prior experience with CD4+ count data suggests that the counts maintain an average of roughly 1,100 before seroconversion, then abate as time passes (Diggle, Liang, and Zeger, 1994). The MRGEEI and LLGEE reflect this suggestion best with the MRGEEI fit having less variability.
Figure 28: CD4+ Data with Linear, Quadratic, and Cubic Linear Predictor Parametric Model Estimates and a LLGEE Model Estimate. The bandwidth is $h = 0.6$, as chosen by the CVnoV selector. All models use the log link and the identity variance function. The parametric models use the CS structure, and the local model uses the IND structure.
Figure 29: The Cubic Linear Predictor Parametric Model, LLGEE, and the MRGEE Models Fitted to the CD4 Data. The bandwidth is $h = 0.6$, as chosen by the CVnoV selector. The MRGEEI $\lambda$ estimate is 0.73 and the MRGEEII $\lambda$ estimate is 0.16. All models use the log link and the identity variance function. The parametric models use the CS structure, and the local model uses the IND structure.
9.2 Summary and Concluding Remarks

9.2.1 Summary

Maximum likelihood for generalized linear models (GLIM) and the technique of generalized estimating equations (GEE), the methods of interest in this work, relax the classical least squares regression assumptions considerably to allow the mean structure to be a nonlinear transformation of a linear combination of the parameters, allow for non-normal distributions, and in the case of GEE, allow for correlation within groups (or clusters) of responses, a data structure called clustered correlated. Generalized estimating equations are used to estimate mean models for cluster correlated data when only the mean model and the variance function (of the mean) have been specified. These weak assumptions are motivated by the difficulty in formulating multivariate extensions of distributions in the exponential family, especially distributions for categorical data. GEEs are often appropriate for modelling the population mean of counts, binary, and continuous responses when the correlation structure is a nuisance rather than the focus. The variance function of the mean can be taken from a distribution in the exponential family, though it need not be. GEEs are hailed for providing consistent parameter estimates, even when the correlation structure is misspecified, if the correlation parameter estimators are consistent\textsuperscript{12}.

Like many statistical models, the GEE mean model and variance function may be misspecified. Robustification against variance function misspecification, and especially correlation structure misspecification, is done through a semiparametric “sandwich” variance estimation of the standard errors. Robustification of the mean function is possible through a kernel-based localization of GEEs, which makes weak assumptions about its functional form. The local GEE model does not require non-identity working correlation structures to improve asymptotically efficiency (Lin and Carroll, 2000). Only in small samples may the bias and variance be improved in some cases by accounting for the correlation structure.

Much emphasis of local estimation is placed on the bandwidth ($h$), which controls the degree of smoothness of the resulting fits to the data. Lin and Carroll (2000) propose a cross-

\textsuperscript{12}One has to be careful that the parameter values for which the correlation estimators are consistent exist given the true structure of the correlation matrix.
validation (CV) method in which Pearson residuals are scaled by the model-based variance, a function that includes a scale parameter. In practice, we found the scale parameter estimator to depend on the bandwidth in a way that distorts the CV objective function and leads to large bandwidth estimates. We explored six bandwidth selectors: CV, PRESS*, and PRESS**, each with and without the estimate of the variance and each with a fixed scale parameter. PRESS* and PRESS** are penalized CV selectors that are extensions of bandwidth selectors from Mays et al. (2001) to clustered correlated data. The simplest bandwidth selector, CV without $V$ (CVnoV) was found to perform best. However, PRESS** with $V$ was very competitive in small samples, performing better than CVnoV in several cases. This is consistent with previous research in MRR (Mays et al. 2001). It seems counter-intuitive that bandwidth can be chosen better by not incorporating the variance of the observations, since doing so is optimal in the least squares context. However, the comparison of the bandwidth selectors was based on nonstandard assumptions; the variance functions and the correlation structures were incorrect.

Our main interest concerned a situation that lies in-between the two extremes of knowing the mean model entirely and not knowing it at all. We imagine that a model builder specifies a parametric model that he or she believes may only be partly correct. Then hybridization of global (parametric) and local (nonparametric) models is used to formulate a semiparametric mean model that draws on the strengths of global and local regression for estimation. These ideas are based on previous work on model robust regression (MRR) by Mays et al. (2001) for the linear model, and model robust regression for quantal regression by Nottingham and Birch (2000). The aforementioned contributions have shown that MRR methods can attain lower bias and variance than both parametric and nonparametric models under slight misspecification of the parametric mean model, especially in small samples. They also showed that MRR is consistent regardless of the parametric specification and can achieve a fast $n^{-1/2}$ rate of convergence when the model is not misspecified. We extended their work by robustifying the GEE mean model specification using local GEEs.

Two MRR-type estimators were proposed, one in which mixing occurs on the mean or $\mu$-scale, called MRGEEI, and one in which mixing occurs on the linear predictor or $\eta$-scale, called MRGEEII. The consequences of estimation on these two scales were considered.
One set of simulation models were formulated so that contamination of the mean function occurred on the $\mu$-scale. A second set induced contamination on the $\eta$ scale. The mean scale mixing model, MRGEEI, was found to perform best under both scales of contamination, but for a few cases it was shown that this phenomenon is a consequence of summarizing the fits in terms of $\mu$-scale AvgMSE rather than $\eta$-scale AvgMSE.

The mixing parameter ($\lambda$) estimators used in the hybridization of the parametric and local models were an extension of the optimal selector from Mays et al. (2001), which minimizes the distance between the fits and the data. An AvgMSE-based estimator for $\lambda$ was also developed and found to be competitive with the minimum distance approach. It was dropped from further consideration because of its computational complexity. Simultaneous estimation of bandwidth and mixing parameter were also considered. Simultaneous estimation does not pose an identifiability problem of the parameters $h$ and $\lambda$ and a case study showed that simultaneous estimation may lead to a slight improvement in AvgMSE over conditional estimation.

The optimal bandwidths and optimal mixing parameter were obtained by computation or simulation for the simulation models. Computational expressions for the MSE at the observed data locations were derived using Taylor series expansions. Both the parametric and local GEE were shown to be special cases of a quasi-likelihood framework, a paradigm that simplified the development of MSE computations.

The parametric and local models were shown to be consistent, a desirable feature for estimation. Consistency was shown and convergence rates were found for the optimal MRGEE estimators. The optimal MRGEE estimators converge to the true model at the parametric rate if the parametric mean model is correct and converge at the local rate when the parametric mean model is misspecified, a result that is valid whether the data are correlated or not.

The optimal MRGEE methods, based on the optimal local linear GEE (LLGEE) were found to outperform both the parametric model and the optimal LLGEE under slight degrees of misspecification. This occurs where the AvgMSE of the parametric model crossed above the AvgMSE of the local model, an effect we termed “cross-over.” The data-driven MRGEE methods did not perform as well compared to the parametric, but did perform as well as
LLGEE in many cases and better than both in several cases. The cross-over effect was apparent in several data-driven cases. Thus, we have shown that the MRGEE methods can be beneficial over both the parametric and local models, both theoretically and in practice.

We recommend that at high degrees of misspecification, the user should employ LLGEE. If there is no misspecification, the user should employ the parametric model. Otherwise, there is a demonstrated advantage in applying the MRGEE methods in cases where the possibility of misspecification exists. If the mean model is formulated on the mean-scale and one wishes to interpret results on the scale of the responses, then MRGEEI is preferred and will tend to perform better than MRGEEII in terms of the mean squared error of the fits. However, if one prefers to transform the response and study effects on the linear predictor or $\eta$-scale only, MRGEEII is preferred and may perform better than MRGEEI in terms of the mean squared error of the linear predictor.

9.2.2 Concluding Remarks

Correlated data are becoming increasingly common as are the use of models that account for dependency in responses. Datasets are also increasing in size as collection methods become more automated and databases are developed to record longitudinal data. Consequently, the number of applications of GEE methods and local GEE will continue to increase. Also, combinations of parametric and local models are likely to become more popular as users of statistics find more need for adaptive models. The MRR method has repeatedly been shown to be successful and we have shown that it also holds promise for clustered correlated data.

9.3 Immediate Extensions of this work

The methods that have been given can be extended to more general model setups than we have considered. First, the local model can be applied to multiple covariates and the same estimation methods may be used if a multidimensional kernel function, such as the spherically symmetric (§3.2.2) or product kernel, is employed, as long as the kernel function depends on a single bandwidth\textsuperscript{13}. The MSE computational expressions continue to hold

\textsuperscript{13}If the covariates are on different scales, they may be standardized so that a single bandwidth is appropriate. Multiple bandwidth parameters may be used if the bandwidth selection methods are extended.
as given. The local model may be constructed as response surface. For example, in the local linear case, it would be a hyperplane. Other local models besides polynomials are possible through localization of the parametric model by wrapping kernel weights about the model variance-covariance matrix. Kernel weights may be selectively applied to some of the covariates, giving the capability of partial localization. Second, the parametric and local models on the \( \mu \)-scale may easily be generalized to nonlinear mean functions. In most places throughout the dissertation, one needs only to replace \( \langle f_{ij} \rangle X \) and \( \langle f_{0ij} \rangle X_0 \) with \( D \) and \( D_0 \) respectively, where the \( ij^{th} \) row of \( D \) is given by

\[
d'_{ij} = \frac{\partial F(x^*_i \hat{\beta})}{\partial \hat{\beta}} \bigg|_{\hat{\beta} = \beta},
\]

for the parametric model and similarly for \( D_0 \) in the local model. In this case, only MRGEEI exists.

The link functions, linear predictor models, variance functions, working correlation structures, regressors, and even the datasets used in the MRGEEI model may differ for the parametric and local model components. If the kernel weights of the local procedure are set to 1, MRGEEI can be used to mix together two different parametric models. The parametric and local components of the MRGEEII model can have similar flexibility, except that they must use the same link function since mixing occurs on the \( \eta \)-scale. For example, using the methods in this work, a mixing parameter could be used to compare a parametric model with log link and identity variance function to a model with identity link and square root variance function for Poisson count data. We have SAS\textsuperscript{®} code that allows such possibilities.

### 9.4 Topics for Further Research

As with any research project, there are usually more unanswered questions at the end of the study than at the beginning. This is certainly true here. Some of the major interests for future work are summarized as follows.

1. We believe that under the limiting conditions posed by Lin and Carroll (2000), the convergence rate properties of MRR developed in Starnes and Birch (2000) carry through to the MRGEE estimators with few additional assumptions. Also, we conjecture that
the limiting distributions for the MRGEE estimators are Gaussian as Fan and Ullah (1999) have shown for a MRR-type estimator. Proving these results would establish the large sample, theoretical properties for our methods. Extensions of the convergence rate properties of local quantal MRR regression, found in Chapter 5 of Starnes (1999) are possible for GLIM models in the context of uncorrelated data. However, his work relies on limiting distribution theorems by Fan, Heckman, and Wand (1995) for local GLIM that assume data are uncorrelated. At this point we are uncertain how to extend their results to the case where local GLIM (LLGEE with IND working structure) is applied to clustered correlated data. Furthermore, Whittle’s inequality (Whittle, 1960) either needs to be extended if possible or a generalization needs to be found for clustered correlated data.

2. There is room for improvement with the bandwidth selectors. We have focused on cross-validation as the means of selection, but this approach has been shown to have large variability (e.g., Figure 16). It may be worthwhile to study more sophisticated bandwidth selectors which employ multiple levels of smoothing to obtain estimates that can be plugged into (99), the asymptotically AIMSE optimal bandwidth. These approaches have been shown to lead to less variable selectors in simpler contexts (Ruppert, Sheather, and Wand, 1995).

3. The data-driven mixing parameters need improvement. They exhibited poorer performance than hoped for in the simulation studies, choosing λ too large, with respect to the optimal λ, in several cases. They should be more sensitive to mean model misspecification, especially at small sample sizes where the gain from using optimal MRGEE methods is the greatest.

4. There is clearly a need for studying generating mechanisms for clustered correlated data, especially categorical and count data. These mechanisms should correspond to phenomena observed in practice.

5. It would be interesting to develop an inference procedure for λ and use this to test two competing GEE models. A nonparametric inference procedure is immediately available
if one uses bootstrapping techniques.

6. The MRR approach could be extended to generalized linear mixed models (GLMM).

7. The efficiency of GEE parametric models, when there is slight mean model misspecification, should be explored. We have found that incorporating a non-identity working correlation structure under these conditions does not improve the efficiency of the estimates, even under high levels of corr/var.

8. It would be interesting to determined under what conditions MRGEE mixing eliminates the curse of dimensionality (§4.2) that occurs when using multiple covariates in a local procedure, if such elimination is possible. Fan and Ullah (1999) have shown that when the specified parametric model converges to the true model fast enough, the rate of convergence of their MRR-type estimator does not depend on the number of covariates used in the local model.

9. Simultaneous estimation of $h$ and $\lambda$ may be worth further investigation. The case study in §5.5 indicates that improvement in MSE may be possible and that $h$ and $\lambda$ are simultaneously identifiable parameters.
A Appendix for Chapter 5

A.1 Unbiasedness of the CV Function

Here we prove Theorem 1 (see page 67).

Let
\[ d(h) = \sum_{i=1}^{s} (m_i - \hat{y}_{i,-i}^{LGE})'W_i(m_i - \hat{y}_{i,-i}^{LGE}) \]
and let
\[ \hat{d}(h) = \sum_{i=1}^{s} (y_i - \hat{y}_{i,-i}^{LGE})'W_i(y_i - \hat{y}_{i,-i}^{LGE}), \]
where \( \hat{y}_{i,-i}^{LGE} \) is the local GEE fit for cluster \( i \) with the responses for cluster \( i \) removed and \( W_i \) is some fixed \( n_i \times n_i \) matrix. We show \( E[\hat{d}(h)] = constant + E[d(h)] \).

First recall that \( \text{Cov}[y_2, y_1] = E[(y_2 - E[y_2])(y_1 - E[y_1])'] \). Then \( E[y_2'y_1'] = \text{Cov}[y_2, y_1] + E[y_2E[y'_1]], \) and

\[
E[y'_1Ay_2] = E[tr(y'_1Ay_2)] \\
= E[tr(Ay_2'y_1)] \\
= tr(AE[y_2'y_1]) \\
= tr(ACov[y_2, y_1]) + tr(AE[y_2]E[y'_1]) \\
= tr(ACov[y_2, y_1]) + E[y'_1]AE[y_2].
\]

Now,

\[
E[d(h)] = \sum_{i=1}^{s} E \left[ (y_i - \hat{y}_{i,-i}^{LGE})'W_i(y_i - \hat{y}_{i,-i}^{LGE}) \right] \\
= \sum_{i=1}^{s} E \left[ ((y_i - m_i) + (m_i - \hat{y}_{i,-i}^{LGE}))'W_i((y_i - m_i) + (m_i - \hat{y}_{i,-i}^{LGE})) \right] \\
= \sum_{i=1}^{s} E \left[ (y_i - m_i)'W_i(y_i - m_i) + (m_i - \hat{y}_{i,-i}^{LGE})'W_i(m_i - \hat{y}_{i,-i}^{LGE}) + 2(y_i - m_i)'W_i(m_i - \hat{y}_{i,-i}^{LGE}) \right].
\]

Evaluate each term in the third line above. The first term has expected value

\[
\sum_{i=1}^{s} E [ (y_i - m_i)'W_i(y_i - m_i) ]
\]
\[
\sum_{i=1}^{s} \text{tr}(W_i V_i) + 0'W_i 0 = \sum_{i=1}^{s} \text{tr}(W_i V_i).
\]

The expectation of the second term is
\[
\sum_{i=1}^{s} E \left[ (m_i - \hat{y}_{i,-i}^{LGE})'W_i (m_i - \hat{y}_{i,-i}^{LGE}) \right],
\]
which is \(E[d(h)]\) by definition. Finally, the expectation of the third term is
\[
2 \sum_{i=1}^{s} E \left[ (y_i - m_i)'W_i (m_i - \hat{y}_{i,-i}^{LGE}) \right] = 2 \sum_{i=1}^{s} \left\{ m_i'W_iE[\hat{y}_{i,-i}^{LGE}] - E[y_iW_i\hat{y}_{i,-i}^{LGE}] \right\} = 2 \sum_{i=1}^{s} \left\{ m_i'W_iE[\hat{y}_{i,-i}^{LGE}] - tr(W_iCov[\hat{y}_{i,-i}^{LGE}, y_i]) - E[y_iW_i\hat{y}_{i,-i}^{LGE}] \right\} = -2 \sum_{i=1}^{s} tr(W_iCov[\hat{y}_{i,-i}^{LGE}, y_i]),
\]
where \(Cov[\hat{y}_{i,-i}^{LGE}, y_i] = \{Cov[\hat{y}_{ij,-i}^{LGE}, y_{ij}]\}\), i.e., a matrix with \((j, j')\) element \(Cov[\hat{y}_{ij,-i}^{LGE}, y_{ij}]\).

Combining these results, we have
\[
E[\hat{d}(h)] = E[d(h)] + \sum_{i=1}^{s} \text{tr}(W_i V_i) - 2 \sum_{i=1}^{s} tr(W_iCov[\hat{y}_{i,-i}^{LGE}, y_i])
\]

Note that \(Cov[\hat{y}_{i,-i}^{LGE}, y_i]\) depends on \(h\). Thus, to prove the theorem we must have
\[
\sum_{i=1}^{s} tr(W_iCov[\hat{y}_{i,-i}^{LGE}, y_i]) \text{ constant over } h.
\]
But, since \(\hat{y}_{i,-i}^{LGE}\) is the fit for cluster \(i\) with cluster \(i\) removed, \(Cov[\hat{y}_{ij,-i}^{LGE}, y_{ij}] = 0\) for all \((j, j')\) and
\[
E[\hat{d}(h)] = E[d(h)] + \sum_{i=1}^{s} \text{tr}(W_i V_i).
\]

This completes the proof. Notice that if \(W_i = V_i^{-1}\), then \(E[\hat{d}(h)] = E[d(h)] + N\).

### A.2 Limits for \(\text{tr}(H^{LGE})\) and \(SSE_h\)

The limits of the trace of the local GEE “hat” matrix, \(\text{tr}(H^{LGE})\), are established (pg. 73) followed by the limits for \(SSE_h\) (pg. 74).
For

\[
H_{LGEE} = \begin{bmatrix}
    x_{011}'(X_0'\hat{W}_{011}X_0)^{-1}x_{011}'\hat{W}_{011} \\
    x_{012}'(X_0'\hat{W}_{012}X_0)^{-1}x_{012}'\hat{W}_{012} \\
    \vdots \\
    x_{0s_0s_1}'(X_0'\hat{W}_{0s_0s_1}X_0)^{-1}x_{0s_0s_1}'\hat{W}_{0s_0s_1}
\end{bmatrix},
\]

a crucial bound on the trace is the upper \(N^*\) bound. (Recall that \(N^*\) is the number of unique prediction locations). Let the rows of \(H_{LGEE}\) be indexed by \(k, k = 1, \ldots, N = \sum_i n_i\), the columns by \(\ell, \ell = 1, \ldots, N\), and its elements be denoted by \(h_{k\ell}\). First we show that the upper bound on \(h_{kk}\), the \(k^{th}\) diagonal element, is \(1/r_k\) where \(r_k\) is the number of responses at \(x_k^*\) (i.e., number of times \(x_k^*\) is repeated in the data). Assume that \(\hat{W}_{0k}\) is diagonal, that is, the local working correlation structure \(R_q\) is \(I_{n_q}\) for this argument.

Since the \(k^{th}\) row of \(H_{LGEE}\) is taken from the \(k^{th}\) row of

\[
X_0(X_0'\hat{W}_{0k}X_0)^{-1}X_0'\hat{W}_{0k},
\]

the \(k^{th}\) diagonal element of \(H_{LGEE}\) can be expressed as

\[
h_{kk} = x_{0k}'(X_0'\hat{W}_{0k}X_0)^{-1}x_{0k}\hat{w}_{0kk} = w_{0kk}^{1/2}x_{0k}'(X_0'\hat{W}_{0k}X_0)^{-1}x_{0k}w_{0kk}^{1/2},
\]

where \(\hat{w}_{0kk}\) is the \(k^{th}\) diagonal element of \(\hat{W}_{0k}\). This is the \(k^{th}\) diagonal element of the projector

\[
H_k = \hat{W}_{0k}^{1/2}X_0(X_0'\hat{W}_{0k}X_0)^{-1}X_0'\hat{W}_{0k}^{1/2}.
\]

Let \(q = 1, \ldots, N\) and \(t = 1, \ldots, N\) index the rows and columns of \(H_k\) and let \(h_{kqt}\) be its \((q, t)^{th}\) element. Since this matrix is idempotent, \(H_k = H_k^2\) and \(h_{kqq} = \sum_{t=1}^N h_{kqt}h_{ktq}\). Since \(H_k\) is symmetric, \(h_{kqq} = \sum_{t=1}^N h_{kqt}^2\). Let \(q = k\) for the remainder of this argument. Then, \(h_{kqq} = h_{k,kk}\) equals the \(k^{th}\) diagonal element of \(H_{LGEE}\), by definition.

Now, \(x_{0k}\) is repeated \(r_k\) times in \(X_0\). Let \(k, q_2, \ldots, q_{r_k}\) be the indices (of the rows) of \(X_0\) that identify the repeated \(x_{0k}\). That is, \(x_{0k} = x_{0q_2} = \cdots = x_{0q_{r_k}}\) are all repeats of \(x_{0k}\) in \(X_0\). This implies that the weights for these points are all equal: \(\hat{w}_{0kk}^{1/2} = \hat{w}_{0q_2}^{1/2} = \cdots = \hat{w}_{0q_{r_k}}^{1/2}\), because the weight matrix \(\hat{W}_{0k}\) gives the same kernel weight, variance function weight and derivative weight at the point \(x_k^*\). Thus,

\[
h_{kk} = h_{k,kk} = \hat{w}_{0kk}^{1/2}x_{0k}'(X_0'\hat{W}_{0k}X_0)^{-1}x_{0k}\hat{w}_{0kk}^{1/2}.
\]
\[ h_{k,q_2} = \hat{w}_{0kk}^2 x_0' \left( X_0' \hat{W}_{0k} X_0 \right)^{-1} x_{0q_2} \hat{w}_{0q_2q_2}^{1/2}, \]
\[ \vdots \]
\[ h_{k,q_r k} = \hat{w}_{0kk}^2 x_0' \left( X_0' \hat{W}_{0k} X_0 \right)^{-1} x_{0q_r k} \hat{w}_{0q_r q_r}^{1/2}, \]

are all equal. Combining these results we find that
\[ h_{kk} = \sum_{t=1}^{N} h_{k,kt}^2 = h_{k,kk}^2 + h_{k,q_2}^2 + \cdots + h_{k,q_r k}^2 + \sum_{t \neq \{k,q_2, \ldots, q_r k\}} h_{k,kt}^2 \]
\[ = r_k h_{kk}^2 + \sum_{t \neq \{k,q_2, \ldots, q_r k\}} h_{k,kt}^2, \]

which implies
\[ h_{kk} (1 - r_k h_{kk}) = \sum_{t \neq \{k,q_2, \ldots, q_r k\}} h_{k,kt}^2 \geq 0. \]

Thus, \( r_k h_{kk} \leq 1 \) or \( h_{kk} \leq \frac{1}{r_k} \).

Finally, since there are \( N^* \) non-overlapping sequences \((h_{k,kk}, h_{k,q_2}, \ldots, h_{k,q_r k})\) of identical diagonal elements in \( H^{LGE} \) and each sequence must have a sum \( \leq r_k \frac{1}{r_k} = 1 \), the trace of \( H^{LGE} \) must be \( \leq N^* \).

Next we show that \( \text{tr}(H^{LGE}) \to N^* \) as \( h \to 0 \). Consider prediction at the point \( x^*_k \). As \( h \to 0 \), the kernel weights approach 0 at all locations not identical to \( x^*_k \). Provided that the square-root kernel weights at the points not equal to \( x^*_k \) limit to 0 more quickly than the square-root kernel weight at \( x^*_k \) increases, the weights in
\[ \hat{W}_{0k} = \text{blockdiag} \left( \left\langle \hat{f}_{0ij} \right\rangle K_{hi}^{1/2} \hat{V}_{0i}^{-1} K_{hi}^{1/2} \left\langle \hat{f}_{0ij} \right\rangle \right) \]
limit to 0 except for \( \hat{w}_{0kk}, \hat{w}_{0q_2q_2}, \ldots, \hat{w}_{0q_r k q_r k} \), i.e., those that correspond to the point \( x^*_k \). The idea is that the square-root cross-product of the kernel weight at \( x^*_k \) with the kernel weight at any other point not equal to \( x^*_k \) must limit to 0. This seems reasonable for some kernels (e.g., kernels with bounded support), but a counter-example may be possible.

For simplicity, assume the kernel function has weights that limit as described above. Then, in the row of weights
\[ [h_{k1} h_{k2} \cdots h_{kk} \cdots h_{kN}] = x_0' (X_0' \hat{W}_{0k} X_0)^{-1} X_0' \hat{W}_{0k}, \]
those elements that correspond to \( x_k^* \), \( h_{kk} = h_{kq_2} = \cdots = h_{kq_{r_k}} \), are increasing while all other \( h_{k\ell} \) in the row are limiting to 0 as \( h \to 0 \). Since

\[
(X_0 (X_0' \hat{W}_{0k} X_0)^{-1} X_0' \hat{W}_{0k}) X_0 = X_0,
\]

we have \( X_0 (X_0' \hat{W}_{0k} X_0)^{-1} X_0' \hat{W}_{0k} 1_N = 1_N \), which implies the rows of \( H_k \) must sum to 1. Thus,

\[
h_{kk} + h_{kq_2} + \cdots + h_{kq_{r_k}} = 1 - \sum_{\ell \neq \{k, q_2, \ldots, q_{r_k}\}} h_{k\ell} \to 1.
\]

By a similar argument as before, \( h_{kk} = h_{kq_2} = \cdots = h_{kq_{r_k}} \). In conclusion, \( h_{kk} \to 1/r_k \) as \( h \to 0 \) for each \( k \) which implies \( \text{tr}(H^{L\text{GEE}}) \to N^* \).

Note that \( \hat{W}_0 \) did not need to be diagonal in order to show \( \text{tr}(H^{L\text{GEE}}) \to N^* \). But this does not imply that \( \text{tr}(H^{L\text{GEE}}) \leq N^* \) for nondiagonal \( \hat{W} \). In fact, we discovered a counter-example in which \( \text{tr}(H^{L\text{GEE}}) > N^* \). The model setup (see §7.2) was as follows: \( h = 0.05 \), \( \mu \)-scale contaminated Poisson, \( \gamma = 1 \), \( \sigma = 0.4 \), \( s = 3 \), \( n_i = 10 \), \( \forall i \), log link, Poisson variance function, and an AR1 correlation structure. This is not a concern for us because our simulation studies were done using working independence in the local model.

The \( \text{tr}(H^{L\text{GEE}}) \) also has a lower limit. As \( h \to \infty \), the kernel weights all become equal. This implies \( \hat{W}_{01} = \hat{W}_{02} = \cdots = \hat{W}_{0N} = \hat{W}_0 \) which implies

\[
H^{L\text{GEE}} \to \begin{bmatrix}
X_0' (X_0' \hat{W}_{0k} X_0)^{-1} X_0' \hat{W}_{0k}
X_0' (X_0' \hat{W}_{0k} X_0)^{-1} X_0' \hat{W}_0
\vdots
X_0' (X_0' \hat{W}_{0k} X_0)^{-1} X_0' \hat{W}_0
\end{bmatrix} = X_0 (X_0' \hat{W}_{0k} X_0)^{-1} X_0' \hat{W}_0.
\]

Thus, as \( h \to \infty \), \( \text{tr}(H^{L\text{GEE}}) \to \text{tr}(X_0 (X_0' \hat{W}_{0k} X_0)^{-1} X_0' \hat{W}_0) = \text{tr}(I_p) = p \), the number of model parameters in the local model.

An alternative “hat” matrix can be proposed that is constructed from rows of the symmetric projection matrices, \( \hat{W}_{0k} X_0 (X_0' \hat{W}_{0k} X_0)^{-1} X_0' \hat{W}_{0k}^{1/2} \). It can be motivated as follows. Consider a transformation of the \( \eta \)-scale model,

\[
W_{0ij}^{1/2} \eta_{L\text{GEE}} = W_{0ij}^{1/2} X_0 \theta(x_{ij}^*),
\]

where \( W_{0ij}^{1/2} \) is the “root” of \( W_{0ij} \) such that \( W_{0ij}^{1/2} W_{0ij}^{1/2} = W_{0ij} \). \( W_{0ij}^{1/2} \) may be obtained by a Cholesky decomposition or by diagonalization of \( W_0 \). Now,

\[
W_{0ij}^{1/2} \eta_{L\text{GEE}} = \hat{W}_{0ij} X_0 \theta(x_{ij}^*) \approx \hat{W}_{0ij} X_0 (X_0' \hat{W}_{0ij} X_0)^{-1} X_0' \hat{W}_{0ij}^{1/2} \hat{W}_{0ij}^{1/2}.
\]
which is analogous to regressing the transformed response \( W_{0ij}^{1/2} \mathbf{F}^{-1}(\mathbf{y}) \) on the variable \( W_{0ij}^{1/2} \mathbf{z}_{0ij} \), i.e., \( \hat{W}_{0ij}^{1/2} \mathbf{X}_0 \left( \mathbf{X}'_0 \hat{W}_{0ij} \mathbf{X}_0 \right)^{-1} \mathbf{X}'_0 \hat{W}_{0ij}^{1/2} \) is a projection matrix. The rows of the local hat matrix may be formed using the \( ij \)th rows of this matrix for all \( ij \).

\[
\mathbf{H}^{LGEE}_2 = \begin{bmatrix}
\hat{W}_{011}^{1/2} \mathbf{X}_0 (\mathbf{X}'_0 \hat{W}_{011} \mathbf{X}_0)^{-1} \mathbf{X}'_0 \hat{W}_{011}^{1/2} \\
\hat{W}_{012}^{1/2} \mathbf{X}_0 (\mathbf{X}'_0 \hat{W}_{012} \mathbf{X}_0)^{-1} \mathbf{X}'_0 \hat{W}_{012}^{1/2} \\
\vdots \\
\hat{W}_{0sn}^{1/2} \mathbf{X}_0 (\mathbf{X}'_0 \hat{W}_{0sn} \mathbf{X}_0)^{-1} \mathbf{X}'_0 \hat{W}_{0sn}^{1/2}
\end{bmatrix},
\]

where \( \hat{W}_{0ij}^{1/2} \) is the \( ij \)th row of \( \hat{W}_{0ij}^{1/2} \) (the row corresponding to the \( j \)th response in the \( i \)th cluster). As \( h \to \infty \), a \( p - 1 \) degree local polynomial GEE model approaches a global \( p - 1 \) degree polynomial GEE model. The trace of \( \mathbf{H}^{LGEE}_2 \) for this “globalized” local model is in fact \( p \). Furthermore, by similar arguments to those above, \( \text{tr}(\mathbf{H}^{LGEE}_2) \to N^* \) as \( h \to 0 \). However, again, if \( \hat{W} \) is not diagonal, it may be possible that \( \text{tr}(\mathbf{H}^{LGEE}_2) > N^* \). We have found counter-examples illustrating this for many positive-definite \( \hat{W} \).

Another possible measure of the “degrees of freedom” of the nonparametric GEE is the rank of \( \mathbf{H}^{LGEE} \). But, since \( \hat{\eta} \neq \mathbf{H}^{LGEE} \mathbf{z} \), the rank of \( \mathbf{H}^{LGEE} \) would not necessarily count the degrees of freedom in \( \hat{\eta} \). We will not consider it here, but it may be worth investigating the properties of \( \text{rank}(\mathbf{H}^{LGEE}) \) in future research.

Limits for \( SSE_h = \sum_{i=1}^s \left( \mathbf{y}_i - \hat{\mathbf{y}}_{i,-i}^{LGEE} \right)' \left( \mathbf{y}_i - \hat{\mathbf{y}}_{i,-i}^{LGEE} \right) \) are now considered. As \( h \to 0 \) the nonparametric fit eventually interpolates the means at the observed prediction locations so \( SSE_h \to SSE_0 \), where \( SSE_0 = \sum_{i=1}^s \left( \mathbf{y}_i - \bar{\mathbf{y}}_i \right)' \left( \mathbf{y}_i - \bar{\mathbf{y}}_i \right) \). On the other hand, as \( h \to \infty \), the local model becomes a global model. It can be argued that for any finite bandwidth, the fit is more flexible than this “globalized” local model and thus \( SSE_h < SSE_0 \). Though this argument does not prove that \( SSE_h < SSE_0 \), it seems reasonable and no counter examples have been found. If the boundary is a serious concern for the user, setting \( SSE_0 \) equal to the largest \( SSE_h \) computed (over \( h \) within the same model and same dataset) will make certain that the bound is maintained.
A.3 Minimization of GD_I and GD_{II}

We show how minimization of (101) leads to (103) (pg. 77). The minimization of (102) is the same and will not be shown.

\[
GD_I(\lambda) = (m - \hat{y}_{MRGEI})' W_I (m - \hat{y}_{MRGEI})
\]

\[
= (m - \lambda \hat{y}_{LGE} - (1 - \lambda)\hat{y}_{GEE})' W_I (m - \lambda \hat{y}_{LGE} - (1 - \lambda)\hat{y}_{GEE})
\]

\[
= (m - \lambda (\hat{y}_{LGE} - \hat{y}_{GEE} - \hat{y}_{GEE})' W_I (m - \lambda (\hat{y}_{LGE} - \hat{y}_{GEE} - \hat{y}_{GEE})
\]

\[
= m' W_I m - 2\lambda m' W_I (\hat{y}_{LGE} - \hat{y}_{GEE}) - 2m' W_I \hat{y}_{GEE}
\]

\[
+ \lambda^2 (\hat{y}_{LGE} - \hat{y}_{GEE})' W_I (\hat{y}_{LGE} - \hat{y}_{GEE}) + 2\lambda (\hat{y}_{LGE} - \hat{y}_{GEE})' W_I \hat{y}_{GEE}
\]

\[
+ \hat{y}_{GEE} W_I \hat{y}_{GEE}.
\]

Only the second row and the second term of the first row in the equality above depend on \(\lambda\), so we can ignore all other terms in the minimization. The remaining terms are a quadratic function of the form \(a\lambda^2 + 2b\lambda\) where

\[
a = (\hat{y}_{LGE} - \hat{y}_{GEE})' W_I (\hat{y}_{LGE} - \hat{y}_{GEE})
\]

and

\[
b = (\hat{y}_{LGE} - \hat{y}_{GEE})' W_I \hat{y}_{GEE} - m' W_I (\hat{y}_{LGE} - \hat{y}_{GEE})
\]

\[
= (\hat{y}_{LGE} - \hat{y}_{GEE})' W_I (\hat{y}_{GEE} - m)
\]

\[
= -(\hat{y}_{LGE} - \hat{y}_{GEE})' W_I (m - \hat{y}_{GEE}).
\]

Since \(W_I\) is assumed to be positive definite, this has a unique minimum at \(\lambda = -b/a\) or

\[
\lambda_{optMRGEI} = \frac{(\hat{y}_{LGE} - \hat{y}_{GEE})' W_I (m - \hat{y}_{GEE})}{(\hat{y}_{LGE} - \hat{y}_{GEE})' W_I (\hat{y}_{LGE} - \hat{y}_{GEE})}.
\]

A.4 The Variance of \(\bar{y}\) and \(F^{-1}(\bar{y})\)

Expressions for the variance of \(\bar{y}\) and \(F^{-1}(\bar{y})\) (pg. 78) are developed here. The elements of \(\bar{y}\) consist of the averages at the unique prediction locations. These may be averages over independent responses from different clusters, dependent responses within clusters, or a combination of these. Covariances between averages at different prediction locations will
involve the responses that have clusters in common at those points. For example, if every pair of prediction points have observations with at least one cluster in common—as they do in the simulation study—then all the averages in $\bar{y}$ are correlated. If all the regressors are cluster-specific (see §2.2.2) then the $\bar{y}_k$ will be uncorrelated. Thus, $Var[\bar{y}]$ and $Var[F^{-1}(\bar{y})]$ are not necessarily block diagonal matrices. In calculating $Var[\bar{y}]$, it is important to identify to which cluster an observation belongs. This will be done with indicator variables.

An example used to illustrate a correlation structure is shown in Figure 30. If the responses are clustered correlated, computing the variance at point $x_2$ involves the $Var[y_{14}]$, $Var[y_{24}]$, and $Var[y_{25}]$, but also $Cov[y_{24}, y_{25}]$. For any cluster that has more than one response at a prediction location, their covariances will enter into the variance calculation. Computing the covariance between averages at $x_1$ and $x_2$ will involve only $Cov[y_{23}, y_{24}]$ and $Cov[y_{13}, y_{14}]$. Hence, computing the covariance of averages at different prediction locations involves only those clusters that have points at both locations.

Let $x^*_1, x^*_2, \ldots, x^*_N$ be the unique prediction locations and let $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_N$ be the respective averages at those locations. Recall that for any given location $x^*_k$ the number of
observations at that location is \( r_k \) (see §A.2). Then the average at \( x_k^* \) is

\[
\bar{y}_k = \frac{\sum_{s=1}^{n_i} \sum_{j=1}^{n_i} I(x_{ij}^* = x_k^*) y_{ij}}{\sum_{s=1}^{n_i} \sum_{j=1}^{n_i} I(x_{ij}^* = x_k^*)} = \frac{1}{r_k} \sum_{i=1}^{n_i} \sum_{j=1}^{n_i} I(x_{ij}^* = x_k^*) y_{ij}.
\]

Since \( \text{Var}[\bar{y}_k] = \text{Cov}[\bar{y}_k, \bar{y}_\ell] \), we only need to compute the covariance to find the elements of \( \text{Var}[\bar{y}] \). The covariance between \( \bar{y}_k \) and \( \bar{y}_\ell \) for \( k = 1, \ldots, N^* \) and \( \ell = 1, \ldots, N^* \) is

\[
\text{Cov}[\bar{y}_k, \bar{y}_\ell] = \text{Cov} \left[ \frac{1}{r_k} \sum_{i=1}^{n_i} \sum_{j=1}^{n_i} I(x_{ij}^* = x_k^*) y_{ij}, \frac{1}{r_\ell} \sum_{i=1}^{n_{i'}} \sum_{j=1}^{n_{i'}} I(x_{i'j'}^* = x_\ell^*) y_{i'j'} \right]
\]

\[
= \frac{1}{r_k r_\ell} \sum_{i=1}^{n_i} \sum_{j=1}^{n_i} \sum_{i'=1}^{n_{i'}} \sum_{j'=1}^{n_{i'}} I(x_{ij}^* = x_k^*) I(x_{i'j'}^* = x_\ell^*) \text{Cov}[y_{ij}, y_{i'j'}].
\]

Since \( \text{Cov}[y_{ij}, y_{i'j'}] = 0 \) for all \( i \neq i' \),

\[
\text{Cov}[\bar{y}_k, \bar{y}_\ell] = \frac{1}{r_k r_\ell} \sum_{i=1}^{n_i} \sum_{j=1}^{n_i} \sum_{i'=1}^{n_{i'}} \sum_{j'=1}^{n_{i'}} I(x_{ij}^* = x_k^*) I(x_{i'j'}^* = x_\ell^*) \text{Cov}[y_{ij}, y_{i'j'}].
\]

(148)

Note that even though \( \text{Cov}[y_{ij}, y_{i'j'}] = \text{Cov}[y_{i'j'}, y_{ij}] \), these observations can not be interchanged in the covariance calculation. This is because in general \( I(x_{ij}^* = x_k^*) I(x_{i'j'}^* = x_\ell^*) \neq I(x_{i'j'}^* = x_\ell^*) I(x_{ij}^* = x_k^*) \).

The \((k, \ell)^{th}\) element of \( \text{Var}[\bar{y}] \) is given by (148). Once \( \text{Var}[\bar{y}] \) has been computed, \( \text{Var}[F^{-1}(\bar{y})] \) can be estimated by making use of a first order multivariate Taylor series approximation,

\[
F^{-1}(\bar{y}) \approx F^{-1}(\mu) + \left\langle \frac{\partial F^{-1}(\bar{y})}{\partial \bar{y}} \right|_{\bar{y}=\mu} (\bar{y} - \mu).
\]

The derivative above can be simplified to

\[
\frac{\partial F^{-1}(\bar{y})}{\partial \bar{y}} = \left[ \frac{1}{f(F^{-1}(\bar{y}))} \right],
\]

(division is elementwise) since the derivative of an inverse function is the reciprocal of the derivative of the function when that function is differentiable (\( F \) was assumed to be differentiable). Substituting \( \bar{y} \) as an estimate for \( \mu \), we have

\[
\text{Var}[F^{-1}(\bar{y})] \approx \left\langle f(F^{-1}(\bar{y})) \right\rangle^{-1} \text{Var}[\bar{y}] \left\langle f(F^{-1}(\bar{y})) \right\rangle^{-1}.
\]

(149)

Estimates of \( \text{Var}[^\bar{y}] \) and \( \text{Var}[F^{-1}(^\bar{y})] \) can be computed by estimating \( \text{Cov}[y_{ij}, y_{i'j'}] \) in (148) with elements from \( \hat{V} \). This is recommended with some caution: if the correlation structure is not correct, neither will the elements of \( \hat{V} \).
A.5 Minimum MSE $\lambda$ Estimation

The calculation details for the $MSE(x^*)$, $\lambda_{AvgMSEopt}$, and $\lambda_{IMSEopt}$ are shown here (see pg. 80).

For $MSE(x^*)$ we have

$$
Bias^2(x^*) = \left[(1 - \lambda)E^P + \lambda E^{NP} - m\right]^2
= \left[\lambda(E^{NP} - E^P) + (E^P - m)\right]^2
= \lambda^2(E^{NP} - E^P)^2 + 2\lambda(E^{NP} - E^P)(E^P - m) + (E^P - m)^2,
$$

and

$$
Var(x^*) = (1 - \lambda)^2V^P + \lambda^2V^{NP} + 2\lambda(1 - \lambda)C
= V^P - 2\lambda V^P + \lambda^2V^P + \lambda^2V^{NP} + 2\lambda C - 2\lambda^2 C
= \lambda^2(V^P + V^{NP} - 2C) - 2\lambda(V^P - C) + V^P
$$

Thus,

$$
MSE(x^*) = Bias^2(x^*) + Var(x^*)
= \lambda^2 \left[(E^P - E^{NP})^2 + V^P + V^{NP} - 2C\right]
- 2\lambda \left[(E^P - E^{NP})(E^P - m) + V^P - C\right] + (E^P - m)^2 + V^P,
$$

or simply $a\lambda^2 - 2b\lambda + c$ for appropriately defined $a$, $b$, and $c$. Since $a > 0$, a unique minimum of $\lambda$ is obtained at $\lambda = b/a$. Hence,

$$
\lambda(x^*) = \frac{(E^P - E^{NP})(E^P - m) + V^P - C}{(E^P - E^{NP})^2 + V^P + V^{NP} - 2C}.
$$

The Average MSE is given by $\frac{1}{N} \sum_{i=1}^{s} \sum_{j=1}^{n_i} MSE(x^*_{ij})$ and has the minimum

$$
\lambda_{AvgMSEopt} = \frac{1}{N} \sum_{i=1}^{s} \sum_{j=1}^{n_i} (E^P(x^*_{ij}) - E^{NP}(x^*_{ij}))(E^P(x^*_{ij}) - m(x^*_{ij})) + V^P(x^*_{ij}) - C(x^*_{ij})
\frac{1}{N} \sum_{i=1}^{s} \sum_{j=1}^{n_i} (E^P(x^*_{ij}) - E^{NP}(x^*_{ij}))^2 + V^P(x^*_{ij}) + V^{NP}(x^*_{ij}) - 2C(x^*_{ij})
$$

or simply

$$
\lambda_{AvgMSEopt} = \frac{\sum_{i=1}^{s} \sum_{j=1}^{n_i} (E^P(x^*_{ij}) - E^{NP}(x^*_{ij}))(E^P(x^*_{ij}) - m(x^*_{ij})) + V^P(x^*_{ij}) - C(x^*_{ij})}{\sum_{i=1}^{s} \sum_{j=1}^{n_i} (E^P(x^*_{ij}) - E^{NP}(x^*_{ij}))^2 + V^P(x^*_{ij}) + V^{NP}(x^*_{ij}) - 2C(x^*_{ij})}
$$
The integrated MSE is $\int MSE(x^*)dx^*$. Because the multiple integral is a linear operator, the minimization steps are the same as those for $MSE(x^*)$. The minimum is

$$\lambda_{IMSE_{opt}} = \frac{\int \left[ (E^P(x^*) - E^{NP}(x^*))(E^P(x^*) - m(x^*)) + V^P(x^*) - C(x^*) \right] dx^*}{\int \left[ (E^P(x^*) - E^{NP}(x^*))^2 + V^P(x^*) + V^{NP}(x^*) - 2C(x^*) \right] dx^*}.$$
B Appendix for Chapter 6

B.1 Covariance of Parametric and Nonparametric Fits

The covariance between \( \hat{\eta}^P \) and \( \hat{\eta}^{NP} \) at \( x^* \) is derived (see pp. 91–92).

First express the covariance as

\[
\text{Cov}[x'\hat{\beta}, x'_0\hat{\theta}] = E[(x'\hat{\beta} - E[x'\hat{\beta}])(x'_0\hat{\theta} - E[x'_0\hat{\theta})]'.
\]

Using the approximate results

\[
x'\hat{\beta} \approx x'\beta^* + x'\left(\sum_{i=1}^s X'_iW_iX_i\right)^{-1}\sum_{i=1}^s X'_iW_i\langle f_{ij}\rangle^{-1}(y_i - F(X_i\beta^*))
\]

\[
x'_0\hat{\theta} \approx x'_0\theta^* + x'_0\left(\sum_{i=1}^s X'_0W_{0i}X_{0i}\right)^{-1}\sum_{i=1}^s X'_0W_{0i}\langle f_{0ij}\rangle^{-1}(y_i - F(X_{0i}\theta^*))
\]

we can approximate the covariance as

\[
\text{Cov}[x'\hat{\beta}, x'_0\hat{\theta}] \approx E[(x'\beta^* + x'\delta - E[x'\beta^* + x'\delta]) (x'_0\theta^* + x'_0\gamma - E[x'_0\theta^* + x'_0\gamma])]'
\]

Using this result, substituting for \( \delta \) and \( \gamma \), and factoring, we have,

\[
\text{Cov}[x'\hat{\beta}, x'_0\hat{\theta}] \approx x'\left( E[\delta\gamma'] - E[\delta]E[\gamma'] \right) x_0.
\]

With the aid of the notation in Table 7, where

\[
L_{ik} = E[(y_i - F(X_i\beta^*)) (y_k - F(X_{0k}\theta^*))]' - (m_i - F(X_i\beta^*)) (m_k - F(X_{0k}\theta^*))'.
\]

\[
= \text{Cov}[(y_i - F(X_i\beta^*), (y_k - F(X_{0k}\theta^*))]
\]

\[
= \text{Cov}[y_i, y_k].
\]
Since the clusters are uncorrelated, this covariance is 0 when $i \neq k$ and equal to $Var[y_i]$ when $i = k$. The double sum in (150) hence reduces to a single sum and we can approximate the covariance term as

$$Cov[x'\hat{\beta}, x'_0\hat{\theta}] \approx x'B \left( \sum_{i=1}^{s} C_i Var[y_i]C_{0i} \right) B_0 x_0.$$

### B.2 Expectation of GEE MOM Estimators

The expected values of the MOM scale and correlation estimators (pp. 94–95) are calculated for the user’s model, keeping in mind that the mean model may be misspecified. It is understood that the expectations, variances, and covariances of the responses $Y_{ij}$ are known (e.g., they may be given by or computed from the true model).

To keep the calculations as simple as possible, many assumptions will be used. Let $n_i = n$ for all $i$. Assume the quantity $F(x'_{ij}\hat{\beta})$ is fixed at $F(x'_{ij}\beta^*)$ and the user’s variance function, $a^*(F(x'_{ij}\hat{\beta}))$, is fixed at $a^*(F(x'_{ij}\beta^*))$ (i.e., condition on $\beta^*$). The expectation of

$$\hat{\phi}^2 = \frac{1}{sn - p} \sum_{i=1}^{s} \sum_{j=1}^{n} \frac{(Y_{ij} - F(x'_{ij}\hat{\beta}))^2}{a^* (F(x'_{ij}\beta^*))},$$

is

$$E[\hat{\phi}^2] \approx \frac{1}{sn - p} \sum_{i=1}^{s} \sum_{j=1}^{n} E \left[ \frac{((Y_{ij} - m(x^*_{ij})) + (m(x^*_{ij}) - F(x'_{ij}\beta^*)))^2}{a^* (F(x'_{ij}\beta^*))} \right]$$

$$= \frac{1}{sn - p} \sum_{i=1}^{s} \sum_{j=1}^{n} a^* (F(x'_{ij}\beta^*)) \left[ E \left[ (Y_{ij} - m(x^*_{ij}))^2 \right] + E \left[ (m(x^*_{ij}) - F(x'_{ij}\beta^*))^2 \right] \right. \right.$$

$$\left. + 2E \left[ (Y_{ij} - m(x^*_{ij}))(m(x^*_{ij}) - F(x'_{ij}\beta^*)) \right] \right]$$

$$= \frac{1}{sn - p} \sum_{i=1}^{s} \sum_{j=1}^{n} Var[Y_{ij}] + \frac{(m(x^*_{ij}) - F(x'_{ij}\beta^*))^2}{a^* (F(x'_{ij}\beta^*))}$$

$$\overset{define}{=} \tilde{\phi}^2.$$

For the expected value of the compound symmetry estimator,

$$\hat{\rho}_{CS} = \left( \frac{1}{2} \frac{sn(n - 1) - p}{SN} \right) \sum_{i=1}^{s} \sum_{j<k}^{n} \frac{(Y_{ij} - F(x'_{ij}\beta^*))(Y_{ik} - F(x'_{ik}\beta^*))}{\sqrt{\hat{\phi}^2 a^* (F(x'_{ij}\beta^*))} \sqrt{\hat{\phi}^2 a^* (F(x'_{ik}\beta^*))}},$$

201
suppose that $\hat{\phi}^2$ is fixed at $\bar{\phi}^2$. Then, $E[\hat{\rho}_{CS}]$

$$
\approx \left(\frac{1}{2} sn(n-1) - p\right)^{-1} \sum_{i=1}^{s} \sum_{j<k}^{n} \frac{1}{\hat{\phi}^2 \sqrt{a^*(F(x'_{ij}\beta^*))} a^*(F(x'_{ik}\beta^*)) \sqrt{a^*(F(x'_{ij}\beta^*))} a^*(F(x'_{ik}\beta^*))} 

\times E \left[ \left( (Y_{ij} - m(x^*_{ij})) + (m(x^*_{ij}) - F(x'_{ij}\beta^*)) \right) \left( (Y_{ik} - m(x^*_{ik})) + (m(x^*_{ik}) - F(x'_{ik}\beta^*)) \right) \right] 

= \left(\frac{1}{2} sn(n-1) - p\right)^{-1} \sum_{i=1}^{s} \sum_{j<k}^{n} \frac{1}{\hat{\phi}^2 \sqrt{a^*(F(x'_{ij}\beta^*))} a^*(F(x'_{ik}\beta^*))} \frac{\text{Cov}[Y_{ij}, Y_{ik}] + (m(x^*_{ij}) - F(x'_{ij}\beta^*)) (m(x^*_{ik}) - F(x'_{ik}\beta^*))}{\bar{\phi}^2 \sqrt{a^*(F(x'_{ij}\beta^*))} a^*(F(x'_{ik}\beta^*))}.
$$

The approximate expectation of

$$
\hat{\rho}_{ARI} = \frac{1}{(n-1)s - p} \sum_{i=1}^{s} \sum_{j=1}^{n-1} \frac{Y_{ij} - F(x'_{ij}\hat{\beta})}{\sqrt{\hat{\phi}^2 a^*(F(x'_{ij}\hat{\beta}))}} \frac{Y_{ij+1} - F(x'_{ij+1}\hat{\beta})}{\sqrt{\hat{\phi}^2 a^*(F(x'_{ij+1}\hat{\beta}))}},
$$

is found using the same technique as above, leading to $E[\hat{\rho}_{ARI}] \approx (n-1)s - p)^{-1} \sum_{i=1}^{s} \sum_{j=1}^{n-1} \frac{\text{Cov}[Y_{ij}, Y_{i,j+1}] + (m(x_{ij}) - F(x'_{ij}\beta^*)) (m(x_{i,j+1}) - F(x'_{i,j+1}\beta^*))}{\bar{\phi}^2 \sqrt{a^*(F(x'_{ij}\beta^*))} a^*(F(x'_{i,j+1}\beta^*))}.$

Fixing $F(x'_{ij}\hat{\beta})$ at $F(x'_{ij}\beta^*)$, $a^*(F(x'_{ij}\hat{\beta}))$ at $a^*(F(x'_{ij}\beta^*))$, and $\bar{\phi}^2$ at $\bar{\phi}^2$ may seem objectionable. Ignoring the variability of $\hat{\beta}$ and $\bar{\phi}^2$ in the calculations is equivalent to using a first order Taylor series expansion of $\bar{\phi}^2$ about $\beta^*$ in the case of computing $E[\bar{\phi}^2]$, or a first order Taylor series expansions of $\hat{\rho}_{CS}$ and $\hat{\rho}_{ARI}$ about $\beta^*$ and $\bar{\phi}^2$, in the case of computing $E[\hat{\rho}_{CS}]$ or $E[\hat{\rho}_{ARI}]$. This is because the expectation of a first order approximation amounts to substitution of $\beta^*$ and/or $\bar{\phi}^2$ if $E[\beta] \approx \beta^*$. In §7.4.1 it is shown that $\beta^*$ is within 0.1%-1% relative error of $E[\hat{\beta}]$ for the simulation models of §7.1 when corr/var is small and within 0.1%-11% relative error for the simulation models when corr/var is high (see Table 12).

The results of this section are summarized in Table 33.

### B.3 Simplification of Parametric GEE Fit Variance in the GLIM Context

It is easy to show that

$$
\langle f_{ij} \rangle^{-1} \text{Var}[Y_i] \langle f_{ij} \rangle^{-1} W_i
$$

202
from the center of expression (128) on pages 98–99 simplifies to
\[ \left\langle A(m(X_i)) \right\rangle = \left\langle \frac{A(F(X_i\beta))}{A(F(X_i\beta)\beta^*)} \right\rangle \]
when the data are assumed to be uncorrelated. Lack of correlation implies that all matrices in this expression are diagonal. Diagonal matrices may be commuted freely and have inverses that are simply reciprocals of the diagonal elements. Thus,
\[
\langle f_{ij} \rangle^{-1} \text{Var}[Y_i] \langle f_{ij} \rangle^{-1} W_i \\
= \langle f_{ij} \rangle^{-1} A(m_i) \langle f_{ij} \rangle^{-1} \langle f_{ij} \rangle A^{-1}(F(X_i\beta^*)) \langle f_{ij} \rangle \\
= \langle f_{ij} \rangle^{-1} A(m_i) A^{-1}(F(X_i\beta^*)) \langle f_{ij} \rangle \\
= \left\langle \frac{A(m(X_i))}{A(F(X_i\beta))} \right\rangle .
\]
C Appendix for Chapter 7

C.1 GLMM Simulation Model Properties

The unconditional expectation and variance of the GLMM simulation model given on page 106 are approximated by viewing \( F(x'_{ij}\beta_S + g(x^*_{ij}) + e_i) \) as a function of \( e_i \) and expanding it about \( E[e_i] = 0 \) with a first order Taylor series for cluster \( i \):

\[
F(\eta_i + g_i + 1e_i) \approx F(\eta_i + g_i) + \left( f(x'_{ij}\beta_S + g(x^*_{ij})) \right)_i 1_n e_i,
\]

where \( \eta_i = \{x'_{ij}\beta_S\}_{n_i \times 1} \) and \( g_i = \{g(x^*_{ij})\}_{n_i \times 1} \). Also, treat

\[
a \left( F(x'_{ij}\beta_S + g(x^*_{ij}) + e_i) + G(x^*_{ij}) \right)
\]

as a function of \( e_i \) and expand it in a first order Taylor series about 0:

\[
a \left( F(x'_{ij}\beta_S + g(x^*_{ij}) + e_i) + G(x^*_{ij}) \right) \approx 
\]

\[
a \left( F(x'_{ij}\beta_S + g(x^*_{ij})) + G(x^*_{ij}) \right) + a' \left( F(x'_{ij}\beta_S + g(x^*_{ij})) + G(x^*_{ij}) \right) f(x'_{ij}\beta_S + g(x^*_{ij}))e_i.
\]

Then,

\[
E[Y_i] = E_{e_i} \left[ E_{Y_i|e_i} [Y_i|e_i] \right]
\]

\[
= E_{e_i} \left[ F(\eta_i + g_i + 1_n e_i) + G_i \right]
\]

\[
\approx E_{e_i} \left[ F(\eta_i + g_i) + \left( f(x'_{ij}\beta_S + g(x^*_{ij})) \right)_i 1_n e_i + G_i \right]. \text{ Thus,}
\]

\[
E[Y_i] \approx F(\eta_i + g_i) + G_i,
\]

where \( G_i = \{G(x^*_{ij})\}_{n_i \times 1} \), and,

\[
Var[Y_i] = E_{e_i} \left[ Var_{Y_i|e_i} [Y_i|e_i] \right] + Var_{e_i} \left[ E_{Y_i|e_i} [Y_i|e_i] \right]
\]

\[
= E_{e_i} \left[ A(F(\eta_i + g_i + 1_n e_i) + G_i) \right] + Var_{e_i} \left[ F(\eta_i + g_i + 1_n e_i) + G_i \right]
\]

\[
\approx E_{e_i} \left[ \left( a \left( F(x'_{ij}\beta_S + g(x^*_{ij})) + G(x^*_{ij}) \right) \right) 
\]

\[
+a' \left( F(x'_{ij}\beta_S + g(x^*_{ij})) + G(x^*_{ij}) \right) f(x'_{ij}\beta_S + g(x^*_{ij}))e_i \right] 
\]

\[
+ Var_{e_i} \left[ F(\eta_i + g_i) + \left( f(x'_{ij}\beta_S + g(x^*_{ij})) \right)_i 1_n e_i + G_i \right]. \text{ Thus,}
\]

\[
Var[Y_i] \approx A(F(X_i\beta + g_i) + G_i) + \left( f(x'_{ij}\beta_S + g(x^*_{ij})) \right) V_{e_i} \left( f(x'_{ij}\beta_S + g(x^*_{ij})) \right),
\]

where \( V_{e_i} = \sigma^2 1_n 1_n' \).
C.2 Exact Moments of the Poisson Simulation Model

Finding the exact moments for the Poisson simulation model with log link and Gaussian random effect (pp. 106–107) is motivated as follows.

First we find \( E[e^{\eta_j + g_{ij} + e_i}] \) and \( Cov[e^{\eta_j + g_{ij}} + e_i, e^{\eta_k + g_{ik}} + e_i] \). Now,

\[
E[e^{\eta_j + g_{ij} + e_i}] = e^{\eta_j + g_{ij}} E[e^{e_i}] = e^{\eta_j + g_{ij}} M_{e_i}(1),
\]

where \( M_{e_i}(t) = e^{\sigma^2 t^2/2} \), the moment generating function for \( N(0, \sigma^2) \), is being evaluated at \( 1 \) (Hoel, Port, and Stone, 1971, pg. 198). Thus,

\[
E[e^{\eta_j + g_{ij} + e_i}] = e^{\sigma^2/2} e^{\eta_j + g_{ij}}.
\]

The covariance can also be found using the moment generating function.

\[
Cov[e^{\eta_j + g_{ij} + e_i}, e^{\eta_k + g_{ik} + e_i}] = e^{\eta_j + g_{ij}} e^{\eta_k + g_{ik}} Cov[e^{e_i}, e^{e_i}] = e^{\eta_j + g_{ij} + \eta_k + g_{ik}} Var[e^{e_i}]
\]

\[
= e^{\eta_j + g_{ij} + \eta_k + g_{ik}} \left( E[e^{2e_i}] - E[e^{e_i}] E[e^{e_i}] \right)
\]

\[
= e^{\eta_j + g_{ij} + \eta_k + g_{ik}} \left( M_{e_i}(2) - M_{e_i}(1) \right)
\]

\[
= e^{\eta_j + g_{ij} + \eta_k + g_{ik}} \left( \sigma^2 - e^{\sigma^2/2} \right)
\]

\[
= e^{\eta_j + g_{ij} + \eta_k + g_{ik}} e^{\sigma^2} \left( e^{\sigma^2/2} - 1 \right).
\]

Thus,

\[
E[Y_i] = E_{e_i} \left[ E_{Y_i|e_i} [Y_i|e_i] \right] = E_{e_i} \left[ e^{\eta_i + G_i + 1_n_i e_i} + G_i \right] = e^{\sigma^2/2} e^{\eta_i + G_i} + G_i
\]

and,

\[
Var[Y_i] = E_{e_i} \left[ Var_{Y_i|e_i} [Y_i|e_i] \right] + Var_{e_i} \left[ E_{Y_i|e_i} [Y_i|e_i] \right]
\]

\[
= E_{e_i} \left[ \left( e^{\eta_j + g_{ij} + e_i} + G_{ij} \right) \right] + Var_{e_i} \left[ e^{\eta_i + G_i + 1_n_i e_i} + G_i \right]
\]

\[
= \left( e^{\sigma^2/2} e^{\eta_j + g_{ij}} + G_{ij} \right) + e^{\sigma^2} \left( e^{\sigma^2/2} - 1 \right) \left\{ e^{\eta_j + g_{ij} + \eta_k + g_{ik}} \right\}_{n_i \times n_i}.
\]

Note that this can be expressed as a function of the mean if there is no \( \mu \)-scale contamination \( (G \equiv 0) \). Let \( \mu_i = e^{\sigma^2/2} e^{\eta_i + G_i} \), then

\[
Var[Y_i] = \langle \mu_i \rangle + \left( e^{\sigma^2/2} - 1 \right) \mu_i \mu_i'.
\]

205
C.3 Expectation Limits of MOM Estimators

For simplicity of the arguments, assume $n_i = n$ for all $i$. Assume that the mean function $m(x^*_ij)$ is known. Recall that $a^*$ is the user’s variance function.

We use the approximate expectations obtained in Appendix B.2. The terms in

$$E[\hat{\phi}^2] \approx \frac{1}{sn - p} \sum_{i=1}^{s} \sum_{j=1}^{n} \text{Var}[Y_{ij}] + \left( m(x^*_ij) - F(x^*_ij\beta^*) \right)^2 \frac{a^* (F(x^*_ij\beta^*))}{a^* (F(x^*_ij\beta^*))},$$

are independent of $i$ because the clusters are identically distributed. Denote by $\gamma$ the inner sum, then

$$E[\hat{\phi}^2] \approx \frac{1}{sn - p} \sum_{i=1}^{s} \gamma = \frac{\gamma}{sn - p}. \quad (151)$$

For the simulation models of §7 this becomes

$$E[\hat{\phi}^2] \approx \frac{1}{sn - p} \sum_{i=1}^{s} \gamma = \frac{\gamma}{sn - p}.$$ 

This reduces to

$$\lim_{s \to \infty} E[\hat{\phi}^2] \approx \lim_{s \to \infty} \frac{s\gamma}{sn - p} = \frac{\gamma}{n}.$$ 

For the simulation models of §7 this becomes

$$\lim_{s \to \infty} E[\hat{\phi}^2] \approx \frac{1}{n} \sum_{j=1}^{n} \frac{a^* (F(x^*_ij\beta^*)) + \sigma^2 f^2(x^*_ij\beta^*) + \left( m(x^*_ij) - F(x^*_ij\beta^*) \right)^2}{a^* (F(x^*_ij\beta^*))},$$

if the user’s variance function $a^*$ is the (incorrect) variance function $a$ from an exponential family member. In this case, the responses may be overdispersed if there is either correlation ($\sigma^2 \neq 0$) or mean model misspecification, or both.

For $E[\hat{\rho}_{CS}] \approx$

$$\left( \frac{1}{2} sn(n - 1) - p \right)^{-1} \bar{\phi}^{-2} \sum_{i=1}^{s} \sum_{j<k} \text{Cov}[Y_{ij}, Y_{ik}] + \left( m(x^*_ij) - F(x^*_ij\beta^*) \right) \left( m(x^*_ik) - F(x^*_ik\beta^*) \right) \frac{a^* (F(x^*_ij\beta^*))}{a^* (F(x^*_ik\beta^*))},$$

denote the inner sum by $\gamma_{\rho_{CS}}$. Then,

$$E[\hat{\rho}_{CS}] \approx \left( \frac{1}{2} sn(n - 1) - p \right)^{-1} \bar{\phi}^{-2} \sum_{i=1}^{s} \gamma_{\rho_{CS}} = \frac{s\gamma_{\rho_{CS}}}{\bar{\phi}^2 (\frac{1}{2} sn(n - 1) - p)},$$

and

$$\lim_{s \to \infty} E[\hat{\rho}_{CS}] \approx \lim_{s \to \infty} \frac{s\gamma_{\rho_{CS}}}{\bar{\phi}^2 (\frac{1}{2} sn(n - 1) - p)} = \frac{\gamma_{\rho_{CS}}}{\bar{\phi} (n - 1)}.$$
For the simulation models, \( \text{Cov}[Y_{ij}, Y_{ik}] \approx \sigma^2 f(x'_{ij} \beta^*) f(x'_{ik} \beta^*) \), can be used in \( \gamma_{pcs} \), if the exact covariance is not available.

Finally, for \( E[\hat{\rho}_{ARI}] \approx \frac{1}{\phi^2((n-1)s-p)} \sum_{i=1}^{s} \sum_{j=1}^{n-1} \text{Cov}[Y_{ij}, Y_{i,j+1}] + \left( m(x^*_{ij} - F(x'_{ij} \beta^*)) \left( m(x^*_{i,j+1}) - F(x'_{i,j+1} \beta^*) \right) \right) \sqrt{a^* \left( F(x'_{ij} \beta^*) \right) a^* \left( F(x'_{i,j+1} \beta^*) \right)} \),

and denote the inner sum as \( \gamma_{\rho_{ARI}} \). Then,

\[
E[\hat{\rho}_{ARI}] \approx \frac{1}{\phi^2((n-1)s-p)} \sum_{i=1}^{s} \gamma_{\rho_{ARI}} = \frac{s \gamma_{\rho_{ARI}}}{\phi^2(s(n-1)-p)} \xrightarrow{s \to \infty} \frac{n \gamma_{\rho_{ARI}}}{\gamma_{\phi}(n-1)}.
\]

In this case, \( \text{Cov}[Y_{ij}, Y_{i,j+1}] \approx \sigma^2 f(x'_{ij} \beta^*) f(x'_{i,j+1} \beta^*) \), can be used in \( \gamma_{\rho_{ARI}} \).

### C.4 Example Correlation Matrices for the Simulation Models

The true correlation matrices for the four simulation models on pages 108–110 under no misspecification are computed at the values of \( \sigma \) given in Table 9. There are \( n_i = 10 = n \) responses per cluster, one at each design point \( x_i \) (or \( x_k, i, k = 1, \ldots, 10 = n \)), so the matrices are \( 10 \times 10 \). Since the matrices are symmetric it is necessary only to give either the lower or upper triangle. In the following tables, the lower triangle corresponds to the correlation structure for the low setting of \( \sigma \), or low corr/var, and the upper triangle corresponds to the correlation structure for the high setting of \( \sigma \), or high corr/var. The Poisson model structures are exact, to two decimal places.
Table 34: $\mu$-Poisson Low Level and High Level Correlation Structures

<table>
<thead>
<tr>
<th>$x_i \backslash x_k$</th>
<th>0</th>
<th>1/9</th>
<th>2/9</th>
<th>3/9</th>
<th>4/9</th>
<th>5/9</th>
<th>6/9</th>
<th>7/9</th>
<th>8/9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.81</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.81</td>
<td>0.79</td>
<td></td>
</tr>
<tr>
<td>1/9</td>
<td>0.24</td>
<td></td>
<td>0.83</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.83</td>
<td>0.82</td>
<td>0.81</td>
</tr>
<tr>
<td>2/9</td>
<td>0.26</td>
<td>0.28</td>
<td></td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.84</td>
<td>0.83</td>
<td>0.82</td>
</tr>
<tr>
<td>$\sigma = 0.12$</td>
<td>3/9</td>
<td>0.27</td>
<td>0.29</td>
<td>0.30</td>
<td></td>
<td>0.86</td>
<td>0.86</td>
<td>0.85</td>
<td>0.84</td>
<td>0.82</td>
</tr>
<tr>
<td>Low</td>
<td>4/9</td>
<td>0.27</td>
<td>0.29</td>
<td>0.31</td>
<td>0.32</td>
<td></td>
<td>0.86</td>
<td>0.86</td>
<td>0.85</td>
<td>0.84</td>
</tr>
<tr>
<td>Corr.</td>
<td>5/9</td>
<td>0.27</td>
<td>0.29</td>
<td>0.31</td>
<td>0.32</td>
<td>0.32</td>
<td></td>
<td>0.86</td>
<td>0.85</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>6/9</td>
<td>0.27</td>
<td>0.29</td>
<td>0.30</td>
<td>0.31</td>
<td>0.32</td>
<td>0.32</td>
<td></td>
<td>0.85</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>7/9</td>
<td>0.26</td>
<td>0.28</td>
<td>0.29</td>
<td>0.30</td>
<td>0.31</td>
<td>0.31</td>
<td>0.30</td>
<td></td>
<td>0.83</td>
</tr>
<tr>
<td>$\sigma = 0.4$</td>
<td>8/9</td>
<td>0.24</td>
<td>0.26</td>
<td>0.28</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
<td>0.28</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.23</td>
<td>0.24</td>
<td>0.26</td>
<td>0.27</td>
<td>0.27</td>
<td>0.27</td>
<td>0.26</td>
<td>0.24</td>
<td></td>
</tr>
</tbody>
</table>

Table 35: $\mu$-Binomial Low Level and High Level Correlation Structures

<table>
<thead>
<tr>
<th>$x_i \backslash x_k$</th>
<th>0</th>
<th>1/9</th>
<th>2/9</th>
<th>3/9</th>
<th>4/9</th>
<th>5/9</th>
<th>6/9</th>
<th>7/9</th>
<th>8/9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.73</td>
<td>0.76</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.76</td>
<td>0.73</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>1/9</td>
<td>0.21</td>
<td></td>
<td>0.81</td>
<td>0.82</td>
<td>0.83</td>
<td>0.83</td>
<td>0.82</td>
<td>0.81</td>
<td>0.78</td>
<td>0.73</td>
</tr>
<tr>
<td>2/9</td>
<td>0.24</td>
<td>0.29</td>
<td></td>
<td>0.85</td>
<td>0.86</td>
<td>0.86</td>
<td>0.85</td>
<td>0.84</td>
<td>0.81</td>
<td>0.76</td>
</tr>
<tr>
<td>$\sigma = 0.25$</td>
<td>3/9</td>
<td>0.27</td>
<td>0.32</td>
<td>0.36</td>
<td></td>
<td>0.88</td>
<td>0.88</td>
<td>0.87</td>
<td>0.85</td>
<td>0.82</td>
</tr>
<tr>
<td>Low</td>
<td>4/9</td>
<td>0.28</td>
<td>0.33</td>
<td>0.38</td>
<td>0.42</td>
<td></td>
<td>0.89</td>
<td>0.88</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td>Corr.</td>
<td>5/9</td>
<td>0.28</td>
<td>0.33</td>
<td>0.38</td>
<td>0.42</td>
<td>0.43</td>
<td></td>
<td>0.88</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>6/9</td>
<td>0.27</td>
<td>0.32</td>
<td>0.36</td>
<td>0.40</td>
<td>0.42</td>
<td>0.42</td>
<td></td>
<td>0.85</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>7/9</td>
<td>0.24</td>
<td>0.29</td>
<td>0.33</td>
<td>0.36</td>
<td>0.38</td>
<td>0.38</td>
<td>0.36</td>
<td></td>
<td>0.81</td>
</tr>
<tr>
<td>$\sigma = 0.8$</td>
<td>8/9</td>
<td>0.21</td>
<td>0.25</td>
<td>0.29</td>
<td>0.32</td>
<td>0.33</td>
<td>0.33</td>
<td>0.32</td>
<td>0.29</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.18</td>
<td>0.21</td>
<td>0.24</td>
<td>0.27</td>
<td>0.28</td>
<td>0.28</td>
<td>0.27</td>
<td>0.24</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Figure 31 illustrates these correlation structures graphically. The correlation matrices of the four simulation models are all roughly compound symmetric, and differ mostly between the Poisson and binomial model types. The Poisson models have structures that are nearly
Table 36: \( \eta \)-Poisson Low Level and High Level Correlation Structures

<table>
<thead>
<tr>
<th>( x_i \backslash x_k )</th>
<th>0</th>
<th>1/9</th>
<th>2/9</th>
<th>3/9</th>
<th>4/9</th>
<th>5/9</th>
<th>6/9</th>
<th>7/9</th>
<th>8/9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>0.77</td>
<td>0.78</td>
<td>0.79</td>
<td>0.80</td>
<td>0.81</td>
<td>0.82</td>
<td>0.83</td>
<td>0.83</td>
<td>0.84</td>
</tr>
<tr>
<td>1/9</td>
<td>0.16</td>
<td>0.80</td>
<td>0.81</td>
<td>0.82</td>
<td>0.83</td>
<td>0.84</td>
<td>0.84</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>2/9</td>
<td>0.17</td>
<td>0.81</td>
<td>0.82</td>
<td>0.83</td>
<td>0.84</td>
<td>0.85</td>
<td>0.86</td>
<td>0.86</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>( \sigma = 0.15 )</td>
<td>3/9</td>
<td>0.18</td>
<td>0.20</td>
<td>0.20</td>
<td>0.84</td>
<td>0.85</td>
<td>0.86</td>
<td>0.87</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>Low Corr.</td>
<td>4/9</td>
<td>0.19</td>
<td>0.20</td>
<td>0.22</td>
<td>0.23</td>
<td>0.87</td>
<td>0.87</td>
<td>0.88</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td>( \sigma = 0.55 )</td>
<td>5/9</td>
<td>0.20</td>
<td>0.22</td>
<td>0.25</td>
<td>0.26</td>
<td>0.88</td>
<td>0.89</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>6/9</td>
<td>0.21</td>
<td>0.23</td>
<td>0.25</td>
<td>0.28</td>
<td>0.30</td>
<td>0.90</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>7/9</td>
<td>0.23</td>
<td>0.24</td>
<td>0.26</td>
<td>0.30</td>
<td>0.32</td>
<td>0.91</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>8/9</td>
<td>0.24</td>
<td>0.26</td>
<td>0.29</td>
<td>0.31</td>
<td>0.33</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>1</td>
<td>0.25</td>
<td>0.27</td>
<td>0.29</td>
<td>0.31</td>
<td>0.33</td>
<td>0.37</td>
<td>0.39</td>
<td>0.41</td>
<td>0.41</td>
<td>0.41</td>
</tr>
</tbody>
</table>

Table 37: \( \eta \)-Binomial Low Level and High Level Correlation Structures

<table>
<thead>
<tr>
<th>( x_i \backslash x_k )</th>
<th>0</th>
<th>1/9</th>
<th>2/9</th>
<th>3/9</th>
<th>4/9</th>
<th>5/9</th>
<th>6/9</th>
<th>7/9</th>
<th>8/9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>0.71</td>
<td>0.74</td>
<td>0.75</td>
<td>0.76</td>
<td>0.76</td>
<td>0.75</td>
<td>0.74</td>
<td>0.71</td>
<td>0.66</td>
</tr>
<tr>
<td>1/9</td>
<td>0.21</td>
<td>0.79</td>
<td>0.80</td>
<td>0.81</td>
<td>0.81</td>
<td>0.80</td>
<td>0.79</td>
<td>0.75</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>2/9</td>
<td>0.24</td>
<td>0.29</td>
<td>0.84</td>
<td>0.85</td>
<td>0.85</td>
<td>0.84</td>
<td>0.82</td>
<td>0.79</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>( \sigma = 0.25 )</td>
<td>3/9</td>
<td>0.27</td>
<td>0.32</td>
<td>0.36</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.84</td>
<td>0.80</td>
<td>0.75</td>
</tr>
<tr>
<td>Low Corr.</td>
<td>4/9</td>
<td>0.28</td>
<td>0.33</td>
<td>0.38</td>
<td>0.42</td>
<td>0.87</td>
<td>0.86</td>
<td>0.85</td>
<td>0.81</td>
<td>0.76</td>
</tr>
<tr>
<td>( \sigma = 0.75 )</td>
<td>5/9</td>
<td>0.28</td>
<td>0.33</td>
<td>0.38</td>
<td>0.42</td>
<td>0.43</td>
<td>0.86</td>
<td>0.85</td>
<td>0.81</td>
<td>0.76</td>
</tr>
<tr>
<td>6/9</td>
<td>0.27</td>
<td>0.32</td>
<td>0.36</td>
<td>0.40</td>
<td>0.42</td>
<td>0.42</td>
<td>0.84</td>
<td>0.80</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>7/9</td>
<td>0.24</td>
<td>0.29</td>
<td>0.33</td>
<td>0.36</td>
<td>0.38</td>
<td>0.38</td>
<td>0.36</td>
<td>0.79</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>8/9</td>
<td>0.21</td>
<td>0.25</td>
<td>0.29</td>
<td>0.32</td>
<td>0.33</td>
<td>0.33</td>
<td>0.32</td>
<td>0.29</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>1</td>
<td>0.18</td>
<td>0.21</td>
<td>0.24</td>
<td>0.27</td>
<td>0.28</td>
<td>0.28</td>
<td>0.27</td>
<td>0.24</td>
<td>0.21</td>
<td>0.21</td>
</tr>
</tbody>
</table>

compound symmetric, while the binomial structure is more complex. The binomial structure can be described as a function of the distance between the prediction locations and a function of how close pairs of observations are to the center of the \( x \) space.
Figure 31: Correlation Structures for the Simulation Models

$\mu$-Scale Poisson

$\eta$-Scale Poisson

$\mu$-Scale Binomial

$\eta$-Scale Binomial
D Appendix for Chapter 8

D.1 Consistency of MOM Estimators

The scale, CS, and AR(1) correlation parameter MOM estimators are shown to be $\sqrt{s}$-consistent under the condition that the design points remain fixed for all clusters (see pg. 171).

Note from Theorem 2 of Liang and Zeger (1986) that $\sqrt{s}$-consistency of the scale MOM estimator is required only when $\beta_{GEE}$ is known, and $\sqrt{s}$-consistency of the correlation parameter estimators is required only when $\phi$ and $\beta_{GEE}$ are known. This greatly simplifies the problem of proving consistency because the true means, which are fixed, can be used in place of the estimated ones.

Consistency of $\hat{\phi}^2$

The scale estimator using the true means is given by

$$\hat{\phi}^2 = \frac{1}{sn_i - p} \sum_{i=1}^{s} U_i,$$

where

$$U_i = \sum_{j=1}^{n_i} \frac{(Y_{ij} - F(x'_{ij}\beta_{GEE}))^2}{a^* (F(x'_{ij}\beta_{GEE}))},$$

and $a^*$ is the user’s variance function. Suppose that the design points are fixed and remain the same for every cluster as $s \to \infty$. Then $n_i = n$, $i = 1, \ldots, s$ and the cluster response vectors, $Y_i$, $i = 1, \ldots, s$ are iid for every $s$. This implies that the $U_i$, $i = 1, \ldots, s$ are iid. Assuming that $Var[U_1] < \infty$, the Lindeberg-Lévy central limit theorem (CLT) implies that

$$\sqrt{s} \left( \frac{1}{s} \sum_{i=1}^{s} U_i - E[U_1] \right) \xrightarrow{d} N(0, Var[U_1]).$$

Treating $1/n$ as a degenerate random variable, Slutsky’s theorem yields

$$\sqrt{s} \left( \frac{1}{sn} \sum_{i=1}^{n} U_i - \frac{1}{n} E[U_1] \right) \xrightarrow{d} N(0, n^{-2} Var[U_1]),$$

or

$$\sqrt{s} \left( \frac{sn}{sn - p} \frac{1}{sn} \sum_{i=1}^{s} U_i - \frac{1}{n} E[U_1] \right) \xrightarrow{d} N(0, n^{-2} Var[U_1]),$$

211
since $\frac{sn}{sn-p} \sum_{i=1}^{s} U_i$ and $\frac{1}{sn} \sum_{i=1}^{s} U_i$ are asymptotically equivalent as $s \to \infty$. Convergence in distribution implies a sequence is bounded in probability, so
\[
\sqrt{s} \left( \frac{1}{sn-p} \sum_{i=1}^{s} U_i - \frac{1}{n} E[U_1] \right) = O_p(1)
\]
or equivalently,
\[
\hat{\phi}^2 - \tilde{\phi}^2 = O_p(s^{-1/2}),
\]
where $\tilde{\phi}^2 = E[U_1]/n$. Thus, $\hat{\phi}$ is $\sqrt{s}$-consistent for $\tilde{\phi}$. The quantity $E[U_1]/n$ is derived in Appendix C.3 (denoted as $\gamma_\phi/n$). Note that the condition $\text{Var}[U_1] < \infty$ cannot be verified in general under the assumptions of GEE, because it requires the moments of $Y_i$ higher than the 2nd order to be known.

\underline{Consistency of the Compound Symmetry MOM estimator}

The $\sqrt{s}$-consistency of the CS MOM estimator, using the true means and $\tilde{\phi}^2$,
\[
\hat{\rho}_{CS} = \left( \frac{1}{2} sn(n-1) - p \right)^{-1} \sum_{i=1}^{s} Q_i,
\]
where
\[
Q_i = \sum_{j<k}^{n} \frac{(Y_{ij} - F(x_{ij}' \beta_{GEE}))(Y_{ik} - F(x_{ik}' \beta_{GEE}))}{\sqrt{\tilde{\phi}^2 a^*(F(x_{ij}' \beta_{GEE}))} \sqrt{\tilde{\phi}^2 a^*(F(x_{ik}' \beta_{GEE}))}},
\]
is based on an argument similar to the one above. Assuming that $\text{Var}[Q_1] < \infty$, the Lindeberg-Lèvy central limit theorem (CLT) implies that
\[
\sqrt{s} \left( \frac{1}{s} \sum_{i=1}^{s} Q_i - E[Q_1] \right) \overset{d}{\to} N(0, \text{Var}[U_1]).
\]
Treating $2/n(n-1)$ as a degenerate random variable, Slutsky’s theorem gives
\[
\sqrt{s} \left( \frac{2}{sn(n-1)} \sum_{i=1}^{s} Q_i - \frac{2}{n(n-1)} E[Q_1] \right) \overset{d}{\to} N(0, 4(n(n-1))^{-2} \text{Var}[Q_1]),
\]
or
\[
\sqrt{s} \left( \frac{2}{sn(n-1)-p} \sum_{i=1}^{s} Q_i - \frac{2}{n(n-1)} E[Q_1] \right) \overset{d}{\to} N(0, 4(n(n-1))^{-2} \text{Var}[U_1]),
\]
since $\frac{2}{sn(n-1)-p} \sum_{i=1}^{s} Q_i$ and $\frac{2}{sn(n-1)} \sum_{i=1}^{s} Q_i$ are asymptotically equivalent as $s \to \infty$. Then
\[
\hat{\rho}_{CS} - \tilde{\rho}_{CS} = O_p(s^{-1/2}),
\]
212
where $\hat{\rho}_{CS} = \frac{2}{n(n-1)} E[Q_1]$. Thus, $\hat{\rho}_{CS}$ is $\sqrt{s}$-consistent for $\tilde{\rho}_{CS}$, which is derived in Appendix C.3, where it is shown to be $\gamma_{\rho_{CS}}/(0.5\gamma_{\phi}(n - 1))$.

**Consistency of the Autoregressive(1) MOM estimator**

Showing the $\sqrt{s}$-consistency of the AR(1) MOM estimator involves nearly identical steps as given above for $\hat{\rho}_{AR(1)}$. The result is

$$\hat{\rho}_{AR} - \tilde{\rho}_{AR} = O_p(s^{-1/2}),$$

where $\tilde{\rho}_{AR} = \frac{1}{n-1} E[Q_1]$ and

$$Q_1 = \sum_{j=1}^{n-1} \frac{(Y_{ij} - F(x'_{ij}\beta_{GEE}))(Y_{i,j+1} - F(x'_{i,j+1}\beta_{GEE}))}{\sqrt{\phi^2 a^*\left(F(x'_{ij}\beta_{GEE})\right)}} \sqrt{\phi^2 a^*\left(F(x'_{i,j+1}\beta_{GEE})\right)}.$$

Thus, $\hat{\rho}_{AR}$ is $\sqrt{s}$-consistent for $\tilde{\rho}_{AR}$. $\tilde{\rho}_{AR}$ is shown in Appendix C.3 to be $n\gamma_{\rho_{AR}}/(\gamma_{\phi}(n - 1))$. 

213
References


214


217


218


Tukey, J. (1977), *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.


Vita

Seth Kenneth Stanley Clark, son of Wilson S. and Susan A. Clark, was born in Vermont in February of 1975. He was raised on a farm where his parents taught him good morals and the value of hard work. Clark studied mathematics at Castleton State College in Vermont from 1993–1997. He graduated summa cum laude, the first in his class, with a 4.0 cumulative GPA. While at Castleton, Clark won numerous awards and scholarships including the Barry Goldwater national scholarship in mathematics, science and engineering. He went on to study statistics at Virginia Tech and obtained his master’s degree in 1999 with the support of a Cunningham Fellowship. After his first year of study at Virginia Tech, Clark won the Boyd Harshbarger award for superior scholarship during the first year of graduate studies in statistics in the spring of 1998. Clark completed his Ph.D. degree, a life-long goal, in April of 2002. He has accepted a biostatistics position in the Biometrics Unit at Merck & Co. in West Point, PA. Seth is married to Sherry Sprouse and has 1.5 children: John H. and Elizabeth M. D., on the way.