CHAPTER 3

Estimation Techniques

3.1 The analysis of count data

Count data describe events that take nonnegative integer values for each observation. Many economic data come in the form of nonnegative integers; among those that have been examined in the past are the number of grocery purchases per period (Gilbert, 1979), the number of patents applied for by a particular firm during a year (Hall et al., 1983; Hausman et al., 1984), the number of ship damage incidents (Lawless, 1987; McCullagh and Nelder, 1989; Winkelmann and Zimmermann, 1991), the number of trips made to a vacation area (Hellerstein, 1991; Englin and Shonkwiler, 1995), and the number of births by a woman and the number of consultations with specialist doctors (Winkelmann, 1995). In addition, count data models have a variety of applications in political science, where, for example, the number of actions that nations or political groups take with respect to each other are recorded (Azar and Sloan, 1975).

Count data usually have a non-negligible probability of zero, which makes the use of log-linear relationships problematic. One possibility for dealing with the impossibility of taking a logarithm of zero is to eliminate all groups of data that include observations of zero, but this requires that the number of these groups is small compared to the whole sample. Another possibility is to add a small value to all zero observations so that the logarithms can be taken. A third possibility is to set all zeros equal to one, and to add a dummy variable to implicitly allow a value different from one. However, none of these devices are entirely satisfactory because a simple OLS analysis does not constrain the expected number of events to be nonnegative, as it obviously must be with count data. As the mean of the censored normal distribution differs from the mean of the uncensored normal distribution, the analysis will suffer from a sample selection bias, and is therefore unlikely to yield correct estimates. King (1988) reviews several of the possibilities for dealing with problems where observations are equal to zero, and concludes that OLS estimates of count data are inefficient with inconsistent standard errors, and that logged OLS estimates on event count data have the same

---

98 See King (1988) for further examples of the use of count data in political science.


100 Pakes and Grilliches (1980).
problems and are also biased and inconsistent.\textsuperscript{101}

The problem of the censored normal distribution can be handled by using a Tobit model, which assumes that the distribution of the data is censored normal, that is, a mixture of a discrete and a continuous part. The normal distribution is applied to all observations which are larger than zero, and the full probability of the negative part of the Normal distribution is assigned to the censoring point at zero.\textsuperscript{102} Yet the Tobit model still assumes a normal distribution for all positive observations, which places a non-negative probability on non-integers. While the error of approximating discrete data with a continuous distribution becomes negligible if the observations have a large range, the approximation is likely to yield biased results if most of the observations fall into a small range.

The Poisson distribution has been widely used to avoid the approximation of count data by a continuous distribution, and to ensure nonnegative predictions. Unfortunately the basic Poisson distribution makes assumptions about the data generating process that do not hold in all cases. In many cases the data is described better by a more generalized Poisson process, or by the negative binomial distribution (see for example Hausman \textit{et al.}, 1984; Cameron and Trivedi, 1986; King, 1989). In the literature, the negative binomial distribution is usually derived as a mixed Poisson distribution, and its additional second parameter is chosen as a straightforward relaxation of the assumption that the mean must equal the variance. However, the negative binomial distribution can be derived in other ways that make an alternative parameterization more obvious. It is acknowledged that the usual parameterization is somewhat arbitrary.\textsuperscript{103} An alternative parameterization seems to yield better results, which makes it worthwhile to explore this other case.

Sections 3.2 and 3.3.1 derive and motivate the Poisson distribution and the negative binomial distribution as they are used in the literature. Section 3.3.2 gives an alternative derivation of the negative binomial distribution, and Section 3.3.4 motivates a different parameterization, which will be used in Chapter 5 for the analysis of the data on building permits in Pennsylvania.

It is convenient to assume that the true distribution can be approximated by a Poisson or a negative binomial distribution, because the estimation of such models with the maximum likelihood technique is straightforward. The estimation procedure with other discrete distribu-

\textsuperscript{101} King (1988), p. 859.


\textsuperscript{103} Cameron and Trivedi (1986), p. 33.
tions is either very cumbersome or even impossible for the cases where the distribution is not available in a closed form. Yet if the distributional assumptions are wrong, then the estimated coefficients will be biased. Markov Chain Monte Carlo (MCMC) methods, which use Monte Carlo integration techniques together with Markov Chains to gain information about the sample distribution and the optimal coefficients, are a convenient tool to examine further distributions. The most frequently used MCMC computational method is the Gibbs-Sampler, because it is easy to implement and applicable to a wide variety of problems. Section 3.4 describes MCMC methods, and motivates their application to count data.

3.2 Poisson Models

The use of the Poisson distribution can be motivated as follows: assume that \( Y_i \) denotes the number of occurrences of a certain event for an individual \( i \) within a given interval of time, where \( Y_i = 0,1,2,\ldots \). As any realization \( y_i \) is observed only at the end of each interval, it is necessary to make certain assumptions about the unobserved process that generates these observable counts. Under the following 4 assumptions (see King, 1988)

1. only one event can occur at every moment in time,
2. zero events have occurred at the start of each period,
3. the length of each observation period is identical,
4. the probability that an event occurs at any point in time is constant within the observation period and independent of all previous events during that period,

the number of events within each observation period is Poisson distributed with the univariate density function

\[
P(Y_i = y_i | \lambda_i) = \frac{e^{-\lambda_i} \cdot \lambda_i^{y_i}}{y_i!} \quad \text{for } \lambda_i > 0 \text{ and } y_i = 0,1,\ldots
\] (3.1)

While the first three assumptions can be relaxed if more general formulations of the Poisson distribution are considered, the fourth assumption of independence is crucial if the model is to be estimated with a maximum likelihood method. If the data are not independent, then the joint distribution of all observations will not be the product of the unconditional distributions of all single observations. Most methods determine the likelihood by summing up the logarithms of the likelihoods of all observations, which is not correct if the observa-

---

104 Cameron and Trivedi (1986), pp. 33-34.

105 For alternative derivations of the Poisson distribution see Redheffer (1953) and Walsh (1955).

The first two moments of the Poisson distribution are equal, and are given by \( \text{E}[Y_i] = \text{VAR}[Y_i] = \lambda_i \). The Poisson regression model is therefore specified by \( \lambda_i = \lambda(x_i, \beta) \), where \( x_i \) is a vector of exogenous variables and \( \beta \) is the parameter vector. Note that \( \lambda \) is a deterministic function, and the randomness in the regression is created only by the Poisson specification. To avoid negative values of \( \lambda_i \), the functional form \( \lambda_i = \exp(x_i \beta) \) is frequently used. If the data are fairly homogenous, this functional form does not cause difficulties, but if some observations are large outliers which cannot be excluded, then \( \lambda_i \) becomes very large and the log-likelihood of this observation becomes extremely small. This can lead to imprecise summation in the estimation routine. Even though some authors acknowledge that this log-linear relationship is only one of many possible specifications,\(^{107}\) almost no one used another functional form in the literature so far.\(^{108}\)

The use of the Poisson distribution has several advantages over an approximation of the data by a continuous distribution (see Hausman et al., 1984): first, the specification has no difficulty dealing with observations \( y_i \) that are equal to zero, because it incorporates directly the integer nature of the data. Especially if the observed values of the events are small, a continuous approximation will lead to incorrect results. Second, the single observation periods can be aggregated as long as the assumption of independece is not violated, because the aggregate data will also be Poisson distributed. Finally, the estimation of a model proceeds straightforwardly, because there is a globally concave loglikelihood function, which leads to fast convergence to a unique maximum.\(^{109}\) This loglikelihood function is given by

\[
\ln L(\beta | y) \propto \sum_{i=1}^{n} (-\lambda_i + y_i \cdot \ln \lambda_i - \ln y_i !) \tag{3.2}
\]

where \( \lambda_i \) will be replaced by the chosen functional form.

This use of maximum likelihood techniques with the Poisson distribution requires two strong assumptions that may not be justified for many data sets. First, the assumption of independence might be violated, as many events do not occur independently over time, but are rather conditioned by previous outcomes. Heckman and Borjas (1980) have modeled this ‘occurrence dependence’ in the context of unemployment. Alternatively, it is also possible

---


\(^{108}\) An exception is Gilbert (1979) who uses a simple linear relationship.

\(^{109}\) Gourieroux et al. (1984), p. 703.
that events occur in ‘spells’, and while the spells themselves follow one probability distribution, the events within these spells occur according to a different distribution.\textsuperscript{110} Second, the assumption that the data are generated by a process with a mean equal to the variance may not be correct; many datasets are characterized by a variance that is larger than the mean (overdispersion).\textsuperscript{111} Only few authors have found a variance that is smaller than the mean (underdispersion) in their datasets.\textsuperscript{112}

The basic Poisson model can be generalized by relaxing the assumption that $\lambda_i$ is a deterministic function, and by replacing it with the assumption that $\lambda_i$ is generated by a stochastic function $\lambda_i = \lambda(x_i, \beta, \xi)$, where $\xi$ is an error term. $x_i$ and $\beta$ are still assumed to be deterministic, so that $\lambda_i$ follows the same distribution as $\xi$. The resulting mixed distribution is described by $E[f(Y|\lambda_i)]$, that is, the expectation taken with respect to the distribution of $\lambda_i$.\textsuperscript{113} If $f(\lambda_i)$ is the density function of the now random parameter $\lambda_i$, the distribution of each $Y_i$ is obtained by integrating over $\lambda_i$, which results in

$$P(Y_i = y_i) = \int_0^\infty P(Y_i = y_i | \lambda_i) f(\lambda_i) \, d\lambda_i. \quad (3.3)$$

Johnson and Kotz (1992) derive several specific mixed Poisson distributions, but it should be noted that it is not possible to obtain a closed form solution of the integral for all parametric forms of $f(\lambda_i)$. In addition, for many mixed Poisson distributions it might be difficult to motivate a suitable parameterization of $f(\lambda_i)$, as the expression of the first moment can be rather complex.\textsuperscript{114} However, if $f(\lambda_i)$ is assumed to follow a gamma distribution, then the

\textsuperscript{110} Cresswell and Frogatt (1963) and Xekalaki (1983).

\textsuperscript{111} Gourieroux et al. (1984), p. 703.

\textsuperscript{112} See for example King (1989) and Winkelmann (1995).

\textsuperscript{113} See Johnson and Kotz (1992), pp. 305-312.

\textsuperscript{114} See Johnson and Kotz (1992, pp. 326-335) and the analysis by Gourieroux \textit{et al.} (1984). For example, the mixture of the Poisson distribution and a generalized inverse-Gaussian distribution yields the Sichel distribution (Sichel, 1971). The mean $\mu_i$ and the variance $\mu_i$ of the Sichel distribution are determined by

$$\mu_i = \frac{\xi K_{\gamma}^{\alpha}(\omega)}{K_{\gamma}^{\alpha}(\omega)} , \quad \mu_i = \mu_i \left( 1 + \frac{\xi K_{\gamma}^{\alpha}(\omega)}{K_{\gamma}^{\alpha}(\omega)} - \mu_i \right)$$

where $\xi$, $\omega$ and $\gamma$ are the three parameters of the distribution, and $K_{\gamma}$ is a modified Bessel function (see Willmot, 1986). Any attempt to link the independent variables to the dependent variable is difficult to motivate, and the optimization, which involves the repeated calculation of the modified Bessel function, is still very time consuming.
mixed Poisson distribution in equation 3.3 is the negative binomial distribution, which has a straightforward interpretation, and is discussed below.

3.3 Negative Binomial Models

3.3.1 Standard derivation and specification

Greenwood and Yule (1920) are credited for first deriving and applying the negative binomial distribution in the literature, even though some special forms of this distribution were already discussed by Pascal (1679). Following the work of Greenwood and Yule (1920), Hausman et al. (1984) derived the negative binomial distribution from a Poisson distribution that is mixed with a gamma distribution; most references in the literature refer to this publication to justify their now commonly used parameterization. The derivation is as follows:

Let \( Y_i \) follow a Poisson distribution with the parameter \( \lambda_i \). Assume that this parameter follows a two-parametric gamma distribution \( f(\lambda_i; \alpha_i, \phi_i) \), whose density function is given by

\[
f(\lambda_i; \alpha_i, \phi_i) = \frac{\lambda_i^{\alpha_i - 1} e^{-\frac{\lambda_i}{\phi_i}}}{\phi_i^{\alpha_i} \Gamma(\alpha_i)}
\]  

(3.4)

For the purpose of finding an interpretation of the parameters of the negative binomial distribution, it is common to redefine the second parameter as \( \phi_i = \mu_i / \alpha_i \),\(^{115}\) which results in the new density function

\[
f(\lambda_i) = \frac{\lambda_i^{\alpha_i - 1} e^{-\frac{\lambda_i}{\mu_i}}}{\mu_i^{\alpha_i} \Gamma(\alpha_i)} \mu_i^{\alpha_i} \Gamma(\alpha_i)
\]  

(3.5)

with a mean equal to \( \mu_i \) and \( \text{Var}[\lambda_i] = \mu_i^{\alpha_i} / \alpha_i \). Using equation (3.3), the marginal density

\[^{115}\text{This is the so-called ‘index-parameterization’. See Cameron and Trivedi (1986), p. 32, and Lawless (1987), p. 210.}\]
\[ P(Y_{i}=y_{i}) \] can now be calculated as\(^{116}\)

\[
P(Y_{i} = y_{i}) = \int_{0}^{\infty} \frac{\lambda_{i}^{y_{i}} e^{-\lambda_{i}}}{y_{i}!} \frac{\alpha_{i}^{-1}}{\mu_{i} \Gamma(\alpha_{i})} e^{-\frac{\alpha_{i}}{\mu_{i}}} d\lambda_{i}
\]

\[
= \frac{\alpha_{i}^{y_{i}}}{y_{i}! \mu_{i}^{\alpha_{i}} \Gamma(\alpha_{i})} \int_{0}^{\infty} \lambda_{i}^{y_{i}+\alpha_{i}-1} e^{-\lambda_{i}} d\lambda_{i}
\]

\[
= \left( \frac{\alpha_{i} + y_{i} - 1}{\alpha_{i} - 1} \right) \left( \frac{\mu_{i}}{\mu_{i} + \alpha_{i}} \right)^{y_{i}} \left( \frac{\alpha_{i}}{\mu_{i} + \alpha_{i}} \right)^{\alpha_{i}}.
\] \hspace{1cm} (3.6)

which implies that \( Y_{i} \) has a negatively binomial distribution with the parameters \( \alpha_{i} > 0 \) and \( \mu_{i} > 0 \).

As a result of the index-parameterization of the gamma distribution, the mean of the negative binomial distribution is equal to the parameter \( \mu_{i} \), and the variance is given by \( \mu_{i} + \mu_{i}^{2}/\alpha_{i} \). The index parameter \( \alpha_{i} \) determines the degree of dispersion, that is, the degree by which the variance differs from the mean. For \( \alpha_{i} \to \infty \) the distribution converges to the Poisson distribution which has a ‘null dispersion’ (the variance equals the mean). As both parameters are positive, the variance of the negative binomial distribution is larger than the mean and the distribution can be used to model data with overdispersion.\(^{117}\)

As \( \alpha_{i} \) can be any rational number, it is necessary to calculate the factorial in the binomial coefficient by using the relationship between factorials and the gamma function \( \Gamma(x-1)=x! \). The probability \( P(Y_{i} = y_{i}) \) can then be calculated as

\[
P(Y_{i} = y_{i}) = \frac{\Gamma(y_{i}+\alpha_{i})}{y_{i}! \Gamma(\alpha_{i})} \left( \frac{\mu_{i}}{\mu_{i} + \alpha_{i}} \right)^{y_{i}} \left( \frac{\alpha_{i}}{\mu_{i} + \alpha_{i}} \right)^{\alpha_{i}}.
\] \hspace{1cm} (3.7)

There are various possibilities for specifying the exact form of heteroskedasticity with

\(^{116}\) See Johnson and Kotz (1992), p. 204, and Cameron and Trivedi (1986), p. 33. For the derivation it is helpful to note that the complete gamma function \( \Gamma(x) \) is defined as \( \Gamma(x) = \int t^{x-1} e^{t} dt \).

\(^{117}\) The Poisson and the gamma distribution have a natural connection; if the Poisson distribution describes the distribution of arrivals at a certain point, then the time that passes between these arrivals is gamma distributed. See Press et al. (1995), p. 293.
the parameters $\alpha_i$ and $\gamma_i$. Cameron and Trivedi (1986) suggest\(^{118}\) the use of $\alpha_i = (1 / \delta_i) \mu_i^k$, with $\delta_i > 0$ and an arbitrary constant $k$, which results in $\text{Var}[Y_i] = \mu_i + \delta_i \cdot \mu_i^{2-k}$. In their paper they test the appropriateness of two different values of $k$. Setting $k=1$ results in $\text{Var}[Y_i] = E[Y_i] + \delta_i \cdot E[Y_i]^2$, which represents a constant ratio of the variance to the mean. Setting $k=0$ yields $\text{Var}[Y_i] = E[Y_i] + \delta_i \cdot E[Y_i]^2$, which is a model with a ratio of the variance to the mean that is linear in the mean. Lee (1986), Cameron and Trivedi (1990) and Winkelmann (1995) describe tests for generalized linear models that can help to determine which relationship between the variance and the mean fits the data best. Both of the above specifications have been used in the literature; McCullagh and Nelder (1983) and Hausman et al. (1984) used the first specification, while the second specification seems to fit most data better, and is now used by most authors as the variance-mean relationship. Clearly there are more possible specifications, and it is not clear why Cameron and Trivedi’s second specification should always yield an adequate fit. Before Cameron and Trivedi proposed the two specifications, other relationships had been used, for example Armitage (1957), Finney (1976) and McCullagh and Nelder (1983) used $\text{Var}[Y_i] = \delta_i \cdot E[Y_i]^k$ with various values of $k$. A different approach to determine the best relationship between the variance and the mean is described and motivated in Section 3.4.

The most widely used estimation technique to estimate the negative binomial model is the maximum likelihood method.\(^{119}\) If $n$ is the number of independent observations, then the likelihood function of the negative binomial distribution can be determined according to

$$L(\beta, \alpha, \mu, y) \propto \prod_{i=1}^{n} \frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1) \Gamma(\alpha)} \left( \frac{\mu_i}{\mu_i + \alpha} \right)^{y_i} \left( \frac{\alpha}{\mu_i + \alpha} \right)^{\alpha_i}. \quad (3.8)$$

For any nonnegative integer $y_i$ and any $\alpha_i > 0$, it is possible to write $\Gamma(y_i + \alpha_i) / \Gamma(\alpha_i) = \alpha_i (\alpha_i + 1) \ldots (\alpha_i + y_i - 1)$, so that the loglikelihood function can be written without using the gamma function as\(^{120}\)

\(^{118}\) At least this is what they seem to suggest, even though they do not use the index $i$ on $\delta$. The index is excluded in almost all publications since then, with the exception of Englin and Shonkwiler (1995), who are more specific and write correctly $\delta_i = \delta / q_i$. According to this formulation, the parameter $\delta$ is estimated, and then transformed into $\delta_i$. Cameron and Trivedi point out that this is only one of many parameterizations, which they chose because it fit their data best. But this parameterization seems to be the exclusive choice in all published estimations ever since (if the authors were aware of it!), including the estimations performed by LIMDEP (where the parameterization is not mentioned in the manual).

\(^{119}\) Other techniques are discussed by Gourieroux et al. (1984), Cameron and Trivedi (1986) and Lawless (1987).

\[ \ln L(\beta_i|\alpha_i, \mu_i, Y_i) = \sum_{i=1}^{n} \left( \sum_{j=0}^{y_i} \ln(\alpha_j) + \ln(y_i!) + \alpha_j \ln(\mu_j + 1) - y_i \ln(\mu_j + 1) \right) , \] (3.9)

which decreases the calculation time, and makes it possible to calculate analytical first derivatives of the loglikelihood function.

### 3.3.2 Alternative derivations of the negative binomial distribution

As an alternative to the negative binomial distribution, which can only be used to handle data with overdispersion, more general count models have been developed. The derivation of the negative binomial distribution as a mixed Poisson and gamma distribution, which is used exclusively in the literature, is useful, because it describes the negative binomial as an overdispersed Poisson distribution; this makes it possible to test for the degree of dispersion, and to decide between the two possible distributions.

But this approach also seems to point towards the two specifications between the variance and the mean that were proposed by Cameron and Trivedi (1986), and does not suggest alternative parameterizations. Boswell and Patil (1970) have combined 15 different derivations of the negative binomial distributions, and identified several underlying chance mechanisms without any apparent relationship. Following the work of Katz (1965), King (1989) developed a general count model which allows for over-, under- and null dispersion, and does not rely on a combination of the Poisson and the Gamma distribution. While the general count model is very useful if the degree of dispersion is unknown, it does not provide any advantages if the data is known to be overdispersed. The following derivation therefore does not follow King’s model, but is used to motivate a different approach to the negative binomial distribution.\(^{121}\)

If one assumes that \(Y_i\) has a distribution from the general binomial family with the mass function

\[ P(Y_i) = \frac{1}{Y_i!} (1-\alpha_i)(1-2\alpha_i)...(1-(Y_i-1)\alpha_i) \cdot \lambda_i^{Y_i} \cdot (1+\lambda_i\alpha_i)^{Y_i-i} \] (3.10)

\(^{121}\)The derivation was pointed out to me by Dr. George Terrell, Statistics Department, Virginia Polytechnic Institute and State University.
then depending on the sign of $\alpha_i$ (the ‘negative’ dispersion parameter) and on appropriate parameterization of $\alpha_i$ and $\lambda_i$, this mass function can coincide with the binomial, the Poisson, and the negative binomial distribution. The first two moments are given by $E[Y_i] = \lambda_i/(1+\lambda_i\alpha_i)$ and $\text{Var}[Y_i] = \lambda_i/(1+\lambda_i\alpha_i)^2$.

If $\alpha_i$ is positive, $\alpha_i = 1/n_i$, and $\lambda_i = n_ip_i/(1-p_i)$ then 3.10 becomes

$$P(Y_i) = \frac{1}{Y_i!} \frac{n_i^{-1} n_i^{-2} \cdots n_i^{-Y_i+1}}{n_i} \frac{p_i^{Y_i}}{(1-p_i)^{Y_i}} (1-p_i)^{n_j-Y_i}$$

(3.11)

which is the binomial distribution for $n_i = 0, 1, 2, 3...$ with $E[Y_i] = n_ip_i > \text{Var}[Y_i] = n_ip_i(1-p_i)$ (underdispersion).

If $\alpha_i$ is equal to zero, equation 3.10 converges to the Poisson distribution as

$$P(Y_i) = \frac{1}{Y_i!} \lambda_i^{Y_i} (\lim_{\alpha_i \to 0} [1+\lambda_i\alpha_i]^{-\alpha_i}) = \frac{\lambda_i^{Y_i}e^{-\lambda_i}}{Y_i!}$$

(3.12)

with $E[Y_i] = \text{Var}[Y_i] = \lambda_i$ (null dispersion).

If $\alpha_i$ is negative, $\alpha_i = -1/k_i$, and $\lambda_i = k_iq_i$, then equation 3.10 describes the negative binomial distribution because

$$P(Y_i) = \frac{1}{Y_i!} \frac{k_i+1 k_i+2 \cdots k_i+Y_i-1}{k_i} \frac{k_i^{Y_i} q_i^{Y_i}}{(1-q_i)^{Y_i}} (1-q_i)^{k_i}$$

(3.13)

$$= \left(\frac{k_i+Y_i-1}{Y_i}\right) q_i^{Y_i} (1-q_i)^{k_i},$$

with $E[Y_i] = k_iq_i/(1-q_i) < \text{Var}[Y_i] = k_iq_i/(1-q_i)^2$ (overdispersion). The variance-mean relationship is therefore $\text{Var}[Y_i] = E[Y_i] \cdot (1-q_i)^{-1}$. This expression is equivalent to equation 3.6 if $q_i = \mu_i/(\mu_i + \alpha_i)$. 

- 49 -
3.3.3 Alternative specification of the two parameters

To find an alternative to the specification of the relationship between the independent variables $X_i$ and the two parameters $q_i$ and $k_i$ as described in Section 3.3.1, it is informative to examine the behavior of the coefficient of variation $CV_i = \sigma_i / \mu_i$, where $\sigma_i$ is the standard deviation and $\mu_i$ is the mean, for various values of $\mu_i$. If the mean $\mu_i$ is an integer, the smallest possible variation of all integer observations is zero (all observations are equal), so that $CV_i$ is equal to zero as well. If the mean is not an integer, then the minimum variance is positive, and reaches its maximum of $\frac{1}{4}$ at half the distance between two integers (half of the observations are equal to the lower integer, the other half are equal to the higher integer). The coefficient of variation at this point becomes $1/(2\mu_i)$, that is, it decreases with an increasing mean. This minimum relationship between $\mu_i$ and $CV_i$ is shown as the curve $AA$ in Figure 3.1; the true relationship between $CV_i$ and the mean must be either on or above this curve. Notice that the slope of the function $CV$ at the maximum variance is $-1/(2\mu_i^2)$, that is, the point midway between two integers is to the right of the maximum of each of the ‘humps’ in the figure.

Observed data are typically noisy, which makes it sensible to bound the coefficient of variation away from 0. Instead of using $AA$, it will be more reasonable to describe $CV_{min}$ as the strictly decreasing curve $BB$ that is given by the function $f(\mu_i) = 1/(2\mu_i)$ for $\mu_i \geq \frac{1}{2}$. This function is tangent to the minimum relationship at the points midway between two integers. However, for values of $\mu_i$ that are smaller than $\frac{1}{2}$, this function becomes very large, and it is unlikely that many examined data sets will have very small observations together with large observations that would cause $CV_{min}$ to become very large. As $\mu_i \to 0$, it is more likely that most data sets will consist of few observations other than 0 and 1, although this is an empirical question. Therefore, for the interval $[0,\frac{1}{2}]$, it may be more promising to use the coefficient of variation of the binomial distribution for $CV_{min}$, which is given by ($(1-\mu_i)/\mu_i$)$^{\frac{1}{2}}$, and which does not increase as fast as $1/(2\mu_i^2)$ as $\mu_i \to 0$. Notice that both functions have the same value and the same first derivative of $-\frac{1}{2}$ at $\mu_i = \frac{1}{2}$, which makes the spliced function particularly attractive. The complete minimum relationship between $CV_{min}$ and $\mu_i$ is then given by

Figure 3.1 Relationship between the mean and the coefficient of variation.
\[
CV_{i}^{\text{min}} = \begin{cases} 
\sqrt{\frac{1-\mu_i}{\mu_i}} & \text{if } \mu_i \leq \frac{1}{2} \\
\frac{1}{2\mu_i} & \text{if } \mu_i \geq \frac{1}{2}
\end{cases}, \quad (3.14)
\]

which is shown as curve BB in Figure 3.1.

The true \( CV_i \) will most likely be above \( CV_{i}^{\text{min}} \), so it will be necessary to add something to \( CV_{i}^{\text{min}} \). How much needs to be added is again an empirical question. A constant might be a good first attempt, but adding a normal curve with 3 parameters to \( CV_{i}^{\text{min}} \) can be expected to yield a better fit. The three parameters of the normal curve determine the height and the location of the maximum, and the distance between the points of inflection. An additional degree of freedom might be spent on the size of the tails of the curve, suggesting a function of the form of the \( t \)-distribution, which is given by

\[
f(t) = \alpha \left(1 + \frac{\delta \left(t - \beta\right)^2}{\gamma}\right)^{-\frac{\delta+1}{2}}.
\]

In this expression, \( \alpha \) determines the height and \( \beta \) determines the location of the maximum, while \( \gamma \) is a measure of dispersion. \( \delta \) determines the thickness of the tails; as \( \delta \) approaches zero, the function approaches the normal curve. A positive \( \delta \) describes a \( t \)-function with tails that are thicker than the tails of the normal curve, and a negative \( \delta \) describes a symmetric beta function with thinner tails than the normal curve and with a bounded domain.

In data sets with widely different observations it may be advantageous to specify the relationship between the mean and the variance of the negative binomial distribution in terms of a transformation of the mean instead of the mean itself, to accommodate skewness. One transformation which can be done without the cost of an additional parameter is the logarithmic transformation. The final relationship between \( \mu_i \) and \( CV_i \) is then given by

\[
CV_i = \begin{cases} 
\sqrt{\frac{1-\mu_i}{\mu_i}} + \alpha \left(1 + \frac{\delta \left(ln\mu_i - \gamma\right)^2}{\beta}\right)^{-\frac{\delta+1}{2}} & \text{if } \mu_i \leq \frac{1}{2} \\
\frac{1}{2\mu_i} + \alpha \left(1 + \frac{\delta \left(ln\mu_i - \gamma\right)^2}{\beta}\right)^{-\frac{\delta+1}{2}} & \text{if } \mu_i \geq \frac{1}{2}
\end{cases}, \quad (3.16)
\]

and is shown as curve CC in Figure 3.1.
The specification of the regression function yields the mean $\mu_i$, and equation 3.16 yields an estimate of the coefficient of variation; the variance can now be determined as $\text{Var}[Y_i] = (CV_i \mu_i)^2$. From the equations for the mean and the variance of the negative binomial distribution in equation 3.13 it is now possible to determine the two parameters $q_i$ and $k_i$ as

$$
q_i = 1 - \frac{\mu_i}{\sigma_i^2},
$$

$$
k_i = \frac{\mu_i^2}{\sigma_i^2 - \mu_i}.
$$

(3.17)

If the parameterization of equation 3.6 is used instead, $\mu_i$ is determined by the regression equation, and $\alpha_i = \mu_i/(\sigma_i^2 - \mu_i)$.

How does this new specification relate to the two approaches by Cameron and Trivedi (1986)? Their first specification, which shows a constant mean-variance relationship, translates into

$$
CV_i = \sqrt{1 + \frac{\alpha_i^{-1}}{\mu_i}},
$$

(3.18)

and the second specification with the linear mean-variance relationship can be expressed as

$$
CV_i = \sqrt{1 + \frac{\alpha_i^{-1} \mu_i}{\mu_i}}.
$$

(3.19)

Both specification allow only for a realization of heterogeneity that is monotonic in the mean; this is not necessarily the case for all data sets, so that the new formulation is more general. The advantages of the new specification ought to improve the estimation by enough to justify the burden of having to estimate three additional parameters.

The likelihood ratio test can give an indication whether the new specification yields a significant improvement. If $\lambda$ is the ratio of the constrained likelihood $L^c$ to the unconstrained likelihood $L^u$, it can be shown that the test statistic $LR = -2 \cdot \ln \lambda = 2 \cdot (\ln L^u - \ln L^c)$ is asymptotically distributed as chi-square with $r$ degrees of freedom, where $r$ is the number of addi-
tional parameters in the unconstrained loglikelihood.\textsuperscript{122} It should be emphasized, however, that the likelihood ratio test enables choice between two models only in the case of nested models, so that an unambiguous null hypothesis exists.\textsuperscript{123} Clearly the new specification does not simply relax some restrictions posed by the previous formulations, which makes the choice problem non-nested. The new specification is used in the empirical investigation in Chapter 5, and the results are compared with those of the second of Cameron and Trivedi’s specifications. Although the new specification is capable of adopting a shape that describes heteroskedasticity in the same way as the traditional specifications, the estimated form is very different from a strictly monotonic relationship. In addition, the new specification results in considerably higher loglikelihoods. This implies that the new relationship is attractive.

3.5 Markov Chain Monte Carlo methods

With maximum likelihood techniques the distributional choice has a large impact on the outcome of the estimation procedure, which makes it crucial to find a distribution that is a sufficiently close approximation of the true distribution of the data.\textsuperscript{124} Yet standard analysis requires a distribution that is available in closed form, and whose parameterization can be motivated. The previous four sections have explained the Poisson and the negative binomial distribution. However, not all count data are Poisson or negatively binomial distributed, and it becomes important to investigate other methods of count data analysis.

For the derivation of the negative binomial distribution in Section 3.3 it was assumed that the data $Y$ follow a Poisson distribution $P(Y|\lambda)$ with the parameter $\lambda$ following a gamma distribution $f$. The marginal distribution of the unknown but observable $Y$ was calculated in equation 3.3 as

$$P(Y) = \int P(Y, \lambda) \ d\lambda = \int f(\lambda) \ P(Y|\lambda) \ d\lambda . \quad (3.20)$$

Yet in the context of the estimation problem, we are interested not so much in the distribution of the data, $P(Y)$, but rather in the distribution of the parameters conditional on the data, $P(\lambda|Y)$, which can be used to calculate the expected values of the coefficients. Following Bayes’ rule, this distribution is given by $P(\lambda|Y) = P(Y, \lambda)/P(Y)$, so that the conditional distri-
bution can be determined as

\[ P(\lambda|Y) = \frac{f(\lambda) \ P(Y|\lambda)}{\int f(\lambda) \ P(Y|\lambda) \ d\lambda} \quad (3.21) \]

To calculate the expected values of the coefficients, assume that the regression equation that connects the explanatory variables to the dependent variable is specified by the function \( g(\lambda) \). The expected values of the function parameters are then given by

\[ E[g(\lambda)|Y] = \frac{\int g(\lambda) \ f(\lambda) \ P(Y|\lambda) \ d\lambda}{\int f(\lambda) \ P(Y|\lambda) \ d\lambda} \quad (3.22) \]

In principle \( P \) and \( f \) can describe any distribution. Yet an algebraic solution of the integrals in equation 3.22 is very often not feasible, because for many prior distributions \( f \) the posterior distribution in the denominator of equation 3.22 is known only up to a normalizing constant. Until recently this technical difficulty made it frequently impossible to use this approach for data analysis. Fortunately, Markov Chain Monte Carlo methods (MCMC) offer a fairly straightforward possibility for calculating the expected value of \( g(\lambda) \) for a wide variety of distributions. MCMC methods use a Markov Chain to perform a Monte Carlo integration of the integrals in equation 3.22.\(^{125}\)

Figure 3.2 explains the principle of Monte Carlo integration for the one-dimensional case. The task is to determine the integral \( G \) of the function \( g \). By sampling \( n \) different points within the area \( A \) from a uniform distribution, the integral \( G \) is estimated as the area \( A \) multiplied by the fraction of the random points that fall below the curve \( g \). As the sample size \( n \) is exogenous, \( G \) can be approximated as closely as desired.

In the context of an estimation problem, Monte Carlo integration can be used to draw samples from a given (posterior) distribution \( \pi(X) \), and to determine the expected value of the parameters of this distribution from these samples.\(^{126}\) However, drawing samples from \( \pi(X) \)

\(^{125}\) See Tanner (1993) and Geweke (1995) for technical introductions to MCMC, and Casella and George (1992) and Gilks et al. (1996) for more intuitive approaches.

\(^{126}\) This description of MCMC closely follows Gilks et al. (1996), pp. 4-12.
will frequently not be feasible, because $\pi(X)$ may be known only up to a normalizing constant. Fortunately the $\{X_t\}$ do not need to be independent, but may follow any mechanism (for example a Markov chain) that uses the state space of $\pi(X)$ to draw samples in the correct proportions.

A Markov chain is a sequence of random variables that are sampled from a distribution $q(X_t | X_{t-1})$ (called ‘transition kernel’), so that each sample $X_t$ depends only on its direct predecessor $X_{t-1}$, but not on the entire history of the chain. After a sufficiently long ‘burn-in’ phase, the samples will no longer depend on the original starting value of the chain, and the Markov Chain will converge to samples that are drawn from a unique stationary distribution $\pi$. These samples can then be used to calculate $E[g(X)]$, where $g(X)$ is any function of $X$.

It is clearly necessary to find an appropriate transition kernel $q(X_t | X_{t-1})$ which guarantees that the stationary distribution $\pi$ is actually the distribution of interest. A general algorithm for such a kernel was developed by Metropolis et al. (1953). First a random vector $X^P$ is generated from a proposed transition density function $q(X_{t-1}, X_t)$. With probability

$$
\alpha(X_{t-1}, X^P) = \min \left\{ \frac{\pi(X^P) q(X_{t-1}, X^P)}{\pi(X_{t-1}) q(X^P, X_{t-1})}, 1 \right\}
$$

the new vector of the Markov chain is set equal to the sampled random vector ($X_t = X^P$); otherwise the new vector is set equal to the previous vector ($X_t = X_{t-1}$), which means that the chain does not move. It can be shown that the transition distribution $q(X_{t-1}, X_t)$ can have any
form, while the chain will still converge to the unique $\pi(X)$. However, the speed of convergence to $\pi(X)$ depends on the choice of $q(X_{i-1}, X_i)$, so that the burn-in phase can be very long with an ill-suited transition distribution.

A special form of the general algorithm, which updates the whole vector $X$ at every run, is the Metropolis-Hastings algorithm, which divides $X$ into $m$ subvectors and updates the subvectors one at a time. Subvector $i$ (denoted by $X_i$) is updated by using a proposal distribution $q_i$ that depends on a proposed update $X_i^p$ and on the values of all subvectors except $i$ (denoted by $X_{-i}$). The updating probability $3.22$ therefore changes to

$$\alpha(X_{-i}, X_i, X_i^p) = \min \left\{ \frac{\pi(X_{-i}^p | X_{-i}) q_i(X_i^p | X_{-i}, X_i)}{\pi(X_i^p | X_{-i}) q_i(X_i | X_{-i}, X_i)}, 1 \right\}$$

(3.24)

where $\pi(X_i | X_{-i})$ is the full conditional distribution of $X_i$ given all the other subvectors $X_{-i}$; $\pi(X_i | X_{-i})$ can be calculated as

$$\pi(X_i | X_{-i}) = \frac{\pi(X_i)}{\int \pi(X) \, dX_i}.$$  

(3.25)

The Gibbs Sampler is a special form of the Metropolis-Hastings algorithm, because it assigns every element of $X$ to a different subvector, so that each element of $X$ is updated separately. The proposal distribution $q_i$ of the Gibbs Sampler is given by $q_i(X_i^p | X_{-i}, X_{-i}) = \pi(X_i^p | X_{-i})$. Use of this proposal distribution in $3.22$ results in an acceptance probability of $1$, that is, the proposed value $X_i^p$ is always accepted. In short, the Gibbs Sampler consists of nothing but a (very long) set of samples from full conditional distributions that will ultimately converge to a set of samples from the desired distribution.$^{128}$

It is relatively straightforward to sample from full conditional distributions, which explains the widespread use of the Gibbs Sampler since 1990. So far, most of its applications are in the areas of biometrics and physics, but the Gibbs Sampler has also been used for unit root analysis in GNP data (McCulloch and Tsay, 1992; Geweke, 1993; Chibb and Greenberg, 1994), and for the estimation of technology utilization by manufacturers (Andrews, Berger

$^{127}$Gilks et al. (1996), pp. 7-8.

$^{128}$The proof that the Gibbs Sampler actually converges to the distribution of interest is given in Casella and George (1992), pp. 169-170, and in Geweke (1995), pp. 45-46.
In Chapter 6 the Gibbs Sampler is used to analyse the number of building permits with distributions other than the Poisson or the negative binomial. The results differ greatly from the results in Chapter 5, which makes one wonder how many previous count data analyses owe their results to their specific limited distributional assumptions.

\footnotesize
\begin{itemize}
  \item For examples of non-economic applications see Gilks, Richardson and Spiegelhalter (1996a).
\end{itemize}

\textit{129} For examples of non-economic applications see Gilks, Richardson and Spiegelhalter (1996a).