JOURNAL OF APPLIED ECONOMETRICS, VOL. 1, 29–53 (1986)

ECONOMETRIC MODELS BASED ON COUNT DATA: COMPARISONS AND APPLICATIONS OF SOME ESTIMATORS AND TESTS

A. COLIN CAMERON

Department of Economics, Ohio State University, U.S.A.

AND

PRAVIN K. TRIVEDI

Department of Economics, Faculty of Economics and Commerce, Australian National University, Canberra, Australia 2600

SUMMARY

This paper deals with specification, estimation and tests of single equation reduced form type equations in which the dependent variable takes only non-negative integer values. Beginning with Poisson and compound Poisson models, which involve strong assumptions, a variety of possible stochastic models and their implications are discussed. A number of estimators and their properties are considered in the light of uncertainty about the data generation process. The paper also considers the role of tests in sequential revision of the model specification beginning with the Poisson case and provides a detailed application of the estimators and tests to a model of the number of doctor consultations.

1. INTRODUCTION

The objective of this paper is to discuss the estimation and testing of *count data models* from the viewpoint of an applied econometrician who may be concerned predominantly with practical aspects of their use. A number of articles on these models, in which the dependent variable takes only non-negative integer values corresponding to the number of events occurring in a given interval, have appeared recently in the econometrics literature. These include those by Gilbert (1979) and Hausman, Hall and Griliches (1984). We believe that, the apparent novelty of count data models notwithstanding, many of the insights and lessons learnt from standard applied econometric research carry over to this area.

In particular, the idea that applied econometric analysis involves an iterative modelling cycle consisting of specification, evaluation, comparison and eventual model revision is now widely accepted (Hendry and Wallis, 1984). It is therefore useful to discuss count data models from the viewpoint of an econometrician who wishes to go through this process. For the normal linear regression model a great deal has been learnt about the process of model estimation and post-estimation model evaluation. Since most count data models can be accommodated within an extended or generalized linear model framework, many of these insights for the normal linear regression model carry over to count data models. McCullagh and Nelder (1983) draw many parallels between the regression model on the one hand and count data models and limited dependent variable models on the other.

The modelling of random counts is widespread and long established in the biometric literature (see Patil (1970) for an extensive introduction). However, as Gilbert (1979) and Hausman, Hall and Griliches (1984), henceforth referred to as HHG (1984), have pointed

0883–7252/86/010029–25$02.50

Received February 1985
Final revision June 1985

out, there is also much scope for application of count data models in econometric research. Examples are the number of purchases per period (Gilbert), the number of patents applied for (HHG, 1984) the number of visits to doctors, admissions to hospital and medicines taken (Cameron, Trivedi, Milne and Piggott (1984), henceforth referred to as Cameron *et al.* (1984)), the number of spells of unemployment, the number of strikes in a month and so forth.

In some cases the nature of data available to econometricians will dictate the use of count data models, regardless of economic theory. In other cases the nature of economic decision processes may actually lead to econometric models of variables naturally measured as counts. An example is provided by models which lead to corner solutions and hence to discrete choice, but where the choice may be made several times in the time interval studied, or the number of times the choice is exercised is itself optimally chosen at the beginning of the period. Yet another possibility is to view counted data as a discrete categorical proxy variable for an unobserved continuous variable. In these contexts it is natural to attempt to model the probability that the economic event will occur $n = 0,1,2,\ldots$ times in a given time interval.

Whatever the ability of economic theory to suggest variables that explain the count data of interest, it will generally offer little guidance in selection of a stochastic model for the count data. Furthermore, the extensive probabilistic literature on modelling discrete random events will usually be inappropriate. This is to be contrasted with biometrics where there may be strong reasons (such as genetics theory) for using a particular stochastic model. Thus, whereas the common practice of using the Poisson distribution as the starting point may be reasonable in biometric applications, the use of this distribution for modelling non-negative integer-valued economic events involves quite strong and empirically questionable assumptions (restrictions). A similar warning could be given to the econometrician using continuous data, who often regards the choice of the normal linear regression model as uncontentious (with effort directed more towards controlling for possible heteroscedasticity, autocorrelation or endogeneity). But when count data are used, considerably more attention needs to be paid to model specification and evaluation. Section 2 of the paper therefore considers the specification question in the context of a small number of empirically interesting count data models.

Many count data models are non-linear, though they can be usefully viewed, *à la* Nelder and Wedderburn (1972), as 'generalized linear models' and their extensions. The discussion of estimation of these models is closely related to the general theory of estimation of possibly misspecified non-linear models. Section 3 of the paper exposits the application of this theory to count data models and draws out its implications for applied econometric work. Section 4 considers how one may evaluate certain aspects of an estimated model using diagnostic tests that will reveal its weaknesses and point to desirable respecifications. Section 5 provides an illustrative application of a sequential modelling strategy that is espoused in the earlier sections of the paper and a comparison of some estimators. Section 6 concludes.

2. SPECIFICATION ISSUES

In the analysis of count data it is natural to consider discrete models, including not only the Poisson and compound Poisson distributions but also a large number of others which are extensively discussed by Patil (1970). However, certain categorical models such as the ordinal probit model (McKelvey and Zavoina, 1975) and even some continuous models may be appropriate in some cases. We concentrate on discrete data models, but do consider the ordinal probit model in Section 5 as well as models based on the normal distribution.

2.1. The Basic Poisson Model

Let Y_i denote the number of occurrences, for the ith of N individuals, of an event of interest in a given interval of time, $Y_i = 0,1,2,\ldots$ Let $y(t, t + \mathrm{d}t)$ denote the number of events observed in the interval $(t, t + \mathrm{d}t)$. If

$$\Pr[y(t,t+\mathrm{d}t) = 0] = 1 - \lambda \mathrm{d}t + O(\mathrm{d}t)$$

and

$$Pr[y(t,t+\mathrm{d}t) = 1] = \lambda \mathrm{d}t + O(\mathrm{d}t)$$

so that

$$\Pr[y(t,t+\mathrm{d}t) > 2] = O(\mathrm{d}t), \text{ as } \mathrm{d}t \to 0$$

then the number of events in an interval of a given length is Poisson distributed with the probability density

$$\Pr(Y_i = y_i) = \mathrm{e}^{-\lambda_i}\lambda_i^{y_i}/y_i!\,, \qquad y_i = 0,1,2,\ldots, \qquad i = 1,2,\ldots,N \tag{1}$$

where y_i is the realized value of the random variable. This is a one-parameter distribution with mean and variance of Y_i equal to λ_i. To incorporate exogenous variables $X_{ij}(j = 1, \ldots, K)$, including a constant, the parameter λ_i is specified to be†

$$\lambda_i = \exp(\mathbf{X}_i\boldsymbol{\beta}) \tag{2}$$

In applied work the Poisson regression model is restrictive in several ways. First, it is based on the assumption that events occur independently over time. The independence assumption may break down in several ways. There may be a form of dynamic dependence between the occurrence of successive events. Prior occurrence of an event, such as an accident or illness, may raise the probability of subsequent occurrence of the same or similar event. In the context of unemployment spells this form of dynamic interdependence has been dubbed *occurrence dependence* by Heckman and Borjas (1980). In the context of biometric literature the *contagion model* considers the possibility that an occurrence of an event, such as an accident or illness, may subsequently modify the probability of occurrence of a similar event. Xekalaki (1983) provides several references to this literature. Yet another mode of dynamic interdependence is inherent in the notion that events occur in 'spells' and the spells themselves occur according to some probability law, whereas the events within a given spell, which occur according to a different probability law, may be dependent. This model is discussed by Cresswell and Froggatt (1963) and Xekalaki (1983). For example, the event that A saw his doctor on Tuesday may not be independent of the event that he also saw the doctor on Monday, if both events arise out of the occurrence of a single spell of illness. Independence implies that Pr (A sees doctor on Tuesday | A saw doctor on Monday) = Pr (A sees doctor on Tuesday | A did not see doctor on Monday). The breakdown of the independence assumption raises fundamental issues which are discussed further later in the section.

Secondly, the assumption that the conditional mean and variance of Y_i given $\mathbf{X}_i$ are equal may also be too strong and hence fail to account for the overdispersion—the variance exceeds the mean—that characterizes many data sets.‡ In applied work it is desirable to test the Poisson restriction and to relax it if appropriate. Inappropriate imposition of this restriction may produce spuriously small estimated standard errors of $\hat{\boldsymbol{\beta}}$.

†The exponential function is used to ensure the non-negativity of y_i. Gilbert (1979) alternatively specified a linear function but this can cause obvious computational difficulties.

‡See Gourieroux, Monfort and Trognon (1984b, p. 703, footnote 3).

2.2. Compound Poisson Models

One way to relax this restriction is to allow for unexplained randomness in λ_i by replacing (2) by the stochastic equation

$$\ln \lambda_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i \tag{3}$$

where the error term could reflect a specification error such as unobserved omitted exogenous variables, as suggested by Gourieroux, Montfort and Trognon (1984a,b), henceforth referred to as GMT (1984–), or more straightforwardly simply intrinsic randomness, as in HHG (1984). Let $g(\varepsilon_i)$ denote the probability density function for ε_i. Then the marginal density of Y_i can be obtained by integrating with respect to ε_i:

$$\begin{aligned}\Pr[Y_i = y_i] &= \int \Pr[Y_i = y_i | \mathbf{X}_i, \varepsilon_i] g(\varepsilon_i) \mathrm{d}\varepsilon_i \\ &= \int \frac{\mathbf{e}^{-\exp(\mathbf{X}_i\boldsymbol{\beta}+\varepsilon_i)} \exp(\mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i)^{yi}}{y_i!} g(\varepsilon_i)\mathrm{d}\varepsilon_i \end{aligned} \tag{4}$$

Expression (4) defines a compound Poisson distribution whose precise form depends upon the specific choice of $g(\varepsilon_i)$. For certain parametric forms, such as the gamma which we shall examine in the next section, a closed form expression for (4) can be obtained; but for other choices, such as the standard normal density, the resultant compound Poisson might not have a closed form and hence be computationally cumbersome.

Compound Poisson distributions provide a natural generalization of the basic Poisson models. Often they are motivated by a desire for greater flexibility, specifically a desire to account for frequently observed overdispersion in data and to provide a better fit. This is achieved essentially by appealing to aggregation over a heterogeneous population as the source of overdispersion. But unless the parametric form of $g(\varepsilon_i)$ can be motivated from fundamental considerations there is some arbitrariness involved here and it becomes relevant to explore the consequences of misspecifying the model in this important respect. GMT (1984a) have provided a theoretical analysis that can be applied to this issue.

2.3. Negative Binomial Models

Although the negative binomial model can be motivated in a number of different ways, see Boswell and Patil (1970), its representation as a specific compound Poisson distribution has been common (HHG, 1984, pp. 921–922). One way is to allow for inter-person heterogeneity by allowing λ_i to vary randomly according to a probability law. For example, if $g(\varepsilon_i)$, or equivalently $f(\lambda_i)$, is assumed to be a gamma distribution, then the integration in (4) leads to the negative binomial. The specific parameterization of the resulting form is determined by the parameterization of the gamma distribution. In the i.i.d. case where no exogenous variables appear in the model the resulting difference is of no consequence, but in the regression case the differences in parameterization deserve some comment. We use the 'index' parameterization of the gamma distribution. Then $\lambda_i \sim$ Gamma (ϕ_i, ν_i) with density (for $\lambda_i > 0$, $\phi_i > 0$, $\nu_i > 0$)

$$f(\lambda_i) = \frac{1}{\Gamma(\nu_i)} \left(\frac{\nu_i\lambda_i}{\phi_i}\right)^{\nu_i} \exp\left(\frac{-\nu_i\lambda_i}{\phi_i}\right) \frac{1}{\lambda_i} \tag{5}$$

$$E[\lambda_i] = \phi_i \text{ and } \mathrm{Var}(\lambda_i) = \frac{1}{\nu_i}\phi_i^2 \tag{6}$$

The parameter ϕ_i is the mean and ν_i is called the index or precision parameter. It is readily shown that

$$\begin{aligned}\Pr[Y_i = y_i] &= \int \Pr[Y_i = y_i \mid \lambda_i] f(\lambda_i)\mathrm{d}\lambda_i \\ &= \frac{\Gamma(y_i + \nu_i)}{\Gamma(y_i + 1)\Gamma(\nu_i)} \left(\frac{\nu_i}{\nu_i + \phi_i}\right)^{\nu_i} \left(\frac{\phi_i}{\nu_i + \phi_i}\right)^{yi}\end{aligned} \tag{7}$$

with

$$E[Y_i] = \phi_i \tag{8}$$

and

$$\mathrm{Var}(Y_i) = \phi_i + \frac{1}{\nu_i}\phi_i^2 \tag{9}$$

(7) is one possible parameterization of the negative binomial distribution. Since $\phi_i > 0$ and $\nu_i > 0$ it is clear that the variance exceeds the mean, so the model allows for overdispersion

Different negative binomial regression models can be generated by linking the parameters ϕ_i and ν_i of the underlying distribution for λ_i to the explanatory variables $\mathbf{X}_i$ in different ways. Again to ensure non-negativity of the mean a natural specification is $E[Y_i] = \exp(\mathbf{X}_i\boldsymbol{\beta})$, obtained by letting $\phi_i = \exp(\mathbf{X}_i\boldsymbol{\beta})$. A wide range of variance–mean relationships can be obtained by letting $\nu_i = (1/\alpha)(\exp(\mathbf{X}_i\boldsymbol{\beta}))^k$, for $\alpha > 0$ and arbitrary constant k. Then $\mathrm{Var}(Y_i) = \exp(\mathbf{X}_i\boldsymbol{\beta}) + \alpha\exp((2 - k)\mathbf{X}_i\boldsymbol{\beta}) = E[Y_i] + \alpha(E[Y_i])^{2-k}$.

The model we call Negbin I is obtained by setting $k = 1$. Then $\mathrm{Var}(Y_i) = (1 + \alpha)E[Y_i]$, implying a constant variance–mean ratio. This model is given by McCullagh and Nelder (1983, p. 132) and is used by HHG (1984, p. 922).[†] The model we call Negbin II is obtained by setting $k = 0$. Then $\mathrm{Var}(Y_i) = E[Y_i](1 + \alpha E[Y_i])$, so that the variance–mean ratio is linear in the mean. This model is given by McCullagh and Nelder (1983, p. 194) and is used by GMT (1984b) and HHG (1984, p. 917). Clearly there are many more possibilities.

The two parameterizations imply different assumptions about the functional form of heteroscedasticity—a point which is not emphasized in the literature—and hence in general will lead to different estimates of the parameter β.

The two alternative specifications of the gamma heterogeneity distribution amount to *different parameterizations* in the univariate model, but where a regression component is present they lead to *different models*. This difference is also relevant when we consider the test of the null hypothesis that the distribution of Y_i is Poisson against the alternative that it is negative binomial.

2.4. Towards More General Count Data Models

The negative binomial model is simply one example of a generalised Poisson model. Generalization can proceed along two routes which are by no means mutually exclusive.

First, the strong independence assumption of the Poisson model may be relaxed by choosing one of the alternative models used in the biometric literature. These include the

[†] HHG use a different parameterization of the gamma distribution to obtain a negative binomial model with $E[Y_i] = (1/\delta)\exp(\mathbf{X}_i\boldsymbol{\theta})$ and $\mathrm{Var}(Y_i) = (1 + 1/\delta)(1/\delta)\exp(\mathbf{X}_i\boldsymbol{\theta})$. We suppose that $\mathbf{X}_i$ includes a constant. Then partitioning $\mathbf{X}_i\boldsymbol{\theta} = \theta_0 + \mathbf{X}_{1i}\boldsymbol{\theta}_1$ and $\mathbf{X}_i\boldsymbol{\beta} = \beta_0 + \mathbf{X}_{1i}\boldsymbol{\beta}_1$, the model of HHG is the same as that in the text with $\alpha = (1/\delta)$, $\beta_0 = \theta_0 - \log\delta$ and $\boldsymbol{\beta}_1 = \boldsymbol{\theta}_1$.

above-mentioned 'true contagion' and 'spells' models. Yet another possibility would be to retain the independence assumption but choose a distribution other than the Poisson.

The negative binomial model as developed and applied in the econometric literature is essentially the 'apparent contagion' model of biometrics. According to this model individuals have constant but unequal probability of experiencing an event. A different model is the 'true contagion' model where all individuals initially have the same probability of experiencing an event but this is modified by prior occurrence of events. Another model is the 'proneness' model according to which individuals are heterogeneous in respect of their proneness to certain events, with this heterogeneity attributed to individual and/or environmental factors. Yet another alternative mentioned earlier is a 'spells' model according to which events occur in clusters and are dependent.

A second alternative approach to generalization consists of retaining the basic Poisson model of events and making alternative more general heterogeneity assumptions. This simply means that one attempts to employ increasingly general compound Poisson distributions to fit the data. The statistical literature is replete with examples of compound Poisson distributions; see Johnson and Kotz (1972) for a survey.

Depending upon which of the two approaches considered above is followed, we are led to problems of model testing and discrimination which might be dealt with differently according to the aims and objectives of the investigator. Using standard econometric terminology one may say that the problem of discriminating between alternative models of events *for a given individual* involves a comparison of alternative structural hypotheses, whereas the problem of discriminating between alternative (compound Poisson) models for a given sample of individuals involves a comparison of different reduced forms. The problem of discriminating between structural hypotheses of the kind mentioned above is essentially a problem of identification and may not be solvable for any one cross-section data set. An example is given below. The problem of choosing between alternative reduced form models essentially on the basis of goodness of fit is a less complex problem. However, from an empirical viewpoint it is desirable to understand the effects of both kinds of misspecification if only to motivate different tests of misspecification. The issue is reconsidered in Section 4.

3. ESTIMATION

3.1. Overview

Leading references on the estimation of count data models specified from a viewpoint congenial to that of the applied econometrician are Nelder and Wedderburn (1972), Wedderburn (1974), McCullagh (1983), McCullagh and Nelder (1983), GMT (1984a, 1984b, 1984c) and HHG (1984). We assume that the applied econometrician will be interested in choosing from a range of feasible estimators on the basis of their theoretical properties, the effect of specification errors on the properties of the selected estimator and computational cost. Our task here is to summarize briefly some of the known results relating to these issues, avoiding excessive detail but providing sufficient background to make comprehensible the discussion in the later sections of this paper. To some extent this task has been fulfilled by GMT (1984c).

The essential choice in estimation is between maximum likelihood methods on the one hand, based on strong distributional assumptions, and on pseudo-maximum-likelihood or maximum quasi-likelihood methods on the other, based on weaker assumptions.

If the probability distribution of the variable y_i is known to belong to a specified parametric family, that is, the data generation process is known, and the likelihood function is well-behaved, maximum likelihood (ML) is the obvious estimation procedure.

If the specified probability distribution is not necessarily the 'true' one and the ML method is applied as if the 'true' distribution had been specified, then following GMT (1984a,b,c), we shall refer to the method as 'pseudo' ML (PML). In general there is no reason to believe that the PML estimator will be consistent—model misspecification can lead to inconsistent estimators. But, in the special case where the specified distribution for y_i lies in the linear exponential family, GMT (1984a) show that, regardless of the true distribution for y_i, the PML estimator is consistent provided only that the mean is correctly specified.†

Some commonly used distributions are members of the linear exponential family with an additional nuisance parameter. If this nuisance parameter can be related to the mean and variance in a certain way, GMT (1984a) propose the quasi-generalized pseudo-maximum-likelihood (QGPML) estimator. This estimator is consistent if both the mean and variance are correctly specified.‡

The results of GMT (1984a) are closely related to those of McCullagh (1983), who derived the asymptotic distribution for the maximum quasi-likelihood (MQL) estimator defined by Wedderburn (1974). The quasi-likelihood is defined implicitly as the sum of N subfunctions, each subfunction having derivative proportional to (y_i minus $E[y_i]$) divided by the variance of y_i. It can be shown that the PMLE is a special case of the MQLE, and the QGPMLE permits more complex mean–variance relationships than does the MQLE.

The main advantages of the PML and QGPML estimators are that they require fewer distributional assumptions and may be computationally less burdensome than ML estimators. The disadvantages are that if the complete distribution for the data can be correctly specified the PMLE will be (usually) less efficient than the QGPMLE which in turn will be (usually) less efficient than the fully efficient MLE, and that for model uses such as prediction knowledge of the mean and (possibly) the variance may be much less informative than knowledge of the complete distribution.

3.2. Properties of Estimators

The presentation here is based on GMT (1984a), who use the results of Burguette, Gallant and Souza (1982) on the properties of extremum estimators which are obtained by maximizing a stochastic objective function.§ For completeness we present results for the multivariate case where $\mathbf{y}_i$ is a G-dimensional vector, though in our applications to count data y_i will be scalar.

ML estimator

The *maximum likelihood* (ML) estimator $\hat{\boldsymbol{\theta}}$ maximizes the likelihood function, or

† A well-known example is PML estimation of $\boldsymbol{\beta}_0$ based on assuming that y_i is distributed as Normal $(\mathbf{X}_i\boldsymbol{\beta}_0, \sigma^2)$ with σ^2 known. Then

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i^{\mathrm{T}} y_i$$

is consistent for $\boldsymbol{\beta}_0$ even if y_i has some other distribution, as long as $E(y_i) = \mathbf{X}_i\boldsymbol{\beta}_0$.

‡ A classic example is regression analysis when the variance is proportional to the square of the mean. Then y_i is specified to be distributed as Normal $(\mathbf{X}_i\boldsymbol{\beta}_0, \alpha(\mathbf{X}_i\boldsymbol{\beta}_0)^2)$, and the QGPMLE for $\boldsymbol{\beta}_0$ is an empirical (or quasi) generalized least squares estimator.

§ Extremum estimators can be viewed as generalizing the results of Jennrich (1969) and are given an excellent exposition by Amemiya (1985). A special feature of the study of Burguette, Gallant and Souza (1982) is that $\{\mathbf{X}_i\}$ are assumed to be i.i.d. random variables, so that asymptotic variance expressions involve the expectation E_x.

equivalently maximized

$$\sum_{i=1}^{N} \log l(\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\theta}) \tag{10}$$

where $l(\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}_0)$ is the specified conditional probability density for $\mathbf{y}$ given $\mathbf{X}$.

If the specified density is the true density, then it can be shown under very general conditions that $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}_0$, and $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_0^{-1})$, where

$$\mathcal{I}_0 = E_x E_0 \left[-\frac{\partial^2 \log l(\mathbf{y},\mathbf{X},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\theta}_0} \right]$$

is the information matrix, and E_0 is the expectation with respect to the density $l(\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}_0)$.

PML estimators

When the specified density is not the true density, Huber (1967) and White (1982) show that $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_*$ which maximizes $E_x E_* [\log l(\mathbf{y}, \mathbf{X}, \boldsymbol{\theta})]$, where E_* denotes the expectation with respect to the true conditional density of $\mathbf{y}$ given $\mathbf{X}$, and $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1} \mathcal{J}\mathcal{I}^{-1})$ Where

$$\mathcal{I} = E_x E_* \left[-\frac{\partial^2 \log l(\mathbf{y},\mathbf{X},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\theta}_*} \right]$$

and

$$\mathcal{J} = E_x E_* \left[\frac{\partial \log l(\mathbf{y},\mathbf{X},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}_*} \frac{\partial \log l(\mathbf{y},\mathbf{X},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\theta}_*} \right]^{\dagger}$$

Clearly $\boldsymbol{\theta}_* \neq \boldsymbol{\theta}_0$ necessarily, and $\hat{\boldsymbol{\theta}}$ may be inconsistent for $\boldsymbol{\theta}_0$.

The fundamental contribution of GMT (1984a) is to give conditions under which $\boldsymbol{\theta}_* = \boldsymbol{\theta}_0$. We partition $\boldsymbol{\theta}$ into $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and suppose that (a) the conditional mean of $\mathbf{y}_i$ is $f(\mathbf{X}_i, \boldsymbol{\beta}_0)$, (b) the remaining parameters $\boldsymbol{\alpha}$ are specified and need not be estimated and (c) the specified density $l(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}_0)$ is of the form $l(\mathbf{y}, f(\mathbf{X}, \boldsymbol{\beta}_0))$, where $l(\mathbf{y}, \boldsymbol{\mu})$ is a member of a *linear exponential family* (LEF) with mean parameterization, so that

$$l(\mathbf{y}, \boldsymbol{\mu}) = \exp\{A(\boldsymbol{\mu}) + B(\mathbf{y}) + \mathrm{C}(\boldsymbol{\mu})\mathbf{y}\} \tag{11}$$

where it can be shown that

$$E[\mathbf{y}] = \boldsymbol{\mu} = -\left(\frac{\partial \mathrm{C}(\boldsymbol{\mu})}{\partial\boldsymbol{\mu}}\right)^{-1} \frac{\partial A(\boldsymbol{\mu})}{\partial\boldsymbol{\mu}} \quad \text{and} \quad \mathrm{Var}(\mathbf{y}) = \left(\frac{\partial \mathrm{C}(\boldsymbol{\mu})}{\partial\boldsymbol{\mu}}\right)^{-1}$$

The *pseudo-maximum-likelihood* (PML) estimator $\hat{\boldsymbol{\beta}}$ based on an LEF maximizes the likelihood function of the model defined by (a), (b) and (c), or equivalently maximizes

$$\sum_{i=1}^{N} [A(f(\mathbf{X}_i, \boldsymbol{\beta})) + \mathrm{C}(f(\mathbf{X}_i, \boldsymbol{\beta}))\mathbf{y}_i] \tag{12}$$

where $E[\mathbf{y}_i \mid \mathbf{X}_i] = f(\mathbf{X}_i, \boldsymbol{\beta}_0)$.

† White (1982) calls $\hat{\boldsymbol{\theta}}$ a 'quasi' MLE. We follow GMT in calling $\hat{\boldsymbol{\theta}}$ a 'pseudo' MLE, and use 'quasi' in a different sense below.

If the conditional mean is correctly specified, i.e. $E_*[\mathbf{y}_i \mid \mathbf{X}_i] = f(\mathbf{X}_i, \boldsymbol{\beta}_0)$, then $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}_0$, and $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{J}^{-1}\mathbf{I}\mathbf{J}^{-1})$, where

$$\mathbf{J} = E_x\left[\frac{\partial f}{\partial \boldsymbol{\beta}}\boldsymbol{\Sigma}^{-1}\frac{\partial f}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right], \quad \mathbf{I} = E_x\left[\frac{\partial f}{\partial \boldsymbol{\beta}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}_*\boldsymbol{\Sigma}^{-1}\frac{\partial f}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right]$$

$$\frac{\partial f}{\partial \boldsymbol{\beta}} = \left.\frac{\partial f(\mathbf{X},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}_0}, \qquad \boldsymbol{\Sigma} = \left(\left.\frac{\partial \mathbf{C}(f(\mathbf{X},\boldsymbol{\beta}))}{\partial f(\mathbf{X},\boldsymbol{\beta})}\right|_{\boldsymbol{\beta}_0'}\right)^{-1}$$

$\boldsymbol{\Sigma}$ is the specified conditional variance–covariance matrix for $\mathbf{y}$, and $\boldsymbol{\Omega}_*$ is the true one.

The essential result is that regardless of other properties of the true conditional distribution of y, provided the conditional mean is correctly specified, the PMLE is consistent. The asymptotic distribution of the PMLE will, however, depend on other properties of the true distribution for y, since **I** depends on $\boldsymbol{\Omega}_*$. If $\boldsymbol{\Omega}_*$ is unknown we can extend the results of Eicker (1967) and White (1980) for regression models based on the normal distribution to those based on an LEF and consistently estimate **I** by

$$\hat{\mathbf{I}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \hat{f}_i}{\partial \boldsymbol{\beta}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - f(\mathbf{X}_i,\hat{\boldsymbol{\beta}}))^2\,\hat{\boldsymbol{\Sigma}}_i^{-1}\frac{\partial \hat{f}_i}{\partial \boldsymbol{\beta}^{\mathrm{T}}}$$

where

$$\frac{\partial \hat{f}_i}{\partial \boldsymbol{\beta}} = \left.\frac{\partial f(\mathbf{X}_i,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} \text{ and } \hat{\boldsymbol{\Sigma}}_i = \left(\left.\frac{\partial \mathbf{C}(f(\mathbf{X}_i,\boldsymbol{\beta}))}{\partial f(\mathbf{X}_i,\boldsymbol{\beta})}\right|_{\hat{\boldsymbol{\beta}}}\right)^{-1}$$

Alternatively, if the true $\boldsymbol{\Omega}_* = \boldsymbol{\Omega}(\mathbf{X}_i, \boldsymbol{\beta}_0)$ we replace $(y_i - f(\mathbf{X}_i, \hat{\boldsymbol{\beta}}))^2$ in $\hat{\mathbf{I}}$ by $\boldsymbol{\Omega}(\mathbf{X}_i, \hat{\boldsymbol{\beta}}_0)$. In either case a consistent estimate for **J** is

$$\hat{\mathbf{J}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \hat{f}_i}{\partial \boldsymbol{\beta}}\hat{\boldsymbol{\Sigma}}_i^{-1}\frac{\partial \hat{f}_i}{\partial \boldsymbol{\beta}^{\mathrm{T}}}$$

QGPML estimators

The LEF includes the normal (with σ^2 specified), Poisson, negative binomial (with ν in (7) specified), binomial (with n given) and gamma (with ν in (5) specified) distributions. Some of these distributions include a nuisance parameter. PML estimators based on these distributions will be consistent for $\boldsymbol{\beta}_0$ regardless of the value specified for the nuisance parameter. However, if the nuisance parameter could be estimated from the data, rather than arbitrarily specified a value, a more efficient estimator for $\boldsymbol{\beta}_0$ may be possible.

GMT (1984a) propose a method to do this. We partition $\boldsymbol{\theta}$ into $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and suppose that (a) the conditional mean of $\mathbf{y}_i$ is $f(\mathbf{X}_i, \boldsymbol{\beta}_0)$; (b) the conditional variance of $\mathbf{y}_i$ is $\boldsymbol{\Omega}(\mathbf{X}_i, \boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)$; (c) the specified density $l(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)$ is of the form $l(\mathbf{y}, f(\mathbf{X}_i, \boldsymbol{\beta}_0), \psi(f(\mathbf{X}_i, \boldsymbol{\beta}_0), \boldsymbol{\Omega}(\mathbf{X}_i, \boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)))$, where $l(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\xi})$ is a member of an *LEF with nuisance parameter* (and mean parameterization), so that

$$l(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\eta}) = \exp\{A(\boldsymbol{\mu}, \boldsymbol{\eta}) + B(\mathbf{y}, \boldsymbol{\eta}) + \mathbf{C}(\boldsymbol{\mu}, \boldsymbol{\eta})\mathbf{y}\} \tag{13}$$

where it can be shown that $E[\mathbf{y}] = \boldsymbol{\mu}$ and $\mathrm{Var}(\mathbf{y}) = (\partial \mathbf{C}(\boldsymbol{\mu}, \boldsymbol{\eta})/\partial \boldsymbol{\mu})^{-1}$; and (d) $\boldsymbol{\eta} = \psi(\boldsymbol{\mu}, \boldsymbol{\Omega})$ is a differentiable function of $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$, and ψ defines, for any given $\boldsymbol{\mu}$, a one-to-one relationship between $\boldsymbol{\eta}$ and $\boldsymbol{\Omega}$.

The *quasi-generalized pseudo-maximum-likelihood* (QGPML) estimator $\hat{\boldsymbol{\beta}}$ based on LEF with nuisance parameter is obtained by maximizing the likelihood of the model defined by (a), (b), (c) and (d), with $\boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta}_0, \boldsymbol{\alpha}_0) \equiv \psi(f(\mathbf{X}_i, \boldsymbol{\beta}_0), \boldsymbol{\Omega}(\mathbf{X}_i, \boldsymbol{\beta}_0, \boldsymbol{\alpha}_0))$ replaced by $\boldsymbol{\eta}(\mathbf{X}_i, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}})$ where $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\alpha}}$ are strongly consistent estimates of $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$ of order $O(N^{-\frac{1}{2}})$. Equivalently, $\hat{\boldsymbol{\beta}}$ maximizes

$$\sum_{i=1}^{N} [A(f(\mathbf{X}_i,\boldsymbol{\beta}), \boldsymbol{\eta}(\mathbf{X}_i,\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\alpha}})) + \mathbf{C}(f(\mathbf{X}_i,\boldsymbol{\beta}), \boldsymbol{\eta}(\mathbf{X}_i,\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\alpha}}))\mathbf{y}_i] \tag{14}$$

where

$$E[\mathbf{y}_i \mid \mathbf{X}_i] = f(\mathbf{X}_i,\boldsymbol{\beta}_0) \text{ and } \mathrm{Var}(\mathbf{y}_i \mid \mathbf{X}_i) = \Omega(\mathbf{X}_i,\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0) = \left(\frac{\partial \mathbf{C}(\mathbf{X}_i,\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)}{\partial f(\mathbf{X}_i,\boldsymbol{\beta}_0)}\right)^{-1}$$

If both the conditional mean and variance of $\mathbf{y}_i$ are correctly specified, then $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}_0$ and $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{K}^{-1})$, where

$$\mathbf{K} = E_x\left[\frac{\partial f}{\partial \boldsymbol{\beta}} \boldsymbol{\Omega}^{-1} \frac{\partial f}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right], \quad \frac{\partial f}{\partial \boldsymbol{\beta}} = \left.\frac{\partial f(\mathbf{x}_i,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}_0} \quad \text{and } \boldsymbol{\Omega} = \boldsymbol{\Omega}(\mathbf{x}_i,\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)$$

A consistent estimate for $\mathbf{K}$ is given by

$$\hat{\mathbf{K}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \hat{f}_i}{\partial \boldsymbol{\beta}} \hat{\boldsymbol{\Omega}}_i \frac{\partial \hat{f}_i}{\partial \boldsymbol{\beta}^{\mathrm{T}}}$$

The initial consistent estimate $\tilde{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$ may be the PMLE based on any LEF. From this a consistent estimator $\tilde{\boldsymbol{\alpha}}$ for $\boldsymbol{\alpha}$ may be formed, see below for examples.

Comparison of estimators

The QGPMLE $\hat{\boldsymbol{\beta}}$ has a variance–covariance matrix which depends only on the functional forms for the conditional mean and variance of $\mathbf{y}$, so a very large number of QGPML estimators based on different objective functions of the form (14) and different initial consistent parameter estimates will have the same asymptotic distribution.

If the specified conditional mean and variance of $\mathbf{y}_i$ are the true ones, the QGPMLE for $\boldsymbol{\beta}_0$ attains the lower bound for the variance–covariance matrix of the PMLE for $\boldsymbol{\beta}_0$. However, although the QGPMLE is then more efficient than the PMLE it is not necessarily efficient. In the special case where the true density is itself a member of an LEF of form (13), with $\boldsymbol{\mu} = f(\mathbf{X}, \boldsymbol{\beta})$ and $\boldsymbol{\eta} = \boldsymbol{\eta}(\mathbf{X}, \boldsymbol{\alpha})$ (equal to those specified for the QGPMLE), the QGPMLE for $\boldsymbol{\beta}_0$ is asymptotically equivalent to the MLE for $\boldsymbol{\beta}_0$ (obtained by maximizing the true likelihood function jointly with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$), and hence is fully efficient. Note that for a generalized linear model $\boldsymbol{\eta}(\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\alpha})$ is of the simpler form $\boldsymbol{\eta}(\mathbf{X}, \boldsymbol{\alpha})$.†

† Generalized linear models (GLM) are given a comprehensive treatment by McCullagh and Nelder (1983). The scalar dependent variable y is assumed to have a density of the form

$$l^{+}(y,\mu,\alpha) = \exp\left\{\frac{1}{a(\alpha)}(y\pi - b(\pi)) + c(y,\alpha)\right\}$$

where it can be shown that $\mu = E[y] = b'(\pi)$. For α given, $l^{+}(y, \pi)$ is a member of a LEF with 'natural' parameterization. To link this with GMT who use the 'mean' parameterization, invert $b'(\pi) = \mu$ to obtain $\pi = d(\mu)$ and hence

$$l^{+}(y,\mu,\alpha) = \exp\left\{-\frac{b(d(\mu))}{a(\alpha)} + c(y,\alpha) + \frac{d(\mu)}{a(\alpha)}y\right\}$$

For α a freely varying parameter this is a special case of the LEF with nuisance parameter defined in (13), with η a function of α alone.

Now define the pseudo joint MLE (PJMLE) to be the estimator obtained by maximizing a possibly misspecified likelihood function with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

The PJMLE for $\boldsymbol{\beta}_0$ is consistent and fully efficient if the specified conditional mean, variance and distribution of y_i are the true ones. If the distribution is misspecified the PJMLE may be inconsistent. One trivial example where it is still consistent is when there is no nuisance parameter, so that maximization of the likelihood is with respect to $\boldsymbol{\beta}$ alone, and the specified likelihood function is a member of an LEF—this is the case for Poisson regression models. Another example of consistency of the PJMLE for $\boldsymbol{\beta}_0$ despite incorrect specification of the distribution is when the specified distribution is a member of a quadratic exponential family (GMT, 1984a, p. 691)—this is the case for the normal distribution but not for the negative binomial distribution. If additionally $\boldsymbol{\eta}(\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{\eta}(\mathbf{X}, \boldsymbol{\alpha})$, it can be shown that in this case the QGPMLE will be asymptotically equivalent to the PJMLE though both may be inefficient.†

4. SOME MODEL EVALUATION PROCEDURES

As in standard regression analysis, in modelling count data it is desirable to supplement estimation with additional tests to determine whether the fitted model is adequate and whether a specific deficiency of any initially entertained model can be removed by progression to a less restrictive model. Alternatively one may wish to begin with a relatively unrestricted model and proceed to a more restrictive, but data coherent, model on grounds of parsimony and/or statistical efficiency. This section considers a number of tests which are intended to help such a specification search.

4.1. Discrimination Between Rival Structural Hypotheses

If the objective of the researcher is to evaluate the adequacy of the independence assumption of the Poisson model against alternative assumptions which allow for interdependence between events, then generally it will not be possible to do so on the basis of goodness of fit tests alone applied to the reduced form compound Poisson models. An example illustrating the fundamental identification problem involved was provided by Irwin (1941) who showed that a univariate negative binomial distribution for accidents can be derived as a compound Poisson distribution with a gamma compounder, or as a true contagion model in which the probability of an individual having an accident depends upon the number of previous accidents sustained. A more recent example is provided by Xekalaki (1983), also in the context of accident theory. The author considers a particular three-parameter distribution for accidents, the 'univariate generalized Waring distribution', UGWD, and shows that it is consistent with 'proneness', 'contagion' and 'spells' models. So a good fit of the UGWD to data is of no help in discriminating between rival structural hypotheses. (Note that both these examples refer to models without a regression component.)

A necessary condition for discriminating between compound Poisson models and rival models which incorporate dependence between events is that we have available more detailed data, possibly panel data. Intuitively this should be obvious, since an assumption about dynamic interdependence between events implies something about the time intervals between successive events. It is well known that if the process generating events is a Poisson process then the time between successive events has an exponential distribution with a constant hazard

† An example is that $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ is asymptotically equivalent to the estimator for $\boldsymbol{\beta}_0$ obtained by; maximizing jointly w.r.t. $\boldsymbol{\beta}$ and σ^2 the likelihood function based on assuming $y_i \sim \mathcal{N}(\mathbf{X}_i, \boldsymbol{\beta}_0, \sigma_0^2)$, even if the normality assumption is incorrect.

function and, furthermore, the times between events are independent. Therefore the Poisson process implies that the times between successive events are i.i.d. exponential random variables. Of course these times may be i.i.d. but not necessarily exponential. It seems appropriate, therefore, to first test whether the process is a renewal process before testing whether it is a Poisson renewal process. (A process in which intervals between events are i.i.d. is a renewal process.) Tests for renewal processes generally and for Poisson renewal process specifically, based on information on intervals between events, are discussed by Cox and Lewis (1966) at length and by Lawless (1982, Chapter 10.2) more briefly. The cases discussed there do not include models with a regression component and do not allow for heterogeneity, but extensions and generalizations may be possible. To implement such tests one requires more data than a single cross-section. For example, one test is based on the first serial correlation coefficient between lengths of successive time intervals. Under the null hypothesis that the process is a renewal process its value would be zero. Clearly, to test this one requires panel data.

That the availability of panel data constitutes a necessary condition for discriminating between 'true' and 'spurious' (or 'apparent') contagion has been recognized at least since the work of Bates and Neyman (1952). In the context of the problem of discrimination between 'true' and 'spurious' (unemployment) duration dependence, Heckman and Singer (1984) have also emphasized the value of panel data and provided conditions for identifiability. Extension of this analysis to count data models with a regression component remains an area for future research.†

4.2. Tests for Poisson Model

The discussion above concerned specification tests. We now consider misspecification tests which are designed to highlight inadequacy of the maintained model is specific directions. In this subsection we consider the problem of choosing between Poisson models and more general models, given that the investigator has correctly specified the mean function $\mu_i \equiv E[y_i \mid \mathbf{X}_i, \boldsymbol{\beta}]$.

A natural basis for testing the adequacy of the Poisson model is the relationship between $\mathrm{Var}(y_i \mid \mathbf{X}_i, \boldsymbol{\beta})$ and $E[y_i \mid \mathbf{X}_i, \boldsymbol{\beta}]$. We propose tests of

$$H_0 : y_i \sim \text{Poisson } (\mu_i = f(\mathbf{X}_i, \boldsymbol{\beta}))$$

against

$$H_{\mathrm{A}} : E[y_i \mid \mathbf{X}_i, \boldsymbol{\beta}] = \mu_i$$

$$\mathrm{Var}(y_i \mid \mathbf{X}_i, \boldsymbol{\beta}) = \mu_i + \alpha\mu_i^l, \text{ for given } l$$

where the distribution for y_i under H_{A} may not necessarily be specified. The model under H_{A} nests many potentially interesting models as special cases. For example, if y_i is specified to be distributed as the negative binomial under H_{A}, the value $l = 1$ corresponds to the Negbin I model and $l = 2$ to the Negbin II model.

Since the variance of y_i equals μ_i when y_i is Poisson distributed, tests of H_0 against H_{A} are

† Models for panel count data are beyond the scope of this paper. HHG (1984) take a similar approach to Mossiman (1970) to control for 'apparent' contagion by allowing an individual specific effect in a panel data analogue of (3)—$\ln \lambda_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + \varepsilon_i$—to have both a random component which is controlled for along the lines of (4) and a fixed component which is controlled for by conditioning for each individual on $\Sigma_t y_{it}$. GMT (1984b) propose an error component model with both individual specific and time specific effects, but this will require a very long panel for consistent estimation of the parameters. Neither set of authors considers 'true' contagion.

based on tests for $\alpha = 0$. We consider tests using the familiar principles of the Wald, the likelihood ratio and score (Lagrange multiplier) tests.

Score tests

We follow Lee (1984) and consider score tests of the Poisson model against the alternative that the distribution belongs to a system of distributions studied by Katz (1963) which contains as special cases the Poisson, negative binomial and binomial distributions.† An advantage of this system of distributions is that it not only allows for overdispersion (in which case it defines the negative binomial distribution) but it also permits underdispersion.

We specify the mean and variance of the Katz distribution under $H_{\rm A}$ to be respectively μ_i and $(\mu_i + \alpha\mu_i^l)$.‡ The Poisson distribution is obtained when $\alpha = 0$, so a natural test for H_0 against $H_{\rm A}$ is a score test for $\alpha = 0$. By similar manipulations to those of Lee we obtain the efficient score under H_0:

$$\sum_{i=1}^{N} \frac{\partial \log \Pr\{Y_i = y_i\}}{\partial \alpha}\bigg|_{H0} = \sum_{i=1}^{N} \frac{1}{2}\mu_i^{l-2}\{(y_i - \mu_i)^2 - y_i\} \tag{15}$$

From this we obtain the test statistic

$$\hat{T}_{\rm L} = \frac{1}{\sqrt{2}}\,\frac{\sum_{i=1}^{N} \hat{\mu}_i^{l-1}\frac{1}{\hat{\mu}_i}\{(y_i - \hat{\mu}_i)^2 - y_i\}}{\sqrt{\left[\sum_{i=1}^{N} (\hat{\mu}_i^{l-1})^2\right]}} \tag{16}$$

where $\hat{\mu}_i = f(\mathbf{X}_i, \hat{\boldsymbol{\beta}})$ and $\hat{\boldsymbol{\beta}}$ is the MLE for $\boldsymbol{\beta}$ in the null hypothesis Poisson model. Under H_0, $\hat{T}_{\rm L} \rightarrow_L \mathcal{N}(0,1)$. When testing specifically against the negative binomial distribution, i.e. against overdispersion, a one-sided normal test based on $\hat{T}_{\rm L}$ should be used.

Since $\mathrm{Var}_{H0}(1/\mu_i\{(y_i - \mu_i)^2 - y_i\}) = 2$, an asymptotically equivalent test statistic under H_0 is

$$\hat{T}'_{\rm L} = \frac{\sqrt{N}\sum_{i=1}^{N} \hat{\mu}_i^{l-1}\frac{1}{\hat{\mu}_i}\{(y_i - \hat{\mu}_i)^2 - y_i\}}{\sqrt{\left[\sum_{i=1}^{N}\left(\frac{1}{\hat{\mu}_i}\{(y_i - \hat{\mu}_i)^2 - y_i\}\right)^2\right]}\sqrt{\left[\sum_{i=1}^{N} (\hat{\mu}_i^{l-1})^2\right]}} \tag{17}$$

† The system of discrete distributions studied by Katz (Katz, 1963; Johnson and Katz, 1972) is defined by the recursive relationship

$$\Pr\{Y = y+1\} = \begin{cases} \left(\dfrac{\lambda + \gamma y}{y+1}\right)\Pr\{Y = y\}, & \lambda + \gamma y \geq 0 \\ 0 & , \ \lambda + \gamma y < 0 \end{cases}$$

where $\lambda > 0$, $\gamma < 1$ and $y = 0,1,2,\ldots$ This system has mean $\lambda/(1-\gamma)$ and variance $\lambda/(1-\gamma)^2$, and includes as special cases the negative binomial $(0 < \gamma < 1)$, Poisson $(\gamma = 0)$, binomial $(\gamma < 0$ and $-\lambda/\gamma$ an integer) and other $(\gamma < 0$ and $-\lambda/\gamma$ not an integer) distributions.

‡ Lee (1984) specifies $\mu_i = \exp(\mathbf{X}_i\boldsymbol{\beta})$ and considers two different specifications for the variance under the alternative hypothesis,

$$\left(\mu_i + \frac{\omega}{1-\omega}\,\mu_i\right) \text{ and } \left(\mu_i + \frac{\delta}{1-\delta}\,\mu_i^2\right)$$

which with the obvious reparameterization correspond to $l = 1$ and $l = 2$ above.

The statistic is easily computed as it is the square root of N times the centred R^2 from the least squares regression of $(1/\hat{\mu}_i)\{(y_i - \hat{\mu}_i)^2 - y_i\}$ against $\hat{\mu}_i^{l-1}$.

Overdispersion tests

The above tests require that the actual distribution of y_i, rather than just the mean and variance, be specified under H_A. Cox (1983) proposes a test statistic which does not require specification of the distribution of y_i under H_A. This is possible by considering only local alternatives to the null hypothesis Poisson distribution of y_i, and Cox calls his test a test for 'modest amounts of overdispersion'.

In the general framework of Cox we consider independently distributed observations $Y_1, \ldots, Y_N$. Each observation Y_i has density $f(y_i, \lambda_i)$, where the scalar parameter λ_i is itself the realization of a random variable Λ_i distributed with mean $\mu_i = f(\mathbf{X}_i, \boldsymbol{\beta})$ and variance τ_i^2. Under H_0, $\tau_i^2 = 0$, whereas the alternative overdispersed model permits $\tau_i^2 > 0$ and possibly a function of μ_i. The restriction that Cox makes is to suppose that τ_i^2 is $O(1/\sqrt{N})$, and it is in this sense that Cox considers modest amounts of overdispersion.

Under certain regularity conditions (Cox and Hinkley, 1974, Chapter 9.2), by taking a second-order Taylor series expansion of $f(y, \lambda)$ about $\lambda = \mu$ and taking expectations with respect to λ Cox obtains the following $O(1/N)$ approximation to the density of the overdispersed model

$$f^*(y,\mu,\tau^2) = f(y,\mu)\exp(\tfrac{1}{2}\tau^2 h(y,\mu))/a(y,\mu,\tau^2) \tag{18}$$

where

$$h(y,\mu) = \left\{\frac{\partial \log f(y,\mu)}{\partial\mu}\right\}^2 + \frac{\partial^2 \log f(y,\mu)}{\partial\mu^2} \tag{19}$$

and $a(y, \mu, \tau^2)$ is a normalizing constant and we have omitted the subscript i on y_i, μ_i and τ_i^2.

For his test Cox considers the case where τ^2 does not depend on μ. More generally we suppose $\tau^2 = (1/\sqrt{N})\alpha\mu^l$, for given l. Then a test for overdispersion is simply a test of the null hypothesis $\alpha = 0$ against the alternative $\alpha > 0$. A natural test is based on the score statistic

$$\sum_{i=1}^{N} \frac{\partial \log f^*(y_i,\mu_i,(1/\sqrt{N})\alpha\mu_i^l)}{\partial\alpha}\bigg|_{\alpha=0} = \frac{1}{2\sqrt{N}}\sum_{i=1}^{N} \mu_i^l h(y_i, \mu), \tag{20}$$

where $\hat{\mu}_i = f(\mathbf{X}_i, \hat{\boldsymbol{\beta}})$ and $\hat{\boldsymbol{\beta}}$ is the MLE for $\boldsymbol{\beta}$ under H_0.

We now apply these results to the case where $y_i \sim$ Poisson $(\lambda_i = \mu_i)$ under H_0. First note that under the alternative hypothesis $\text{Var}(y) = E_\lambda[\text{Var}(y \mid \lambda)] + \text{Var}_\lambda(E[y \mid \lambda]) = E_\lambda[\lambda] + \text{Var}_\lambda(\lambda) = \mu + (1/\sqrt{N})\alpha\mu^l$. This is the same as H_A given earlier, except now only local alternatives $((1/\sqrt{N})\alpha)$ are considered. Secondly, for the Poisson $h(y, \mu) = (1/\lambda^2)\{(y - \lambda)^2 - y\}$. Therefore (20) becomes

$$\frac{1}{2\sqrt{N}}\sum_{i=1}^{N} \mu_i^l \frac{1}{\mu_i^2}\{(y_i - \mu_i)^2 - y_i\}$$

which is equivalent to (15) and clearly yields the test statistic $\hat{T}_L$ given in (16). So the same test statistic is obtained by considering either local alternatives to the Poisson or alternatives to the Poisson in the Katz system of distributions.†

† See Cameron and Trivedi (1985) for more detailed analysis of the commonalities between various test procedures. In addition to the above tests, White's information matrix test and its obvious similarity with Cox's test and regression based tests are discussed.

Wald test

The advantage of the score test is that the model need only be estimated under H_0. However, if the model is estimated under the alternative hypothesis, a Wald test may be easy to compute. For example, if the Negbin I model is estimated to yield the MLE $\hat{\alpha}$ and asymptotic variance var($\hat{\alpha}$) for α, the Wald test of H_0: Poisson (μ_i) against H_A: negative binomial with mean μ_i and variance ($\mu_i + \alpha\mu_i$) is based on the usual T-statistic $\hat{\alpha}/\sqrt{[\text{var}(\hat{\alpha})]}$, which is asymptotically distributed as $\mathcal{N}(0, 1)$. Note that a one-sided test is appropriate.

Likelihood ratio test

If both the model under H_0 and the model under H_A are estimated by maximum likelihood, and the model under H_0 can be derived from the model under H_A by a single parametric restriction (on α), the quantity 2 times log-likelihood ratio is asymptotically distributed as $\chi^2(1)$ under H_0. For example, we can form log-likelihood ratio tests for the Negbin I model or Negbin II model against the Poisson model.

4.3. Implications for Sequential Modelling

In common with the use of score tests in many other areas, several of the above tests have a certain drawback from an applied viewpoint: they are helpful in rejecting the maintained model but less so in directing the user to a better class of models. Cox's approach indicates that, regardless of the alternative considered, the test will be the same whenever the $O(1/N)$ approximation to the density is the same. Misspecification tests do not therefore indicate a unique alternative model. The question then arises: what should an investigator do if the Poisson model is found to be data incoherent? It is clear that tests based on specific parametric alternatives are more helpful in this regard. If the test indicates overdispersion, the negative binomial seems an obvious alternative, whereas if underdispersion is the problem then the binomial or truncated Poisson could be appropriate. Even if a specific parametric alternative to the Poisson is not adopted, it is useful to find out whether the Poisson model fits the data adequately to guard against the invalid imposition of strong restrictions which will produce artifically low variances.

Suppose, as is likely in many empirical situations, the negative binomial model is preferred. Then, of course, it too should be subjected to additional tests against other more flexible alternatives. A random effects negative binomial is potentially more flexible and can be generated beginning with (7) and treating the quantity v_i as a random variable with a beta density, $\mathrm{d}B(a, b)$, with parameters a and b. This is the approach followed in HHG (1984, p. 927); also see Ord (1972, pp. 74 and 85). We do not consider such models here. It is clear from section 2.3 that the negative binomial already allows great flexibility for modelling the relationship between Var(y_i) and $E(y_i)$.

It is more convenient to proceed as follows. Following the rejection of the Poisson model, re-estimate the model under the negative binomial assumption and by QGPML method. If the two sets of results differ substantially, then this probably indicates the failure of the negative binomial assumption and the investigator should prefer QGPMLE. However, this still leaves open the question of whether a random effects negative binomial may be preferred to a Poisson model with specification errors. We do not pursue this question here.

5. APPLICATION

5.1. The Data and The Model

This section contains an application of the sequential modelling strategy advocated in the earlier sections. It is based on and related to a much more extensive study of the demand for health insurance and the use of health care services in Australia. See Trivedi *et al.* (1984) and Cameron *et al.* (1984) for a discussion of the substantive economic issues. Here we shall confine discussion to matters of econometric methodology leaving the interested reader to pursue other details in the previous papers.

The data for the application are derived from the Australian Health Survey 1977–78 which contains information on the number of consultations with a doctor or specialist, denoted NOCNSLT, in the two-week period before an interview. No information was obtained on intervals between consultations. The sample consists of 5190 individuals (heads-only) over 18 years old who answered all the questions essential to our analysis. The explanatory variables in the model comprise, in addition to a constant denoted ONE, thirteen variables, namely SEX, AGE, AGESQ (age-squared), INCOME, ILLNESS (recent illnesses), ACTDAYS (number of reduced activity days), HSCORE (general health questionnaire score), CHCOND1, CHCOND2 (number and type of chronic conditions) and dummy variables for four levels of health insurance cover, denoted, respectively, as MEDLEVY, LEVYPLUS, FREEPOOR and FREEREPAT. The omitted dummy variable is MEDLEVY (medibank levy). LEVYPLUS represents a higher level of insurance cover, whereas FREEPOOR and FREEREPAT represent a basic level of insurance cover supplied free of charge to certain individuals on grounds of, respectively, low income and pensioner or repatriation status.

The major focus of the empirical studies mentioned above was on the nexus between insurance level and health care use, that is whether a high insurance level caused individuals to 'consume' more health care services, and on the role of income. The NOCNSLT variable for which we report the results here is only one of seven variables that were analysed using count data models.

Some of the characteristics of the raw data on NOCNSLT are as follows: 79·8 per cent of 5190 interview respondents had zero consultations, 15·1 per cent had one consultation and the remainder had up to a maximum of nine consultations. The sample mean is 0·302, the sample variance is 0·637, indicating substantial overdispersion in raw terms. The frequency distribution of NOCNSLT is unimodal.

5.2. Estimators for Count Data Models

A number of models for count data are defined in Table I. These models vary along three dimensions: different functional forms may be specified for the objective function (i.e. the 'assumed' distribution), the mean and the variance of y.

First, the objective functions are based on the normal, Poisson or negative binomial distributions.

Secondly, the mean is almost always parameterized as $f(\mathbf{X}, \boldsymbol{\beta}) = \exp(\mathbf{X}\boldsymbol{\beta})$ to ensure that it is positive. The exception is the OLS model where we let $f(\mathbf{X}, \boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$ for the normal

Table I. Some models for count data

	Model	Distribution	Mean	Variance
1.	OLS	Normal	$\mathbf{X}\boldsymbol{\beta}$	α
2.	Normal	Normal	$\exp(\mathbf{X}\boldsymbol{\beta})$	α
3.	Poisson	Poisson	$\exp(\mathbf{X}\boldsymbol{\beta})$	$\exp(\mathbf{X}\boldsymbol{\beta})$
4.	Negbin I	Negative binomial	$\exp(\mathbf{X}\boldsymbol{\beta})$	$(1+\alpha)\exp(\mathbf{X}\boldsymbol{\beta})$
5.	Negbin II	Negative binomial	$\exp(\mathbf{X}\boldsymbol{\beta})$	$\exp(\mathbf{X}\boldsymbol{\beta})(1+\alpha\exp(\mathbf{X}\boldsymbol{\beta}))$

Table II. Some estimators for count data

	Model		Estimator	$\hat{\boldsymbol{\beta}}$	Var($\hat{\boldsymbol{\beta}}$)
1.	OLS	1a	ML	1a	1a
		1b	QGPML	Same as 1a	Same as 1a
		1c	PML	Same as 1a	1c
2.	Normal	2a	ML	2a	2a
		2b	QGPML	Same as 2a	Same as 2a
		2c	PML	Same as 2a	2c
3.	Poisson	3a	ML	3a	3a
		3b	QGPML	Does not exist	—
		3c	PML	Same as 3a	3c
4.	Negbin I	4a	ML	4a	4a
		4b	QGPML	4b	4b
		4c	PML	4c	4c
5.	Negbin II	5a	ML	5a	5a
		5b	QGPML	5b	5b
		5c	PML	Same as 4c	Same as 4c

distribution to include OLS estimation. A positive mean is required for ML (though not PML and QGPML) estimation based on the Poisson and negative binomial distributions.

Thirdly, the variance varies according to the distribution specified and the ways in which overdispersion might be modelled. For OLS and Normal models the customary specification of constant variance is made. For the Poisson model, by definition the variance equals the mean. Negbin I and Negbin II have been discussed in section 2.3.

For each model three types of estimators for $\boldsymbol{\beta}$ may be possible. The MLE assumes that the distribution, mean and variance are correctly specified. The PMLE assumes that only the mean may be correctly specified, and sets the nuisance parameter α equal to an arbitrary constant. The properties of these estimators have been discussed in section 3.2. However, some of these estimators overlap, leading to a considerable economy in computation and reporting of results. This is summarized in Table II.

First, note that a QGPMLE cannot be constructed for the Poisson model because there is no nuisance parameter. Secondly, OLS, Normal and Poisson are all generalized linear models, see section 3.2. It can be shown that for these models the solution for $\boldsymbol{\beta}$ does not depend upon α, so ML, QGPML and PML estimators of $\boldsymbol{\beta}$ for any one model are equivalent and the asymptotic variance–covariance matrices of the MLE and QGPMLE will be equivalent. Thirdly, Negbin I and Negbin II are not generalized linear models, so there is no overlap in estimation. Fourthly, the PML estimators for Negbin I and Negbin II are equivalent since only the knowledge of the specified mean is used.

For the QGPML estimators we need consistent estimators of the nuisance parameters for two reasons: (1) to form the actual QGPMLE for $\boldsymbol{\beta}$ for those models which are not generalized linear models in the sense of McCullagh and Nelder (1983), and (2) to form the estimated variance–covariance matrix of the QGPMLE for $\boldsymbol{\beta}$. We proceed as follows. First, obtain the PMLE $\hat{\boldsymbol{\beta}}$ for the model under consideration. Secondly, since $E[(y_i - f(\mathbf{X}_i, \boldsymbol{\beta}))^2] = \boldsymbol{\Omega}(\mathbf{X}_i, \boldsymbol{\beta}, \alpha)$, we can estimate α by regression based on the relation $(y_i - f(\mathbf{X}_i, \hat{\boldsymbol{\beta}}))^2 = \boldsymbol{\Omega}(\mathbf{X}_i, \hat{\boldsymbol{\beta}}, \alpha) + \varepsilon_i$, where ε_i is an independent error with zero mean. For $\boldsymbol{\Omega}(\mathbf{X}_i, \boldsymbol{\beta}, \alpha) = \alpha$ we obtain

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(\mathbf{X}_i, \hat{\boldsymbol{\beta}}))^2$$

and specializing to the case $f(\mathbf{X}_i, \boldsymbol{\beta}) = \exp(\mathbf{X}_i\boldsymbol{\beta})$, for $\boldsymbol{\Omega}(\mathbf{X}_i, \boldsymbol{\beta}, \alpha) = \exp(\mathbf{X}_i\boldsymbol{\beta})(1 + \alpha \exp(\mathbf{X}_i\boldsymbol{\beta}))$ we obtain

$$\hat{\alpha} = \frac{\sum_{i=1}^{N} (\exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}))^2\{(y_i - \exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}))^2 - \exp(\mathbf{X}_i\hat{\boldsymbol{\beta}})\}}{\sum_{i=1}^{N} (\exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}))^4}$$

whereas for $\boldsymbol{\Omega}(\mathbf{X}_i, \boldsymbol{\beta}, \alpha) = (1 + \alpha) \exp(\mathbf{X}_i\boldsymbol{\beta})$ we obtain

$$\hat{\alpha} = \frac{\sum_{i=1}^{N} \exp(\mathbf{X}_i\hat{\boldsymbol{\beta}})\{(y_i - \exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}))^2 - \exp(\hat{\mathbf{X}}_i\boldsymbol{\beta})\}}{\sum_{i=1}^{N} (\exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}))^2}$$

When an estimate of α is needed for the asymptotic variance–covariance matrix of the QGPMLE, we again use the above formulae for the estimator for α, but base the actual estimate on the QGPMLE for $\boldsymbol{\beta}$ rather than the PMLE for $\boldsymbol{\beta}$.[†]

Column (1) of Table III gives OLS estimates and standard errors corresponding to 1a and 1b of Table II. Column (2) corresponds to estimators 3a and 3b of Table II; columns (3) and (4) to 4a and 5a and column (5) to the QGPML estimator 5b. Column (7) gives the standard errors of the PML estimator for the OLS model. These are the same as Whites (1980) heteroscedasticity consistent standard errors. Column (8) corresponds to estimators 2a and 2b. Column (6) refers to the ordinal probit model which is discussed later in this section. We do not report results for estimators 2c, 3c and 4c.

5.3. Analysis of Results

The estimates in column (1) of Table III are of interest because the first step in the analysis is often estimation by ordinary least squares. Though this regression is significant as a whole it shows a negligible role for either INCOME or the three insurance level dummies. In the spirit of residual analysis consider the least squares residuals $\hat{u}_i = y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}$ which are used to estimate the following regressions (standard errors in parenthesis are obtained using Eicker–White heteroscedastic consistent estimator):[‡]

$$\hat{u}_i^2 = \underset{(0{\cdot}2034)}{2{\cdot}2158}\,(\mathbf{X}_i\hat{\boldsymbol{\beta}}) \tag{21}$$

$$\frac{\hat{u}_i^2}{(\mathbf{X}_i\hat{\boldsymbol{\beta}})} = \underset{(0{\cdot}2737)}{1{\cdot}1962} + \underset{(0{\cdot}3996)}{0{\cdot}6717}\,(\mathbf{X}_i\hat{\boldsymbol{\beta}}) \tag{22}$$

[†] The estimators here follow suggestions by GMT (1984b). Alternatively when the model is a GLM we could follow the method of McCullagh and Nelder (1983) and use an estimate of α based on the generalized Pearson χ^2 statistic. When $\boldsymbol{\Omega}(\mathbf{X}_i, \boldsymbol{\beta}, \alpha) = \alpha$ we get the same estimate as above, but when $\boldsymbol{\Omega}(\mathbf{X}_i, \boldsymbol{\beta}, \alpha) = (1 + \alpha)\exp(\mathbf{X}_i\boldsymbol{\beta})$, we obtain

$$\widehat{(1 + \alpha)} = \frac{1}{N - K} \sum_{i=1}^{N} \frac{(y_i - f(\mathbf{X}_i, \hat{\boldsymbol{\beta}}))^2}{f(\mathbf{X}_i, \hat{\boldsymbol{\beta}})}$$

[‡] These tests serve only as a crude guide. $\mathbf{X}_i\hat{\boldsymbol{\beta}}$ possibly less than zero causes obvious problems. Even abstracting from this, the use of $\hat{\boldsymbol{\beta}}$ rather than $\boldsymbol{\beta}$ in (21) and (22) will lead to disturbances which are not only heteroscedastic but also correlated. For a more rigorous treatment see Cameron and Trivedi (1985).

Table III. Alternative estimates for NOCNSLT equation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
					QGPML			
			Negbin	Negbin	based on	Ordinal	PMLE	
	OLS	Poisson	I	II	Negbin II	Probit†	for OLS	Normal
One	0·0276	−2·2244	−0·6270	−2·1902	−2·1958	−1·390	0·0276	−2.23
		(0·1443)	(0·2253)	(0·2224)	(0·1833)	(0·1480)	(0·0733)	(0·1056)
SEX	0·0338	0·1570	0·1638	0·2164	0·1992	0·1301	0·0338	−0.0569
	(0·0216)	(0·0406)	(0·0602)	(0·0659)	(0·0541)	(0·0436)	(0·0229)	(0·0185)
AGE	0·2032	1·0547	0·2769	−0·2207	0·2535	−0·5784	0·2032	3·5651
	(0·4100)	(0·7499)	(1·1257)	(1·2334)	(1·0057)	(0·8226)	(0·4470)	(0·4320)
AGESQ	−0·0621	−0·8466	0·0223	0·6137	0·0680	0·9159	−0·0621	−3·6120
	(0·4587)	(0·8092)	(1·1190)	(1·3801)	(1·1152)	(0·9009)	(0·5143)	(0·4285)
INCOME	−0·0573	−0·2048	−0·1345	−0·1422	−0.1613	−0·0537	−0·0573	−0·3933
	(0·0331)	(0·0619)	(0·0957)	(0·0976)	(0·807)	(0·0663)	(0·0348)	(0·0388)
LEVYPLUS	0·0352	0·1230	0·2127	0·1191	0·1124	0·1418	0·0352	0·2188
	(0·0249)	(0·0560)	(0·0842)	(0·0849)	(0·0699)	(0·0558)	(0·0217)	(0·0392)
FREEPOOR	−0·1033	−0·4412	−0·5379	−0·4978	−0·4731	−0·3347	−0·1033	0·2067
	(0·0525)	(0·1163)	(0·2093)	(0·1750)	(0·1448)	(0·1257)	(0·0476)	(0·0821)
FREEREPAT	0·0332	0·0799	0·2086	0·1458	0·1139	0·1874	0·0332	0·0016
	(0·0382)	(0·0701)	(0·1038)	(0·1174)	(0·0948)	(0·0757)	(0·0433)	(0·0901)
ILLNESS	0·0599	0·1870	0·1959	0·2145	0·2035	0·1523	0·0599	0·1391
	(0·0084)	(0·0142)	(0·0206)	(0·0257)	(0·0202)	(0·0163)	(0·0099)	(0·0066)
ACTDAYS	0·1032	0·1268	0·1123	0·1437	0·1359	0·0998	0·1032	0·1209
	(0·0036)	(0·0035)	(0·0056)	(0·0075)	(0·0053)	(0·0055)	(0.0097)	(0·0032)
HSCORE	0·0170	0·0301	0·0358	0·0381	0·0354	0·0821	0·0170	0·0229
	(0·0052)	(0·0074)	(0·0105)	(0·0143)	(0·0110)	(0·0091)	(0·0077)	(0·0028)
CHCOND1	0·0044	0·1142	0·1326	0·0997	0·1013	0·0624	0·0044	0·0819
	(0·0237)	(0·0515)	(0·0746)	(0·0766)	(0·0644)	(0·0499)	(0·0222)	(0·0359)
CHCOND2	0·0416	0·1417	0·1742	0·1905	0·1691	0·1402	0·0416	−0·0507
	(0·0359)	(0·0586)	(0·0890)	(0·0948)	(0·0771)	(0·0657)	(0·0463)	(0·0378)
Variance			0·4551	1·0766	0·4899			
parameter α			(0·0405)	(0·0984)				
$-\log L$		3355·542	3226·589	3198·744	3024·437	3140·6		
Iter; Time		12;467	10;592	12;563		20;686		19;1093
R^2	0·20							

Estimated ancillary parameters and their standard errors for the ordinal probit model are: $\hat{\mu}_2 = 0{\cdot}9417$ (0·0313); $\hat{\mu}_3 = 1{\cdot}512$ (0·0518); $\hat{\mu}_4 = 1{\cdot}700$ (0·0627); $\hat{\mu}_5 = 1{\cdot}9420$ (0·0806); $\hat{\mu}_6 = 2{\cdot}079$ (0·0942); $\hat{\mu}_7 = 2{\cdot}330$ (0·1029); $\hat{\mu}_9 = 3{\cdot}299$ (0·4376). †To obtain slopes of the conditional mean function at regressor means, multiply slopes by 1·3526.

The first regression clearly rejects the hypothesis that the coefficient of $(\mathbf{X}_i\hat{\boldsymbol{\beta}})$ is unity which is what we would expect under the Poisson model and suggests that variance is a multiple of the mean which is consistent with the negative binomial model. The second regression is a rough test of whether variance is linear in the mean to suggest whether Negbin I or Negbin II is preferable. The coefficient of $(\mathbf{X}_i\hat{\boldsymbol{\beta}})$ is now positive and just significant on a one-sided test at 5 per cent significance level and the intercept is not significantly different from unity, indicating that there is some support for the second over the first parameterization.

Next the Poisson model was estimated notwithstanding that preliminary analysis favoured the negative binomial. These results are given in column (2). The estimated Poisson model was used to apply Lee's score test of the null hypothesis of Poisson versus the alternatives of a number of the general Katz family of frequency distributions (including the negative binomial) with variance $\exp(\mathbf{X}\beta)(1 + \alpha \exp(\mathbf{X}\boldsymbol{\beta}))$. The estimated T-statistic was 29·89 compared with the critical value at 5 per cent significance level of 1·96, which strongly rejects the Poisson.

Which parameterization of the negative binomial should one choose? Though we recognize that issues of power of tests need to be investigated, proceeding in a data-analytic spirit, we used Poisson estimates to calculate parallel regression equations to (21) and (22). Define $\hat{E}(y_i) = \hat{\lambda}_i$ and $\hat{V}(y_i) = (y_i - \hat{\lambda}_i)^2$. We obtained

$$\hat{V}(y_i) = \underset{(0\cdot069)}{2\cdot2180}\,\hat{\lambda}_i \tag{23}$$

$$\frac{\hat{V}(y_i)}{\hat{\lambda}_i} = \underset{(0\cdot1051)}{1\cdot0569} + \underset{(0\cdot212)}{0\cdot888}\,\hat{\lambda}_i \tag{24}$$

(23) again confirms that the negative binomial is preferred to the Poisson and (24) suggests, more positively than did (22), that the second rather than the first parameterization of the negative binomial is appropriate. We have indicated already our preference for the second parameterization on grounds of computational efficiency.

All tests which we have applied lead to the rejection of the Poisson model. Although this does not imply that the Negbin I and II are automatic alternative choices, they are favoured as preferred alternatives to the Poisson model.

The negative binomial model estimates are given in columns (3) and (4). The Wald statistics ($\mathcal{N}(0, 1)$) for testing Poisson against Negbin I and Negbin II are respectively 11·23 (=0·4551/0·0405) and 10·84. The corresponding likelihood ratio test statistics ($\chi^2(1)$) are, respectively, 257·90 (=2 × (3355·54 − 3226·59)) and 313·58. Again the Poisson model is strongly rejected. Note that squaring the score and Wald tests (so that $\chi^2(1)$) yields score > LR > Wald, which is the reverse of the well-known analytical results for the normal linear regression model. The log-likelihood associated with Negbin II is higher. Negbin II is also slightly more efficient computationally, something that was independently confirmed on other samples. We think this favours the second over the first parameterization of the negative binomial and we return to this issue later.

We now compare the actual parameter estimates in Columns (1)–(4). We are particularly interested in whether a higher insurance level is associated with more visits to a doctor. Although both sets of estimates yield similar answers, and are notably different from those produced by the OLS results, there are differences of detail. The LEVYPLUS dummy coefficient is almost twice as large for Negbin I as for Negbin II.

The similarity of the Poisson point estimates to those in the Negbin models is rather striking, but note that the estimated variances under the Poisson assumption are generally substantially smaller, reflecting the consequences of imposing on the data the equality of conditional mean and variance. This example confirms that the major impact of the distributional assumption is on estimated variances rather than point estimates of parameters. Comparison of OLS with Poisson and Negbin estimates is different because the former model has mean $\mathbf{X}\boldsymbol{\beta}$ and the latter models have mean $\exp(\mathbf{X}\boldsymbol{\beta})$.

We next relax the negative binomial assumption in the Negbin II model to the assumption that only the mean and variance are correctly specified. QGPMLE based on Negbin II uses the 2-step procedure given in section 5.2. This leads to a higher value of 'likelihood' and the estimated standard errors are somewhat smaller than under the Negbin assumptions. However, most of the substantive inferences, for example, regarding the role of insurance status, would be the same in either case. So in some respects the choice between the two estimators is not critical.

In column (7) we give the PMLE for the OLS model as an example in which the MLE may be consistent despite missspecification—the parameter estimates in columns (1) and (7) are identical—but the standard errors are inconsistently estimated—all but two standard errors in column (1) are understated (by roughly 10–20 per cent).

The estimates obtained under the assumption of a Normal model, given in column (8), are in certain respects markedly different from those for Negbin II. This is especially so in the case of coefficients of INCOME, FREEPOOR and CHCOND2 variables.

In terms of computing time, estimation is cheapest under the Poisson assumption and least so under the Normal assumption. Savings in computer time resulting from the Poisson assumption could be greater still if an efficient special purpose computer program such as GLIM were to be employed (see the Appendix). The overall impression obtained from this example is that, at least with 5190 observations, relatively few coefficients are sensitive to the choice of estimator and they are the ones with relatively small '*t*-ratios'.

5.4. Ordinal Probit as an Alternative to Count Data Models

Although we could have stopped at this stage, we felt that the sensitivity of our conclusions to the choice of the model needed further probing. Our motivation for this is somewhat specific to our example but will extend to other situations in which count data models are used. This is that even though NOCNSLT seems a natural cardinally measured variable it could be validly treated as an ordinal measure of the use of doctor's services.† Thus, for example, three doctor-consultations represents a higher level of doctor usage than two consultations, but not necessarily 50 per cent more. In the context of the HHG (1984) paper, *n* patents may represent a higher level of R & D activity but not necessarily double that associated with $n/2$ patents. Inverting an argument due to McKelvey and Zavoina (1975, p. 103), we might regard an observed variable which is of count form as reflecting a methodological limitation in collecting data, and being no more than a proxy, measured on a crude ordinal scale, for the true unobserved variable which the model is intended to predict. (In our case we are interested in the use of doctor's services but not the number of consultations *per se*). From this viewpoint the emphasis on embedding the model in a parametric family of discrete distributions appears somewhat misplaced. It may be better to use a statistical model suitable for analysing ordinal level dependent variables. McKelvey and Zavoina (1975) have developed such a model, known as the ordinal probit model, as an extension of the dichotomous probit model. See Maddala (1983) for an exposition.

† This suggestion was made to us by James Heckman.

In using this model for our purposes we treat the observed counted variable Y_i, as a proxy for the variable of theoretical interest, Y_i^*, which by assumption is assumed to be distributed as $\mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2)$. $Y_i(i = 1, \ldots, N)$ is treated as a categorical variable with M response categories $R_1, \ldots, R_M$, related to the unobserved variable, Y_i^*, as follows. Let $\mu_0, \mu_1, \ldots, \mu_M$ denote $M + 1$ real numbers with $\mu_0 = -\infty$, $\mu_M = +\infty$ and $\mu_0 \leqslant \mu_1 \leqslant \ldots \leqslant \mu_M$, such that

$$Y_i \in R_k \Leftrightarrow \mu_{k-1} < Y_i^* < \mu_k$$

for $1 \leqslant k \leqslant M$.

Since Y is treated as ordinal (categorical) it can be represented as a series of dummy variables by defining

$$Y_{ik} = \begin{cases} 1 \text{ if } Y_i \in R_k \\ 0 \text{ otherwise} \end{cases}$$

where $i = 1, \ldots, N$; $k = 1, \ldots, M$.

Denoting the cumulative standard normal density by $\Phi(t)$ and imposing the identifying restrictions $\mu_1 = 0$, $\sigma = 1$, the ordinal probit model leads to the probability function

$$\Pr[Y_{ik} = 1] = \Phi[\mu_k - \mathbf{X}\boldsymbol{\beta}] - \Phi[\mu_{k-1} - \mathbf{X}\boldsymbol{\beta}]$$

which forms the basis of maximum likelihood estimation of parameters $\mu_1, \ldots, \mu_{M-1}$ and the vector $\boldsymbol{\beta}$.

Finally, therefore, it is interesting to compare the earlier estimates of our model for NOCNSLT with those based on maximum likelihood estimation of the ordinal probit model. The estimate of $\boldsymbol{\beta}$ is in column (6) and the estimate of μ is at the bottom of Table III. As they stand, columns (2)–(5) are not directly comparable with column (6). However, the inferences regarding the qualitative influence of sex, income, health insurance status, health status and chronic conditions on NOCNSLT would be almost exactly the same as those based on the estimated count data models. This does increase our confidence in the robustness of our substantive conclusions and hence seems a worthwhile final check of the results. But note also that the ordinal probit estimation has provided additional interesting information on the μ parameters, which all have rather small standard errors. The observed counts Y_i range from 0 to 9 whereas the corresponding unobserved theoretical values Y_i^* are calibrated by the estimated $\mu_k(k = 0, \ldots, 9)$ which range between 0 (which is a normalized value) and 3·30 That is, individuals falling below $\mu_1 = 0$ record a zero count, those between $\mu_1 = 0$ and $\mu_2 = 0{\cdot}9417$ record one count, those between $\mu_2 = 0{\cdot}9417$ and $\mu_3 = 1{\cdot}512$ record two counts and so forth. This provides a further useful calibration of differential propensities of individuals for consultations with doctors. (Since the ordinal probit model involves an additional eight parameters, comparison of fit between it and other alternatives should be based on a modified log-likelihood ratio such as that proposed by Akaike.)

6. CONCLUSIONS

In common with many econometric problems, the specification of count data models raises fundamental questions about the underlying objectives of the analysis; that is, whether a reduced form or a structural interpretation is desired. Within the context of single period cross-section data, only reduced form type models would appear to be feasible. From the viewpoint of estimation a number of available models can be usefully treated as non-linear regression models with specific conditional mean–variance or heteroscedasticity structures. To some extent flexibility in specification can be achieved through the specification of such heteroscedasticity structures. If the matter is viewed this way, then it seems possible to develop a sequential modelling strategy based on some simple tests, in which one can proceed to increasingly flexible and data-coherent models, beginning with the basic Poisson model. We have provided a detailed application of this modelling strategy to illustrate both its simplicity and feasibility. Our results are broadly supportive of the QGPMLE procedure advocated by GMT but we also advocate further exploration of suitable categorical variable models as an alternative approach. Additional work remains to be done on the power of various test procedures and the extension of the results on non-nested model comparisons to count data models.

ACKNOWLEDGEMENTS

This paper originated in an applied econometric project on the demand for health insurance and health care in Australia. We are grateful to the Commonwealth Department of Health and the A.N.U. Faculties Research Fund for financial support and to James Heckman, Tony Lancaster, Lung-Fei Lee, Grayham Mizon, Hashem Pesaran and anonymous referees for this journal for helpful comments and suggestions for improvements.

APPENDIX

The negative binomial models were estimated using the optimization routine for user-defined functions in the LIMDEP package of Greene (1983). This package also allowed estimation by NLLS and ordinal probit methods. The iterative method of Berndt *et al.* (1974) was used with analytical first derivatives provided. Considerable computational savings arise by using the recursion relations for the gamma and digamma functions, rather than the power series expansions referred to by Gilbert (1979), Hausman *et al.* (1984) and Gourieroux *et al.* (1984). For example, the ratio $\Gamma(\nu_i + y_i)/\Gamma(\nu_i)$ in (7) equals $\Pi_{j=1}^{y_i}(\nu_i + j - 1)$. Note also that the remaining gamma term $\Gamma(y_i + 1)$ is easily computed since $y_i + 1$ is an integer, and does not need to be recalculated at each iteration. A number of other models were also estimated under the negative binomial assumption (Cameron *et al.*, 1984). Computational time on a Univac 1100/82 for $N = 5190$ and 13 parameters was between 480 and 640 CPU seconds, with I/O time one to two times CPU time. Computational time per iteration doubled as N doubled and roughly doubled as M quadrupled. When ordinal probit was used instead CPU time was between 25 and 40 per cent greater.

Program LIMDEP was also used for the Poisson model. The Poisson model took fewer iterations and about two-thirds the CPU time of comparable negative binomial models. In addition, the Poisson model was estimated using program GLIM. Computational time was considerably less—about one-third that of LIMDEP—because GLIM takes advantage of the special structure of exponential family distributions such as the Poisson to estimate the model by iterative weighted least squares (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1983). Although GLIM macros exist for the negative binomial model, they are unfortunately not computationally efficient because the negative binomial distribution is only a member of the exponential family when ν is known.

REFERENCES

Amemiya, T. *Advanced Econometrics* (1985), Harvard University Press, Cambridge.

Bates, G. E. and J. Newman (1952), 'Contributions to the theory of accident proneness', *University of California Publications in Statistics*, **1**, 215–275.

Berndt, E. R., B. H. Hall, R. E. Hall and J. A. Hausman (1974), 'Estimation and Inference in Nonlinear Structural Models', *Annals of Economic and Social Measurement*, **3**, 653–666.

Boswell, M. T. and G. P. Patil (1970), 'Chance mechanisms generating the negative binomial distributions', in G. Patil (ed.), *Random Counts in Models and Structures: Volume 1*, The Pennsylvania State University Press, Pennsylvania.

Burguette, J., R. Gallant and G. Souza (1982), 'On unification of the asymptotic theory of nonlinear econometric models', *Econometric Reviews*, **1**, 151–190.

Cameron, A. C. and P. K. Trivedi (1985), 'Regression based tests for overdispersion', Technical Report No. 9, Econometric Workshop, Stanford University.

Cameron, A. C., P. K. Trivedi, F. Milne and J. Piggott (1984), 'Microeconometric models of the demand for health insurance and health care in Australia: II', Australian National University, *Working Papers in Economics and Econometrics*, No. 106.

Cox, D. R. (1983), 'Some remarks on overdispersion', *Biometrika*, **70**, 269–274.

Cox, D. R. and D. V. Hinkley (1974), *Theoretical Statistics*, Chapman and Hall, London.

Cox, D. R. and P. A. W. Lewis (1966), *The Statistical Analysis of Series of Events*, Methuen, London.

Cresswell, W. L. and P. Froggatt (1963), *The Causation of Bus Drivers Accidents*, Cambridge University Press, London.

Eicker, F. (1967), 'Limit theorems for regressions with unequal and dependent errors', *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley and Los Angeles.

Gilbert, C. L. (1979), 'Econometric models for discrete economic processes', *Discussion Paper*, University of Oxford, presented at the Econometric Society European Meeting, Athens.

Gourieroux, C., A. Monfort and A. Trognon (1984), 'Pseudo maximum likelihood methods: theory', *Econometrica*, **52**, 681–700.

Gourieroux, C., A. Monfort and A. Trognon (1984), 'Pseudo maximum likelihood methods: applications to Poisson models', *Econometrica*, **52**, 701–720.

Gourieroux, C., A. Monfort and A. Trognon (1984), 'Pseudo-likelihood methods: a Survey', *Document de travail ENSAE/INSEE*, No. 8406.

Greene, W. H. (1983), 'LIMDEP: a program for estimating the parameters of qualitative and limited dependent variables', *The American Statistician*, **37**, 170.

Hausman, J., B. H. Hall and Z. Griliches (1984), 'Econometric models for count data with an application to the patents—R & D relationship', *Econometrica*, **52**, 909–938.

Heckman, J. and B. Singer (1984), 'Econometric models of duration', *Journal of Econometrics*, **24**, 63–132.

Heckman, J. and G. Borjas (1980), 'Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model for heterogeneity and state dependence', *Econometrica*, **47**, 247–283.

Hendry, D. D. and K. F. Wallis, (eds) (1984), *Econometrics and Quantitative Economics*, Basil Blackwell, Oxford and New York.

Huber, P. (1967), 'The behaviour of maximum likelihood estimates under nonstandard conditions', in *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, University of California Press, Berkeley and Los Angeles.

Irwin, J. O. (1941), 'Discussion on Chambers and Yule's paper', *Journal of the Royal Statistical Society*, Supplement 7, 101–109.

Jennrich, R. (1969), 'Asymptotic properties of nonlinear least squares estimators', *The Annals of Mathematical Statistics*, **40**, 633–643.

Johnson, N. L. and S. Kotz (1972), *Discrete Distributions*, Wiley, New York.

Katz, L. (1963), 'Unified treatment of a broad class of discrete probability distributions', *Proceedings of the International Symposium on Discrete Distributions*, Montreal, pp. 172–182.

Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*, Wiley, New York.

Lee, L-F. (1984), 'Specification tests for Poisson regression models', Department of Economics, University of Minnesota, unpublished paper.

Maddala, G. S. (1983), *Limited-dependent and Qualitative Variables in Econometrics*, Cambridge University Press.

McCullagh, P. (1983), 'Quasi-likelihood functions', *The Annals of Statistics*, **11**, 59–67.

McCullagh, P. and J. A. Nelder (1983), *Generalised Linear Models*, Chapman and Hall, London.

McKelvey, R. D. and W. Zavoina (1975), 'A statistical model for the analysis of ordinal level dependent variables', *Journal of Mathematical Sociology*, **4**, 103–120.

Mossiman, J. E. (1970), 'Compound multinomial distributions', in G. P. Patil (ed.), *Random Counts in Scientific Work*, **3**, Pennsylvania State University Press, Pennsylvania, Chapter 1.

Nelder, J. A. and R. W. Wedderburn (1972), 'Generalised linear models', *Journal of the Royal Statistical Society, Series B*, **135**, 370–384.

Ord, J. K. (1972), *Families of Frequency Distributions*, Griffin, London.

Patil, G. P. (1970), *Random Counts in Models and Structures: Volume 1*, The Pennsylvania State University Press, Pennsylvania.

Trivedi, P. K., C. Cameron, F. Milne and J. Piggott (1984), 'Microeconometric models of the demand for health insurance and health care in Australia: I', Australian National University, *Working Papers in Economics and Econometrics*, No. 105.

Wedderburn, R. W. M. (1974), 'Quasi-likelihood functions, generalised linear models and the Gauss–Newton method', *Biometrika*, **61**, 439–447.

White, H. (1980), 'A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity', *Econometrica*, **48**, 817–838.

White, H. (1982), 'Maximum likelihood estimation of misspecified models', *Econometrica*, **50**, 1–25.

Xekalaki, E. (1983), 'The univariate generalised Waring distribution in relation to accident theory: proneness, spells or contagion?', *Biometrics*, **39**, 887–895.