New Econometric Models for Longitudinal Count Data with an Excess of Zeros: Two Applications in Health Economics.

Jean-Eric Tarride

A Thesis in the

Department of Economics

Presented in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy at

Concordia University

Montreal, Quebec, Canada

March 2004

Your file Votre référence
ISBN: 0-612-90405-9
Our file Notre référence
ISBN: 0-612-90405-9

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: **Jean-Eric Tarride**

Entitled: **New Econometrics Models for Longitudinal Count Data with an Excess of Zeros: Two Applications in Health Economics**

and submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY (Economics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. R. Hall _____Chair

Dr. J. LeLorier _____ External Examiner

Dr. A. Hochstein _____ External to Program

Dr. M. Sampson _____ Examiner

Dr. J. Hansen _____Examiner

Dr. J. McIntosh _____Thesis Supervisor

Approved by _____
Chair of Department or Graduate Program Director

Feb 2 2004

_____
Dean of Faculty

ABSTRACT

New Econometric Models for Longitudinal Count Data with an Excess of Zeros: Two Applications in Health Economics.

Jean-Eric Tarride, Ph.D.

Concordia University, 2004

The purpose of this doctoral thesis is to provide new econometric models to analyze longitudinal count data characterized by a high proportion of zeros in the data. Previous econometric studies have dealt with many characteristics such as the discrete and longitudinal aspects of the dependent count variable or the presence of covariates and unobserved individual heterogeneity. However, none have taken into account the issues associated with an excess of zeros in a longitudinal framework. While it is well known in the univariate case that when an excess of zeros is significant, the mean has to be corrected to take into account this feature of the data, this issue has often been ignored in the longitudinal case. An excess of zeros in the data may lead to important modeling issues associated with the analysis of longitudinal count data.

Two new econometric models are presented to address the six following characteristics: 1) count outcome 2) a limited number of repeated measurements, 3) presence of covariates, 4) unobserved heterogeneity, 5) presence of correlation due to the repeated nature of the data and 6) an excess of zeros.

The first model, a Quadrivariate Negative Binomial Hurdle model, was developed to analyze the number of doctor visits made by a panel of more than 4,000 German

followed over 4 years. In the second example, a Quadrivariate Negative Binomial Zero-Inflated model was used to analyze an unpublished subset of a longitudinal clinical trial in which the treatments were very effective in reducing the number of occurrences of one variable collected over time in this trial. These two new models were nested to the Quadrivariate Negative Binomial model, allowing us to test for an excess of zeros.

The main result is that the excess of zeros was significant in our two examples and assuming that only one process generates the data is incorrect. As such, the Multivariate Negative Binomial Hurdle and Zero-Inflated models are superior than standard Univariate Negative Binomial model, Quadrivariate Negative Binomial model and Generalized Estimating Equations model. These new models performed well in predicting the mean counts and the mean proportion of zeros in the data at each time period. This thesis demonstrated that caution should be taken in analyzing longitudinal count data in the presence of a high proportion of zeros in the data and correlation over time. Models ignoring these features may yield inconsistent estimates.

ACKNOWLEDGMENTS

I am grateful to my director, Professor James McIntosh, who always took the time to guide and encourage me throughout this doctoral thesis.

I would like to thank my wife Lori for accompanying me everyday over the last four years in this adventure.

A big smile to my daughter Madeleine for humoring me everyday! All of this would have not been possible without my grandparents Madeleine and Marcel Tarride.

TABLE OF CONTENTS

LISTS OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Longitudinal count data or panel count data refers to a series of non-negative integers measured over several time periods for individual units such as persons, households, firms and regions. Economic examples include the annual number of patents generated over time by a sample of firms (Haussman et al. 1984; Cincera 1997; Crepon and Duguet 1997; Montalvo, 1995), the number of days of absence of a cohort of workers followed over several years (Ruser, 1991; Wagner et al., 1993) or the number of insurance claims during a certain period of time (Pinquet, 1998). Applications are common in Health Economics and include the analysis of the number of physician or hospital visits generated by a panel of households (Geil et al., 1997; Winkelman, 2001) or the analysis of the number of occurrences of a specific event over a certain period of time (Diggle, 1995; Albert, 2000). By following firms or individuals over time, longitudinal count data offers a richer framework than cross-section data.

The purpose of this doctoral thesis is to provide new econometric models to analyze longitudinal count data characterized by a high proportion of zeros. Previous econometric studies have dealt with many characteristics associated with longitudinal count data such as the discrete and longitudinal aspects of the dependent count variable or the presence of covariates and unobserved individual heterogeneity, but none have taken into account the issues associated with an excess of zeros in a longitudinal case. However, a zero value is a natural outcome for firms who do not patent (voluntary or involuntary) or for individuals who do not seek medical care. This may lead to important modelling issues associated with the analysis of longitudinal count

data characterized by a high proportion of zeros in the data. Although it is well recognized in the univariate case that not taking into account the presence of extra zeros will result in inconsistent estimates because the mean function would not be correctly specified (Mullhahy 1997; Cameron and Trivedi 1998; Greene, 2000), this issue is typically ignored in longitudinal count data. Solving this problem by "choosing companies so as to minimize the problem ("only 8 per cent of the observations were zeros in any one year") as in Haussman et al. (1984, page 910), has several limitations such as introducing a selection bias. For example, such methods may not be suitable in analyzing the number of patents awarded to small or medium firms which are less likely to patent innovations than larger firms.

Instead it is important to develop new models for longitudinal count data which are able to test and treat for an excess of zeros (i.e., a high proportion of "zero" outcome in the data) in the longitudinal context. In some cases this excess of zeros is significant and should be taken into account by assuming for example that two processes generates the data instead of one. These models must also address the characteristics inherent in longitudinal count data such as repeated observations of a count outcome, the presence of covariates, correlation due to the repeated nature of the data and unobserved heterogeneity. In order to generalize the findings of this research, these new models will be compared with traditional approaches used for the analysis of longitudinal count data in health economics through two applications.

Chapter 2 reviews the traditional econometric approaches used to analyze longitudinal count data. Parametric and non-parametric models which address the discrete

and repeated aspects of the data are presented. In these models the covariates are introduced through the mean function and the correlation is treated by deriving fixed effects or by conditioning. Non-parametric models for longitudinal count data assume that while not specified, the a-priori "distribution" of the dependent variable is from the linear exponential family and/or that the mean function is correctly specified. In some cases these two assumptions may be violated. Generalized Auto-Regressive (GAR) models or time series models have also been introduced to model correlation in time series of count data. When only a limited number of repeated observations is available for analysis it is not known in Generalized Auto-Regressive models how to treat the initial value. Before this problem is resolved, it is not possible to test these models in the same way as a Prais-Winsten approach in the linear regression model.

This survey of the literature indicates that the problem associated with an excess of zeros (i.e. high proportion of zeros) in longitudinal count data is traditionally ignored. Recently a model was developed by Chin-Shang et al. (1999) to take into account an excess of zeros in the univariate case in which three count variables were observed at the same time. Unfortunately this methodology is not generalizable when the number of repeated measurements increases over time. The review of the literature supports evidence of the need for new econometric models to address all the characteristics associated with longitudinal count data and to allow for the testing and treatment of an excess of zeros in the longitudinal case.

A well known area of research in health economics is the analysis of the usage of medical services for which count data models are widely used. Several studies have

been conducted over the last two decades to identify the determinants of the number of physician visits or to assess the impact of a health policy reform aimed at reducing or controlling health expenditures. The model of reference to explain the determinants of the demand for medical care was developed by Grossman (1972). In most empirical studies which use the Grossman's framework, the demand for medical care is generally defined as the number of medical services consumed (e.g., the number of doctor or hospitals visits) by a sample or panel of individuals. Because some individuals do not visit their physician or do not go to a hospital, a "zero" value is a common outcome. The specific nature of count data with excess zeros is well documented in the analysis of the demand for medical care in the univariate case. Two-part models have been shown to be superior to standard (one-part) models in several applications in health economics using cross-section data.

A review of several recent econometric studies indicated that the vast majority of empirical work conducted in this area of research used cross-section surveys to analyze the usage of medical services. This may be due to the paucity of panel data integrating detailed health related questions. In Canada, the only source of information is the Canadian Health Survey (CHS) which is an annual cross-section of a representative sample of Canadian households. Rochon (2003) used two cross sections (1994 and 1998) of the CHS database to examine the effect of the health care system reforms implemented in Canada during the mid-1990s. However, even when longitudinal data is available for analysis, it is common to find models pooling the data and applying a univariate count distribution to the pooled data. This may be

undesirable if the data is correlated over time. Following a panel of households over time is likely to provide more accurate information on the dynamics of the utilization of medical services rather than pooling different cross sections.

In Chapter 3, a new approach is proposed to analyze the demand for physicians using a panel of German individuals followed over four years. The dependent count variable - the number of doctor visits - is characterized by the presence of a high proportion of zeros and correlation over time. After comparing in the presence of correlation univariate count models applied to a pool of cross-sections and longitudinal count data models, a new methodology is presented based on an extension of the Hurdle model to accommodate an excess of zeros in the longitudinal case. This chapter will also document if determinants of the number of physician visits are different between men and women as shown by Geil et al. (1997) in their analysis of hospital visits in Germany. A separate analysis will also be conducted to model the demand for general practitioners and specialists as they may differ as found in Pohlmeier and Ulrich (1995) in their analysis of the number of physician visits in 1984 in Germany.

Chapter 4 presents a new way of looking at the analysis of longitudinal clinical trial count data. While this area of research is typically conducted by biostatisticians, economists and especially econometricians have a growing role to play in the determination of drug efficacy in order to conduct economic drug evaluations. Pharmacoeconomic or the economic evaluation of pharmaceutical products is a growing area of research which is fueled by increased public regulations aimed at controlling

health expenditure increases. Since 1994, Canada has developed guidelines to conduct pharmacoeconomic studies. It is now mandatory in Canada to provide economic evaluations for any new pharmacotherapy in order to get it approved for public reimbursement. Over the last decade, guidelines, publications and textbooks have flourished in the area of pharmacoeconomics providing guidelines in the conduct of pharmacoeconomic studies. The reader is referred to Drummond et al. (1998) for a comprehensive review of the theory and methods for economic evaluations of health care programs. Unfortunately, the literature on pharmacoeconomics has generally ignored the statistical issues associated with the determination of the efficacy of drug treatments. However, if the design of the statistical analysis is incomplete or incorrect, parameter estimates may be inconsistent. In a worst-case scenario, no treatment differences between a new and old pharmacotherapy will be detected while one treatment is more effective than the other. With this scenario (i.e., no superior clinical profile), the economic profile may not demonstrate the superiority of this new treatment even if the new product is launched at a discount price. From a societal point of view, the economic and humanistic impacts are considerable. The government will lose money because it refused a cost-effective treatment. Patients will be deprived of an improved treatment. The pharmaceutical company which developed the compound will not make a positive return on its investment because this new treatment may be not accepted by public or private formularies for reimbursement. In countries where the public health system is predominant, public listing (i.e. reimbursement) is key in order to offer to the afflicted population free access to a new drug treatment.

In this context, Chapter 4 concentrates on the comparison of the efficacy of treatments which are very effective in preventing the occurrence of a health count outcome (e.g., number of seizures, number of episodes of asthma, etc.). High treatment efficacy translates into a high number of zeros episodes observed among cured patients. This is illustrated through an analysis of an unpublished subset of a clinical trial. The analysis concentrates in a count variable which was one of the secondary endpoints collected in this trial. This dependent count variable is characterized by overdispersion and correlation between consecutive time periods. A new model based on a Quadrivariate Negative Binomial Zero-Inflated distribution, is presented to analyze this clinical trial. Results are compared to the standard methods of analysis to analyze longitudinal count data in clinical trials, including Generalized Estimation Equations models. The last chapter concludes by summarizing the main results and implications of this research in the area of health economics. Limitations associated with the two new models presented in chapters 3 and 4 are discussed before identifying future areas of research.

2 ECONOMETRIC MODELS FOR LONGITUDINAL COUNT DATA

Conventional linear regression models may not be appropriate for longitudinal count data as some basic assumptions such as the normality of the residuals are violated. Instead the discreteness of the count variable and the repeated aspect of the data have to be taken into account by the use of appropriate models for analyzing longitudinal count data. These models also need to accommodate two additional characteristics associated with longitudinal count data, which are the presence of unobserved heterogeneity and the correlation over time arising from the repeated aspect of the data. In addition, in some cases the data can be characterized by a high proportion of zeros which need to be taken into account as well.

This chapter reviews the econometric literature on longitudinal count data with an emphasis on longitudinal count data characterized by a small number of repeated measurements on the count dependent variable. Short longitudinal count data sets are common in the real world and present special features which make them different from pure time series count data. Four types of models to analyze longitudinal count data were identified in the economic literature: 1) Parametric models which extend the Poisson univariate or Negative Binomial univariate distribution-based models to the longitudinal case; 2) Non-parametric models relaxing the assumptions on the distribution of the dependent variables such as Pseudo Maximum Likelihood models or models based on generalized estimations equations; 3) Generalized Auto-Regressive models which include the past value of the dependent count variable in the covariates

and 4) A Multivariate Zero-Inflated Poisson model which has recently been developed and applied to the bivariate and trivariate case. A summary of recent applications of these models is presented before identifying important features of longitudinal count data which have traditionally been underreported in the literature.

2.1 Parametric Models

2.1.1 Poisson Models

Basic Poisson Model The starting point of dealing with discrete non-negative integer values is to consider the simple Poisson regression model. Let y_{it} be the dependent variable which represents the count outcome of individual i at time t, where i=1,...,N indexes individuals and t=0,...,T, indexes time periods. The y_{it}'s are assumed to be independent across individuals and across time and have a Poisson distribution conditional on parameters λ_{it}. To ensure non-negativity, the parameters λ_{it} depend on a set of k explanatory variables, x_{it}, such as:

$$\lambda_{it} = \exp(x_{it}\beta). \tag{1}$$

The term $x_{it}\beta$ is the inner product of the covariate vector for individual i at period t and a parameter vector β which is assumed to be the same for all individuals. The Poisson distribution is given by:

$$P(y_{it} \mid \lambda_{it}) = \frac{\exp(-\lambda_{it})\lambda_{it}^{y_{it}}}{y_{it}!}. \tag{2}$$

The dependent variable is related to the set of explanatory variables through the parameter of the Poisson model λ_{it}. The parameter β can be estimated by the maximum likelihood method. The log-likelihood function $l(\beta)$ to maximize with respect to β is:

$$l(\beta) = -\sum_{i=1}^{N}\sum_{t=0}^{T}\exp(x_{it}\beta) + \sum_{i=1}^{N}\sum_{t=0}^{T} y_{it}x_{it}\beta - \sum_{i=1}^{N}\sum_{t=0}^{T}\ln(y_{it}!). \tag{3}$$

In this equation, $Y = (y_{i1}, ..., y_{iT})$ represents the number of counts from t=0 to T for individual i, and i=1...N. This function $l(\beta)$ is β globally concave, hence uniqueness of the global maximum is ensured. The First-Order Condition (FOC) is given by:

$$\sum_{i=1}^{N}\sum_{t=0}^{T} x_{it}\left[y_{it} - \exp(x_{it}\beta)\right] = 0. \tag{4}$$

The mean and the variance of the Poisson distribution are given by:

$$E(y_{it} \mid x_{it}, \beta) = Var(y_{it} \mid x_{it}, \beta) = \lambda_{it} = \exp(x_{it}\beta). \tag{5}$$

The Poisson distribution assumes a mean-variance ratio equal to unity. However, most of the longitudinal count data sets which have been analyzed in the economic area suggest that the dependent count variable is typically overdispersed (i.e. the variance of the dependent variable is greater than its mean). As is well known, overdispersion is associated with unobserved heterogeneity among individuals.

Unobserved heterogeneity in count data can be modelled by the introduction of an individual specific effect in the parameter of the Poisson distribution. A more general formulation of the Poisson mean considers that conditional on λ_{it} and random parameters α_i, y_{it} is Poisson distributed as:

$$y_{it} \sim P(\alpha_i \lambda_{it}). \tag{6}$$

In this specification, the individual specific effects, α_i, are multiplicative. As previously λ_{it} is defined as $\lambda_{it} = \exp(x_{it}\beta)$. Based on this specification, two approaches have been developed to integrate unobserved heterogeneity in the Poisson model. Fixed effects models assume that the α_i's are unknown parameters to be estimated. In Fixed Effects models, the α_i's are eliminated in the maximization process either by concentrating the likelihood function or by conditioning on the sum of the number of counts over time. Random effects models assume that the α_i's are independently identical distributed (iid) random variables generated by a specific distribution. If a Gamma distribution is taken, this leads to the Multivariate Negative Binomial model as a Poisson-Gamma mixture.

Poisson Fixed Effects (PFE) Model The Poisson Fixed Effects (PFE) model considers that conditional on λ_{it} and parameter α_i, y_{it} is Poisson distributed with mean:

$$E\left(y_{it} \mid \lambda_{it}, \alpha_i\right) = \mu_{it} = \alpha_i \, \lambda_{it} = \alpha_i \exp(x_{it}\beta). \tag{7}$$

α_i is the individual fixed effect, λ_{it} is a specific function of covariates x_{it} and β, such as $\lambda_{it} = \exp(x_{it}\beta)$. The conditional joint density for the i^{th} observation is given by:

$$\begin{aligned}\Pr(y_{i1}, y_{i2}, ..., y_{it} \quad | \quad x_{it}, \beta, \alpha_i) &= \prod_{t=0}^{T} \left[\frac{\exp(-\alpha_i \lambda_{it})(\alpha_i \lambda_{it})^{y_{it}}}{y_{it}!} \right] \\ &= \exp(-\alpha_i \sum_{t=0}^{T} \lambda_{it}) \frac{\prod_t \alpha_i^{y_{it}} \prod_t \lambda_{it}^{y_{it}}}{\prod_t y_{it}!}. \end{aligned} \tag{8}$$

It follows that the log-density function for individual i is defined by:

$$\ln \Pr(y_{i1}, y_{i2}, ..., y_{it} \mid \alpha_i, \beta) = -\alpha_i \sum_{t=0}^{T} \lambda_{it} + \ln \alpha_i \sum_{t=0}^{T} y_{it} + \sum_{t=0}^{T} y_{it} \ln \lambda_{it} - \sum_{t=0}^{T} \ln y_{it}!. \tag{9}$$

Differentiating with respect to α_i and setting the resulting equation equal to zeros defines the First-Order Condition (FOC) of the Poisson Fixed Effects model given by:

$$\hat{\alpha_i} = \frac{\sum_{t=0}^{T} y_{it}}{\sum_{t=0}^{T} \lambda_{it}}. \tag{10}$$

This First-Order Condition is substituted back in the log-density function for individual i given in equation (9) to yield the concentrated log-likelihood function for all the sample,

$$
\begin{aligned}
l(\beta) &= \prod_{i=1}^{N}\left[\exp(-\sum_{t=0}^{T} y_{it}) \prod_{t}\left(\frac{\sum_{t=0}^{T} y_{it}}{\sum_{t=0}^{T}\lambda_{it}}\right)^{y_{it}} \frac{\prod_{t=0}^{T}\lambda_{it}^{y_{it}}}{\prod_{t=0}^{T} y_{it}!}\right] \\
&\propto \prod_{i=1}^{N}\left[\prod_{t=0}^{T}\left(\frac{\lambda_{it}}{\sum_{s=1}^{T}\lambda_{is}}\right)^{y_{it}}\right]. \qquad (11)
\end{aligned}
$$

In this result, the individual effects α_i's have disappeared. Differentiating this expression with respect to β yields the second First-Order-Condition which is independent of the α_i, the fixed effects parameters:

$$
\sum_{i=1}^{N}\sum_{t=0}^{T} x_{it}(y_{it} - \lambda_{it}\frac{\sum_{t=0}^{T} y_{it}}{\sum_{t=0}^{T}\lambda_{it}}) = 0. \qquad (12)
$$

Estimates of β are consistent for fixed T and n $\rightarrow \infty$. This result holds despite the presence of the incidental parameter α_i as discussed by Cameron and Trivedi (1998).

Another approach to estimate the fixed effects, α_i's, in a Poisson model, is the conditional maximum likelihood method which was proposed by Haussman et al. (1984) based on the specification of Anderson (1970). As in the preceding approach, the conditional maximum likelihood method considers that the unobserved heterogeneity is the result of an unobserved fixed effect. The conditional maximum likelihood method allows for the fixed effects to be correlated with the regressors by conditioning on the sum over time of the counts for a given individual. In models with multiplicative effects such as $\mu_{it} = \alpha_i\lambda_{it}$, the conditional maximum likelihood method is straightforward since the sum over time of the counts, $\sum_t y_{it}$, is distributed as Poisson with parameter $\alpha_i \sum_t \lambda_{it}$. Conditioning to the sum of counts over the whole period allows

removal of the individual specific effects from the distribution of the dependent variable. The conditional joint density of the conditional Poisson Fixed Effects model is given by:

$$\Pr(y_{i1}, y_{i2}, ..., y_{it} \mid \sum_{t=0}^{T} y_{it}) = \frac{\Pr(y_{i1}, y_{i2}, ..., y_{iT-1}, \sum_{t=0}^{T} y_{it} - \sum_{t=0}^{T-1} y_{it})}{\Pr(\sum_{t=0}^{T} y_{it})}$$

$$= \frac{\frac{\exp(-\sum_t \alpha_i \lambda_{it}) \Pi_t (\alpha_i \lambda_{it})^{y_{it}}}{\prod_t y_{it}!}}{\frac{\exp(-\sum_t \alpha_i \lambda_{it}) (\sum_t \alpha_i \lambda_{it})^{(\sum_t y_{it})}}{\sum_t y_{it}!}}$$

$$= \frac{(\sum_t y_{it})!}{\prod_t y_{it}!} \prod_{t=0}^{T} \left(\frac{\lambda_{it}}{\sum_t \lambda_{it}} \right)^{y_{it}}. \tag{13}$$

Again the individual fixed effects, α_i's, have disappeared in equation (13) because the α_i's simplify to in this equation. In the special case $\lambda_{it} = \exp(x_{it}\beta)$ the preceding expression becomes:

$$\Pr(y_{i1}, y_{i2}, ..., y_{it} \mid \sum_{t} y_{it}) = \frac{(\sum_t y_{it})!}{\prod_t y_{it}!} \prod_{t=0}^{T} \left(\frac{\exp(x_{it}\beta)}{\sum_s \exp(x_{it}\beta)} \right)^{y_{it}}. \tag{14}$$

It follows that the conditional maximum likelihood estimator of the Poisson Fixed Effects model maximizes with respect to β the following conditional log-likelihood

function:

$$l(\beta) = \sum_{i=1}^{N} \left[\ln(\sum_{t=0}^{T} y_{it})! - \sum_{t=0}^{T} \ln(y_{it}!) + \sum_{t=0}^{T} y_{it} \ln \left(\frac{\exp(x_{it}\beta)}{\sum_s \exp(x_{it}\beta)} \right) \right]. \qquad (15)$$

Equation (15) is proportional to the concentrated log-likelihood function given in equation (11) implying that the concentrated maximum likelihood estimate equals the conditional maximum likelihood estimate (Cameron and Trivedi, 1998). The First-Order Condition associated with the Poisson Fixed Effects model is obtained by differentiating equation (15) with respect to β, yielding:

$$\sum_{i=1}^{N} \sum_{t=0}^{T} x_{it} \left(y_{it} - \exp(x_{it}\beta) \left[\frac{\frac{1}{T} \sum_t y_{it}}{\frac{1}{T} \sum_t \exp(x_{it}\beta)} \right] \right) = 0. \qquad (16)$$

Rewriting this equation by setting $\bar{y}_{it} = \frac{1}{T} \sum_{t=0}^{T} y_{it}$ and $\bar{\lambda}_{it} = \frac{1}{T} \sum_{t=0}^{T} \lambda_{it}$ with $\lambda_{it} = \exp(x_{it}\beta)$ yields the following expression which is equivalent to equation (12):

$$\sum_{i=1}^{n} \sum_{t=0}^{T} x_{it} (y_{it} - \lambda_{it} \frac{\bar{y}_{it}}{\bar{\lambda}_{it}}) = 0. \qquad (17)$$

Because $(y_{i1}, y_{i2}, ..., y_{it} \mid \sum_t y_{it})$ is multinomial distributed with probabilities $p_{it}, ..., p_{iT}$ (Cameron and Trivedi, 1998; Winkelman, 2000), with $p_{it} = \lambda_{it} / \sum_t \lambda_{it}$, it follows that in the conditional fixed effects model, y_{it} is Poisson with mean $p_{it} \sum_t \lambda_{it}$. The fixed effects α_i are again estimated by $\sum_t y_{it} / \sum_t \lambda_{it}$.

The Poisson Fixed Effects model allows the introduction of individual unobserved effects which have to be estimated jointly with the coefficients of the covariates. By

concentrating the likelihood function or by conditioning in the total number of counts, the fixed effects can be estimated. The conditional maximum likelihood method allows the control of correlation because the events at a particular point of time are conditional on the total number of counts.

2.1.2 Multivariate Negative Binomial (MNB) Models

In the preceding Poisson models, the individual specific effects were assumed to be fixed over time. It is possible however that the unobserved effects are still invariant over time but random across individuals according to a specific distribution. In a Poisson Random Effects model for longitudinal count data, the Poisson' s parameter becomes:

$$\mu_{it} = \exp(x_{it}\beta + \epsilon_i) = \exp(x_{it}\beta)\gamma_i = \lambda_{it}\gamma_i. \text{ with } \gamma_i = \exp(\epsilon_i) \tag{18}$$

The random term ϵ_i takes into account heterogeneity in the data or possible specification errors of $\lambda_{it} = \exp(x_{it}\beta)$. These misspecifications may result, for example, from the omission of non-observable explanatory variables or from measurement errors of these variables. This review of Multivariate Negative Binomial models starts with a basic description of the different Univariate Negative Binomial models since they are sometimes used for the analysis of longitudinal count data in health economics (i.e. by pooling the data). The Multivariate Negative Binomial model (Johnson, 1997) and its random effects (Haussman et al. 1984) are presented thereafter.

Univariate Negative Binomial Models For any given parametric family of distributions $f(y \mid \gamma)$ conditional on γ, where γ is a random variable with distribution G, the unconditional distribution F_U can be written as:

$$F_U(y) = \int f(y \mid \gamma) dG(\gamma). \tag{19}$$

This integral has an explicit solution when y is distributed as Poisson with parameter $\theta = \lambda\gamma$ and γ is Gamma distributed. Differentiating with respect to γ yields the Univariate Negative Binomial distribution as a Poisson-Gamma mixture. For example, in Cameron and Trivedi (1986), γ_i has a Gamma distribution (ϕ_i, α) such as $g(\gamma_i) = \frac{1}{\Gamma(\alpha)}(\frac{\alpha\gamma_i}{\phi_i})^\alpha \exp(\frac{-\alpha\gamma_i}{\phi_i})\frac{1}{\gamma_i}$, with $\mathrm{E}(\gamma_i) = \phi_i$ and $\mathrm{Var}(\gamma_i) = \frac{1}{\alpha}\phi_i^2$. Integrating the probability mass function with respect to γ_i yields the following Univariate Negative Binomial (UNB) distribution:

$$UNB(Y_i = y_i) = \frac{\Gamma(\alpha + y_i)}{\Gamma(\alpha)\Gamma(y_i + 1)} \left[\frac{\phi_i}{(\alpha + \phi_i)}\right]^{y_i} \left[\frac{\alpha}{(\alpha + \phi_i)}\right]^{\alpha}. \tag{20}$$

The mean and the variance are $E(y_i) = \phi_i$ and $Var(y_i) = \phi_i + \frac{1}{\alpha}\phi_i^2$. Several formulations of the Univariate Negative Binomial distribution can be formulated based on the relation between the underlying parameters of the gamma distribution and the covariates as discussed in Cameron and Trivedi (1986). In the Negative Binomial Type 2 (NB2) model, a quadratic form is assumed for the mean-variance relationship

such as:

$$Var(y_i) = E(y_i) + \alpha\left[E(y_i)\right]^2. \tag{21}$$

The mean, $E(y_i) = \phi_i = \exp(x_i\beta)$, corresponds to the parameter λ_i in equation (18). The Negative Binomial Type 1 (NB1) model has mean and variance given by:

$$E(y_i) = \exp(x_i\beta) \tag{22}$$

$$Var(y_i) = (1+\alpha)E(y_i). \tag{23}$$

Equations (22) and (23) implies a variance to mean ratio constant across individuals and independent of the mean. Another specification of the Univariate Negative Binomial (UNB) distribution is given by Winkelman (1994) as follows:

$$UNB(Y_i = y_i) = \frac{\Gamma(\alpha + y_i)}{\Gamma(\alpha)\Gamma\left(y_i + 1\right)}\left[\frac{\lambda_i}{(1+\lambda_i)}\right]^{y_i}\left[\frac{1}{(1+\lambda_i)}\right]^{\alpha}. \tag{24}$$

In this specification the mean and the variance are given by $E(y_i) = \alpha\lambda_i$ and $Var(y_i) = \alpha\lambda_i(1+\lambda_i)$.

Multivariate Negative Binomial Distribution Mixture representations can be extended to the longitudinal case when y is a vector $Y = (y_0, \ldots y_t)$, θ a scalar and $f(. \mid \theta)$ is the product of its T marginal distributions. Assuming independence,

equation (19) is still a mixture given by:

$$F_u(Y) = \int \prod_{t=0}^{T} f_t(y_t \mid \gamma) dG(\gamma). \tag{25}$$

In this expression, G and each f_t are univariate distributions. The interpretation of γ is that of a common heterogeneity component affecting all counts. It has been shown that multivariate distributions generated in this way have univariate marginals in the same family (Marshall and Olkin, 1990; Kocherlakota and Kocherlakota, 1993). For example, Marshall and Olkin (1990) have generated a bivariate Negative Binomial distribution with f_1 and f_2 being Poisson distributed with parameter $\lambda_1\gamma$ and $\lambda_2\gamma$ and γ has Gamma distribution with parameter $1/\alpha$. Cameron and Trivedi (1998) observed that there was no application or computational experience of multivariate distributions generated in this way.

In the longitudinal case it is assumed that the counts observed at each time period are independently distributed Poisson conditional on γ_i with mean $\mu_{it} = \lambda_{it}\gamma_i$, and γ_i, the unobserved heterogeneity variable, has a Gamma distribution. The term λ_{it} is related to the set of covariates by the exponential function as $\lambda_{it} = \exp(x_{it}\beta)$ in which x_{it} is the vector of covariates of unit i at time t and β, the associated parameter vector, is assumed to be the same for all individuals. For simplicity unknown parameters and strictly exogenous variables are suppressed in the following equations without loss of generality.

The Multivariate Negative Binomial (MNB) probability mass function of $Y_{it} = (y_{i0}, ...y_{it})$ is generated by integrated out γ_i in:

$$MNB\ (Y_{it}) = \int_0^{\infty} \prod_{t=0}^{T} \left[\frac{(\lambda_{it}\gamma_i)^{y_{it}} \exp(-\lambda_{it}\gamma_i)}{y_{it}!} \right] \left[\frac{\gamma_i^{\alpha_i - 1} \exp(-\gamma_i)}{\Gamma(\alpha_i)} \right] d\gamma_i. \quad (26)$$

After collecting the terms independent of γ_i and using the properties of the Gamma distribution, it can easily be shown that the Multivariate Negative Binomial (MNB) density function can be written as:

$$MNB(Y_{it}) = \frac{\Gamma(\alpha + \Sigma_{t=0}^{T} y_{it})}{\Gamma(\alpha)\Pi_{t=0}^{T}\Gamma\left(y_{it}+1\right)} \left[\prod_{t=0}^{T} \left\{ \left(\frac{\lambda_{it}}{(1+\Sigma_{t=0}^{T}\lambda_{it})} \right)^{y_{it}} \right\} \right] \left[\frac{1}{(1+\Sigma_{t=0}^{T}\lambda_{it})} \right]^{\alpha}. \quad (27)$$

A complete description of the properties of the Multivariate Negative Binomial distribution can be found in Johnson et al. (1997). In particular, the mean and the variance of y_{it} at period t of this distribution are given by:

$$E(y_{it}) = \alpha\lambda_{it} \quad (28)$$

$$Var(y_{it}) = \alpha\lambda_{it}(1+\lambda_{it}). \quad (29)$$

The variance is greater than the mean thus modelling overdispersion and the variance-mean ratio increases with the mean. Using the distribution defined by equation (27), the likelihood of the Multivariate Negative Binomial probability distribution, $l(\beta, \alpha)$, has a simple form that is tractable and is given by:

$$l(\beta,\alpha)=\sum_{i=1}^{N}\left[\begin{array}{c}\ln\Gamma(\alpha+\Sigma_{t=0}^{T}y_{it})-\ln\Gamma(\alpha)-\Sigma_{t=0}^{T}\ln\Gamma(y_{it}+1)+\\ \Sigma_{t=0}^{T}\left\{y_{it}\ln\lambda_{it}\right\}-(\alpha+\Sigma_{t=0}^{T}y_{it})\ln(1+\Sigma_{t=0}^{T}\lambda_{it})\end{array}\right]. \tag{30}$$

After parametrization of $(\lambda_{i0},...,\lambda_{it})$, estimation can be performed by maximizing the likelihood function given in equation (30) with respect to β, the vector of parameters, and α, the coefficient of overdispersion which is assumed to have the same value for all individuals, using the Newton-Raphson algorithm.

The Multivariate Negative Binomial model offers several advantages over the preceding Poisson models. The Multivariate Negative Binomial model accounts for unobserved heterogeneity at the unit level while allowing the data on the observed count to be correlated among individuals over time. The coefficient of correlation between two time periods r and s is given by (Johnson et al., 1997):

$$\rho_i(r,s)=\sqrt{\frac{\lambda_{ir}\lambda_{is}}{(1+\lambda_{ir})(1+\lambda_{is})}}. \tag{31}$$

Multivariate Negative Binomial Fixed Effects (MNBFE) Model Similar to the approach of the Poisson Fixed Effects model, Haussman et al. (1984) derived a fixed effects version of the Multivariate Negative Binomial model in order to add individuals specific effects to the Multivariate Negative Binomial model. In this new specification, the individuals specific effects are derived conditional on the total number of counts observed over the period of analysis. Haussman et al. (1984) used a Univariate Negative Binomial Type 1 distribution because the sum of independent Negative Binomial random variables is again Negative Binomial distributed only if

the distribution is Negative Binomial Type I (Winkelman, 2000). This model assumes that the counts are independent over time and that the individual fixed effects are constant over time for each individual.

In the case of two repeated measurements at t= 1 and t= 2, $y = y_1 + y_2$ is distributed as Negative Binomial with parameter $(\lambda_1 + \lambda_2, \delta)$, the following can be written:

$$\Pr(y_1, y_2 \mid y = y_1 + y_2) = \frac{\Pr(y_1)\Pr(y_2)}{\Pr(y)}$$

$$= \frac{\frac{\Gamma(\lambda_1+y_1)}{\Gamma(\lambda_1)\Gamma(y_1+1)}(1+\delta)^{-(y_1+y_2)}(\frac{\delta}{1+\delta})^{\lambda_1+\lambda_2}\frac{\Gamma(\lambda_2+y_2)}{\Gamma(\lambda_2)\Gamma(y_2+1)}}{\frac{\Gamma(\lambda_1+\lambda_2+y)}{\Gamma(\lambda_1+\lambda_2)\Gamma(y+1)}(1+\delta)^{-(y_1+y_2)}(\frac{\delta}{1+\delta})^{\lambda_1+\lambda_2}}. \tag{32}$$

The ratio $(\frac{\delta}{1+\delta})^{\lambda_1+\lambda_2}$ as well as $(1+\delta)^{-(y_1+y_2)}$ cancels out in expression (32) yielding:

$$\Pr(y_1, y_2 \mid y = y_1 + y_2) = \frac{\Gamma(\lambda_1 + y)\Gamma(\lambda_2 + y)\Gamma(\lambda_1 + \lambda_2)\Gamma(y+1)}{\Gamma(\lambda_1 + \lambda_2 + y)\Gamma(\lambda_1)\Gamma(\lambda_2)\Gamma(y_1+1)\Gamma(y_2+1)}. \tag{33}$$

More generally, the Multivariate Negative Binomial Fixed Effects ($MNBFE$) distribution for longitudinal count data is given by:

$$MNBFE(y_{i0}, y_{i1}, \ldots y_{it} \mid \sum_{t=0}^{T} y_{it}) = (\prod_t \frac{\Gamma(\lambda_{it} + y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it}+1)}) \left[\frac{\Gamma(\sum_t \lambda_{it})\Gamma(\sum_t y_{it}+1)}{\Gamma(\sum_t \lambda_{it} + \sum_t y_{it})}\right]. \tag{34}$$

The parameters of the underlying model are $(\lambda_{it},\ \delta_i) = \left(exp(x_{it}\beta), \frac{\phi_i}{\exp(\mu_i)}\right)$. The unobserved heterogeneity is individual specific rather than constant across individuals because ϕ_i and μ_i vary across individuals in the fixed effects specification. ϕ_i is

the individual specific effects in the conditional expectation function and μ_i is an individual specific fixed dispersion parameter. Only the ratio $\frac{\phi_i}{\exp(\mu_i)}$ is identified in this model, but it cancelled out when conditioning on the individual specific sum of counts $\sum_{t=0}^{T} y_{it}$ and only $\lambda_{it} = exp(x_{it}\beta)$ appears in the final equation. The mean and the variance of the Multivariate Negative Binomial Fixed Effects distribution are given by:

$$E(y_{it}) = \exp(x_{it}\beta + \mu_i)/\phi_i \tag{35}$$

$$Var(y_{it}) = \left(\frac{\exp(x_{it}\beta + \mu_i)}{\phi_i}\right)\left(\frac{1 + \exp(\mu_i)}{\phi_i}\right). \tag{36}$$

The mean-variance ratio is different from the original Negative Binomial Type I model. The log-likelihood function, $l(\beta)$, associated with the Multivariate Negative Binomial Fixed Effects model is:

$$\begin{aligned} l(\beta) \;=\; & \sum_{i=1}^{N}\sum_{t=0}^{T}[\ln(\Gamma(\lambda_{it} + y_{it})] - \sum_{i=1}^{N}\sum_{t=0}^{T}\ln[\Gamma(\lambda_{it})\Gamma(y_{it}+1)] + \\ & \sum_{i=1}^{N}\sum_{t=0}^{T}\ln\left[\Gamma(\sum_{t}\lambda_{it})\right] + \sum_{i=1}^{N}\sum_{t=0}^{T}\ln\left[\Gamma(\sum_{t}y_{it}+1)\right] - \\ & \sum_{i=1}^{N}\sum_{t=0}^{T}\ln\left[\Gamma(\sum_{t}\lambda_{it} + \sum_{t}y_{it})\right]. \end{aligned} \tag{37}$$

This Multivariate Negative Binomial Fixed Effects model accounts for the longitudinal aspect of the data, the presence of covariates and the between subject heterogeneity by introducing a fixed effect factor. It allows a variance to mean ra-

tio increasing with the mean but it assumes that the random effects are constant across individuals which may be a limitation of this model. Another weakness of this model is that the probabilities and the marginal effects cannot be computed because the estimation of the fixed effects model requires the fixed effects estimators to be conditioned out (Greene, 1998). Because it is not possible to decompose the Negative Binomial Fixed Effects distribution according to a product of conditional distributions it is not possible to define the distribution for the initial observation for modeling purposes as it will be discussed in Chapter 3.

Multivariate Negative Binomial Random Effects (MNBRE) Model In the preceding fixed effects model, the parameters of the underlying model are:

$$(\lambda_{it}, \delta_i) = \left(exp(x_{it}\beta), \frac{\phi_i}{\exp(\mu_i)} \right). \tag{38}$$

In this equation ϕ_i and μ_i can vary across individuals. Haussman et al. (1984) have integrated random effects in the Multivariate Negative Binomial Fixed Effects distribution. In this new specification, it is assumed that $\delta_i/(1+\delta_i)$ with $\delta_i = \frac{\phi_i}{\exp(u_i)}$, is randomly distributed across individuals, independent of the covariates, as a beta random variable with parameters (a, b) with density function, mean and variance given by:

$$f\left[\delta_i/(1+\delta_i)\right] = \left[B(a,b)\right]^{-1} \frac{\delta_i}{1+\delta_i}^{a-1} (1 - \frac{\delta_i}{1+\delta_i})^{b-1} \tag{39}$$

$$E(\frac{\delta_i}{1+\delta_i}) = \frac{a}{a+b} \tag{40}$$

$$Var(\frac{\delta_i}{1+\delta_i}) = \frac{ab}{(a+b+1)(a+b)}. \tag{41}$$

Haussman et al. (1984) showed that the probability mass function of the Multivariate Negative Binomial Random Effects (MNBRE) distribution can be written as:

$$MNBRE(y_{i0}, y_{i1}, ...y_{it}) = \frac{\Gamma(a+b)\Gamma(a+\sum_{t=0}^{T}\lambda_{it})\Gamma(b+\sum_{t=0}^{T}y_{it})}{\Gamma(a)\Gamma(b)\Gamma(a+b+\sum_{t=0}^{T}\lambda_{it}+\sum_{t=0}^{T}y_{it})}\prod_{t}\frac{\Gamma(\lambda_{it}+\sum_{t=0}^{T}y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it}+1)} \tag{42}$$

Estimation can be done by maximizing the associated log-likelihood function with respect to β and the parameters of the Beta distribution, a and b. The log-likelihood function of the Multivariate Negative Binomial Random Effects model, $l(\beta, a, b)$, is given by:

$$l(\beta,a,b) = \sum_{i=1}^{N}\left[\begin{array}{l} \ln\Gamma(a+b) + \ln\Gamma(a+\sum_{t=0}^{T}\lambda_{it}) + \ln\Gamma(a+\sum_{t=0}^{T}y_{it}) - \\ \ln\Gamma(a) - \ln\Gamma(b) - \ln\Gamma(a+b+\sum_{t=0}^{T}\lambda_{it}+\sum_{t=0}^{T}y_{it}) + \\ \sum_{t=0}^{T}\left\{\ln\Gamma(\lambda_{it}+\sum_{t=0}^{T}y_{it}) - \ln\Gamma(\lambda_{it}) - \ln\Gamma(y_{it}+1)\right\} \end{array}\right] \tag{43}$$

Overall, the Multivariate Negative Binomial Random Effects model accounts for over-dispersion, serial correlation and heterogeneity at the individual level modelling. This specification is referred to by Haussman et al. (1984) as conditional or within-firm model as opposed to the Multivariate Negative Binomial Fixed Effects model or

between-firm model. Between-firm models or marginal models estimate the marginal (between) dimension of the data by averaging the counts over the time period (Haussman et al., 1984).

2.2 Non-Parametric Models

The models presented so far can all be estimated by maximum likelihood techniques. However, longitudinal count data have also been estimated with non-parametric models. The starting point of this methodology is the linear multivariate linear exponential family for which a representative probability mass function is given by:

$$f(y_i, \mu_i) = \exp\left[A(\mu_i(\beta)) + B(y_i) + C(\mu_i(\beta))y_i\right] \tag{44}$$

where y_i is a K dimensional and random count variable with mean and variance $E(y_i) = \mu_i$ and $Var(y_i) = V_i$. The functions, A and B, are real valued functions, C is a K dimensional vector valued function; μ_i depends on a parameter vector, β. This family includes specific distributions for count data such as the Poisson and the Negative Binomial (α given) distributions, as well as the gamma (α given), normal (σ given), binomial (n given), multinomial (n given) and the normal multivariate ($\sum$ given) distributions (Gourrieroux et al., 1984). This family of distribution has desirable properties that are discussed in the following sub-sections.

2.2.1 Pseudo Maximum Likelihood (PML) and Quasi Generalized Pseudo Maximum Likelihood (QGPML) Methods

The Pseudo Maximum Likelihood (PML) method consists in taking a distribution which is a member of the family of linear exponential distribution. Although this distribution may not belong to the true one, when a distribution pertaining to the linear exponential distribution is taken and provided that the mean is correctly specified, consistent and asymptotically normal estimates can be achieved with the Pseudo Maximum Likelihood method as demonstrated by Gourrieroux et al. (1984). Any linear exponential family will yield consistent estimates of the parameters of a correctly specified mean function, regardless of the true model. For longitudinal count data the log pseudo-likelihood to maximize is:

$$\sum_{i=1}^{N}\sum_{t=0}^{T}\left[A(\mu_{it}(\beta)) + C(\mu_{it}(\beta))y_{it}\right]. \tag{45}$$

This expression depends on the chosen pseudo distribution. For a Poisson distribution with mean $E\left[y_{it} \mid x_{it}\right] = \mu_{it} = \exp(x_{it}\beta)$, the objective function to be maximized is (Gourieroux et al, 1984):

$$\sum_{i=1}^{N}\sum_{t=0}^{T}\left[-\exp(x_{it}\beta) + y_{it}x_{it}\beta\right]. \tag{46}$$

The First-Order Condition resulting from maximizing equation (46) is similar to the First-Order Condition obtained from the Poisson likelihood specification defined in equation (4). If the distribution is assumed to be Negative Binomial Type 2 with

parameter δ, the corresponding Pseudo Maximum Likelihood estimates of β can be obtained (Blundell et al., 1995) by maximizing the following equation:

$$\sum_{i=1}^{N}\sum_{t=0}^{T}\left[-y_{it}x_{it}\beta-(\frac{1}{\delta}+y_{it})\ln(1+\delta\exp(x_{it}\beta))\right]. \tag{47}$$

The asymptotic variance-covariance matrix of the Pseudo Maximum Likelihood estimator is given by $J^{-1}IJ^{-1}$ where J and I are:

$$J = E\left[\frac{\partial\mu}{\partial\beta}\sum\nolimits_0^{-1}\frac{\partial\mu'}{\partial\beta}\right], \tag{48}$$

$$I = E\left[\frac{\partial\mu}{\partial\beta}\sum\nolimits_0^{-1}\Omega_0\frac{\partial\mu'}{\partial\beta}\right]. \tag{49}$$

The term $\sum_0$ is the variance of the gamma distribution and Ω_0 is the variance of the true distribution. Because the Pseudo Maximum Likelihood model assumes that only the mean is correctly specified, the Pseudo Maximum Likelihood estimates of the Negative Binomial Type 1 and Type 2 models are similar.

The variance-covariance matrix can be estimated in different ways. Gourrieroux et al. (1984) developed a two-step method which provides better asymptotic estimators than the estimates obtained by Pseudo Maximum Likelihood methods assuming that the conditional moment of second order is known. This method is referred in the economic literature to the Quasi-Generalized Pseudo-Maximum Likelihood (QGPML)

method . In the first step, a consistent estimator of δ is computed by:

$$\widehat{\delta} = \frac{\sum_i \sum_t \left[\left(y_{it} - \exp(y_{it}\widehat{b}) \right)^2 - \exp(y_{it}\widehat{b}) \right] \exp(2x_{it}\widehat{b})}{\exp(4x_{it}\widehat{b})}. \tag{50}$$

In this equation $\widehat{b}$ are the first estimates of β estimated by Pseudo Maximum Likelihood. In the second step, estimates of parameter β are performed by Pseudo Maximum Likelihood with the second order moment taken into account using the estimated consistent value of δ, $\overset{\wedge}{\delta}$. For the gamma pseudo-distribution, the objective function to be maximized is:

$$\sum_{i=1}^{N} \sum_{t=0}^{T} \left\{ \frac{\exp(x_{it}\widehat{b})}{1 + \widehat{\delta} \exp(x_{it}\widehat{b})} (-x_{it}\beta - y_{it} \exp(-x_{it}\beta)) \right\}. \tag{51}$$

The asymptotic variance-covariance matrix is equal to:

$$\left[E_x \left(\frac{x_{it} x_{it}' \exp(x_{it}\widehat{b})}{1 + \widehat{\delta} \exp(x_{it}\widehat{b})} \right) \right]^{-1}. \tag{52}$$

The matrix of the Quasi Generalized Pseudo Maximum Likelihood estimator is smaller than the matrix of the Pseudo Maximum Likelihood estimator $J^{-1}IJ^{-1}$ defined by equation (48) and (49). Both Pseudo Maximum Likelihood and Quasi Generalized Pseudo Maximum Likelihood models assume that the mean function is correctly specified and the distribution is a member of the linear exponential family.

2.2.2 Generalized Estimating Equations (GEE) Model

Liang and Zeger (1988) and Zeger and Liang (1988) have proposed quasi-likelihood longitudinal count data methods that describe the correlation structure among the responses while also taking overdispersion into account. Generalized Estimating Equations models are an extension of the Generalized Linear Model from the independent case to repeated measurements by extending the concept of quasi-likelihood to correlated observations. Generalized Estimating Equations models adjust for correlation on observations of the same individual or firm by the introduction of an arbitrary covariance matrix in the score equations of generalized linear models. Similar to Pseudo Maximum Likelihood and Quasi Generalized Pseudo Maximum Likelihood specifications, Generalized Estimating Equations models assume that the distribution governing the counts over time is from the multivariate linear exponential family. For this family, Gourieroux et al. (1984) showed that:

$$\partial A/\partial \mu_{it} + (\partial C/\partial \mu_{it})\mu_{it} = 0 \text{ and} \tag{53}$$

$$\partial C/\partial \mu_{it} = V^{-1}. \tag{54}$$

With respect to β, the First-Order Condition for maximizing a likelihood function based on this distribution is:

$$\sum_{i=1}^{N}\sum_{t=0}^{T}(\partial \mu_{it}/\partial \beta)' V^{-1}(y_{it} - \mu_{it}) = 0. \tag{55}$$

This system of equations is called Generalized Estimating Equations (GEE). Liang

and Zeger (1986) have proposed various models for the correlation structure including an independent model, an exchangeable or an unspecified correlation structure. In Generalized Estimating Equations models, the mean is related to the set of covariates by a link function h such as $h(\mu_{it}) = h(E(y_{it})) = x_{it}\beta$.

The functional form between the variance and the mean is specified as $Var(y_{it}) = \alpha g(\mu_{it})$, α being the so called dispersion parameter. Through selection of h and g, a general class of responses can be modelled (continuous, binary and count). For count data the link function h is the natural logarithm, whereby $h(\mu_{it}) = log(\mu_{it}) = x_{it}\beta$ and g is the identity function specified as $g(\mu_{it}) = \mu_{it}$.

Generalized Estimating Equations estimates are asymptotically consistent even if the covariance matrix is misspecified as long as the regression equation for the mean is correctly specified and the marginal distribution of the counts is from the exponential family (Liang and Zeger, 1988). However, this may be a poor assumption since the analysis may indicate that a model outside the linear exponential family is required. For example, if the Negative Binomial distribution happens to be the true distribution describing the data, applying a Generalized Estimating Equations model is only correct if the value of the coefficient of overdispersion is known (Lindsey, 1999). If not, the Negative Binomial distribution does not belong to linear exponential families. In a Generalized Estimating Equations model, the score equation for many covariance matrix cannot be integrated back to a likelihood function making comparisons difficult among different models. This is also true for all non-parametric models. Albert (1999) noted that although likelihood-based methods are less robust

than Generalized Estimating Equations models, they are generally more efficient and offer well-defined means of assessing model adequacy.

2.2.3 Generalized Method of Moments (GMM) Model

All the models presented so far assume strict exogeneity of the covariates. However, this may not always be the case and other models have to be sought. The Generalized Method of Moments model allows for correlated fixed effects but relaxes the strict exogeneity assumption of the regressors. Similar to the Poisson multiplicative effect distribution, the Generalized Method of Moments model (Hansen, 1982) assumes the following set of conditional mean restrictions:

$$E(y_{it} \mid z_{is}, \varepsilon_i) = \exp(x_{it}\beta + \varepsilon_i), \forall\ s \leq t. \tag{56}$$

Here z_{is}, s=0, 1, 2,...,t, represents any set of instruments such as equation (56) holds. The unobserved fixed effect, ε_i, can be removed by a quasi transformation proposed by Chamberlain (1992 a):

$$E\left[y_{it} - y_{it+1} \exp\left\{(x_{it} - x_{it+1})\beta\right\} \mid z_{is}\right] = 0, \forall\ s \leq t. \tag{57}$$

These orthogonality conditions remain valid under weak exogeneity of the regressors because the conditioning set is dated at period t or earlier (Blundell et al., 1995). Generalized Method of Moments models allow for heteroskedacity and any serial correlation pattern of the errors terms while relaxing the assumption of the

strict exogeneity of the regressors. However, as Generalized Estimating Equations models, Generalized Method of Moments models can not be inflated to accommodate an excess of zeros.

2.2.4 Simulated Maximum Likelihood (SML) Model

Another non-parametric approach which has been recently used for panel count data analysis (Cincera, 1997) is based on simulated likelihood methods. The Simulated Maximum Likelihood approach assumes that the distribution of the random ε_i is associated with a random vector which is generated by a known distribution, F, as follows:

$$\varepsilon_i = F(u_i, \theta). \tag{58}$$

Here θ is a vector of parameters and the distribution of u_i is given. Gourieroux et al. (1991 a, b) showed that when the ε_i represents a univariate continuous variable, the function $F(u_i, \theta)$ can be chosen as the inverse of the cumulated distribution function of the ε_i and in such a case, the distribution of the u_i is the uniform distribution on $[0, 1]$. With these assumptions, the log-likelihood function can be written as:

$$\sum_{i=1}^{N}\sum_{t=0}^{T} \ln\left(\int_{-\infty}^{+\infty} \frac{\exp(-\exp(-x_{it}\beta + u_i))(\exp(-x_{it}\beta + u_i))^{y_{it}}}{y_{it}!} dy(u_i)\right). \tag{59}$$

The integral of this log-likelihood function can be computed by simulation on the u_i and by taking the sample mean. These simulations are quite simple given that the

distribution of the u_i is known, yielding:

$$\sum_{i=1}^{N}\sum_{t=0}^{T}\ln\left(\frac{1}{H}\sum_{h=1}^{H}\frac{\exp(-\exp(-x_{it}\beta+u_{ih}))(\exp(-x_{it}\beta+u_{ih}))^{y_{it}}}{y_{it}!}\right). \quad (60)$$

This method is referred to the Simulated Maximum Likelihood method. When H is fixed and tends to infinity, the estimator is biased and the asymptotic bias is of the order of $\frac{1}{H}$. The Simultated Maximum Likelihood estimator is consistent and asymptotically efficient when H and N tend to infinity such as $\frac{\sqrt{n}}{H}$ tends to zero.

2.3 Generalized Auto-Regressive (GAR) Model

Generalized Auto-Regressive (GAR) models have been proposed to analyze longitudinal count data. In these models, the analysis is conditional on past outcomes as well as current and past values of exogenous variables. In the first Auto-Regressive model the conditional mean follows an AR(1) process of the form:

$$E(y_{it}) = \exp(x_{it}\beta + \rho y_{it-1}). \quad (61)$$

Here y_{it-1} appears as a regressor in the conditional mean. Cameron (1998) notes that this model is explosive for $\rho > 0$. Instead, he recommends the following approach (Cameron and Trivedi, 1998):

$$E(y_{it}) = \mu_{it|t-1} = \exp(x_{it}\beta + \rho \ln y^*_{it-1}) = \exp(x_{it}\beta)(y^*_{it-1})^{\rho}. \quad (62)$$

In this equation, y^*_{it-1} is a transformation of y_{it-1} such as:

$$\begin{aligned} y^*_{it-1} &= \max(c, y_{it-1}), 0 < c < 1 \\ y^*_{it-1} &= y_{it-1} + c,\ c > 1. \end{aligned} \tag{63}$$

Cameron and Trivedi (1998) indicated that this transformation was required because if $y_{it-1} = 0$ then the conditional mean $\mu_{t|t-1} = 0$ then $y_{it} = 0$. Once the process for determining the evolution of the mean and the functional form of the conditional distribution has been chosen, then parameter estimation can be implemented. For example if the conditional density $f(y_{it} \mid x_{it}, y_{it-1})$ is Poisson or Negative Binomial distributed, estimation can be done by maximizing the likelihood function, $l(\beta, \rho)$, given by:

$$l(\beta, \rho) = \sum_{i=1}^{N} \sum_{t=1}^{T} f(y_{it} \mid x_{it}, y_{it-1}). \tag{64}$$

It is not clear, however, how to treat the initial value in short longitudinal count data sets in which the number of repeated measurements is small. In the normal case, it has been shown that treating the initial value as a covariate may lead to inconsistent estimates (Greene, 1998; Greene, 2000). First-Auto-Regressive models have also been estimated using Generalized Estimating Equations models (Diggle, 1996; Liang and Zeger, 1988; Zeger, 1988) which included past values of the dependent variable as regressors. Markov models have also been used to model a time series of epilepsy seizures (Le et al., 1992; Zeger and Liang 1991).

Using patent data, Blundel et al. (1995) proposed a Linear Feedback model which determined whether the presence of serial correlation could be viewed as an issue of dynamic specification in the patenting process. The Linear Feedback model is defined by the following quasi-differenced orthogonality conditions:

$$E\left[\left(y_{it}-y^{*}_{it-1}\rho\right)-\left(y_{it+1}-y^{*}_{it}\rho\right)\exp\left\{\left(x^{**}_{it}-x^{**}_{it+1}\right)\beta^{*}\right\}\mid z_{is}\right]=0,\forall\ s\leq t \qquad (65)$$

where:

$$Y^{*}_{it}=(y_{it},y_{it-1},y_{it-2}),\ \rho^{'}=(\rho_1,\rho_2,\rho_3),\ x^{**}_{it}=(k_{it},s_{it})\text{ and }\beta^{*'}=(\beta_k,\beta_s).$$

In equation (65), lagged values of the count variable among the regressors are present. Fixed effects can be removed by first or quasi differencing. If the t-2 lagged and higher values of Y_{it} are used as instruments for $(y_{it-1}-y_{it-2})$, then consistent estimates can be obtained as long as the residuals are not serially correlated (Cincera, 1997).

Other models cited in the econometric literature on time series count data are First-Order Integer-Valued Autoregressive, INAR(1), (Al-Osh and Alzaid, 1987), First-Order Integer-Valued Autoregressive Moving Average, INARMA(1), (McKenzie, 1986; Al-Osh and Alzaid, 1987) or serially correlated models (Zeger 1988). These models have been developed for pure time series of count data when T goes to infinity. INAR(1) and INARMA(1) are parametric models that consider that y_{it} is the sum of an integer whose value is determined by past outcomes and independent innovation

whose value does not depend on the past outcome. Serially correlated errors models consider that the serial correlation is introduced in y_{it}, via serial correlation in a multiplicative latent variable. Covariates have recently been integrated in mixed INAR(1) models by Bockenholt (1999). However, these models are not suitable for short longitudinal count data because they assume that T goes to infinity which is a strong assumption in the case of short longitudinal count data sets.

2.4 Multivariate Zero-Inflated Poisson (MZIP) Model

To account for an excess of zeros in the data, an extension of the Univariate Zero-Inflated model developed by Lambert (1992) has been recently extended to the multivariate case by Chin-Shang et al. (1999). In this concept, the term "multivariate" does not refer to longitudinal count data but to the observation of m discrete outcomes $(y_1, y_2, ..., y_m)$ at one point in time. The authors have extended the concept of the Bivariate Zero-Inflated Poisson distribution to the multivariate case. With parameters $(\lambda_{10}, \lambda_{20}, ..., \lambda_{m0}, \lambda_{00})$, the Multivariate Zero-Inflated Poisson (MZIP) distribution is defined as follows:

$$(y_1, y_2, .., y_m) \sim (0, 0, ..., 0) \text{ with probability } p_0$$

$$\sim (Poisson(\lambda_1), 0, ..., 0) \text{ with probability } p_1$$

$$\sim (0, Poisson(\lambda_2), 0, ..., 0) \text{ with probability } p_2$$

...

$$\sim (0, 0, ..., Poisson\ (\lambda_m)) \text{ with probability } p_m$$

$$\sim Multivariate\ Poisson(\lambda_{10}, \lambda_{20}, ..., \lambda_{m0}, \lambda_{00}) \text{ with probability } p_{11}. \quad (66)$$

In this equation $p_0 + p_1 + ... + p_m + p_{11} = 1$. The terms $\lambda_{10}, \lambda_{20}, ..., \lambda_{m0}, \lambda_{00}$ are the means of independent Poisson variables and $\lambda = \lambda_{10} + \lambda_{20} + ... + \lambda_{m0} + \lambda_{00}$.

The probability mass function associated with the distribution given in equation (66) is:

$$\begin{aligned} \Pr(y_1 &= 0, y_2 = 0, ..., y_m = 0) = p_0 + p_1 \exp(-\lambda_1) + p_2 \exp(-\lambda_2) + ... \\ &+ p_m \exp(-\lambda_m) + p_{11} \exp(-\lambda) \end{aligned}$$

$$\Pr(y_1, y_2 = 0, ..., y_m = 0) = \frac{p_1 \lambda_1^{y_1} \exp(-\lambda_1) + p_{11} \lambda_{10}^{y_1} \exp(-\lambda)}{y_1!}$$

$$\Pr(y_1 = 0,\ y_2,\ y_3 = 0, ..., y_m = 0) = \frac{p_2 \lambda_2^{y_2} \exp(-\lambda_2) + p_{11} \lambda_{20}^{y_2} \exp(-\lambda)}{y_2!}$$

...

$$\Pr(y_1 = 0, y_2 = 0, ...y_{m-1} = 0,\ y_m) = \frac{p_m \lambda_m^{y_m} \exp(-\lambda_m) + p_{11} \lambda_{m0}^{y_m} \exp(-\lambda)}{y_m!}. \quad (67)$$

The special case in which at least two of the y_i's are not 0 is defined by:

$$\Pr(y_1,...,y_m) = p_{11} \sum_{j=0}^{\min(y_1,y_2,...,y_m)} \frac{\lambda_{10}^{y_1-j}\lambda_{20}^{y_2-j}...\lambda_{m0}^{y_m-j}\lambda_{00}^{j}}{(y_1-j)!...(y_m-j)!j!} \exp(-\lambda). \tag{68}$$

Chin-Shang (1999) showed that when $(y_1, y_2, ..., y_m)$ has a Multivariate Zero-Inflated Poisson (MZIP) distribution with Univariate Zero-Inflated Poisson (ZIP) distributions as marginals. The marginal distribution of y_i is given by:

$$y_i \sim 0 \text{ with probability } 1 - p_i - p_{11}$$

$$y_i \sim \ Poisson(\lambda_i) \text{ with probability } p_i + p_{11}. \tag{69}$$

The mean and variance are defined by the two following equations:

$$E(y_i) = (p_i + p_{11})\lambda_i \text{ and} \tag{70}$$

$$Var(y_i) = (p_i + p_{11})\lambda_i \left[1 + (1 - p_1 - p_{11})\lambda_i\right]. \tag{71}$$

The covariance matrix between y_i and y_j is given by:

$$Cov(y_i,y_j) = p_{11}\lambda_{00} + \left[p_{11}(1 - p_{11} - p_i - p_j) - p_i p_j\right] \lambda_i\lambda_j. \tag{72}$$

The Maximum Likelihood estimates , can be obtained by maximizing the following

log-likelihood with respect to $(\lambda_{10}, ..., \lambda_{m0}, p_0, ..., p_m)$:

$$l(\lambda_{10}, ..., \lambda_{m0}, p_0, ..., p_m) = \sum_{l=1}^{n} \log \Pr(y_{1l}, ..., y_{ml}; \lambda_{10}, ..., \lambda_{m0}, \lambda_{00}, p_0, ..., p_m). \quad (73)$$

In this equation, $y_1 = (y_{11}, ..., y_{m1}), ..., y_n = (y_{1n}, ..., y_{mn})$ are n independent random vectors each having the m-dimensional Multivariate Zero-Inflated Poisson distribution. The probability mass distribution of the Multivariate Poisson Zero-Inflated,Pr, is defined in equation (68).

The main limitation of this Multivariate Zero-Inflated Poisson distribution given in equation (73) is related to the complicated form of the maximum likelihood function. Nonetheless, it is possible to compute estimates in the bivariate and trivariate cases as shown by Chin-Shang (1999). Chin-Shang (1999) applied a Trivariate Zero-Inflated model to analyze the number of defects in a Nortel manufacturing plant of electronic equipment when each item can have three categories of defect at the same time. Results indicated that the Trivariate Zero-Inflated distribution is the preferred distribution over Univariate Zero-Inflated or Trivariate distributions in predicting the proportion of defects.

Although it may be possible in theory to adapt this model for longitudinal count data, several practical limitations were associated with this distribution. In terms of computation, there are eight joint probabilities and three of them are non zeros in the three-dimensional case (Chin-Shang, 1999). When the number of types of defects per item (as in Chin-Shang, 1999) or the number of repeated measurements (in the longitudinal case) is greater than 3, the computation may be challenging as

the number of joint probabilities increases. In addition, in this model, each defect is expressed as a binary outcome (defect versus no defect) which is different from classical longitudinal count data in which the outcome can take several discrete values. Finally, this model does not include any covariate in the mean function.

2.5 Applications

Most economic applications of longitudinal count data models are found in the longitudinal analysis of the number of patents awarded to firms over a certain period of time. Several authors (Haussman et al., 1984; Pakes and Griliches, 1984; Montalvo, 1993; Crepon and Dugret, 1993; Jaffe, 1986; Blundel et al., 1995; Cincera 1997) have studied the dynamics of the structure of patent research and development (R&D) by considering the annual number of patent applications generated over time as a function of present and lagged levels of research and development expenditures. Haussman et al. (1984) have developed benchmark econometric models to analyze longitudinal count data in the context of panel data. Starting with the basic Poisson model of equation (2), Haussman et al. (1984) have developed Poisson and Negative Binomial conditional (within firms) and marginal (between firm) models. The authors derived in particular the Multivariate Negative Binomial Random Effects model which models the disturbance in the within and between dimensions. The authors applied the models given in equation (2), (14), (34) and (42) to a sample of 128 firms followed over 7 years (1968-1974). The results indicated that the data wanted both a disturbance in the conditional (within) dimension and a disturbance in the marginal

(between) dimension. Technological dummies and a variable reflecting the pool of spillovers arising from the technological activity of other firms have been introduced by Jaffe (1986). Research activity, technological spillovers and sector based dummies have been considered as explanatory variables by Crepon and Duguet (1997) and Cincera (1997). Dynamic specifications of the patenting process were examined by Blundel et al. (1995), Crepon and Duguet (1997) and Cincera (1997). In these models, past values of patents in the explanatory variables were used to test the impact of those variables in the current patenting. Crepon and Duguet (1997) showed that the past number of patents has a non-linear fixed effect in the production of current patented innovations. The assumption of the exogeneity of the variables was relaxed by Montalvo (1993) who used a Generalized Method of Moments estimator.

In health economics, applications of longitudinal count data models have concentrated in two areas. The analysis of the use of health resources such as the annual number of days in hospital or the number of doctor consultations among a particular population followed over several years, is critical to determine the determinants of health care utilization or to measure the effect of a reform of the health care system. Geil et al. (1997) examined for a panel of German households followed over 8 years the annual number of days spent in the hospital in function of the status of insurance, income, and work occupation. Winkelman (2001) analyzed the impact of a reform of the health care system in Germany using a panel of households followed for several years. Chiappori et al. (1998) investigated the presence of moral hazard in the demand for health care services using a controlled natural experiment carried over

2 years. Another area of research closely related to health economics is the analysis of the recurrence of a particular event over a certain period of time. For example, the analysis of the efficacy of a drug treatment in clinical trials in which a count health outcome is measured at several points in time was studied by Diggle (1993), Liang and Zeger (1995), Albert (1999) and Thall (1996) using Generalized Estimating Equations models. This area of research is very important since the estimates of treatment efficacy are used to conduct economic evaluations of new drug treatments.

Longitudinal count data models have also been used by Ruser (1991) in labor economics to model the number of days of absenteeism in 2,788 manufacturing establishments from 1979 through 1984. The Poisson Fixed Effects model was also used by Page (1995) to model the number of housing units shown by housing agents to each of two paired auditors who differed by minority status. Several models including Haussman et al.'s Multivariate Negative Binomial Random Effects model (1984) were considered by Pinquet (1997) to model the number and severity of insurance claims to determine bonus malus coefficients used in experience-rated insurance. All the data sets analyzed in these studies have in common that they are characterized by a limited number of observations over time, outlining again the importance of the treatment of the initial value. In addition, the majority of the publications do not report the correlation of the dependent count variable nor the proportion of zeros in the data. This is important, for example, if a two-part decision approach should be considered to explain the patenting of small and medium firms. For example, the data used by Cincera (1997) had 23% of firms with zeros patents but no attempt was

made to test for an excess of zeros.

Other models identified in this survey of the literature are Generalized Event Count models and Markov models for time series of counts. Generalized Event Count models are based on the Katz family (Winkelman and Zimmermann, 1995) which allows for the modelling of both underdispersion and overdispersion in count data. Because all the empirical work identified in this review of the literature studies reported overdispersion, the Generalized Event Count model was not considered in this chapter. In addition, Cameron (1998) reported that the Negative Binomial model and the Generalized Event Count models provided similar results in the estimation of the number of doctor visits using a cross-section of Australian data. As noted by the authors, the Negative Binomial model is easier to compute than the Generalized Event Count model.

2.6 Discussion

The discussion centered on two aspects that are under-reported in the econometric literature, namely the treatment of the initial value and the treatment of zeros in longitudinal count data. The different approaches presented so far have focussed on modelling a dependent count variable characterized by unobserved heterogeneity and correlation over time. It is clear from the literature on parametric models that the standard Poisson model is too limited to accommodate data characterized by overdispersion. To fix this shortcoming, individual fixed or random specific effects Poisson models have been introduced to characterize unobserved heterogeneity in

the data. Non-parametric Pseudo Maximum Likelihood, Quasi Generalized Pseudo Maximum Likelihood, Generalized Method of Moments and Generalized Estimating Equations models, while not specifying the density function, assume that the density function is a member of the exponential family, which for count data applies only to the Poisson distribution. In addition, if the data is characterized by many zeros, the mean function may not be correctly specified and the assumptions of the Pseudo Maximum Likelihood theory are violated.

Time series for longitudinal count data have also been developed to incorporate the time series aspect of the data, assuming a sufficient number of repeated observations. However, with longitudinal count data for which the number of repeated measurements is small, the initial value may be very important. For example, it is not known what the role of the initial value is in the generalized autoregressive model suggested by Cameron and Trivedi (1998) in equation (64). In econometrics, it has been shown that dropping the initial value(s) in a time series context could be misleading if T is small. The resulting loss in efficiency in small sample may be a problem if the regressors have a trend component (Greene, 1998). This is why methods for maximizing the likelihood function for the last t-1 observations (Cochrane-Orcut) or for all t observations (Prais-Winsten) are generally contrasted in a time series approach using normal data. In addition, when the initial value is one of the regressors, the conditional definition of any multivariate distribution does not hold anymore. This is only the case when a joint distribution is specified and when the conditional distribution does not contain the initial value as a regressor.

All but one of the studies reviewed in this chapter have not considered the problem arising from a high proportion of zeros in the data in the longitudinal case. This is important for several reasons. While it is well known that overdispersion is caused by unobserved heterogeneity, it is less known that an excess of zeros can be a strict implication of unobserved heterogeneity (Mullhaly, 1997). Since unobserved heterogeneity and an excess of zeros are not mutually exclusive of each other, it is important to have tools to deal with both.

The approach for extra zeros in the univariate case is to inflate the distribution given to the dependent variable by assuming that two processes generate the zeros instead of one (Zero-Inflated models) or by assuming a dichotomous process (Hurdle models). In Zero-Inflated models, an additional route of zeros is added. The Hurdle approach to the problem associated with an over representation of zeros is to capture the differences between zeros and at least one occurrence by assuming that different processes govern the zeros and the positive counts. As such, the Hurdle model is a two-part model in which the first part is a binary outcome modelling the possibility of crossing the Hurdle. The second part is a truncated count model which describes positive observations arising from crossing the zero Hurdle.

In a model in which the probability mass function is never specified, it is not possible to inflate the distribution to accommodate the zeros. In this sense, non-parametric models are unable to take into account an excess of zeros in the data. If the Hurdle or the Zero-Inflated specification is the true specification governing the data, then a model not taking into account this characteristic may lead to inconsistent results

since the mean function is not correctly specified (Cameron and Trivedi, 1998; Greene 2000). For example, Generalized Estimating Equations or other non-parametric models based on Pseudo Maximum Likelihood theory may lead to inconsistent estimates in the presence of an excess of zeros because the assumption on the mean does not hold anymore. Following the benchmark study of Lambert (1992) in the univariate case, Chin-Shang et al. (1999) developed a Multivariate Zero-Inflated Poisson model. Unfortunately, this model is not practically computable for longitudinal count data in which the outcome can take several discrete values and/or characterized by the presence of covariates and a large number of repeated measurements on the dependent count variable.

The next two chapters present two new models to analyze longitudinal count data characterized by a high proportion of zeros. These models were designed to fill a gap in the literature on longitudinal count data which traditionally ignores the problem of excess zeros in the longitudinal framework.

3 MULTIVARIATE NEGATIVE BINOMIAL HURDLE MODEL: AN APPLICATION TO THE LONGITUDINAL ANALYSIS OF THE NUMBER OF PHYSICIAN VISITS

An important area of research in health economics is the study of the determinants of the use of medical resources (e.g., number of physician visits). In most industrialized countries, health expenditures as a percent of the gross domestic product have increased during the last three decades. The three major reasons cited to explain this growth are the ageing of the population and more expensive treatment alternatives coupled with a large public health sector where the incentive structure does not promote economic use of resources.

Confronted with this enormous rise in health expenditures, several countries have put in place cost-containment strategies aimed at reducing or controlling health expenditures growth. Some strategies are directed at the supply side (e.g., reform of the public health insurance scheme, increased ambulatory care, generic prescribing, economic evaluations of any new pharmaceutical product prior to public reimbursement, licensing and pricing) and/or at the demand side (e.g., listing defining the drugs eligible for reimbursement, patient cost-sharing such as caps or co-payment for prescriptions). Health policies are multidimensional affecting the pharmaceutical industry, wholesalers and retailers, consumers and prescribers of pharmaceuticals.

In this cost-containment environment, the number of econometric studies to identify the determinants of the use of medical services has increased considerably over the last twenty years. The work of Grossman (1972) provided the theoretical framework to model the demand for medical services. In this framework, the demand for health is seen as an individual investment decision similar to standard human capital theory. The demand for medical care is derived from the demand for health which is seen as a durable good that depreciates over time. Individuals maximize utility from health given the production function of health and a budget constraint.

In studying the use of health services, count data models have a wide applicability since most of the units are non-negative integers or count data such as the number of hospitalizations or the number of physician visits over a certain period of time. Due to the presence of excess zeros as a result of no visiting a physician or not spending a night in a hospital, "in modelling the usage of medical services, the two-part model has served as a methodological cornerstone of empirical analysis" (Deb and Trivedi, 1999).

Surprisingly, a review of literature indicated that almost all methods to analyze the utilization of health services were based on univariate count data distributions applied to cross-sectional data or by pooling several cross-sections. This may be due to the absence of panel data sets following individuals over time and collecting medical information (e.g., utilization of health care services, self-reported status of health) along with demographic and socio-economic characteristics of these individuals. In many countries such as Canada, only cross-sectional household surveys are available

for analyzing the demand for medical care. This is not the case in Germany for which a large public panel database (German Socio-Economic Panel, GSOEP) has been available since 1984. Using this data set, three studies were recently carried out in this country to analyze the number of physician or hospitalization visits. The most recent study was conducted by Winkelman (2001) and reported in a discussion paper. As such this discussion paper presents preliminary work. In this paper, Winkelman (2001) pooled data from 1995 to 1999 to assess the impact of the 1997 German health reform using the number of doctor visits over the last three months as a primary outcome. This independence assumption (i.e. pooling data) may be a serious limitation if in fact the dependent variable exhibits dependence over time. If repeated observations on individuals are available over time, longitudinal count data methods presented in Chapter 2 should be used to analyze the determinants of the demand for medical care. Since some individuals may not consult their doctor or do not go to the hospital, the methodology should also address the problem of excessive zeros in the longitudinal case.

This chapter is organized as follows. Section 3.1 recalls the economic theory behind the analysis of the number of physician visits, through the framework of Grossman (1972). Econometric models retrieved from 10 studies conducted between 1988 to 2001 to analyze the number of physician or hospital visits are presented. A discussion driven by the paper of Winkelman (2001) will focus on the appropriateness of pooling various cross-sections and applying a univariate distribution to the pooled data rather than exploiting the longitudinal aspect of the data through Multivariate

distributions for longitudinal count data. Section 3.2 describes a longitudinal sub-set of the German Socio-Economic Panel that is analyzed in this paper. The data contains information on more than 4,000 individuals followed over 4 years and for whom information was available each year. The dependent count variable, the number of physician visits over the last three months, is observed each year for the period 1984-1987 along with several individual covariates. The dependent variable shows signs of overdispersion and correlation over time and is characterized by a high proportion of zeros as almost half of the population did not visit a physician for a particular year. Section 3.3 presents the results of the estimation of standard models currently used to analyze the number of physician visits: the Univariate Negative Binomial model applied to the pooled data set, a Quadrivariate Negative Binomial model and the Multivariate Negative Binomial Random Effects model developed by Haussman et al. (1984) for panel count data. A new methodology is presented in Section 3.4 based on an extension of the Univariate Negative Binomial Hurdle model to the longitudinal framework. Applied to the data, the resulting model is a Quadrivariate Negative Binomial Hurdle model which is nested to the Quadrivariate Negative Binomial model. An interesting feature of this model is that correlation is introduced in each stage of the Hurdle process. The Quadrivariate Negative Binomial Hurdle model is compared with the standard Univariate Negative Binomial Hurdle model applied to the pooled data in order to assess the independence assumption in two-part models (i.e. Hurdle models). Section 3.5 presents the results of the analysis of the number of specialist visits as in Pohlmeier and Ulrich (1995) who showed important differences

in the analysis of the number of generalist and specialists visits. Section 3.6 concludes by summarizing the main contributions of this chapter which are to have identified and proposed alternative ways of dealing with important methodological gaps in the economic literature relative to the analysis of the number of physician visits.

3.1 The Analysis of the Number of Physician Visits

Grossman (1972) derived a structural demand for health which has been the benchmark for the analysis of health care use. In Grossman's model, the demand for medical services is derived from the demand for health as individuals maximize their utility from health. This theory of rational choice over health care is defensible on several grounds because many health care options leave room for some thoughtful consideration or at least some planning. This theory also implies that the physician serves as an agent for patient-consumers and can make rational choices on their behalf even in urgent situations, which is reasonable to assume.

Over the last twenty years count data models have been widely applied in health economics to study the use of health services (e.g. number of physician/hospital visits) in the context of Grossman. However, the majority of these studies have analyzed cross-section data from household surveys. Even when the data was available for longitudinal analysis, two recent studies pooled several cross-sections assuming independence over time of the dependent count variable. A discussion motivated by the recent working paper of Winkelman (2001) will concentrate on this methodological aspect.

3.1.1 The Model of Grossman

Grossman's model of the demand for health is the most common economic approach to analyze the number of physician visits. In Grossman's model, the demand for health is seen as an individual investment decision similar to standard human capital theory. Health is considered as a durable good that depreciates over time. The stock of health capital can be accumulated by combining medical services and other inputs to improve health. In this context, the demand for medical care is derived because services serve to maintain or improve health capital. In this model, individuals maximize their utility from health given the production function of health and a budget constraint. The structural form of the demand for medical services can be modelled as:

$$D(t) = \beta_0 + \beta_1 H(t) + \beta_2 W(t) + \beta_3 P(t) + \beta_4 A(t) + \beta_5 Env(t) + \beta_6 E(t) + u(t). \tag{74}$$

D(t) represents the individual demand for medical services at time t (e.g., number of physician visits with realizations 0, 1, 2,...) which is a function of the existing stock of health capital H, the wage rate W, the price for medical services P, a time trend A to reflect the effect of age, a vector of environmental effects Env, education E and a stochastic error u. In this specification the coefficient associated with health, β_1, enters the specification with a coefficient different from unity. Poor health will result in a higher demand for medical care and we expect β_1 to be negative and different from unity. According to this theory, β_2 is expected to be negative because a high wage rate implies better return from work leading to a substitution of time for medical services.

Since higher prices for medical services should decrease the demand for medical care, it is expected that β_3, the coefficient associated with the price of medical services, is negative. In reality, the nominal price of medical services (exclusive of time costs) is almost zero due to public insurance schemes in most Western countries. According to moral hazard theory, insurance coverage should decrease the cost of medical care and therefore insurance coverage increases the demand. The time trend, A, measures the positive impact of age on the rate of health capital depreciation implying β_4 is positive. The impact of environmental controls on $D(t)$ should be positive if environmental factors (Env) are damaging such as jobs with a high risk of injury. The coefficient associated with education is expected to be positive because education (E) is associated with more efficient health production in Grossman's model. Using this framework, various authors have analyzed the number of physician or hospital visits.

3.1.2 Econometric Applications

This review presents the econometric models retrieved from 10 studies analyzing the number of physician or hospital visits along with the main findings of these papers. These models address the main characteristics of health care utilization data: 1) count outcomes such as the number of doctor consultations; 2) overdispersion as a result of unobserved heterogeneity; 3) the presence of covariates as identified in Grossman and 4) presence of a high proportion of zeros as individuals may not seek medical care. In general, two-part models such as Hurdle or Mixture models have been shown to be superior to Poisson or Negative Binomial models (Pohlmeier and Ulrich, 1995; Winkelmann, 2001; Gurmu, 1997; Deb and Trivedi, 1995). While it is difficult

to compare studies conducted in different health care settings and among different populations, health status has always been identified as the major determinant of the use of for medical services. Various conclusions were yielded with respect to the impact of private insurance, income and education. The following presents a review of the econometric models used in these 10 studies as well as their major findings.

Univariate Negative Binomial (UNB) Models Even when longitudinal count data was available, the Univariate Negative Binomial distribution has been used in almost every study analyzing the number of physician or hospital visits. For example, Geil et al. (1997) used a Univariate Negative Binomial Type 2 model to study hospitalization in Germany. They analyzed a subset of a German panel data set in which the waves 1984 to 1989, 1992 and 1994 were retrieved. The dependent variable was the number of hospital visits. Their data was unbalanced due to missing data but all individuals had to be observed at least twice. The analysis was based on 30,590 observations on 5,180 individuals aged 25 to 64. Their methods of analysis included an Univariate Negative Binomial model (equation 20) applied to the pooled data. In addition, to account for the panel aspect of the data, Geil et al. (1997) estimated a Multivariate Negative Binomial Random Effects model (equation 42). The results indicated small differences between the univariate and longitudinal models in terms of significance, signs and estimates of the explanatory variables. Contrary to Grossman's theory, age, income and education were found in their study to have a limited and insignificant impact on the demand for hospital visits.

Another contribution of the authors is to have shown that men and women react

differently to economic incentives. Geil et al. (1997) found that the kind of insurance coverage did not play an important role for the hospitalization decision but that women are more likely to react to economic incentives than men. For example, having a private insurance is only associated with a significant reduction in the number of hospital trips made by women. Estimation results indicated that a high level of private insurance coverage in Germany was not significant to curb the demand for hospital trips. This result was demonstrated using various dummy variables representing the different types of public and private health insurance contracts in Germany. For both genders, having children and working were significant variables but being married had a different impact on men and women.

Univariate Negative Binomial Hurdle (UNBH) Model Another important potential source of overdispersion is due to an excessive number of zeros (e.g., high proportion of zeros). Overdispersion may also induce an excess of zero counts than would otherwise be the case (Mullahy, 1986) and the excess of zeros may be a strict implication of unobserved heterogeneity (Mullahy, 1997). Since these two possibilities are not exclusive, it is important to test and to treat for an excess of zeros. This is very important in heath economics since depending on the characteristics of the population or the health care system studied, the proportion of individuals not consulting a doctor may be important. In studying the number of doctor visits, Hurdle or two-part models were preferred to their parental model in Pohlmeier and Ulrich (1995), Deb and Trivedi (1997), Gurmu (1995) and Winkelman (2001).

The Hurdle approach to the excess zeros problem captures the differences be-

tween zeros and at least one occurrence of physician visits by assuming that different processes govern the zeros and the positive counts. As such, the Hurdle model is a two-part model in which the first part is a binary outcome modelling the possibility of crossing the Hurdle. The second part is a truncated count model which describes positive observations arising from crossing the zero Hurdle. The rationale for using Hurdle models in the analysis of physician visits comes from the principal-agent theory (Pohlmeier and Ulrich, 1995) which suggests that a decision has to be made by the individual about whether to contact a physician or not. Then a decision is made about the frequency of trips according to a different process which may be influenced by the provider of medical care. This is captured in Hurdle models by explanatory variables, (β_1, β_2), allowed to have different impacts at each stage of the decision-making process. In a Hurdle model, two distributions $f_1$and f_2 generate the data. For a dependent count variable y_i, the probability distribution of the Univariate Hurdle model is given by:

$$\Pr(y_i = 0) = f_1(y_i = 0) \tag{75}$$

$$\text{and } \Pr(y_i = k) = f_2(y_i = k) * \frac{1 - f_1(y_i = 0)}{1 - f_2(y_i = 0)} \; with \; k = 1, 2, ..., . \tag{76}$$

The distributions f_1 and f_2 are univariate probability distributions for non-negative integers respectively, governing the Hurdle part and the process once the Hurdle has been passed, respectively. The distribution f_1 represents the distribution for the contact probability. Here $1 - f_1(y_i = 0)$ is the probability of crossing the Hurdle and

$1 - f_2(y_i = 0)$ is a normalization for f_2 . When $f_1 = f_2$, the Hurdle model collapses to the parent distribution (e.g. Univariate Negative Binomial). If f_1 and f_2 are from the same family of distributions, an excess of zeros can be tested with Likelihood Ratio Tests. If f_1 and f_2 are different univariate distributions the fitness of the model can be tested with the Vuong (1984) test or the Aikake Information Criteria (AIC). The log-likelihood function of the univariate Hurdle, $l(\beta_1, \beta_2)$, model can be written as:

$$\begin{aligned} l(\beta_1, \beta_2) \;=\; & \sum_{i=1}^{N} 1_{(y_i=0)} \ln f_1 (y_i = 0) + \sum_{i=1}^{N} 1_{(y_i>0)} \ln [1 - f_1 (y_i = 0)] + \\ & \sum_{i=1}^{N} 1_{(y_i>0)} \ln f_2 (y_i > 0) - \sum_{i=1}^{N} 1_{(y_i>0)} \ln [1 - f_2 (y_i = 0)] . \end{aligned} \tag{77}$$

Maximization is done with respect to the set of parameters β_1 and β_2 defining f_1 and f_2. The term $1_{(y_i=0)}$ is a binary indicator equal to 1 if the number of physician visits for patient i, y_i, is zero, 0 otherwise. The mean and the variance of the Hurdle model are determined by the probability of crossing the hurdle and by the moments of the truncated densities.

$$\mathrm{E}(y_i) = \Pr [y_i > 0] \, \mathrm{E}_{y_i>0} [y_i \mid y_i > 0] = \frac{1 - f_1(y_i = 0)}{1 - f_2(y_i = 0)} \mathrm{E}(y_i) \tag{78}$$

$$\mathrm{Var}(y_i) = \Pr [y_i = 0] \, \mathrm{E}_{y_i>0} [y_i \mid y_i > 0] + \Pr [y_i > 0] \, \mathrm{Var}_{y_i>0} [y_i \mid y_i > 0] \tag{79}$$

It follows from equation (78) that the mean of the Hurdle specification differs from the parental distribution (i.e., $f_1 = f_2$) by $\frac{1-f_1(y_i=0)}{1-f_2(y_i=0)}$. If the Hurdle model is the true model governing the data (i.e. 2 distributions generate the data: $f_1 \neq f_2$), estimating

the mean independently of the dispersion structure by assuming that only one process generates the data leads to a loss of efficiency and a loss of consistency since the mean function is not correctly specified. Therefore, in the presence of a significant excess of zeros, Generalized Estimating Equations, Pseudo Maximum Likelihood models or models not taking into account the potential over-representation of zeros should not be used.

Unobserved heterogeneity in the univariate Hurdle framework is integrated by taking a censored Negative Binomial specification for the binary process and a truncated at zero Negative Binomial distribution for the behavior once the Hurdle has been passed.

Pohlmeier and Ulrich (1995) analyzed the number of visits to general practitioners and to specialists made by 5,096 German employees in 1985. The authors developed a Negative Binomial distributed Hurdle model to model separately the decision to contact a physician and how often to contact one. Specification tests indicated that the Hurdle specification was preferred to the Univariate Negative Binomial Type 1 model suggesting that a different view of health care is necessary. The results indicated that the contact and frequency decisions followed different processes, which need to be modelled separately. The authors concluded that "ignoring these differences leads to inconsistent parameter estimates and to economic misinterpretations" (Pohlmeier and Ulrich, 1995). By using a variable physician density per 100,000 habitants, the authors also supported evidence of a supplier-induced demand in ambulatory services provided by general practitioners. The variable representing physician density was

insignificant for the contact decision but significant in explaining the frequency of visits once the contact had been made.

Their main results indicated that age, gender, income, private insurance, education and health status were significant for the contact decision for general practitioners, which supports Grossman's theory. However, once the Hurdle was crossed only age, health status and physician density were significant in explaining the number of visits. Education, income and private insurance had not significant impact on the frequency of visits, contradicting Grossman's theory. The results indicated that the decision to contact a physician differed by gender but once the Hurdle was crossed, gender was not longer significant.

The authors also provided evidence to perform separate analyses of the number of visits to GPs and specialists. For example, private insurance was not significant for the contact decision with a specialist. Although significant in both stages of the Hurdle, higher income led to more first contacts with specialists and fewer contacts with generalists. The univariate Hurdle Negative Binomial model was also used by Winkelman (2001) using German data and by Gurmu (1997) using American data.

Gurmu (1997) developed a semi-parametric estimation of a Hurdle model to evaluate the impact of managed care programs for Medicaid eligible on utilization of health care services using the number of doctors and health care centre visits during a period of four months in 1986 in the United States. The sample for analysis was constituted of 511 individuals. In this semi-parametric model, the distribution of the unobserved heterogeneity was approximated at each stage using Laguerre se-

ries expansion. The semi-parametric univariate Hurdle model provided a better fit for the data than the univariate Poisson and Negative Binomial models. The authors concluded that health status measures and age seemed to be more important in determining health care utilization than other socio-economic variables.

Mixture Models Mixture models were developed by Deb and Trivedi (1999) to avoid the strong dichotomy between the population of non-users and users in the Hurdle model. When the number of individuals not seeking medical care is low or if healthy individuals routinely seek health care, Hurdle models may be too restrictive. Mixture models discriminate between low and frequent users of medical services. Deb and Trivedi (1999) compared mixture versus Hurdle models using a cross-sectional sub-sample of the 1987-1988 National Medical Expenditure Survey to study the demand for medical care by the elderly in the United States. In their sample of 4,406 individuals aged 66 and over, 85% of individuals had one visit or more.

Assuming only two populations (low and high users), the density of a Two-Point Finite Mixture is given by:

$$f(y_i) = \pi f_1(y_i) + (1 - \pi) f_2(y_i). \tag{80}$$

Here f_1 and f_2 are Univariate Poisson or Negative Binomial distributions governing low and high user populations. If $\pi = 0$, this expression resumes to the parental distribution and $f_1 = f_2$. As with the Hurdle modelling, the set of explanatory variables can be the same with different parameters to incorporate differences between

a low or high user population. Several measures of medical utilization (the number of physician visits, the number of hospitalizations or the number of non-physicians visits) were used in this study. Based on Monte Carlo experiments and specification tests, the finite mixture Negative Binomial model was preferred over the Univariate Negative Binomial model and its Hurdle specification. Deb and Trivedi (1999) concluded that health measures were more important determinants of the demand for medical care by an elderly population than socio-economic measures.

Panel Probit Model Chiappori et al. (1998) investigated the presence of moral hazard in the demand for medical care in France following a health care reform. In 1993, the national French health insurance scheme reduced the coverage of ambulatory care and pharmaceutical products by 5%. Chiappori et al. (1998) used French insurance data to identify two groups in order to reproduce a controlled natural experiment.

A co-payment rate of 10% for physician visits was introduced in one group in 1994 while no change occurred for the other group. This allowed the authors to test if the number of visits per individual was modified by the co-payment rate. The longitudinal data set included information on 4,578 bank or insurance workers and their relatives followed during two years, 1994 and 1995 respectively.

The number of physician and specialist office visits were analyzed as well as the number of home visits. The authors used a panel probit model in which $y_{it} = 1$ if

individual i had a least one visit in year t and 0 otherwise as:

$$\begin{aligned} y_{it} &= 1 \text{ if } y_{it}^* > 0 \\ y_{it}^* &= f(x_{it}) + \epsilon_{it}. \end{aligned} \tag{81}$$

To test the presence of moral hazard the following formulation was used by Chiappori et al.(1998):

$$\begin{aligned} y_{it}^* = & \left(\alpha + \beta_0' \Gamma_i + \sum_{k=1}^{K} \beta_k X_{ik} + \sum_{k=1}^{K} \beta_k' \Gamma_i X_{ik} \right) + \\ & \gamma_t \left(\Delta\alpha + \Delta\beta_0' \Gamma_i + \sum_{k=1}^{K} \Delta\beta_k X_{ik} + \sum_{k=1}^{K} \Delta\beta_k' \Gamma_i X_{ik} \right) + \varepsilon_{it}. \end{aligned} \tag{82}$$

The term Γ_i is a dummy variable equal to 1 if individual i belongs to the control population (no change between 1994 and 1995) and 0 if i belongs to the group for which a co-payment of 10% was introduced. $\gamma_t = 1$ corresponds to the reference year 1994. Covariates included age, gender and location of work (represented by X_{ik}) and the products of Γ_i and γ_t with these observables. The error term ε_{it} is defined as the sum of two independent, central, normal errors u_i and v_{it}, with $\text{Var}(u_i) = \sigma^2$, $\text{Var}(v_{it}) = 1$, and $\text{Cov}(v_{i1}, v_{i2}) = 0$.

In this controlled natural experiment, it is expected that $\Delta\beta$=0 in the control group. If the introduction of the co-payment has no effect on consumption of medical services, $\Delta\beta_0$=0, $\Delta\beta_k$=0 and $\Delta\beta_k'$=0, $\forall$ k. If the introduction of co-payment has a impact on consumption, it should only affect the tested population ($\Gamma_i = 1$) then $\Delta\beta_k$=0, $\forall\ k$. If the $\Delta\beta_k \neq 0\ \forall\ k$, then the changes shouldn't be only explained by

the reform but by some other unobserved structural changes or by a misspecificaton of the model.

Their results indicated that the presence of moral hazard is not supported in the demand of doctor office visits suggesting a price elasticity close to zero. However a moderate decrease was observed in the number of home visits suggesting the presence of other non-monetary indirect costs such as transportation and time costs. Because of these additional costs, a decrease of 10% in the price of office visits may be not sufficient to impact the demand for office visits. On the other hand, the same 10% reduction in the price of home visits may have a bigger impact because non-monetary costs are zero for home visits. Their findings are in line with results derived from the well-known health insurance experiment conducted by the RAND corporation in the United States in the 1970s. In this experiment, among all physician services, the highest elasticity was observed in the demand for home visits. Other important results from the RAND experiment is that the elasticity of demand for physician services was positive a least for large co-payments (Newhouse, 1996). Chiappori et al. (1998) concluded that more work is guaranteed in the future such as using simulation equations or multinomial probit models to estimate the change in the whole distribution of number of visits per patient year. However, a major limitation of any probit specification is that while it allows to model the contact decision with the physician, it does not provide any information of the frequency of visits.

Probit-Poisson Log Normal Hurdle Model A new model, the Probit-Poisson log normal model with correlated errors, was recently proposed by Winkelman (2001)

to assess the impact of the 1997 reform of public health insurance in Germany on the number of doctor visits using German panel household data. The sample was constituted of Germans aged 20 to 60 years old benefiting from public health insurance and for whom information was available from 1995 to 1999 through the German Socio-Economic Panel database.

The new feature of this model is that the correlation between the zero-Hurdle and the positive part of the distribution is captured which represents a serious advantage over the univariate Hurdle model or the probit model presented previously. This model combined a probit model for the Hurdle part with a truncated Poisson-log normal model for the positive counts as follows. The model considers a latent indicator, z_i, variable such that:

$$y_i = 0 \text{ if } z_i \geq 0 \text{ where } z_i = x_i\gamma + \epsilon_i \tag{83}$$

$$y_i \mid y_i > 0 \sim \text{ truncated } Poisson(\lambda_i). \tag{84}$$

In this formulation, $\lambda_i = \exp(x_i\beta + u_i)$ and ϵ_i and u_i are bivariate normal distributed with mean 0 and covariance matrix

$$\Sigma = \begin{bmatrix} 1 & \rho\sigma \\ \rho\sigma & \rho^2 \end{bmatrix} \tag{85}$$

and $\epsilon_i \mid u_i \sim N(\rho u_i/\sigma, 1-\rho^2)$. The model is defined by writing:

$$P(y_i = 0 \mid u_i) = P(\varepsilon_i \geq -x_i\gamma \mid u_i) \tag{86}$$

$$= \Phi\left(\frac{x_i\gamma + \rho u_i/\sigma}{\sqrt{1-\rho^2}}\right) = \Phi_i^*(u_i). \tag{87}$$

This expression gives the following density function:

$$f(y_i \mid u_i) = \Phi_i^*(u_i)^{d_i} * \left[(1 - \Phi_i^*(u_i)) \frac{\exp(-\lambda_i(u_i))(\lambda_i(u_i))^{y_i}}{[1 - \exp(-\lambda_i(u_i))]\, y_i!}\right]^{1-d_i}. \tag{88}$$

The associated log-likelihood was then evaluated using Gauss-Hermite integration. Based on specification tests (i.e., Aiwake Information Criteria and Vuong's test), Winkelman (2001) concluded that the Probit-Poisson log-normal Hurdle model offered a better fit of the data than the Univariate Negative Binomial model and its Hurdle version applied to the pooled data. The model was also preferred to the Univariate Two-Point Mixture model of Deb and Trivedi (1997).

The dependent variable was the pooled number of doctor visits (general practitioners and specialists) from 1995 to 1999. The covariates of Winkelman's model were age, education, marital status, household size, active sports, health status, employment status and income. Privately insured individuals were excluded from the analysis. Taking 1995 as the reference year, a dummy was associated for each year (here 1995 is the reference year) to assess the impact of the 1997 health reform, which was their primary research objective.

Due to its Hurdle structure, the Probit-Poisson log normal model allows us to assess the effect of the reform at different parts of the distribution. The results indicated that the overall effect of the reform was a 10 percent reduction in the number

of doctor visits. The reform affected the Hurdle part much more than the positive part of the distribution. For example, the probability of being a user decreased by 6.7 percent between 1996 and 1998, almost three times the decreased frequency in the number of visits (2.6 percent). As a result, the reform may discourage "healthy individuals" from visiting their physician which in some cases will have prevented more serious illnesses. These findings were validated by the results of the Two-Point Mixture model for which the low user proportion exhibited the larger response.

Generalized Method of Moments Model The endogeneity of explanatory variables has been investigated in the univariate case by Windemeijer and Santos-Silva (1997) in a model explaining the number of visits to doctors with a self-reported binary health index as endogeneous variable.

In their multiplicative errors model, the conditional mean of the count variable is specified as:

$$E(y_i \mid x_i) = \mu_i = \exp(x_i\beta) \tag{89}$$

This equation defines the following regression model:

$$y_i = \exp(x_i\beta + \tau_i) = \exp(x_i\beta) * \varepsilon_i = \mu_i\varepsilon_i. \tag{90}$$

The Poisson estimator will be inconsistent if some of the k elements of vector of covariates are endogenous. If this is the case $E(\varepsilon_i \mid x_i) \neq 1$. Generalized Method of Moments techniques are available if instruments z_i are available such that E(ε_i |

$z_i) = 1$. In particular, the Generalized Method of Moments estimator minimizes the following objective function:

$$(y-\mu)'M^{-1}Z(Z'\widetilde{\Omega}^*Z)^{-1}Z'M^{-1}(y-\mu). \tag{91}$$

Here $\widetilde{y}$, $\widetilde{\mu}$ are column vectors of the observations and conditional means, respectively. M=diag(μ_i) and $\mu_i = \exp(x_i\beta)$, Z is an N*g matrix of instruments, and $Z'\widetilde{\Omega}^*Z$ is the asymptotic variance of $Z'M^{-1}(y-\mu)$. The optimal instruments are given by:

$$Z^* = \mathrm{E}(\Omega^{*-1}WX \mid Z) \tag{92}$$

and W=diag(y_i/μ_i). In the case of no endogenous regressors, Z=X and $\Omega^{*-1} = M^{-1}$. The Generalized Method of Moments estimator can be applied when there is more than one endogenous regressor.

Windemeijer and Santos Silva (1999) applied this model to a cross section of the British Health and Lifestyle Survey 1991-1992 using a sample of 4,814 individuals. A self-reported binary health index ($H_i = 1$ if health is fair or poor; 0 otherwise) acted as a possible endogenous regressor. Their model was specified as:

$$\begin{aligned} y_i &= \exp(\alpha H_i + x_i\beta) + u_i \\ H_i &= 1 \text{ if } H_i^* = z_i\delta + \varpi_i \geq 0;\ H_i = 0 \text{ otherwise.} \end{aligned} \tag{93}$$

In this equation, z_i represent the instruments which apart from x_i include variables

that explain health but have presumably no impact on the demand for doctor, other than via health (i.e., socio-economic status, alcohol consumption, smoking status, etc.). Based on a Haussman test, the authors rejected the endogeneity of the self-reported health measure in explaining the number of doctor visits. A similar result was found by Lahini and Xing (2002) who reported that the endogeneity of health status was not accepted in a recent cross-sectional analysis of veterans' health care utilization in the United States. In addition, Lahini and Xing (2002) tested and rejected the endogeneity of income and insurance. This was consistent with Cameron (1988) who reported that the endogeneity of health insurance status was accepted in some categories of health care services but not in others. As in Windemeijer and Santos Silva (1997), Cameron's results indicated that the health indicators were the most important factors in determining the number of doctor visits.

3.1.3 Discussion

The discussion is driven by the recent discussion paper developed by Winkelman (2001) as several limitations were associated with this paper. The major drawback of this preliminary work is the method of analysis which relies on pooling the data set. While five years of data were available for each individual, the author did not exploit the full richness of the panel structure of the data which allows us to follow individuals over time. "The basic empirical strategy is to pool the data over the five years and estimate the effects of the reforms by comparing the expected number of visits in 1998 and 1996 ceteris paribus, i.e., for an individual with given characteristics" (Winkelman, 2001). Pooling the data may be convenient because the Vuong and

AIC tests, two "univariate tests", can be used to discriminate between the Negative Binomial model and the new probit-Poisson normal model developed by Winkelman (2001), both non-nested models. Currently, there are no tests which allow us to discriminate among non-nested multivariate count data models for longitudinal count data.

Assuming that the data is independent over time is a strong assumption which may lead to inconsistent estimates if in fact the dependent counts are correlated over time. To demonstrate the importance of this methodological aspect consider the normal case in which two correlated random normal variables y_1 and y_2 are observed at time t=1 and t=2. It is assumed that μ_1, σ_1 and μ_2, σ_2 are respectively the means and standard deviations of the marginal distributions of y_1 and y_2. The bivariate normal distribution, the joint density of y_1 and y_2, is defined as follows:

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-1/2\left[\left(\epsilon_1^2 + \epsilon_2^2 + 2\rho\epsilon_1\epsilon_2\right)/(1-\rho^2)\right]\right\} \tag{94}$$

with $\epsilon_i = \left(\frac{y_t - \mu_t}{\sigma_t}\right)$, t=1,2. The correlation between y_1 and y_2 is modelled by the parameter ρ. The covariance of the bivariate normal distribution is given by $\sigma_{12} = \rho\sigma_1\sigma_2$. If $\rho = 0$, the two variables y_1 and y_2 are independent and the bivariate normal distribution is then equal to the product of the two marginal distributions of y_1 and y_2. In this case,

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-1/2\left[\left(\epsilon_1^2 + \epsilon_2^2\right)\right]\right\} = f(y_1)f(y_2). \tag{95}$$

Table 1: Monte Carlo Simulations. Bivariate versus Pooled Univariate Normal Models

		Bivariate	Pooled data
	Log-Likelihood	-5,399.63	-4,142.10
a_0	Mean	1.055	0.961
	(Std. Dev.)	(0.1890)	(0.115)
a_1	Mean	0.885	0.978
	(Std. Dev.)	(0.135)	(0.060)
σ_1	Mean	2.645	4.143
	(Std. Dev.)	(0.042)	(0.078)
σ_2	Mean	2.298	3.128
	(Std. Dev.)	(0.036)	(0.059)
ρ	Mean	0.652	
	(Std. Dev.)	(0.017)	

Using Monte Carlo simulations, equations (95) and (96) were estimated by maximum likelihood methods. Two random normal variables, y_1 and y_2 were generated with mean and standard deviations respectively equal at 1.96 and 5.04 for y_1 and 1.93 and 3.83 for y_2. The correlation between y_1 and y_2, was 0.915 and only a constant and one covariate were introduced in the model as $y_t = a_0 + a_1 x_{1t} + \varepsilon_t$ and $\varepsilon_t \sim N(\mu_t, \sigma_t)$, t=1,2. The values of a_0 and a_1 were 1.1 and 0.9, respectively.

As indicated in Table 1, which reproduces the results of the Monte Carlo simulations, if the variables are correlated over time, pooling the data gives inconsistent estimates while ignoring the correlation arising from the longitudinal aspect of the data. The values of the parameter estimates, a_0 and a_1 are different between the two approaches. The parameter estimates of a_0 and a_1 given by the estimation of the bivariate normal distribution are closer to the true values than those resulting from a pooled approach.

While these results are well-recognized in the normal case, pooling the data is sometimes used to analyze longitudinal count data even if the dependent variable is

correlated over time. For example, this is the case for the data set that Winkelmann (2001) used in his analysis of the impact of the 1997 German reform using data from 1995 to 1999. Using a subset derived from the same source of data as Winkelman (2001), information was retrieved for 7,000 individuals followed from 1995 to 1998. The year 1999 was not available in our version of the German Socio-Economic Panel database. The correlation between the number of doctor visits over the years varied between 0.3 to 0.4. Consistency should require that if the dependent count variable is correlated over time, correlation should be taken into account by the method of analysis.

Table 2 summarizes the studies presented in section 3.1.2 by authors and type of models used to analyze the number of physician or hospital visits. As shown in this table, recent empirical studies have concentrated on cross-section analyses and have generally ignored the richness of longitudinal panel count data. Only Geil et al. (1997) used a longitudinal count data model to analyze the determinants for hospitalization in Germany. In this study, the authors estimated a Univariate Negative Binomial model applied to the pooled data of various cross-sections of the German Socio-Economic Panel database and the Multivariate Negative Binomial Random Effects model in order to exploit the longitudinal aspect of the data. Results indicated that pooling the data or using the random effects model for panel count data provided similar parameter estimates in signs and magnitudes, which is in disagreement with the findings of the literature on patent data. The authors did not mention that this situation may have happened because the number of trips to a hospital were not

Table 2: Count Models in Health Economics

Author (Year)	Dependent Variable / Country / Year(s) of Analysis	Counts Models
Winkelman (2001)	# Dr. Visits Germany, 1995-1999	Univariate Negative Binomial Univariate Negative Binomial Hurdle Finite Mixture Probit-Poisson-log-normal
Lahiri and Xing (2002)	# Dr. Visits USA, 1992	Univariate Negative Binomial Univariate Negative Binomial Hurdle Generalized Method of Moments
Pohlmeier and Ulrich (1995)	# Dr. Visits Germany, 1995	Univariate Negative Binomial Univariate Negative Binomial Hurdle
Chiappori et al., 1998	# Dr. Visits France, 1993, 1994	Panel Probit Model
Geil et al. (1997)	# Hosp. Visits Germany, 1984-1989, 1992, 1994	Univariate Negative Binomial Multivariate Negative Binomial Random Effects Univariate Poisson Zero-Inflated Univariate Poisson Hurdle
Gurmu (1997)	# Dr. Visits USA, 1986	Univariate Negative Binomial Univariate Negative Binomial Hurdle Semi-parametric Hurdle
Deb and Trivedi (1997)	# Dr. Visits USA, 1987	Univariate Negative Binomial Univariate Negative Binomial Hurdle Finite Mixtures Negative Binomial
Windmeijer and Santos Silva (1997)	# Dr. Visits UK, 1991	Generalized Method of Moment Pseudo Maximum Likelihood
Lopez Nicolas (1998)	# Dr. Visits Spain, 1986	Univariate Negative Binomial Univariate Negative Binomial Hurdle
Cameron and Trivedi (1986)	# Dr. Visits Australia 1977-8 (Cross-section)	Ordinary Least Square Poisson Univariate Negative Binomial Quasi Generalized Pseudo Maximum Likelihood

correlated over years, which may justify pooling the data over time. For example, an analysis of a GSOEP subset, presenting data on 7,000 individuals followed from 1995 to 1998, indicated that the coefficients of the correlation of the annual number of hospital trips were less than 0.1 for the different time periods. In comparison to the data used in Haussman et al. (1984), a correlation of almost 0.9 for all years of analysis was observed for the annual number of patents awarded. It is therefore expected that parameter estimates may be different between univariate and longitudinal approaches in the presence of correlation as well known for patent data.

Another limitation associated with the methodology used by Winkelman (2001) to study the impact of the 1997 German health reform is that the analysis did not provide any insights into the disparity between men and women. Men and women may react differently to economic incentives as found in labor economics (Winkelman, 1994) and as shown by Geil et al. (1997) in their study of the hospital visits. This distinction may be very important in assessing a health care reform. For example, if women do not react to economic incentives, any reform encouraging private insurance coverage would not affect the use of health services by women. Of importance, for recent years the German Socio-Economic Panel database does not allow us to perform a separate analysis of the visits to general practitioners or specialists. As shown by Pohlmeier and Ulrich (1995), modelling the demand for generalists versus specialists is very important since the determinants are different.

3.2 The Data

The source of the data analyzed in this chapter is the German Socio-Economic Panel (GSOEP), a representative sample from Western Germany that has been collected since 1984. The GSOEP is a wide-ranging representative longitudinal study of private households, providing information on many indicators such as household composition, occupational biographies, employment, earnings, health and satisfaction. The Panel was started in 1984 and in 2000, there were more than 12,000 households and more than 20,000 persons sampled. The German Socio-Economic Panel is maintained by the Deutsche Institut für Wirtschaftsforschung (DIW, German Institute of Economic Research). The public use file of the German Socio-Economic Panel is provided free of charge to universities and research centres. An English version of the GSOEP is made available to non-German users who sign a data transfer contract with the German Institute of Economic Research. The version used in this chapter is the German Socio-Economic Panel 1984-1998. Since the time period ended in 1998, our data set cannot be used to reproduce the work of Winkelman (2001) who used data from 1995 to 1999.

Some limitations were identified with this data source. The English version does not contain personal information such as the address of the people surveyed. Therefore, it was impossible to incorporate a variable aimed at reflecting a supplier-induced demand effect such as physician density per size of community of residence as in Pohlmeier and Ulrich (1995). Secondly, the German Socio-Economic Panel does not always allow for a longitudinal analysis because some variables may be omitted from

some waves. For example, the most recent waves giving longitudinal information on the number of visits to a physician are from 1994, 1995, 1996, 1997 and 1998. Unfortunately, it is not possible for this period of time to distinguish between visits to a general practitioner or to a specialist, which may be important from an analytical point of view as shown by Pohlmeier and Ulrich (1995). The only GSOEP waves which provide this type of information on a longitudinal basis are those from 1984 to 1987 which are analyzed in this paper. For this reason, the data analyzed in this chapter was retrieved from the waves 1984, 1985, 1986 and 1987 of the German Socio-Economic Panel by selecting those individuals who provided information for each year of the period of analysis. The data is balanced and our sample contains information on 4,342 individuals (2,183 women and 2,150 men) followed from 1984 to 1987. The sample was composed of individuals aged 25 to 60 in 1984, thus excluding the majority of students and retired people. The data is presented by gender to highlight potential differences that may influence the determinants of the demand for doctor consultations as found by Geil et al. in their analysis of hospital visits in Germany. Tables 4 and 5 present the descriptive statistics of the sample by gender and per year.

3.2.1 Dependent Variable

The dependent variable is the number of physician consultations over the last three months. The main analysis concentrates on the analysis of the number of visits to a general practitioner (GP) in the last quarter prior to the surveys conducted in 1984, 1985, 1986 and 1987. Following Pohlmeier and Ulrich (1995), a separate analysis will be conducted to analyze the number of visits to specialists where specialties

are defined as all specialties in medicine except dentistry and radiology. The mean number of generalist practitioner and specialist visits and their associated standard deviations as well as the proportion of women who did not consult a GP or a specialists are given in rows 2 and 3 of Table 3. Similar information is reported in Table 4 for German men. The description of the covariates are also given in these tables.

Over time, the mean number of physician visits tends to increase from 1.39 to 1.76 for the female population and from 1.10 to 1.48 for men which may be seen as a consequence of the ageing of our sample. Consistent with this assumption, the health index decreases over the years for the two populations and the proportion of individuals having chronic conditions increases with time. No major health reforms were identified during this period of time which could have explained an increase in the number of consultations between 1984 and 1987.

The analysis of the mean and the variance of the dependent variables indicates that the data is characterized by overdispersion. Clearly the Poisson regression will not fit this data since the variance is greater than the mean. Overdispersion in count data can be caused by the presence of random effects such as unobservable individual characteristics, which can be modelled by the addition of a random effect in the Poisson conditional mean.

In our sample, between 50% (female) to 60% (male) did not have any contact with a general practitioner (GP) and 75% and 50% of men and women, respectively, did not consult a specialist for any given year of the period of the analysis. Therefore, it may be also incorrect to assume that a single process generates the data may be

incorrect due to the presence of a high proportion of zeros in the data.

Another important characteristic of the dependent variable is the presence of correlation over time as shown in Tables 5-8. The correlation between the number of GP visits during the period 1984 to 1987 ranges from 0.29 to 0.40 for women and from 0.19 to 0.44 for men, depending on the years. Similar ranges are observed for the specialist visits over time. The coefficients of correlation are generally higher for consecutive observations but the correlation between non-consecutive years is not negligeable. For example, the coefficient of correlation between the number of GP visits in 1984 and in 1987 is 0.29 (women) and 0.19 (men) while the correlation between two consecutive years range from 0.34 to 0.40 (women) and from 0.19 to 0.40 (men).

Not taking into account the correlation may lead to inefficient estimates since the repeated or time series aspect of the data is not recognized. Due to the high presence of zeros in our sample, the methodology should also be able to deal with unobserved heterogeneity and an excess of zeros since they are not mutually independent, while taking into account the correlation present in the data.

3.2.2 Independent Variables

Several variables were retrieved from the GSOEP to model the demand for medical care as in Grossman. It is expected that the use of health care services will increase with advancing age as indicated by our data. As indicated in Tables 3 and 4, the

Table 3: Descriptive Statistics: Means and (Standard Deviations). Female.

Female		1984	1985	1986	1987
GP Visits	Mean	1.40	1.52	1.67	1.76
	(Std. Dev.)	(3.29)	(3.39)	(3.50)	(3.95)
	Zero visits	57.34%	55.32%	50.09%	51.60%
Specialist Visits	Mean	1.74	1.64	1.82	1.74
	(Std. Dev.)	4.07	3.62	4.42	3.84
	Zero visits	53.25%	55.04%	53.11%	52.70%
Age	Mean	41.13	42.13	43.13	44.13
	(Std. Dev.)	(9.92)	(9.92)	(9.92)	(9.92)
Health Status	Mean	6.93	6.79	6.64	6.53
	(Std. Dev.)	(2.53)	(2.37)	(2.38)	(2.35)
Chronic Complaints	Mean	0.31	0.30	0.33	0.34
	(Std. Dev.)	(0.46)	(0.46)	(0.47)	(0.47)
Private Insurance	Mean	0.08	0.08	0.09	0.10
	(Std. Dev.)	(0.27)	(0.28)	(0.29)	(0.29)
Income	Mean	2,896.70	3,056.69	3,242.03	3,354.77
	(Std. Dev.)	(2,070.97)	(2,054.90)	(2,445.08)	(2,334.91)
Married	Mean	0.82	0.82	0.82	0.81
	(Std. Dev.)	(0.38)	(0.38)	(0.39)	(0.39)
Children	Mean	0.55	0.53	0.50	0.47
	(Std. Dev.)	(0.50)	(0.50)	(0.50)	(0.50)
Education in years	Mean	10.45	10.46	10.47	10.49
	(Std. Dev.)	(2.25)	(2.26)	(2.27)	(2.53)
Labor	Mean	0.54	0.52	0.51	0.52
	(Std. Dev.)	(0.50)	(0.50)	(0.50)	(0.50)
Non West German	Mean	0.24	0.24	0.24	0.24
	(Std. Dev.)	(0.43)	(0.43)	(0.43)	(0.43)

Table 4: Descriptive Statistics: Means and (Standard Deviations). Male.

Male		1984	1985	1986	1987
GP Visits	Mean	1.10	1.40	1.51	1.49
	(Std. Dev.)	(3.02)	(3.70)	(3.96)	(3.71)
	Zero visits	63.27%	58.61%	56.44%	57.30%
Specialist Visits	Mean	0.94	1.04	1.19	1.13
	(Std. Dev.)	(3.26)	(3.52)	(4.22)	(3.65)
	Zero visits	73.63%	75.67%	73.97%	72.14%
Age	Mean	41.55	42.55	43.55	44.55
	(Std. Dev.)	(9.74)	(9.74)	(9.74)	(9.74)
Health Index	Mean	7.18	7.01	6.94	6.80
	(Std. Dev.)	(2.51)	(2.34)	(2.31)	(2.28)
Chronic Complaints	Mean	0.29	0.28	0.30	0.31
	(Std. Dev.)	(0.45)	(0.45)	(0.46)	(0.46)
Private Insurance	Mean	0.11	0.11	0.13	0.14
	(Std. Dev.)	(0.32)	(0.32)	(0.33)	(0.34)
Income	Mean	3,043.87	3,137.30	3,302.54	3,426.49
	(Std. Dev.)	(1,842.13)	(2,083.55)	(2,374.59)	(2,305.21)
Married	Mean	0.82	0.83	0.83	0.84
	(Std. Dev.)	(0.38)	(0.38)	(0.37)	(0.37)
Children	Mean	0.54	0.52	0.50	0.48
	(Std. Dev.)	(0.50)	(0.50)	(0.50)	(0.50)
Education in years	Mean	11.24	11.26	11.26	11.27
	(Std. Dev.)	(2.55)	(2.56)	(2.57)	(2.57)
Labor	Mean	0.95	0.90	0.89	0.88
	(Std. Dev.)	(0.22)	(0.30)	(0.31)	(0.33)
Non West German	Mean	0.30	0.30	0.30	0.30
	(Std. Dev.)	(0.46)	(0.46)	(0.46)	(0.46)

Table 5: Correlation Matrix. Number of GP Visits, Female.

Female	GP visits 84	GP visits 85	GP visits 86	GP visits 87
GP visits 84	1	0.40	0.30	0.29
GP visits 85	0.40	1	0.34	0.38
GP visits 86	0.30	0.34	1	0.37
GP visits 87	0.29	0.38	0.37	1

Table 6: Correlation Matrix. Number of GP Visits, Male.

Male	GP visits 84	GP visits 85	GP visits 86	GP visits 87
GP visits 84	1	0.19	0.22	0.19
GP visits 85	0.19	1	0.40	0.42
GP visits 86	0.22	0.40	1	0.44
GP visits 87	0.19	0.42	0.44	1

Table 7: Correlation Matrix. Number of Specialist Visits, Female.

Female	Specialist visits 84	Specialist visits 85	Specialist visits 86	Specialist visits 87
Specialist visits 84	1	0.29	0.29	0.24
Specialist visits 85	0.29	1	0.28	0.27
Specialist visits 86	0.29	0.28	1	0.37
Specialist visits 87	0.24	0.27	0.37	1

Table 8: Correlation Matrix. Number of Specialist Visits, Male.

Male	Specialist visits 84	Specialist visits 85	Specialist visits 86	Specialist visits 87
Specialist visits 84	1	0.32	0.24	0.25
Specialist visits 85	0.32	1	0.30	0.26
Specialist visits 86	0.24	0.30	1	0.35
Specialist visits 87	0.25	0.26	0.35	1

mean age of our sample was 41 in 1984 for both women and men. The impact of age will be modelled as $age * 10^{-1}$ but also as $age * 10^{-1} + age * 10^{-3}$ to investigate non-linearities as supported by the findings of Pohlmeier and Ulrich (1995), Gurmu (1997), Geil et al. (1997) and Winkelman (2001).

Short term health status is reflected by a self perceived health index ranging from 0 to 10, where 10 corresponds to excellent health and 0 to the worst possible, as perceived by the respondent. This variable corresponds to a self-appraisal of physical and mental well-being. Individuals with the worst self-perceived index may use medical care more intensively. Similarly, respondents with a chronic condition are expected to exhibit a greater tendency to seek care and to use more physician services. The presence of chronic conditions is represented by a dummy variable equal to 1 if the individual surveyed has reported any chronic conditions and 0 otherwise. Approximately 30% of our sample suffer from chronic conditions. Self-perceived health scores decrease over time for women and men and the proportion of our sample with chronic

conditions increases over time.

Insurance status is represented by a dummy variable equal to 1 if the individual has private insurance and 0 otherwise. Distinguishing between several forms of public insurance as in Geil et al. (1997) didn't seem relevant since in Germany all members of the public health insurance plan have the same benefits for outpatient care. 90% of our sample have public insurance and this proportion is stable for the two first years and increases thereafter. In Germany, having a private insurance has been associated with a smaller number of physician visits than public insurance (Pohlmeier and Ulrich, 1995).

In most empirical studies, gender has always been a significant determinant of the use of health services, men using medical care services less often than women. As reflected in the data, a higher number of physician visits is associated with women. The proportion of women who have seen a GP is 10% to 20% higher than men, depending on the year. This is more apparent for specialists where women consulted on average 50% to 80% more than men. Geil et al. (1997) provided evidence that a separate analysis for men and women is preferable to analyze hospitalization visits in Germany. This approach seems relevant since in our sample, women are less privately insured than men (8% versus 11% in 1984) and they are more frequently out of the labor force than men (46% versus 5% in 1984). Another difference associated with gender is a higher proportion of non-Germans among men (30%) than women (24%).

The presence of children may also play an important role in the demand for doctor consultations. In order to control for this factor, the presence of children below 16 in

the household is represented by a dummy variable for which a value of 1 corresponds to the presence of children below the age of 16 in the household and zero otherwise. More than 50% of our population have children below the age of 16. In Grossman's model, the education level influences the behavior of individuals towards utilization of health care services as better education is associated with better health. Education, represented in terms of number of years of school, was on average greater for men (11.24 in 1984) than for women (10.50 in 1984).

Income and working status are used as proxies in the analysis to determine if economic factors have a role to play in the decision to use health services. According to Grossman theory, people with higher incomes should consume fewer health care services than lower income respondents due to opportunity costs or a substitution effect. The status of employment (1 if in the labor force, 0 otherwise) can represent the environmental factor of the Grossman model. The occupational status of the individual can be seen as a surrogate for the opportunity costs that are incurred when health care is consumed. Employment status may also measure differences in preferences. This aspect was studied in Germany by Pohlmeier and Ulrich (1995) who investigated the impact of satisfaction with work on the consumption of physician services through several variables related to work satisfaction. To be paid for sickness, German employees need to present sickness certificates from a doctor. Unsatisfied employees may visit their GP in order to be absent and paid for sickness. Moreover, opportunity costs are usually higher for employed than unemployed people and as a result, unemployed respondents may exhibit a greater tendency to seek medical care.

95% of men and 54% of women were working in 1984 in our sample. Curiously, the number of people in the labor force drops sharply from 1984 to 1985 and is steady thereafter but below the employment rate level of 1984. This was not expected since the retirement age is 65 in Germany and that unemployment statistics for Germany did not support a drop in the labor force for this time period.

Due to the limitations of the English version of the German Socio-Economic Panel database, it was not possible to integrate variables to model physician influences such as physician density as in Pohlmeier and Ulrich (1995) and Geil et al. (1997). It was also impossible to benefit from the variable "distance to main city" to model travelling costs as in Geil et al (1997). This variable was only available for the year 1994 in the English version of the German Socio-Economic Panel.

3.3 Standard Estimations

In this section the data set is estimated with the following models: the Univariate Negative Binomial (UNB) model applied to the pooled data, the Quadrivariate Negative Binomial (QNB) model and the Quadrivariate Negative Binomial Random Effects (QNBRE) model. Mixture models which split the population into low and high users of medical care were not considered because this approach is more appropriate for a population composed of older individuals who consult physicians regularly. However, for a younger population as in our sample, this does not seems reasonable since in any year more than half of the population did not consult a doctor. Generalized Method of Moments models were excluded since the endogeneity of self-reported health, in-

surance and income was not supported in recent studies (Lahini and Xing, 2002; Windemeijer and Santos Silva, 1999; Cameron 1988). Moreover, Generalized Method of Moments models cannot accommodate an excess of zeros because no distribution is specified.

The three models presented in this section are one-part models as they assume that only one process generates the data. The standard Hurdle version of the Univariate Negative Binomial model applied to the pooled data (Winkelman, 2001; Geil et al., 1999) will be presented in the next section and compared with a new two-part model for longitudinal count data designed to model overdispersion, correlation and an excess of zeros.

The objectives of this section are: 1) To compare standard univariate and multivariate approaches in the presence of correlation to determine to which extent conclusions should be based on estimations from methods applying a univariate distribution for count data to the pooled data; 2) To document if a separate analysis by gender should be considered in explaining the number of doctor visits and 3) To evaluate the different models in terms of predicting the mean counts and the percentage of zero visits at each time period.

3.3.1 Model Specifications

Univariate Negative Binomial (UNB) Model on Pooled Data The first model to be estimated is the Univariate Negative Binomial model applied to the pooled data set as in Winkelman (2001) and Geil et al. (1999). The specification of the Univariate Negative Binomial (UNB) distribution as given in as given in equation

(24) (Winkelman, 1994) was considered for the analysis. λ_i =exp($x_i\beta$) was defined as:

$$\lambda_i = \exp\left[\begin{array}{c}\beta_0 + \beta_1 Age_i + \beta_3 Gender_i + \beta_4 Ins_i + \beta_5 Edu_i + \beta_6 Lab_i + \beta_7 Chi_i + \\ \beta_8 Inc_i + \beta_9 Mar_i + \beta_{10} Hlt_i + \beta_{11} Chr_i + \beta_{12} NWG_i\end{array}\right]. \tag{96}$$

In this equation and from left to right, the covariates for individual i correspond to age (age*10^{-1} or age*10^{-1} + age^2 * 10^{-3} to investigate non-linearities), status of insurance (1 if private, 0 if public), the number of years of education, being in the labor force (1 if employed, 0 otherwise), having children below the age of 16 (1 if children, 0 otherwise), net household income*10^{-4}, presence of chronic complaints (1=if chronic complaints, 0 otherwise), self-reported health status measure (from 0 to 10, 10 being the best health self-appraisal, 0 the worst health self-appraisal), being married (1=married, 0 otherwise), not being West-German (1=non West-German, 0 otherwise) and gender (1 = male, 0 = female). Once λ_i is parametrized, estimation can be performed by maximizing with respect to β the set of parameters and α, the coefficient of overdispersion, the log-likelihood function, $l(\beta, \alpha)$, associated with the Univariate Negative Binomial distribution:

$$l(\beta,\alpha) = \sum_{i=1}^{N}\left[\begin{array}{c}\ln\Gamma(\alpha+y_i) - \ln\Gamma(\alpha) - \ln y_i! + \\ y_i \ln\lambda_i - (\alpha+y_i)\ln(1+\lambda_i)\end{array}\right]. \tag{97}$$

Quadrivariate Negative Binomial (QNB) Model Because the dependent variable, the number of physician visits, is correlated over time, it is hypothesized that

a multivariate approach is preferable to a univariate approach consisting of pooling the data. The Multivariate Negative Binomial model was introduced in section 2.1.2. In the following specification, $Y_{it} = (y_{i0}, y_{i1}, y_{i2}, y_{i3})$ represents the vector of the number of visits observed for individual i over the t periods when t = 0, 1, 2 and 3 representing the years 1984, 1985, 1986 and 1987, respectively. The counts observed at each time period are assumed to be independently distributed Poisson conditional on γ_i with mean $\theta_{it} = \lambda_{it}\gamma_i$. In this formulation, $\lambda_{it} = \exp(x_{it}\beta)$ in which x_{it} is the vector of covariates for individual i at period t. The variable γ_i, the unobserved heterogeneity variable, has a Gamma distribution. For simplicity, unknown parameters and strictly exogenous variables are suppressed in the following equations without loss of generality.

From equation (27) the Quadrivariate Negative Binomial (QNB) probability mass function of $Y_{it} = (y_{i0}, y_{i1}, y_{i2}, y_{i3})$ can be written as:

$$QNB(Y_{it}) = \frac{\Gamma(\alpha + \Sigma_{t=0}^{3} y_{it})}{\Gamma(\alpha)\Pi_{t=0}^{3}(y_{it}!)} \left[\prod_{t=0}^{3} \left\{ \left(\frac{\lambda_{it}}{(1 + \Sigma_{t=0}^{3}\lambda_{it})} \right)^{y_{it}} \right\}\right] \left[\frac{1}{(1 + \Sigma_{t=0}^{3}\lambda_{it})}\right]^{\alpha} \qquad (98)$$

The mean, the variance of y_{it} at period t and the coefficient of correlation between two time periods r and s of this distribution were given in equations (28), (29) and (31), respectively. Applied to our data, the log-likelihood of the Quadrivariate Negative Binomial probability distribution defined in equation (98) has a simple form given by:

$$l(\beta,\alpha)=\sum_{i=1}^{N}\left[\begin{array}{c}\ln\Gamma(\alpha+\Sigma_{t=0}^{3}y_{it})-\ln\Gamma(\alpha)-\Sigma_{t=0}^{3}(\ln y_{it}!)+\\ \Sigma_{t=0}^{3}\{y_{it}\ln\lambda_{it}\}-(\alpha+\Sigma_{t=0}^{3}y_{it})\ln(1+\Sigma_{t=0}^{3}\lambda_{it})\end{array}\right]. \tag{99}$$

In order to maximize equation (99) with respect to β, the vector of parameters and α, the coefficient of overdispersion, the following parametrization used for individual i at time t is:

$$\lambda_{it}=\exp\left[\begin{array}{c}\beta_0+\beta_1 Age_{it}+\beta_3 Gender_i+\beta_4 Ins_{it}+\beta_5 Edu_{it}+\beta_6 Lab_{it}+\beta_7 Chi_{it}+\\ \beta_8 Inc_{it}+\beta_9 Mar_{it}+\beta_{10} Hlt_{it}+\beta_{11} Chr_{it}+\beta_{12} Nwg_{it}\end{array}\right]. \tag{100}$$

Here t = 0, 1, 2, 3 represent the years 1984, 1985, 1986 and 1988. In this parametrization, all explanatory variables are individually time-variant. They correspond from left to right to the age, gender, the status of insurance, the number of years of education, being in the labor force, having children below the age of 16, the net household income, being married, having chronic complaints, the health status measure and being not West-German of a particular individual i at time t.

The associated coefficients β_0 to β_{12} are not time specific since this was not justified in our analysis. However, in the case of a study developed to measure the impact of a reform of the health care system (e.g., increase of private insurance coverage) on the consumption of medical care over time, it is possible to assign to each year a different coefficient for the insurance variable (β_{4t}, t=1984, 1985, 1986, 1987). This will allow us to determine the time-shape of the impact of the health care reform on the consumption of medical services. Similarly, a dummy variable can be introduced

for each year of the analysis as in Winkelman (2001).

Quadrivariate Negative Binomial Random Effects (QNBRE) Model Haussman et al. (1984) have developed a Negative Binomial Random Effects model for panel count data in order to allow for individual specific unobserved effects. This approach was used by Geil et al. (1997) in their analysis of the determinants of the demand for hospitalization generated by a panel of German individuals followed over years. When equation (42) is applied to our data and with λ_{it} given by equation (100), the log-likelihood function, $l(\beta, a, b)$, to maximize with respect to β the vector of parameters and to a and b, the coefficients associated with the Beta distribution, is:

$$l(\beta,a,b) = \sum_{i=1}^{N} \left[\begin{array}{l} \ln\Gamma(a+b) + \ln\Gamma(a+\sum_{t=0}^{3}\lambda_{it}) + \ln\Gamma(a+\sum_{t=0}^{3}y_{it}) - \\ \ln\Gamma(a) - \ln\Gamma(b) - \ln\Gamma(a+b+\sum_{t=0}^{3}\lambda_{it}+\sum_{t=0}^{3}y_{it}) + \\ \sum_{t=0}^{3}\left\{\ln\Gamma(\lambda_{it}+\sum_{t=0}^{3}y_{it}) - \ln\Gamma(\lambda_{it}) - \ln\Gamma(y_{it}+1)\right\} \end{array} \right]. \quad (101)$$

Overall the Negative Binomial Random Effects model accounts for over-dispersion, serial correlation and heterogeneity at the individual level. Two major limitations were identified in this model. Because individual effects are conditioned out of this model, it is not possible to generate either marginal effects or predicted values making it difficult to appreciate the fitness of the model (Greene, 1995; Haussman et al., 1984). In addition, as will be shown in section 3.4, it is not possible to Hurdle or inflate a Negative Binomial Random Effects distribution because by construction, individual-specific random effects are not identifiable for the initial count in Negative Binomial Random Effects models.

3.3.2 Results

The results of the analysis using one-part models are three-fold. First, it is shown that results based on a univariate analysis may lead to erroneous conclusions because the longitudinal aspect of the data is not considered. Pooling various cross sections does not address this problem because assuming independence of the data over time is not true in our case as shown in Tables 5-8. Secondly, specification tests indicated that a separate analysis by gender is preferable to explain the demand for medical care. Finally, it is shown that one-part models for longitudinal count data do not provide a good fit of the data. They generally fail to predict the number of zeros at each time period which is a serious limitation of these models.

Parameter Estimates Table 9 presents for the whole sample the maximum likelihood parameter estimates and their standard errors for each of the three models estimated in this section. Column A of Table 9 corresponds to the Univariate Negative Binomial (UNB) model when the waves 1984, 1985, 1986 and 1987 are pooled as in the German studies of Geil et al. (1997) and Winkelman (2001). Column B presents the results from the estimation of the Quadrivariate Negative Binomial (QNB) model and column C presents the results from the estimation of the Quadrivariate Negative Binomial Random Effects (QNBRE) model of Haussman et al. (1984) as in Geil et al. (1997). The results presented in Table 9 concern only one-part models (i.e., one distribution governs the data).

Table 9: Maximum Likelihood Parameter Estimates: Negative Binomial Models.

Maximum Likelihood Estimation		UNB Pooled data	QNB 84-87	QNBRE 84-87
Number of Observations (N*T):		4342*4	4342*4	4342*4
Log of Likelihood:		-25,795.33	-28,923.019	-25,028.921
Variable	Parameter	A	B	C
Constant	β_0	1.838*	0.889*	0.462*
		(0.125)	(0.146)	(0.139)
Age*10^{-1}	β_1	0.176*	0.257*	0.182*
		(0.016)	(0.019)	(0.018)
Gender	β_3	-0.134*	-0.116*	-0.168*
		(0.030)	(0.039)	(0.034)
Private Insurance	β_4	-0.319*	-0.232*	-0.232*
		(0.051)	(0.048)	(0.054)
Education	β_5	-0.048*	-0.064*	-0.052*
		(0.007)	(0.009)	(0.008)
Working	β_6	0.121*	0.073*	0.078*
		(0.033)	(0.025)	(0.033)
Children	β_7	0.039	0.085*	-0.021
		(0.031)	(0.030)	(0.033)
Income	β_8	-0.485*	-0.208*	-0.230*
		(0.086)	(0.063)	(0.081)
Married	β_9	0.050	-0.116*	0.0538
		(0.038)	(0.040)	(0.042)
Health	β_{10}	-0.192*	-0.136*	-0.118*
		(0.006)	(0.004)	(0.005)
Chronic conditions	β_{11}	0.554*	0.390*	0.439*
		(0.031)	(0.004)	(0.028)
Non West German	β_{12}	0.152*	0.156*	0.003
		(0.032)	(0.043)	(0.038)
Overdispersion	α	-0.693*	0.894*	
		(0.007)	(0.012)	
	a			1.650*
				(0.030)
	b			1.538*
				(0.039)
* indicates significant at the 5 % level.				

Age, Education, Income, Private Insurance, Working Status and Health Status Globally, the variables representing age, education, income, private insurance, working and health status are significant in all three models with the expected signs, which support the theory of Grossman. The rate of depreciation of health capital stock increases with age. The health status variables (self-reported health and the presence of chronic complaints) are an important determinant of the demand for medical care. Education has a significant negative impact on the demand for medical care because education contributes to a more efficient production of health. Income has a significant negative impact for the demand of medical care due to a substitution effect. According to the theory, private insurance has a significant negative effect on the demand for medical care. Being in the labor force has a significant positive impact on the demand for GP visits according to the three models. Working could represent opportunity costs but also measures dissatisfaction with the job as investigated by Pohlmeier and Ulrich (1995). In Germany, employees who are sick for more than two days need to present a certificate of illness issued by a physician in order to receive sick leave payments. The coefficients of overdispersion are always significant, indicating unobserved heterogeneity in the data.

The models have different implications regarding the impact of the remaining variables (having children, being married and being non west German). These differences are highlighted by a comparison of the results of the different models.

Univariate versus Longitudinal models Having children and being married are significant in the Quadrivariate Negative Binomial model but not in the univariate

model. According to the Quadrivariate Negative Binomial model, being married significantly decreases the number of GP visits while having children significantly increases the number of visits to General Practitioner.. Another difference between the Univariate Negative Binomial and Quadrivariate Negative Binomial estimations is the value on the parameter estimates of private insurance, working status, income, health and chronic conditions which are slightly greater in the univariate model. The Quadrivariate Negative Binomial and Quadrivariate Negative Binomial Random Effects parameter estimates are very close for these variables.

Except for the differences observed in the impact of children, being married but also being non-west German, the different models describe similar patterns. This result was also reported by Geil et al. (1997) whose analysis indicated that the pooled cross-section estimates were similar to the Quadrivariate Negative Binomial Random Effects model when studying the determinants of hospitalization in Germany. However as it will be shown later on, this is not anymore true in our example when the analysis is done by gender.

Quadrivariate Negative Binomial Model versus Quadrivariate Negative Binomial Random Effects Model If we compare the two longitudinal models, there is a significant increase in the log-likelihood function by adding specific-individual effects in the Quadrivariate Negative Binomial specification. However, the estimates of private insurance, education, being in the labor force, income, and to a lesser extent, age, gender and the two health variables are similar in magnitude, sign and significance in the two specifications of the Multivariate Negative Binomial

model. The only differences between the two models are again in the impact of being married and having children.

According to previous empirical results, having children significantly increases the number of visits in the standard Quadrivariate Negative Binomial model but this effect is negative and not significant in the Quadrivariate Negative Binomial Random Effects model. Being married has a significant negative impact in the Quadrivariate Negative Binomial model but a positive non significant impact in the Quadrivariate Negative Binomial Random Effects model. Lastly, not being West-German has a positive impact in both models but this effect is only significant in the Quadrivariate Negative Binomial model and the Univariate Negative Binomial model.

Gender Analysis In all models, gender was found to be significant and a reduced number of visits was associated with the male population. These results have been reported in almost all empirical studies on the determinants of medical utilization. However, including a dummy variable reflecting the gender as a covariate does not allow us to know if the explanatory variables have a different impact in modelling the demand for GP visits as a function of gender. As shown in labor and health economics, women and men respond differently to economic incentives (Zimmerman, 1993; Geil et al., 1997). Likelihood Ratio Tests were therefore conducted to determine if splitting the sample by gender was justified from a statistical point of view.

The Likelihood Ratio Test (LRT) values are displayed in Table 10. Splitting the data by gender is justified in the Univariate Negative Binomial (UNB) and the Quadri-

Table 10: Likelihood Ratio Test Values for Splitting the Sample by Gender.

	UNB	QNB	QNBRE
	Pooled data	84-87	84-87
All observations	40.78	78.79	21.65
	$(\chi^2_{13} = 22.36)$	$(\chi^2_{13} = 22.36)$	$(\chi^2_{14} = 23.69)$

variate Negative Binomial (QNB) models but not with the Quadrivariate Negative Binomial Random Effects (QNBRE) model. This is an interesting result which suggests that allowing individual-specific overdispersion by the introduction of Random Effects may be sufficient to explain differences between men and women. However, in Quadrivariate Negative Binomial Random Effects models, no information is given on any disparities between men and women.

According to the results of Quadrivariate Negative Binomial (QNB) model estimations reported in column A (female) and B (male) of Table 11, women and men differ in terms of income, being married, having children and occupational status. Income, being married and having children are significant in explaining the demand of men for GP visits but these variables are not significant for women. Working has a positive significant effect for women and a negative non-significant impact for men. All other parameters are similar between the two populations in the Quadrivariate Negative Binomial model.

Bigger differences between univariate and multivariate models appear when the analysis is conducted by gender. The results of the Univariate Negative Binomial (UNB) model estimation are given in column C (female) and D (male). For example, income has a significant negative impact for women in the Univariate Negative Bino-

Table 11: Maximum Likelihood Parameter Estimates: Negative Binomial Models, Gender Analysis.

Maximum Likelihood Estimation		QNB 84-87	QNB 84-87	UNB Pooled data	UNB Pooled data
GPs visits		Female	Male	Female	Male
Number of Observations (N*T)		2183*4	2159*4	2183*4	2159*4
Log of Likelihood:		-14900.957	-13982.667	-13574.118	-12200.817
Variable	Parameter	A	B	C	D
Constant	β_0	0.705*	1.001*	1.713*	1.917*
		(0.226)	(0.209)	(0.186)	(0.181)
Age*10^{-1}	β_1	0.254*	0.268*	0.171*	0.177*
		(0.028)	(0.028)	(0.023)	(0.023)
Private Insurance	β_4	-0.226*	-0.229*	-0.316*	-0.296*
		(0.070)	(0.066)	(0.074)	(0.072)
Education	β_5	-0.051*	-0.077*	-0.038*	-0.060*
		(0.013)	(0.013))	(0.010)	(0.010)
Working	β_6	0.156*	-0.038	0.122*	0.090
		(0.03)	(-0.043)	(0.039)	(0.066)
Children	β_7	0.067	0.132*	-0.037	0.026
		(0.042)	(0.0428)	(0.045)	(0.047)
Income*10^{-4}	β_8	-0.123	-0.339*	-0.325*	-0.700*
		(0.077)	(0.100)	(0.105)	(0.136)
Married	β_9	-0.068	-0.182*	-0.005	0.121**
		(0.053)	(0.062)	(0.050)	(0.061)
Health	β_{10}	-0.136*	-0.135*	-0.191*	-0.192*
		(0.005)	(0.005)	(0.008)	(0.009)
Chronic conditions	β_{11}	0.269*	0.521*	0.468*	0.636*
		(0.027)	(0.028)	(0.042)	(0.046)
Non West German	β_{12}	0.149*	0.154*	0.135*	0.168*
		(0.063)	(0.059)	(0.045)	(0.045)
Overdispersion	α	0.901*	0.888*	0.719*	0.668*
		(0.017)	(0.018)	(0.010)	(0.009)
	r				
	s				

* indicates significant at the 5 % level.

mial model but this impact is not significant for women in the Quadrivariate Negative Binomial model. Having children and being married are not significant variables according to the Univariate Negative Binomial estimation for men but these variables are significant in the Quadrivariate Negative Binomial model for this population. While the impact of income, being married and having children was different among men and women in the Quadrivariate Negative Binomial model, in the Univariate Negative Binomial model, the only difference between males and females was related to working status. All other parameters were similar in the Univariate Negative Binomial model.

These results suggest that in models supporting a splitting of the sample by gender and in the presence of correlation, important differences appear between the estimations of the univariate and the Quadrivariate Negative Binomial models. However, this finding is attenuated because Likelihood Ratio Tests did not support the splitting of the sample when a Negative Binomial random effect model was used.

Predictions Once the parameters were estimated, the predicted mean number of GP visits were determined by gender and per year as well as the predicted mean number of zeros using the Quadrivariate Negative Binomial model. It was impossible to predict the mean counts and the number of zeros from the Quadrivariate Negative Binomial Random Effects model because the marginals are not defined in this model (Haussman et al., 1984; Greene, 1998). The difference between the observed and predicted values of the mean number of GP visits is less than 10% as it can be seen in Figure 1. Generally, the predicted mean values are smaller than the observed values.

The omission in our structural equations of important variables such as a physician density variable to represent a supply-side induced effect may explain this result.

Table 12 presents the predicted number of zeros using the Quadrivariate Negative Binomial model. In the calculations, the number of zeros at each time period was determined jointly when the count at time t was set equal to 0 according to the joint distribution of the number of visits from 1984 to 1987. For example, the number of zero visits in 1985 was calculated as $QNB(y_{84}, y_{85} = 0, y_{86}, y_{87})$ according to the density function of the Quadrivariate Negative Binomial given in equation (98). The results indicate that the Quadrivariate Negative Binomial model is unable to predict the number of zeros at each time period offering a poor fit for the data. However, the probability for a particular individual of having zero visits each year of the period 1984 to 1987 was calculated as $QNB(0, 0, 0, 0)$. The results are presented in Table 12 and indicates that the Quadrivariate Negative Binomial model predicts that 21% and 26% of women and men did not consult a GP over the full time period (versus observed values of 23% and 24% for women and men, respectively).

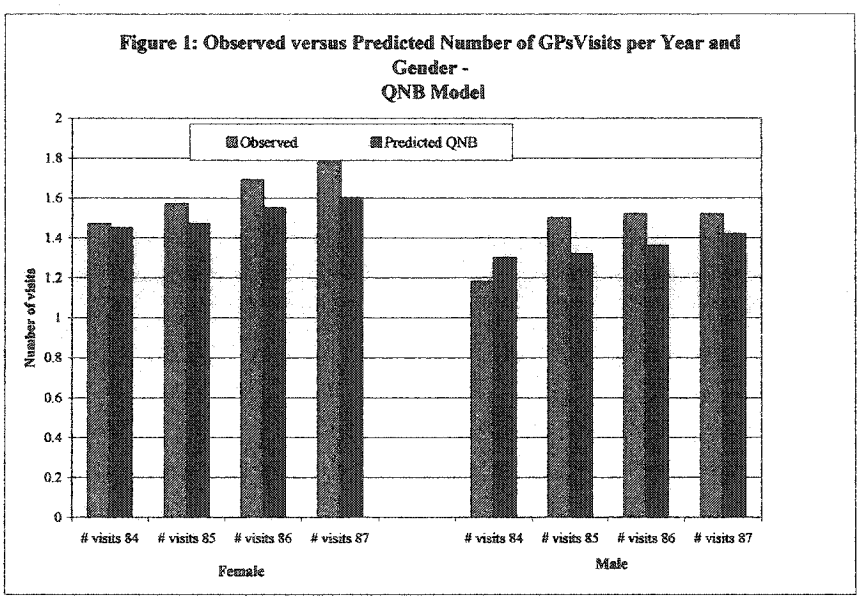
Results also indicate that the Univariate Negative Binomial model applied to the pooled data performs well in predicting the number of zeros and the mean counts but the longitudinal aspect of the data is not integrated as the dependent variable is assumed to be independent over time. In comparison, the Quadrivariate Negative Binomial model takes into account the repeated nature of the data. The predicted coefficient of correlation between two consecutive years, calculated according to equation (31), were approximately 0.6 for women and 0.5 for men which compares to ac-

Table 12: Zeros: Observed and Predicted Values - QNB Model - Analysis by Gender

	QNB Female		QNB Male	
	1984-1987	1984-1987	1984-1987	1984-1987
	Observed	Predicted	Observed	Predicted
0 visits in 1984	0.56	0.06	0.63	0.08
0 visits in 1985	0.55	0.06	0.58	0.09
0 visits in 1986	0.50	0.07	0.56	0.09
0 visits in 1987	0.52	0.07	0.58	0.09
0 visits 1984-1987	0.23	0.21	0.24	0.26

tual values of 0.4 and 0.5. The correlation between non-consecutive observations is generally overestimated. .

However, the Quadrivariate Negative Binomial model assumes that only one process generates the data which may be not true due to the high proportion of zero visits in our sample. Due to these limitations, other alternatives have to be looked into to obtain a better fit for the data and to draw final conclusions.



3.4 Quadrivariate Negative Binomial Hurdle (QNBH) Model

The models presented so far assumed that only one process generates the data which may not be true to the high proportion of zeros in our data as a result of 0 consultations. The Quadrivariate Negative Binomial model was unable to correctly predict the mean number of zeros for a particular year and the construction of the Quadrivariate Negative Binomial Random Effect model makes it impossible to calculate predicted values. To fix these shortcomings, a Quadrivariate Negative Binomial Hurdle model, nested to the Quadrivariate Negative Binomial distribution, is presented in this section. This Multivariate Negative Binomial Hurdle model addresses five characteristics that are commonly found in health care utilization data sets: repeated observations on a count used to analyze the use of medical services, the presence of covariates, unobserved heterogeneity, a count dependence structure and an excess of zeros. Results are compared with the standard Univariate Negative Binomial Hurdle model applied to the pooled data.

3.4.1 Model Specification

Multivariate distributions treat the longitudinal aspect of the data by conditioning to previous events. By definition, any Multivariate Negative Binomial distribution (MNB) can be written as the product of conditional distributions as:

$$MNB(Y_{it}) = NB_0(y_{i0})NB_1(y_{i1} \mid y_{i0}), ..., NB_t(y_{it} \mid y_{i0}, y_{i1}, ..., y_{it-1}). \qquad (102)$$

In this expression $Y_{it} = (y_{i0}, y_{i1}, ...y_{it})$ is the vector of counts observed for indi-

vidual i over the t periods. y_{it} is the count observed at time t for individual i and $t = 1...T$. It follows from this equation that the initial value at $t = 0$, y_{i0}, should be treated as an endogenous variable following an Univariate Negative Binomial distribution (NB_0). The expression $NB_1(y_{i1} \mid y_{i0})$ represents the conditional Negative Binomial distribution of the counts at t=1 conditional on the counts observed at t=0. This conditional distribution, $NB_1(y_{i1} \mid y_{i0})$, is calculated as the ratio of the Bivariate Negative Binomial distribution of (y_{i0}, y_{i1}) and the Univariate Negative Binomial distribution of y_{i0}. For each subsequent time periods, the counts are similarly calculated conditional on the previous observed counts.

As revealed by the literature review, applications of Hurdle models in health economics have so far concentrated on the univariate case and have not been extended to the longitudinal case. In the following the concept of the Hurdle model for univariate count data is extended to the longitudinal case to address the problems of excessive zeros in the longitudinal case and correlation over time. This methodology consists to hurdle each conditional distribution defined in equation (102) so that the model deals with the problem of zeros as well as with the correlation across time. The construction of the Hurdle model implies that at each time period, the zero counts and the positive counts are generated according to two different processes.

A parallel to the principal-agent theory is to consider that each year individuals have to make a choice about visiting a physician and how many times. This is not unrealistic since sick (i.e. sick enough to go to the physician) is not generally an absorbent state but rather a temporary one. This is reflected by the fact that each

year almost 80% of the sample have two or fewer visits over the three-month period prior to the survey. However, this assumption may not hold for those severely sick individuals who require continuous follow-up examinations over time. It is believed that this problem will not invalidate our results since only 3% to 5% of the sample have more than 6 visits (assuming an average of 1 visit every 2 weeks) over the three-month period prior to the survey. Similarly, this model does not consider multiple illness cells since no information was available for the analysis.

The following presents a Quadrivariate Negative Binomial Hurdle distribution which was designed to analyze the number of physician visits made by individuals followed over the period 1984-1987. The Quadrivariate Negative Binomial distribution given in equation (98) was decomposed into a product of conditional probabilities as in equation (102). The Quadrivariate Negative Binomial Hurdle was defined as the product of three conditional Negative Binomial Hurdle distributions (t = 1985, 1986, 1987) and the Univariate Negative Binomial Hurdle distribution for the first observation in 1984.

$$QNBH(y_{i84}, y_{i85}, y_{i86}, y_{i87}) = NBH_{84}(y_{i84})NBH_{85}(y_{i85} \mid y_{i84})$$
$$NBH_{86}(y_{i86} \mid y_{i85}, y_{i84})NBH_{i87}(y_{i87} \mid y_{i86}, y_{i85}, y_{i84})$$

Since there is no previous information available for the first observation, the Hurdle specification of the initial observation y_{i84} for individual i corresponds to a standard Univariate Negative Binomial Hurdle (NBH_{84}) model. Each subsequent time period (t = 1985, 1986 and 1987) was hurdled conditional on the previous

counts. In this equation, $NBH_{85}(y_{85} \mid y_{84})$ is the Negative Binomial Hurdle distribution for the counts in 1985 conditional on the counts observed in 1984. Similarly $NBH_{86}(y_{86} \mid y_{85},y_{84})$ is the Negative Binomial Hurdle distribution for the counts in 1986 conditional on the counts observed in 1985 and 1984 and $NBH_{87}(y_{87} \mid y_{86},y_{85},y_{84})$ is the Negative Binomial Hurdle distribution for the counts in 1987 conditional on the counts observed in 1985 and 1984. The log-likelihood function of the Quadrivariate Negative Binomial Hurdle distribution ($LnLQNBH$) generating the counts $(y_{i84},y_{i85},y_{i86},y_{i87})$, was decomposed into the sum of four log-likelihood functions:

$$LnLQNBH(y_{i84},y_{i85},y_{i86},y_{i87}) = \sum_{i=1}^{N} LnNBH_{84}(y_{i84}) + \sum_{i=1}^{N} LnNBH_{85}(y_{i85} \mid y_{i84}) + \sum_{i=1}^{N} LnNBH_{86}(y_{i86} \mid y_{i85},y_{i84}) + \sum_{i=1}^{N} LnNBH_{87}(y_{i87} \mid y_{i86},y_{i85},y_{i84}). \quad (103)$$

In this expression $\sum_{i=1}^{N} LnNBH_{84}$, is the log-likelihood function associated with the Univariate Negative Binomial Hurdle distribution applied to the counts observed in the first year of observation in 1984. Here $\sum_{i=1}^{N} LnNBH_{84}$ is written as:

$$\sum_{i=1}^{N} LnNBH_{84}(\alpha_1,\beta_1,\alpha_2,\beta_2) = \sum_{i=1}^{N} \begin{bmatrix} 1_{(y_{i84}=0)} \ln(NB_1(y_{i84}=0))+ \\ 1_{(y_{i84}>0)} \ln(1-NB_1(y_{i84}=0))+ \\ 1_{(y_{i84}>0)} \ln(NB_2(y_{i84}))+ \\ 1_{(y_{i84}>0)} \ln(NB_2(y_{i84}=0)) \end{bmatrix}. \quad (104)$$

The term $1_{(yi84=0)} = 1$ if $y_{i84} = 0$ and 0 otherwise. The distribution NB_1 is the Univariate Negative Binomial distribution governing the zeros in 1984 and NB_2 is

the Univariate Negative Binomial distribution governing the positive counts in 1984. Maximization is conducted with respects to the parameters α_1 and $\beta_1$defining the distribution NB_1 and α_2 and $\beta_2$defining NB_2. The terms α's are the coefficients of overdispersion and the β's the coefficients associated with the mean functions of NB_1 and NB_2.

For the subsequent years each conditional distributions was hurdled. For example, $\sum_{i=1}^{N} LnNBH_{85}$, corresponds to the log-likelihood function of the Negative Binomial Hurdle model in 1985 conditional on the counts observed in 1984. Then $\sum_{i=1}^{N} LnNBH_{85}$ was decomposed as:

$$\sum_{i=1}^{N} LnNBH_{85} = \sum_{i=1}^{N} \begin{bmatrix} 1_{(y_{i85}=0)} \ln(NB_1(y_{i85} = 0 \mid y_{i84}))+ \\ 1_{(y_{i85}>0)} \ln(1 - NB_1(y_{i85} = 0 \mid y_{i84}))+ \\ 1_{(y_{i85}>0)} \ln(NB_2(y_{i85} \mid y_{i84}))+ \\ 1_{(y_{i85}>0)} \ln(NB_2(y_{i85} = 0 \mid y_{i84})) \end{bmatrix} \tag{105}$$

In this formulation, NB_1 is the conditional Negative Binomial distribution generating the zeros in 1985 conditional on the counts observed in 1984 witch is the ratio of a Bivariate Negative Binomial distribution and the Univariate Negative Binomial distribution as

$$NB_1(y_{i85} = 0 \mid y_{i84}) = \frac{BNB_1(y_{i85} = 0, y_{i84})}{UNB_1(y_{i84} = 0)} \tag{106}$$

Similarly, the log-likelihood $\sum_{i=1}^{N} LnNBH_{86}$ is calculated by conditioning over

the counts observed in 1984 and 1985 and $\sum_{i=1}^{N} LnNBH_{87}$ by conditioning over the counts observed in 1984, 1985 and 1986. Estimation is obtained by maximizing the sum of these four log-likelihood functions.

The Quadrivariate Negative Binomial Hurdle model assumes that in each stage the analysis is done conditionally on the number of previous physician visits observed the years before. Therefore, correlation is introduced in the contact decision and in the frequency of the visits. For t =1984, 1985, 1986 and 1987, the mean and variance of this model are given by:

$$\mathrm{E}(y_{it}) = \Pr\left[y_{it} > 0 \mid Y_{it-1}\right] E_{y_{it}>0}\left[y_{it} \mid y_{it} > 0\right] \tag{107}$$

$$\mathrm{Var}(y_{it}) = \Pr\left[y_{it} = 0 \mid Y_{it-1}\right] \mathrm{E}_{y_{it}>0}\left[y_{it} \mid y_{it} > 0\right] + \Pr\left[y_{it} > 0\right] \mathrm{Var}_{y_{it}>0}\left[y_{it} \mid y_{it} > 0\right]. \tag{108}$$

By construction, the Quadrivariate Negative Binomial Hurdle model is nested to the Quadrivariate Negative Binomial model. When NB_1=NB_2, the likelihood function of the Quadrivariate Negative Binomial Hurdle model given in equation (104) resumes to equation (99). Therefore, using Likelihood Ratio Tests, it is possible to know if the excess of zeros is significant when analyzing longitudinal count data.

By definition, the likelihood for the Univariate Negative Binomial Hurdle density for the first observation has to be added to the quadrivariate likelihood function in order to respect the conditional definition of any multivariate distribution given in equation (102). Therefore, a random effects version of a Multivariate Negative Binomial Hurdle model in which the random effects would be individual-specific,

cannot be derived. In Haussman et al.'s model (1984), it is impossible to define for the initial observation a Univariate Negative Binomial Fixed Effects distribution because ϕ_i and $\exp(u_i)$, which defines the individual random effects $\delta_i = \frac{\phi_i}{\exp(u_i)}$, are not identifiable in the univariate case.

3.4.2 Results

The likelihood functions of the Quadrivariate Negative Binomial and Quadrivariate Negative Binomial Hurdle (QNBH) distributions were used to test nested hypotheses. The parameter estimates of the Quadrivariate Negative Binomial Hurdle model were also compared to the Univariate Negative Binomial Hurdle model applied to the pooled data to examine differences between two-part models assuming independence of the count dependent variable over time and two-part models treating the longitudinal nature of the data.

Results indicate that two-part models are preferred over one-part models and that an analysis by gender is preferable. However, when the dependent variable is correlated over time, two-part models using pooled data may lead to inconclusive results as suggested by different parameter estimates between the Univariate Negative Binomial Hurdle and Quadrivariate Negative Binomial Hurdle models. The Quadrivariate Negative Binomial Hurdle model offers a good fit of the data by correctly predicting the mean counts and the zeros for each year of the analysis which is a great advantage versus the standard Quadrivariate Negative Binomial model.

Parameter Estimates The results of the Quadrivariate Negative Binomial Hurdle (QNBH) model estimation are presented in Column 2 of Table 13. For the whole sample, the log-likelihood functions increased significantly from –28,923.019 (Quadrivariate Negative Binomial estimation) to -28,032.271 (QNBH) since twice the difference in the log-likelihood function is 90.75 which is greater than 21.03, the 95% critical value for the χ^2_{12}. The excess of zeros is significant in our longitudinal count data set and not addressing this feature of the data will result in inconsistent estimates since the mean function would not be correctly specified. This result supports previous findings suggesting that a different view of health care should be considered. The demand for doctor visits follows a two-step process in which the decision to consult a physician is taken and once this Hurdle is passed, the frequency of visits can be modelled.

Private Insurance, Income and Health Variables According to Grossman's theory, the results of the Quadrivariate Negative Binomial Hurdle model indicated that private insurance and income are significant in determining the demand for medical care. Having private insurance and a high income have a negative impact on the contact decision (1st stage) and on the frequency of GP visits (2nd stage). Health and chronic conditions are also significant in both stages with the expected signs.

The results of the impact of other variables are not fully supported by Grossman's theory. For example, education and being in the work force have a significant impact in the first stage only but not on the frequency of visits. Having children and being married are not significant variables in the decision to contact a GP but are significant

in the frequency of visits.

Quadrivariate Negative Binomial Hurdle Model versus Univariate Negative Binomial Hurdle Model Column 3 of Table 13 presents for comparison the results of the Univariate Negative Binomial Hurdle (UNBH) model applied to the pooled data. According to the Univariate Negative Binomial Hurdle model, the excess of zeros is also significant. If we compare the two models, the longitudinal Hurdle specification and the univariate Hurdle specification have similar interpretations in terms of significance of the explanatory variables. The only difference between the two models is again the role of having children and being married. Having children does not significantly impact the decision to consult a physician in both models but it has a positive impact on the frequency of consultations in the Quadrivariate Negative Binomial Hurdle model and a negative non significant impact in the Univariate Negative Binomial Hurdle model. Finally, comparing the estimates of the parameter indicates that their values and their standard errors are generally smaller using a longitudinal approach rather than pooling various cross-sections.

Gender In both models, gender is significant only in the decision to contact a physician. Being male has a significant negative impact on the probability of contacting a GP. Once the Hurdle is crossed, gender is not significant in explaining the frequency of the visits. This result was also found by Pohlmeier and Ulrich (1995) in their analysis of the number of physician visits generated by a population composed of employees using the 1984 wave of the German Socio-Economic Panel data set.

Table 13: Maximum Likelihood Parameter Etimates: QNBH and UNBH Models.

Maximum Likelihood Estimation		QNBH 84-87		UNBH Pooled data	
GPs visits		Total Sample		Total Sample	
Number of Observations (N*T):		4,342*4		4,342*4	
Log of Likelihood:		-28,032.271		-25,739.404	
Variable	Parameter	1st step	2nd step	1st step	2nd step
Constant	β_0	-0.723*	0.418*	2.637*	1.665*
		(0.242)	(0.143)	(0.859)	(0.172)
Age*10^{-1}	β_1	0.018	0.164*	0.190*	0.178*
		(0.042)	(0.018)	(0.039)	(0.022)
Sex	β_3	-0.189*	0.016	-0.269*	-0.059
		(0.056)	(0.033)	(0.064)	(0.040)
Private Insurance	β_4	-0.262*	-0.144*	-0.321*	-0.329*
		(0.104)	(0.051)	(0.087)	(0.076)
Education	β_5	-0.052*	0.007	-0.095*	-0.010
		(0.013)	(0.008)	(0.017)	(0.010)
Working	β_6	0.168*	0.007	0.234*	0.050
		(0.064)	(0.025)	(0.067)	(0.045)
Children	β_7	-0.039	0.101*	-0.005	0.058
		(0.060)	(0.030)	(0.054)	(0.044)
Income*10^{-4}	β_8	-0.378*	-0.125*	-0.370*	-0.640*
		(0.180)	(0.061)	(0.132)	(0.125)
Married	β_9	0.067	-0.103*	0.123	0.010
		(0.072)	(0.038)	(0.067)	(0.053)
Health	β_{10}	-0.111*	-0.116*	-0.217*	-0.192*
		(0.011)	(0.004)	(0.040)	(0.008)
Chronic conditions	β_{11}	0.403*	0.258*	0.802*	0.463*
		(0.061)	(0.020)	(0.159)	(0.042)
Non West German	β_{12}	-0.085	0.192*	0.012	0.260*
		(0.061)	(0.037)	(0.056)	(0.044)
Overdispersion	α	1.851*	1.275*	0.642*	0.594*
		(0.048)	(0.036)	(0.097)	(0.025)
* indicates significant at the 5 % level.					

However, including a variable gender does not allow us to test if other differences exist by gender. In fact, Likelihood Ratio Tests indicated that the sample should be separated according to gender since twice the difference in the log-likelihood is 118.12 which is greater than 33.89 the χ^2_{26} value for the Quadrivariate Negative Binomial Hurdle model. As reported previously in section 3.3 for one-part models, splitting the model by gender indicates major differences with respect to an aggregated and a longitudinal approach.

Gender Analysis The results of the estimation of the Quadrivariate Negative Binomial Hurdle model by gender are presented in Table 14. One of the main differences when the sample is split by gender is that income is not longer significant at any stage for both genders. This result is not in accordance with Grossman's theory nor with Pohlmeier and Ulrich (1995) for which income was significant in explaining the contact decision. The impact of private insurance is different across the two populations according to each of the two stages of the Hurdle. For women, private insurance is not significant in the decision to consult a physician but private insurance has a significant negative effect on the frequency of visits. For men it is the contrary; having private insurance significantly decreases the probability of contacting a GP but this effect is not significant for the frequency of visits. Differences between genders also appear in the impact of age, being married and having children.

A non-linear relation between age and the number of consultations was found for the male population in modelling the probability decision to contact a general practitioner. This result was not supported in one-part models (Univariate Nega-

tive Binomial, Quadrivariate Negative Binomial and Quadrivariate Negative Binomial Random Effects models) nor for the female population in the Quadrivariate Negative Binomial Hurdle model . Interestingly, for women, age was not found to be significant in explaining the contact decision. For both genders, being married does not play a significant role in explaining the decision to consult a GP. However, being married is significantly associated with a smaller number of GP visits generated by men once the decision has been taken. Having children does not affect the probability of contact but significantly increases the frequency of the visits for both genders. The same pattern is observed for nationality: it does not matter when it is time to first contact a GP, but significantly impacts the number of visits once the first contact has been made. Finally, men and women behave similarly in terms of education. Higher education significantly reduces the probability of contacting a GP for both men and women. Education does not affect the frequency of visits for both genders once the contact has been made.

Gender Analysis: Quadrivariate Negative Binomial Hurdle Model versus Univariate Negative Binomial Hurdle Model In order to compare in the presence of correlation a two-part model for longitudinal count data with the standard univariate Hurdle model, Table 16 presents the estimation by gender of the Univariate Negative Binomial Hurdle model applied to the pooled data. The results are presented by gender since this was justified according to a Likelihood Ratio Test value of 50.540 which is greater than 33.885 the χ^2_{26} value. The results are contrasted

Table 14: Maximum Likelihood Parameter Estimates: QNBH Model. Gender Analysis.

Maximum Likelihood Estimation		QNBH 84-87		QNBH 84-87	
GPs visits		Female		Male	
Number of Observations:		2183*4		2159*4	
Log of Likelihood:		-14501.889		-13471.3204	
Variable	Parameter	1st step	2nd step	1st step	2nd step
Constant	β_0	-0.627	0.190	0.693	0.718*
		(0.356)	(0.206)	(0.513)	(0.215)
Age*10^{-1}	β_1	0.047	0.168*	-0.810*	0.172*
		(0.042)	(0.026)	(0.165)	(0.028)
Age^2*10^{-3}	β_2	-	-	1.005*	-
		-	-	(0.201)	-
Private Insurance	β_4	-0.138	-0.160*	-0.389*	-0.117
		(0.145)	(0.072)	(0.152)	(0.073)
Education	β_5	-0.064*	0.020	-0.044*	-0.006
		(0.019)	(0.012)	(0.019)	(0.013)
Working	β_6	0.154*	0.068*	0.183	-0.068
		(0.074)	(0.033)	(0.157)	(0.041)
Children	β_7	0.011	0.089*	-0.049	0.146*
		(0.083)	(0.042)	(0.093)	(0.046)
Income*10^{-4}	β_8	-0.194	-0.140	-0.403	-0.120
		(0.227)	(0.077)	(0.265)	(0.102)
Married	β_9	0.042	-0.046	0.078	-0.194*
		(0.096)	(0.049)	(0.119)	(0.062)
Health	β_{10}	-0.115*	-0.121*	-0.117*	-0.112*
		(0.015)	(0.005)	(0.017)	(0.006)
Chronic conditions	β_{11}	0.301*	0.140*	0.561*	0.398*
		(0.083)	(0.005)	(0.093)	(0.031)
Non West German	β_{12}	-0.164	0.207*	-0.033	0.174*
		(0.087)	(0.052)	(0.091)	(0.053)
Overdispersion	α	-1.770*	1.341*	1.843*	1.170*
		(0.063)	(0.049)	(0.071)	(0.053)
* indicates significant at the 5 % level.					

against the Quadrivariate Negative Binomial Hurdle model. Important differences in the impact of private insurance, income and employment status appear between the Univariate Negative Binomial Hurdle and Quadrivariate Negative Binomial Hurdle models when the analysis is conducted by gender.

In the Univariate Negative Binomial Hurdle model, private insurance is significant for women in reducing the decision to contact a physician while this effect was insignificant in the Quadrivariate Negative Binomial Hurdle model. For men, private insurance is not longer significant in the contact decision while this variable was significant in the Quadrivariate Negative Binomial Hurdle model. Income and employment status are significant in explaining the frequency of GP visits according to the Univariate Negative Binomial Hurdle model while these variables did not play a significant role in the Quadrivariate Negative Binomial Hurdle model. Income significantly decreases the frequency of visits according to the Univariate Negative Binomial Hurdle but not in the Quadrivariate Negative Binomial Hurdle estimations. In light of these differences, a lot of caution should be taken when interpreting results from univariate Hurdle models pooling count data.

Predictions Once the parameters of the Quadrivariate Negative Binomial Hurdle model were estimated, the predicted mean counts at each time period t, were derived from equation (108). As can be seen in Figure 2, which presents the predicted means for the two populations, the Quadrivariate Negative Binomial Hurdle model gives overall a good fit of the data. The predicted number of zeros at each time period was

Table 15: Maximum Likelihood Parameter Estimates: UNBH model. Gender Analysis.

Maximum Likelihood Estimation		UNBH 84-87		UNBH Pooled data	
GPs visits		Female		Male	
Number of Oservations (N*T):		2,183*4		2,159*4	
Log of Likelihood:		-13,543.7447		-12,170.39	
Variable	Parameter	1st step	2nd step	1st step	2nd step
Constant	β_0	-0.119	1.418*	11.848*	1.973*
		(0.875)	(0.259)	(5.183)	(0.269)
Age*10^{-1}	β_1	0.127*	0.167*	-1.39	0.197*
		(0.036)	(0.031)	(0.785)	(0.033)
Age*10^{-2}	β_2	-	-	2.154*	-
		-	-	(1.074)	-
Private Insurance	β_4	-0.228*	-0.322*	-0.549	-0.309*
		(0.087)	(0.108)	(0.345)	(0.110)
Education	β_5	-0.066*	0.007	-0.221*	-0.034*
		(0.008)	(0.015)	(0.096)	(0.016)
Working	β_6	0.119*	0.064	0.865	-0.012
		(0.049)	(0.053)	(0.573)	(0.090)
Children	β_7	-0.002	0.045	-0.009	0.080
		(0.053)	(0.061)	(0.196)	(0.067)
Income*10^{-4}	β_8	-0.106	-0.514*	-1.364	-0.854*
		(0.109)	(0.160)	(0.750)	(0.120)
Married	β_9	-0.014	0.022	0.634	-0.007
		(0.058)	(0.067)	(0.357)	(0.009)
Health	β_{10}	-0.128*	-0.188*	-0.575*	-0.197*
		(0.019)	(0.011)	(0.280)	(0.013)
Chronic conditions	β_{11}	0.419*	0.375*	2.424	0.558*
		(0.077)	(0.055)	(1.349)	(0.064)
Non West German	β_{12}	-0.005	0.245*	0.158	0.265*
		(0.053)	(0.065)	(0.211)	(0.064)
Overdispersion	α	1.404*	0.656*	0.332*	0.527*
		(0.536)	(0.032)	(0.090)	(0.038)
* indicates significant at the 5 % level.					

calculated conditional on the previous number of physician visits. Due to its Hurdle structure, the Quadrivariate Negative Binomial Hurdle model is also able to predict the number of zero visits at each time period. As presented in Figure 3, this feature of the Quadrivariate Negative Binomial Hurdle model represents a significant improvement over Quadrivariate Negative Binomial models which are unable to predict any zero at all for a particular year.

Figure 2: Observed versus Predicted Mean Number of GPsVisits per Year and Gender. QNBH Model.

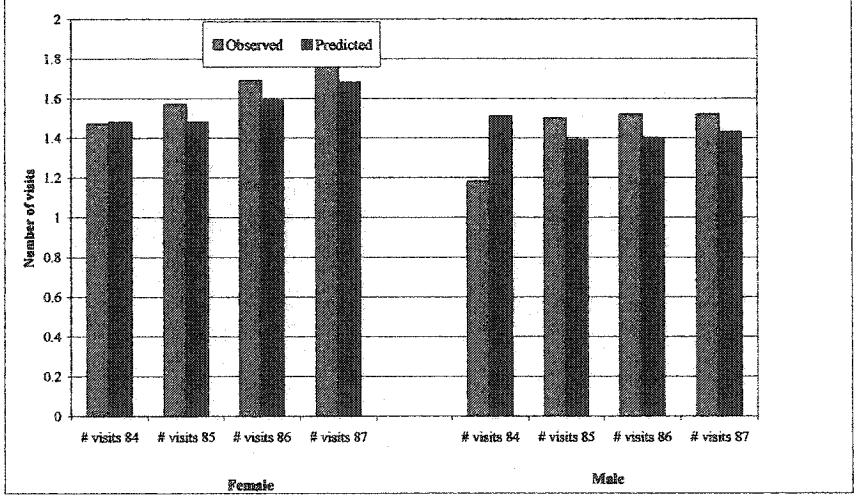
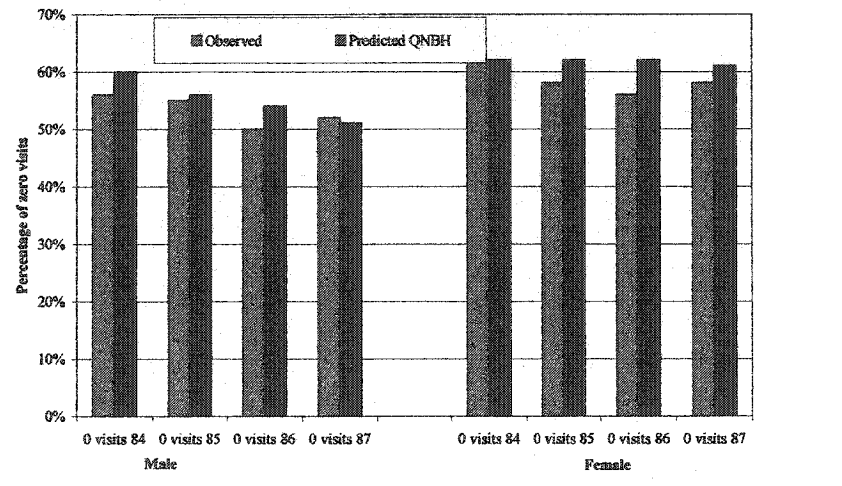


Figure 3: Observed versus Predicted Percentage of Zero GP Visits per Year and per Gender - QNBH Model



3.5 The Analysis of the Number of Specialist Visits

The same methodology was applied for analyzing the of the number of specialist visits to examine if differences exist in the demand for specialists and GPs as found in Pohlmeier and Ulrich (1995). Results indicate that the analysis of GPs and specialists should be done separate as several variables have different impacts in each analysis. It is also shown that an analysis by gender is preferable and that the Quadrivariate Negative Binomial Hurdle model is the preferred model since the excess of zeros is significant. The results from univariate estimations are not presented in this section which concentrates on highlighting the differences between the demand for general practitioners and the demand for specialists when the analysis is conducted by gender using count data models for longitudinal count data.

Likelihood ratio tests supported to conduct a separate analysis for the female and male population for the Quadrivariate Negative Binomial model but also for the Quadrivariate Negative Binomial Random Effects model. This is the first difference between the analysis of the demand for specialists and the demand for GPs. Splitting the model by gender was not justified when a Negative Binomial Random Effects model was used to explain the number of GP visits (Table 10).

3.5.1 Quadrivariate Negative Binomial and Quadrivariate Negative Binomial Random Effects Estimations

Table 16 reports the Quadrivariate Negative Binomial (QNB) and Quadrivariate Negative Binomial Random Effects (QNBRE) estimation results for the specialist equa-

tion after splitting the sample by gender. Results indicate that similar to the analysis of GPs, there is a significant increase in the log-likelihood function when individual-specific random effects are added to the Quadrivariate Negative Binomial Random Effects modelling the demand for specialists. However, contrary to the findings of the GP analysis, the results of the Quadrivariate Negative Binomial and the Quadrivariate Negative Binomial Random Effects models are quite different in terms of significance of the variables and magnitude of the parameter estimates. Other differences appear in the impact of private insurance, education, working status and having children in explaining the demand for specialist and GP. In the GP analysis, private insurance, income, education and the health variables were all significant according to the Quadrivariate Negative Binomial and Quadrivariate Negative Binomial Random Effects models. This is not anymore the case in the longitudinal analysis of the number of specialist visits.

Private insurance is not significant in explaining the number of specialist visits done by women in both models (Quadrivariate Negative Binomial and Quadrivariate Negative Binomial Random Effects models) and for men in the Quadrivariate Negative Binomial Random Effects model. This is different from the GP analysis for which having a private insurance was associated with a significant reduction in the number of GP visits.

Income, which was associated with a significant reduction for men in GP visits, now has a positive and significant impact on the demand of women for specialist visits in the Quadrivariate Negative Binomial model. However, this is not true according to

Table 16: Maximum Likelihood Parameter Estimates: QNB Models. Specialists. Gender Analysis.

Maximum Likelihood Estimation		QNB 84-87	QNB 84-87	QNBRE 84-87	QNBRE 84-87
Specialists visits		Female	Male	Female	Male
Number of Observations (N*T):		2,183*4	2,159*4	2,183*4	2,159*4
Log of Likelihood:		-16,665.479	-11,771.207	-13,858.188	-8,583.099
Variable	Parameter				
Constant	β_0	1.989*	0.751*	1.662*	2.006*
		(0.442)	(0.274)	(0.400)	(0.596)
Age*10^{-1}	β_1	-0.565*	0.198*	-0.815*	-1.669*
		(0.193)	(0.037)	(0.181)	(0.276)
Age^2*10^{-3}	β_2	0.580*	-	0.714*	2.338*
		(0.223)	-	(0.212)	(0.319)
Private Insurance	β_4	-0.022	0.187*	0.060	0.028
		(0.058)	(0.076)	(0.067)	(0.088)
Education	β_5	0.045*	0.027	0.044*	0.167*
		(0.013)	(0.016)	(0.010)	(0.013)
Working	β_6	-0.011	-0.101	0.039	-0.443*
		(0.031)	(0.045)	(0.040)	(0.068)
Children	β_7	0.053	0.024	-0.100*	0.150*
		(0.040)	(0.054)	(0.047)	(0.065)
Income	β_8	-0.309*	0.229*	-0.037	0.156
		(0.082)	(0.052)	(0.071)	(0.096)
Married	β_9	0.092	-0.192*	0.165*	-0.864*
		(0.050)	(0.080)	(0.053)	(0.071)
Health	β_{10}	-0.105*	-0.192*	-0.082*	-0.244*
		(0.005)	(0.006)	(0.008)	(0.010)
Chronic conditions	β_{11}	0.557*	0.579*	0.563*	0.704*
		(0.026)	(0.033)	(0.038)	(0.053)
Non West German	β_{12}	-0.010	0.249*	-0.253*	0.715*
		(0.065)	(0.087)	(0.053)	(0.063)
Overdispersion	α	0.889*	0.617	-	-
		(0.016)	(0.013)	-	-
	a	-	-	1.520*	-1.098*
		-	-	(0.038)	(0.026)
	b	-	-	1.724*	-1.063*
		-	-	(0.062)	(0.038)
* indicates significant at the 5 % level.					

the Quadrivariate Negative Binomial Random Effects model for which income does not play a significant role in explaining the demand for women.

Education has a positive impact on the number of specialist visits generated by women according to the two models and in the Quadrivariate Negative Binomial Random Effects model for men. In comparison, education was associated with a decrease in GP visits.

Working only has a significant negative impact in explaining the demand for specialists in the Quadrivariate Negative Binomial Random Effects model and only for men. According to the Quadrivariate Negative Binomial and Quadrivariate Negative Binomial Random Effects models, working status does not play a significant role in the decision to consult a specialist for women. This is also different from the GP analysis in which the two models found that working was associated with an increase in the number of GP visits for women. Other differences between gender and/or models appear with respect to the impact of having children, being married or not being German.

While these results indicate that the analysis should be done by gender as in Geil et al. (1997) and that the demand for GPs and specialists is different as in Pohlmeier and Ulrich (1995), the fact that almost 60% of the population did not consult a specialist may suggest that the data should be tested for an extra zeros.

3.6 Quadrivariate Negative Binomial Hurdle (QNBH) Estimation

The Quadrivariate Negative Binomial Hurdle (QNBH) model given in equation (103) was used to analyze the demand for specialists. Again the excess of zeros is significant in explaining the number of visits to the specialists. With respect to the Quadrivariate Negative Binomial model estimation, the log-likelihood function increased from −16,665.479 to -15,710 for women and from -11,771.207 to -10,812 for men. Not taking into account the zeros would lead to inconsistent estimates. In addition, specification tests supported again that the determinants of the demand for specialist are different for men and women.

While health and chronic conditions are significant with the expected signs in explaining the demand for specialists and GPs, the results support the assumption that the determinants for the demand of generalists and the demand of specialists are different when two-part models are used to treat for an excess of zeros.

Private insurance is not anymore significant in explaining the frequency of visits to specialists by women once the Hurdle is passed while this variable was reducing significantly the demand for generalists. For men, private insurance has a positive significant impact in the frequency of the visits to specialists while it was not significant for the demand for GPs. In fact, the probability decision to contact a GP was reduced by having a private insurance.

Income was not significant in explaining the demand for GPs in both stages (i.e., contact and frequency) but it is significant in explaining the frequency of visits to

Table 17: Maximum Likelihood Parameter Estimates: QNBH Models. Specialists. Gender Analysis.

Maximum Likelihood Estimation		QNBH 84-87		QNBH 84-87	
Specialist visits		Female		Male	
Number of observations:		2,183		2,159	
Log of Likelihood:		-15,710.077		-10,811.654	
Variable	Parameter	1st step	2nd step	1st step	2nd step
Constant	β_0	0.668	0.514*	-0.118	1.379*
		(0.444)	(0.201)	(0.566)	(0.251)
Age*10^{-1}	β_1	-1.296*	0.092*	-1.357*	0.094*
		(0.150)	(0.026)	(0.203)	(0.034)
Age^2*10^{-3}	β_2	1.309*	-	1.711*	-
		(0.179)	-	(0.244)	-
Private Insurance	β_4	0.051	0.021	0.148	0.163*
		(0.128)	(0.058)	(0.146)	(0.079)
Education	β_5	0.032	0.024*	0.068*	-0.012
		(0.017)	(0.011)	(0.019)	(0.013)
Working	β_6	0.152*	-0.037	0.221	-0.198*
		(0.073)	(0.031)	(0.162)	(0.445)
Children	β_7	0.001	0.073	0.057	-0.011
		(0.081)	(0.041)	(0.105)	(0.056)
Income*10^{-4}	β_8	0.022	-0.194*	0.178	0.163*
		(0.168)	(0.070)	(0.209)	(0.054)
Married	β_9	0.187**	-0.032	0.115	-0.071
		(0.095)	(0.047)	(0.137)	(0.075)
Health	β_{10}	-0.069*	-0.090*	-0.149*	-0.151*
		(0.015)	(0.005)	(0.019)	(0.007)
Chronic conditions	β_{11}	0.567*	0.405*	0.673*	0.377*
		(0.081)	(0.027)	(0.106)	(0.035)
Non West German	β_{12}	-0.327*	0.241*	0.079	0.218*
		(0.089)	(0.057)	(0.106)	(0.070)
Overdispersion	α	2.292*	1.257*	1.598*	1.308*
		(0.089)	(0.046)	(0.063)	(0.061)
* indicates significant at the 5 % level.					

specialists. Higher income reduces significantly the demand for specialist for women which can represent opportunity costs. It is the contrary for men for whom a high income is significantly associated with a higher number of visits to specialists once the contact decision has been made.

Education which was decreasing the contact decision for generalists for both sex, increases the probability of contact with specialist for men. For female, education is only significant in explaining the frequency of visits but not the contact decision.

Working which was increasing significantly the frequency of visits to GPs for women is insignificant in explaining the frequency of specialists. Still working explains the contact decision of specialists for women. Working which did not play a role for men in explaining the demand for GPs, decreases significantly the frequency of specialist visits for men.

Having children significantly increased the frequency of GP visits but this effect was not significant in explaining the demand for specialists. In both cases this variable is not significant for the probability discussion. Being married was insignificant in both stages in the GP analysis but for women, being married increases significantly the probability of contacting a specialists. Finally and not least, while a non-linear form of age was only supported for men in the demand for GPs, non linearities were found for both genders for specialists.

3.7 Conclusions

In conclusion, the analysis of longitudinal count data characterized by an excess of zeros and correlation over time should be conducted carefully as illustrated through the longitudinal analysis of the number of doctor visits in Germany. The contributions of this chapter are threefold:

1. To have identified an important weakness in the current models used to analyze the use of health services such as in Winkelman (2001) and Geil et al (1997). It was demonstrated that in the presence of correlation, econometric models for longitudinal count data should be used instead of univariate models applied to the pooled data.

2. To have proposed a new alternative to deal with an excess of zeros in the longitudinal context while taking into account the other characteristics associated with panel count data such as the presence of correlation due to the repeated nature of the data.

3. To have provided evidence using a longitudinal subset of the German Socio-Economic Panel that a) pooling data may result in inconsistent estimates if the dependent count variable is correlated over time, b) the presence of excess zeros in the longitudinal context can be tested and accounted for by Multivariate Negative Binomial Hurdle models, c) two-part models do not fully support the Grossman's theory and d), analyses by gender or by speciality status should be conducted when necessary and aggregating the data may result in a sub-optimal utilization of the data. This is extremely important in setting health care policies since men and women may react differently to economic or other incentives and the determinants to explain the

demand for generalists and specialists are different.

More specifically, a multivariate framework to analyze longitudinal count data in the presence of covariates, overdispersion, correlated counts and excess of zeros was presented in this chapter. Suspicion of an excess of zeros was tested using Likelihood Ratio Tests because the Quadrivariate Negative Binomial Hurdle model and the Quadrivariate Negative Binomial model are nested. Due to the conditional specification of the Quadrivariate Negative Binomial Hurdle model, correlation is introduced in both stages (i.e., contact and frequency) by conditioning over the previous counts.

In addition to consider a two-part process generating the data, Multivariate Negative Binomial Hurdle models allow us to follow individuals over time and therefore take into account any changes in the characteristics of these individuals. The Multivariate Negative Binomial conditional Hurdle model complements standard multivariate count models which have ignored the problem with zeros in panel count data. This methodology is also preferable to univariate methods for analyzing longitudinal count data because it accounts for the correlation arising from the longitudinal aspect of the data while taking into account the excess of zeros.

It was first shown that the analysis of the number of doctor visits should be done by splitting the sample by gender as found in Geil et al. (1997) in their analysis of hospital visits. This result confirms Winkelman's findings in labor economics (1994) and supports the assumption that men and women react differently to economic incentives. Our results indicated that men and women differ considerably and not taking this characteristic into consideration in assessing health care policies may give an

incomplete interpretation of the data. In our gender analyses, important differences in terms of parameter values and significance of the variables were observed between longitudinal count data models and univariate negative models pooling cross-section data.

This finding differs from Geil et al. (1997) who observed similar patterns in the estimations of longitudinal and univariate negative binomial models in their analysis of hospital visits. The authors noted that "there is greater unobserved firm-specific heterogeneity in the patents case than individual-specific heterogeneity in the hospital visits case". This is certainly true since our specification includes more information in explaining the number of doctor visits than a general R&D variable in the patent case. It is also important to recall that the correlation of the annual number of patents over time is 0.9 versus a correlation of less than 0.1 in the case of the number of hospital visits per year. Therefore, using a longitudinal approach or a pooled approach for analyzing the number of hospital trips is expected to give similar results due to the lack of correlation of the dependent variable over time. On the other hand, the correlation has to be taken into account in the analysis of the number of patents or the number of physician visits.

Our results differ due to the presence of correlation in the number of doctor visits while the number of hospital visits at any given year is not correlated to previous visits. When the analysis was conducted by gender, the analysis of the number of visits to GPs confirmed that univariate and multivariate count models' estimations are different in many regards. This is an important result because most of the literature

in health economics has concentrated on cross section analyses or analyses applying a Univariate Negative Binomial distribution to the pooled data (Winkelman, 2001 and Geil et al., 1997). This important feature of the data would have been missed if splitting the sample by gender had not been investigated. When a dummy variable was used for gender, the interpretation of the Univariate Negative Binomial, Quadrivariate Negative Binomial and Quadrivariate Negative Binomial Random Effects models were almost similar!

In light of these results, it is extremely important to be prudent in analyzing panel or household longitudinal count data. Pooling various cross sections should be used only if the dependent variable is not correlated. In addition, splitting the sample by gender should be tested because men and women may differ in various regards. Based on our results, it is not clear if the analysis of the impact of the 1997 German reform on the number of GP visits (Winkelman, 2001) would have led to different results if the authors had used in their preliminary work longitudinal count data models to explain the number of visits by gender instead of pooling cross-sections for the whole sample.

The most important contribution of this chapter is to have developed a new methodology to deal with longitudinal count data characterized by an excess of zeros, thereby filling an important gap in the economic literature. When the data is correlated and the dependent variable is characterized by a high proportion of zeros, it is not reasonable to use two-part univariate models (Univariate Negative Binomial Hurdle) on pooled data as in Geil et al. (1997) and Winkelman (2001). Instead, the

methodology should simultaneously account for the longitudinal aspect of the data and the presence of zeros. To answer this dilemma, a new model was proposed based on an extension of the Univariate Negative Binomial Hurdle model to the longitudinal case. This model presents several advantages over existing models. Firstly, it is nested to the Multivariate Negative Binomial model and due to its simple form, it can be easily used to test for the presence of extra zeros in longitudinal count data. If the presence of zeros is significant, non-parametric models for longitudinal count data based on Pseudo Maximum Likelihood and General Linear Model theory are not applicable since the mean function is not correctly specified due to this excess of zeros. Secondly, the Quadrivariate Negative Binomial Hurdle model takes into account the correlation because at each time period the decision to contact a physician and the frequency of visits once the contact is made, are determined conditionally on the previous counts. This represents a significant improvement over univariate Hurdle models applied to pooled data assuming independence of the data or over the Panel Probit model used by Chiappori et al. (1998). This chapter demonstrated how important it is to have instruments which allow for the detection and treatment of an excess of zeros in longitudinal count data.

In the analysis of the number of doctor visits generated by a panel of German households followed over 4 years, the excess of zeros was significant, which support the idea that a different view of health care utilization should be considered when analyzing longitudinal count data on utilization of health services. The Quadrivariate Negative Binomial Hurdle model correctly predicted the mean counts but also the

number of zeros for each time period. The findings and implications of this research support the development of panel household data sets in Canada and other countries to fully understand the dynamics of health care services utilization. This methodology is especially appropriate for testing the presence of moral hazard in the demand for medical services when using natural experiments as Chiappori et al. (1998) in France.

The Multivariate Negative Binomial Hurdle model for longitudinal count data with an excess of zeros is generalizable to other fields of economic research such as the analysis of the number of patents generated by small or medium firms or the number of loans generated by a panel of students over a certain period of time.

4 MULTIVARIATE NEGATIVE BINOMIAL ZERO - INFLATED MODEL: AN APPLICATION TO THE LONGITUDINAL ANALYSIS OF CLINICAL TRIAL COUNT DATA

Clinical trials are designed to determine whether treatments are safe and effective. The analysis of clinical trials provides estimates of drug efficacy that are submitted to health regulators by pharmaceutical companies to market a new drug. Economic evaluations of drug treatments are developed based on the efficacy reported from the clinical trial, to evaluate the costs and the benefits of the treatment of interest over a certain period of time. Several validated techniques such as cost effectiveness analyses are used and recommended for this purpose. Economic evaluations are critical to obtain public reimbursement of a new drug treatment. If the new treatment is not cost-effective versus standard treatment, the payer (e.g., government, private insurers) has no economic incentive to reimburse this treatment. Pharmacoeconomics or economic evaluations of new pharmaceutical products are mandatory in several countries including Canada, Australia, the United Kingdom and the United States.

However, it may happen that the efficacy of a drug treatment is not correctly assessed due to improper choice of statistical methods. This was recently illustrated by McIntosh (2001) in a re-analysis of a clinical trial on bladder cancer in which the primary endpoint was the number of tumor recurrences. Previous research (Lawless,

1987; Kanifiard and Gallo, 1995 and Dean and Balshaw, 1997) reported that only one treatment was effective. However the methods yielding these results were not optimal as they did not fully consider the longitudinal aspect of the data and/or the intervals between recurrences or the number of tumors per interval. McIntosh (2001) analyzed the data using a bivariate Negative Binomial model, an auto-regressive Weibull model and an auto-regressive truncated Negative Binomial model. The estimations indicated that both treatments were in fact effective. It is obvious that the results of any pharmacoeconomic study of this clinical trial will be different if one or two drug treatments are effective.

Therefore, it is crucial for the pharmaceutical industry and all the players to have the most efficient and reliable estimates of drug efficacy. Traditionally, this area of research is led by biostatisticians. As economic evaluations of drugs are required for any new drug submission in Canada and several countries, health economists should be concerned about the reliability of the estimates of drug treatment efficacy.

The objective of this chapter is three-fold. Firstly, to review important statistical issues associated with the analysis of longitudinal clinical trials in which the efficacy variable is a count and to show why traditional approaches used in biostatistics may be unsatisfactory. Secondly, a methodology that improves on the shortcomings of these models is proposed and applied to an unpublished longitudinal clinical trial count data set. Results are compared with standard analyses. Finally, the implications of these findings to pharmacoeconomics are illustrated through a hypothetical decision tree cost-effectiveness example.

It is often the case in medicine that the health outcome used as the efficacy variable is a discrete variable or a count. Examples include the number of epilepsy seizures or asthma episodes per time period, carcinoma tumors or the incidence of polio episodes for any given years. In many clinical trials, or other experimental studies, repeated observations of one or more efficacy variables are measured and collected on every participant at several time intervals. It is the change in these variables which measures the efficacy of treatments. However, dealing with their initial value may be problematic when the outcome is a count. Socio-demographic characteristics such as sex, age, weight and height as well as pre-determined variables related to the medical history of each subject are typically recorded in clinical studies.

In order to obtain the most efficient and reliable estimators of the treatment, all the available information should be included in the analysis of longitudinal clinical trial count data. This implies that the discrete and repeated nature of the clinical trial is recognized in the analysis, especially when the efficacy variables are correlated over time periods. Consistency requires that the initial efficacy variable be treated as a random variable in longitudinal analysis rather than an explanatory variable. In addition, the analysis should allow for the presence of random effects or unobserved subject heterogeneity, which when ignored, can distort the assessment of treatment effects. For example, small treatment differences may hide differential effects due to heterogeneity (Lindsey, 1998). It is also very important to differentiate between treatment effect and trend effect. This is especially important if at the end of the clinical trial, the placebo group is getting better. Nonetheless, it is common to find

analyses of longitudinal clinical trials focussing on the endpoint of the trial.

Another important feature of longitudinal clinical trial count data is the possible over-representation of zeros in the data due to the effectiveness of the treatment. If the treatments are very effective in preventing the event of interest, a high number of zeros (i.e. 0 episodes or events) will be present at the end of the trial. As a consequence, the functional form of the distribution may be distorted due to this excess of zeros. Not taking this into account could lead to serious inference problems. As Lindsey (1999) observed, "restricting comparisons of responses under different treatment to differences in means can be extremely misleading if the shape of the distribution is changing. This may involve changes in dispersion, or in other shape parameters such as skewness in a stable distribution, with the treatments or covariates". It is therefore very important to have a methodology that tests and treats for an excess of zeros in the longitudinal discrete case. This is particularly relevant because the standard multivariate models used to analyze longitudinal clinical trial count data are currently unable to deal with an excess of zeros. While models such as Hurdle and Zero-Inflated models are used to treat for excess of zeros in count data in the univariate case, these methods have not been adapted to longitudinal count data.

New methods for analyzing clinical trials are proposed to address six common characteristics of clinical trials: multiple observations on each subject, non-normal integer valued responses, the presence of covariates, the presence of unobservable random effects, correlated responses, and distorted distributions due to an excess of zeros. These innovative methods are based on maximum likelihood estimation of

Multivariate Zero-Inflated distributions for longitudinal count data. Likelihood ratio tests are used to test for the presence of zeros in the longitudinal setting.

The remainder of the chapter is organized as follows. Section 4.2 presents a quick overview of important statistical issues associated with the analysis of longitudinal clinical trial count data as well as the standard methods of analysis used in biostatistics. The importance of treatment efficacy derived from clinical trials in economic evaluations of medicines is outlined in a discussion. Section 4.3 presents the data of an unpublished subset of a longitudinal clinical trial. The data to be analyzed is characterized by four repeated observations on a count health outcome - corresponding to one of the secondary endpoints collected in this trial -, the presence of covariates, overdispersion, correlated counts, and a high proportion of zeros. A Quadrivariate Negative Binomial model and a Generalized Estimating Equations model are presented in Section 4.4 to analyze the data. The Quadrivariate Negative Binomial model accounts for the discrete and repeated nature of the data and considers the baseline efficacy variable as endogenous but fails to predict the number of cured patients (i.e.: 0 episodes) at each time period. This is also the case for the standard Generalized Estimating Equations model used by Diggle (1993) for longitudinal clinical trial count data for which not a single zero outcome was predicted. To take into account the presence of excessive zeros associated with very effective treatments, a Quadrivariate Negative Binomial Zero-Inflated model nested to the Quadrivariate Negative Binomial model is presented in Section 4.5. In this specification zeros are generated according to two regimes. Section 4.6 concludes by summarizing the find-

ings and their implications for economic evaluations of pharmaceutical products when the efficacy variable is a count. This chapter supports the need for the development of guidelines for the statistical analysis of clinical trial count data.

4.1 Clinical Trials and Pharmacoeconomics

The following presents a quick overview of important statistical issues associated with the analysis of longitudinal clinical trial count data characterized by an excess of zeros and shows why traditional methods fail to address these issues. The implications of these findings in economic evaluations of pharmacotherapies are discussed through a fictive cost-effectiveness example.

4.1.1 Statistical Issues

Longitudinal clinical trials provide very rich and detailed data sets. Baseline data on individual characteristics (e.g. sex, age, etc.) are collected as well as other predetermined variables related to the medical history of the subjects (e.g. blood pressure, severity of disease, prior medications, etc.). In addition, a common response in clinical trials involves the occurrence of events (e.g. the number of seizures). In this case, the efficacy variables are represented by counts. The main feature of longitudinal clinical trial count data is that the efficacy variable is measured at starting therapy and thereafter at several times during the administration of the treatments. In this sense, longitudinal clinical trial count data sets are rich in information to determine whether treatments are effective.

Since responses are available for each individual subject, longitudinal clinical tri-

als allow us to compare the treatment effects with respect to their trends over time, to determine the presence of heterogeneity among individuals (between-subject heterogeneity) and the presence of a contagion effect (within-subject time dependence) when the expected number of events at one time depends on the realized number of events at some previous time. For example, if the longitudinal data is aggregated in a single endpoint, contagion can not be detected. In order to obtain the best estimates of the treatments including the time shape of the responses, the complete multivariate sequence needs to be compared simultaneously across treatments, especially if the responses or efficacy variables, are correlated over time. One way of modelling longitudinal clinical trial count data is to assume that the repeated measurements have a multivariate discrete distribution (Lindsey, 1998). However, it is common to find univariate analysis of longitudinal clinical trials relying on mean changes or proportion differences from baseline to end of the study by applying classical theory. Other type of analysis concentrates in the endpoint efficacy variable and treats the baseline efficacy variable as a regressor along with other covariates. Such analyses do not recognize the discrete and/or repeated nature of the data, and as a consequence may yield inconsistent estimates.

The treatment of the baseline efficacy variable is an important methodological issue which has received few attention in the literature on the analysis of longitudinal clinical trials. This is important because many clinical trials are conducted over a short time period (e.g., 12 weeks) and the repeated measurements are done weekly or twice a month. In this case, the initial value may contain much of the information on

parameter values relevant to the analysis. In econometrics, it has been shown that dropping the initial value(s) in a time series context could be misleading when T is small as discussed in section 2.6. However, longitudinal analysis (Diggle, 1993; Albert, 1999) of short longitudinal clinical trial count data considering the baseline efficacy variable as a regressor is frequent. There are strong arguments showing that such a practice could be misleading. Intuitively, for subjects enrolled in the placebo group (i.e.: no treatment), it seems illogical to consider the repeated measurements on the efficacy variable as random variables and the baseline value (t = 0) as exogenous. Since these participants do not receive any treatment along the trial, the baseline value of the efficacy variable has to come from the same distribution as the repeated measures. Of course, the mean values of the counts for the placebo group may change over time due to a natural trend effect but the distribution of the baseline and the repeated observations of the efficacy variable should be the same for the placebo group. More formally, treating the initial value of the efficacy variable y_{i0} as exogenous is wrong because by definition any multivariate distribution is derived by conditioning on the previous observations on the efficacy variable, including the initial value as shown in equation (104).

There is also another practical argument for not using the baseline efficacy variable as a regressor. In the analysis of clinical trials, it is very important to distinguish between trend effect and treatment effect to not over-estimate the efficacy of the treatments. For example, if the placebo group is getting better by end of the trial, the trend effect and treatment effect need to be identified. One disadvantage of

univariate methods which concentrate on the analysis of the endpoint is that they are unable to identify the two effects. However, even in a longitudinal analysis the trend effect wouldn't be identifiable if the analysis does not include a covariate for the trend or when the baseline value is treated as exogenous. In particular, using a multivariate model to explain the number of counts at each time period by a constant, covariates, treatment dummies and a trend variable requires that the baseline value be treated as endogenous to identify the constant at time 0 (i.e. no treatment), and the trend in subsequent observations. If the baseline is used as a covariate, the constant for "no treatment" would not be identifiable at $t = 0$ and consequently the trend effect is not identifiable at $t = 1$ because it is mixed with the constant.

Another important and less well-known methodological issue arises when the treatments are very effective in preventing the occurrence of the event of interest. Treatments are designed to alleviate symptoms, which translate into a reduction in the efficacy variables or in the case of a complete cure, by 0 values of efficacy variables. Therefore, if the treatments are very effective in reducing the event of interest, a high percentage of 0 events would be reported at the end of the trial or slightly before. Cured patients have no episodes or zero counts.

This constitutes a major problem because standard distributions become less representative of the data-generating process. More zeros are observed than any parametric distribution used for the analysis of count data (Poisson or Negative Binomial distributions) would or could predict. For example, a Poisson model will underpredict the true frequency of zeros and large counts while overpredicting the frequency

of small counts. A Negative Binomial model tends to increase the frequency of zeros and high counts but does not yield a good fit of the data if the excess of zeros is unaccounted for (Gurmu, 1996). In economics, Hurdle and Zero-Inflated models have been developed when excessive zeros are present. Hurdle models which allow for a systematic difference in the statistical process governing individuals with zero counts and individuals with one or more counts have been studied in Chapter 3. Zero-Inflated models allow two regimes to generate the zeros (Lambert, 1992). As discussed in Chapter 3, it has been shown that if the excess of zeros is significant, estimating the mean function independently of the dispersion structure governing the Hurdle/Zero-Inflated model could lead to a loss of consistency and efficiency. If a significant excess of zeros is present in the data, the mean function is not correctly specified and models relying on Pseudo Maximum Likelihood theory or on Generalized Estimating Equations theory cannot be implemented.

It appears therefore essential to develop new models to take into account the problem of an excess of zeros. Other methodological problems arising in longitudinal clinical trials count data, but not treated in this chapter, are related to dose effect, missing data or modelling unbalanced repeated observations. For further related discussion on these topics, the reader is referred to Wijenska (1995), Albert (1999), Lambert (1996) and Cnaan (1997).

4.1.2 Standard Analyses

The standard models used to analyze longitudinal clinical trial count data can be classified as Univariate models or Multivariate models relying on Generalized Esti-

mating Equations models. Their main characteristics and drawbacks are summarized below. For further details, a comprehensive review of the literature of the standard practices in biostatistics can be found in Diggle (1993), Albert (1999) and Lindsey (1998, 1999, 2000).

Univariate Models An approach that is sometimes used in the analysis of longitudinal clinical trial count data is to specify a univariate distribution for the endpoint efficacy variable and to include the baseline value as a regressor. For example, in analyzing the total number of tumor recurrences over a 8-week time period in patients suffering from Stage I bladder carcinoma, Kanifiard and Gallo (1993) specified a Poisson distribution for the endpoint defined as the sum of the number of tumors observed over 8 weeks. The model includes the baseline number of tumors as a covariate along with two treatment dummies, the largest baseline tumor size and the length of the follow-up examinations. As stated in equation (102) in a longitudinal context, when the baseline value is one of the regressor, the distribution of the endpoint is not the conditional distribution of the endpoint variable given the baseline value. This is only the case when a joint distribution is specified and when the conditional distribution does not contain the baseline value as a regressor. In addition, univariate models concentrate on the endpoint of the clinical trial, thus discarding valuable information on the treatment effects. Consequently, the set of parameters that can be estimated in the univariate case is smaller than in the longitudinal case. Trend effect and treatment effect over time cannot be identified by any univariate analysis nor the time shape of the parameters. This is inconvenient since knowing when a treatment

reaches its maximum efficacy is very important from a medical point of view.

It should be noted that even when the outcome is not normally distributed it is frequent to find analysis of treatment effect based on means changes or proportion differences. Two-way analysis of variance (ANOVA) and an analysis of covariance (ANCOVA) with the baseline response as covariate are sometimes considered to analyze longitudinal count data. As outlined by Lindsey (1999), conclusions based on such models could be misleading especially if the shape of the distribution of responses changes over time due to the high efficacy of the treatments. In addition, this type of analysis does not recognize the discrete nature of the data and it assumes that the treatment effect is constant among treatment groups, thus ignoring heterogeneity within subjects. Furthermore, the covariates are generally ignored, which may lead to inconsistent coefficient estimates.

Generalized Estimating Equations Models

Marginal Models Marginal models or population-averaged models are perhaps the most widely used models in the analysis of longitudinal discrete data from clinical trials. Marginal models estimate the effect of a set of covariates on the marginal expectation of the response. The dependence/correlation structures among the responses are assumed to be that of a nuisance parameter of secondary importance. Parameter estimation is generally done by Generalized Estimating Equations models which were presented in Section 2.2.2. Generalized Estimating Equations estimation is now included as a macro routine in various software (SAS, Stata) which may explain why

these models are so popular. Diggle (1993) used a Generalized Estimating Equations model to analyze the number of seizure occurrences in a longitudinal epilepsy trial to show little effect of progabide over placebo in reducing the number of seizures. Several limitations associated with Generalized Estimating Equations models were identified in Section 2.2.2.

Random Effects Models Random-effects models or subject-specific models assess the effect of a treatment on an average subject. The subject variability (heterogeneity) is modelled through the addition of random effects in the mean regression relationship. The mean response μ_{it} for the i^{th} response at time t depends now on a set of covariates x_{it} and random effects variables z_{it} such as $\mu_{it} = log(E(y_{it} \mid \gamma_i) = x_{it}\beta + z_{it}\gamma_i$ for count data. The analysis of a clinical trial where the dependent variable was the number of seizures suggested that regression estimates of a random effects model are generally different from marginal model estimates (Diggle, 1993). Estimation of random effects models is sometimes performed by Generalized Estimating Equations models as in Diggle (1993). Therefore, the mean and the distribution assumptions will be valid only if the data does not display a significant excess of zeros, and if the marginal distribution of the counts is a member of the exponential family.

4.1.3 Discussion

For health care interventions such as new pharmacotherapies, many countries have formal requirements for provision of safety and efficacy data prior to product licensing. Randomized controlled clinical trials provide the assessment of the efficacy of a new

drug treatment versus placebo or usual care. However, several global trends in health care are contributing to interest in drug treatment beyond the historic hurdles of efficacy, safety and manufacturing quality. Rising health care costs are pressuring purchasers of health care to increasingly ask whether health care interventions they currently pay for are efficient compared to standard treatments. In Canada, economic evaluations have been mandatory for any new pharmacotherapy to be considered for formulary approval or reimbursement since 1994. The same requirements are shared by some managed care organizations in the United States and other governments such as Australia and the United Kingdom.

There is a growing literature on economic evaluation in health care. Economic evaluations, the comparative analysis of medical therapies in terms of both their cost and consequences (Drummond et al., 1999), provide important information to decision-makers regarding the effectiveness of a medical intervention. Economic evaluations of drug treatments are being conducted under a range of pharmacoeconomic labels, the most common being a cost-effectiveness analysis. Cost-effectiveness evidence at launch time is critical for gaining public or private reimbursement. If the new drug treatment is not cost-effective versus standard care, payers may decide not to reimburse this product. As such, economic evaluations are playing a key role in marketing new pharmaceutical products. However, a pharmacoeconomic evaluation of a health care program is only as good as the effectiveness data it is built upon (Drummond et al., 1999).

A reason often expressed to explain the increase in health expenditures in de-

veloped countries is the increase in the price of new medications marketed by pharmaceutical companies. However, a high price for a very effective treatment may be justified from an economic point of view.

Consider, for example, the following cost-effectiveness scenario in which two treatments, A and B, are compared over a 3-month time period in terms of proportion of patients cured (i.e., 0 episodes). Treatments A and B cost $20 and $10 respectively over the period of analysis. Two visits to a general practitioner are assumed for cured patients at a cost of $40 per visit. Four GP visits are associated with non-cured patients. The cost of the visit for patients not cured is $50 due to an additional test required to be performed. The three-month efficacy rates of treatments A and B are derived from a clinical trial. Results indicate that 80% of the subjects are cured (zero episodes) with treatment A and 60% with treatment B. The statistical analysis indicated that treatment A is statistically superior to treatment B. For modelling purposes, it is also assumed that only two outcomes are observed at the end of the three-month time period: cured and not cured. Using this information, it is possible to calculate a cost-effectiveness ratio in order to compare these two treatments using the decision tree model reported in Figure 4. Decision tree analysis is commonly used in pharmacoeconomics to represent the pattern of care of a particular disease. In this example, the efficacy of each treatment is expressed in terms of percentage of patients cured by the treatment. Other outcomes commonly used in pharmacoeconomics are the number of events avoided by the treatment or the number of lives saved by the treatment.

In our example, the expected cost of treatment A and treatment B are calculated at $124 and $138, respectively. Dividing the expected cost by the number of patients cured yields cost-effectiveness ratios of $180 per cured subject for treatment A and $247 per cured subject for treatment B as indicated in Figure 4. Based on these results, it is concluded that treatment A, even at a premium price, is more cost-effective than treatment B since it is associated with the lowest cost to cure one subject. From an economic point of view, it is rational to reimburse treatment A because more patients are cured at a lower cost which will in turn save money to be reinjected into the health care system.

Decision tree analyses are of course more complex in reality with multiple branches representing different potential outcomes over a certain period of time. Chance modes are evaluated based on success and failure rates given by clinical trials. Failure or dropout rates due to an adverse effect reaction are also collected during clinical trials and may be introduced for modelling purposes if significant. Decision nodes represent different treatment alternatives as a function of the therapeutic response over a certain period of time. For example, following an initial failure at the usual dose, therapeutic options may be to increase the dosage of the treatment to its maximum or to switch to another medication. Resources utilization such as the number of doctor visits, laboratory tests, procedures and medications are identified and costed for each branch of the decision tree to populate the model. Drug prices are introduced in the model to determine the cost effectiveness profile of the treatments of reference.

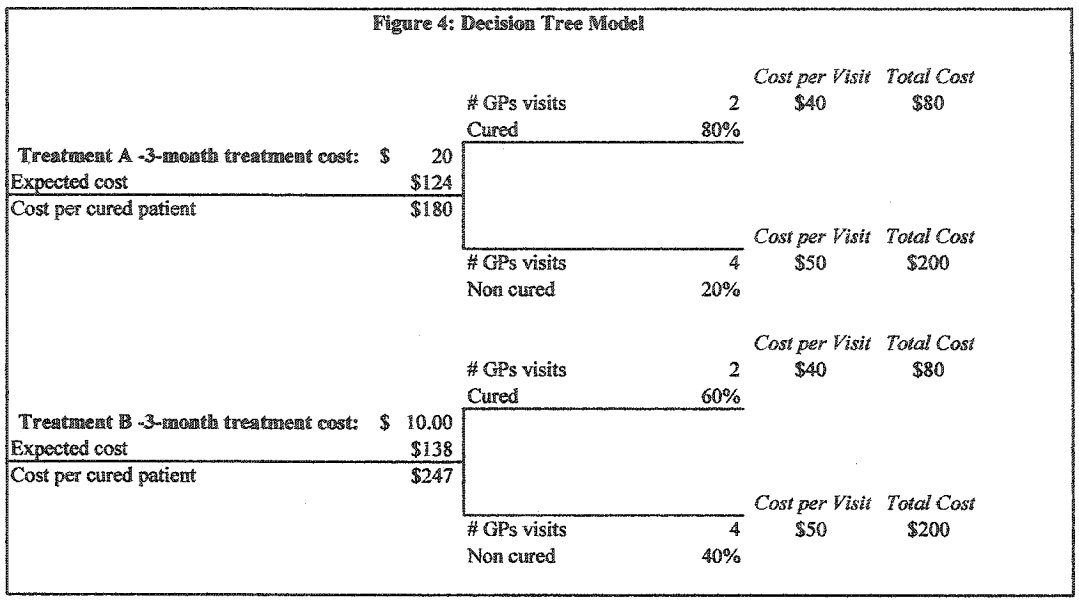
Two pillars of any pharmacoeconomic study are the price of the treatment and its

efficacy. The price of a new treatment may be determined by the economic evaluation to fall under acceptable cost-effectiveness cut-off values. In general, higher prices will be associated with products with better efficacy. Unfortunately, health economists do not have any control over the determination of the efficacy of the treatment since this analysis is conducted by biostatisticians. This may be unfortunate if the statistical analysis is not well designed. In the worst case scenario, the analysis will find a difference between two treatments where there is no difference or vice-versa. In this scenario where both treatments are in fact similar in efficacy, treatment A is not any more cost-effective versus treatment B. In our example, simple calculations indicate that the cost per subject cured is $280 with treatment A if the efficacy is similar to treatment B. This simple example demonstrates the importance of the determination of drug treatment's efficacy to be used in economic evaluations.

Section 4.1.1 and 4.1.2 have reviewed the statistical issues associated with longitudinal count data derived from clinical trials and the major limitations associated with the standard methods of analysis. Analysis of the variance such as the ANOVA procedure does not consider the discrete aspect of the data. Univariate analysis does not consider the longitudinal aspect of the data. Generalized Estimating Equations models assume that the data is from the linear exponential family and that the mean is correctly specified and these models are unable to account for a potential excess of zeros. All these limitations outline the importance of developing methods of analysis which integrate all the characteristics associated with longitudinal clinical trial count data including the presence of extra zeros. This is extremely important since clinical

data is used by health economists to perform economic evaluations of pharmaceutical products for decision-makers.

Figure 4: Decision Tree Model



4.2 The Data

The data that is analyzed in this chapter is an unpublished subset of a twelve-week randomized placebo-controlled clinical trial comparing the efficacy of two treatments and a placebo. Subjects were recruited for the clinical trial and randomized to one of the three groups. Participants were followed for several weeks after initiation of the trial and one assessment was done every two weeks. During these encounters, one primary endpoint and several secondary endpoints were collected to assess the efficacy of the treatments. In the following analysis, one of the secondary endpoints corresponding to the number of events or episodes that subjects were experiencing, was used as efficacy variable. For illustration, subjects with zero episodes at the end of the period of observation will be referred in the following text as cured subjects.

Access to a subset of this clinical trial data was given on condition that the name of the illness, the manufacturer(s), the treatments and the data set would not be disclosed.

This data subset included information on 116 participants for the 6-week period following initiation of the treatments. Thirty received placebo, 54 received treatment A and 32 received treatment B. For our efficacy variable, the baseline value and the number of events at week 2, 4 and 6 were available for each subject as well as the following information: age, height, weight and "duration" as a health indicator. This health indicator reflected how long the subject had been suffering from this particular illness. The mean age of our sample was 59.

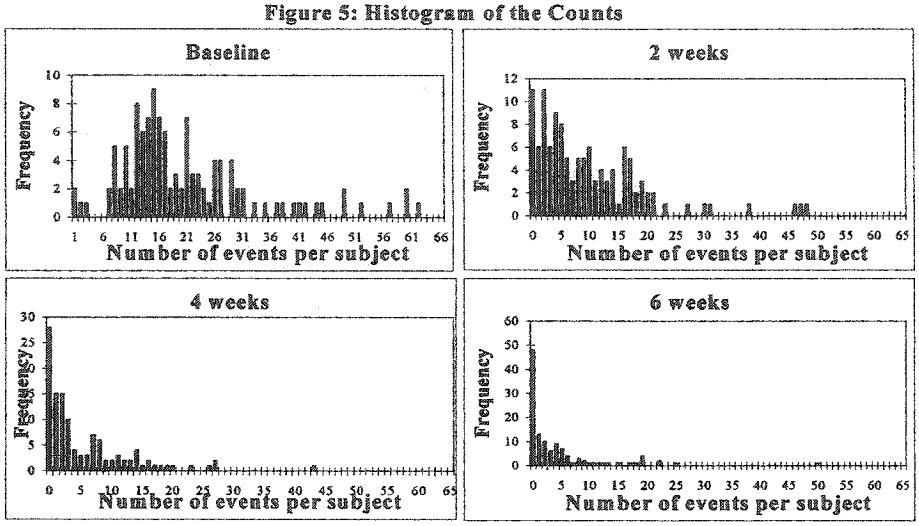
Table 18 reports the descriptive statistics of the counts at baseline and at weeks 2, 4 and 6. As can be observed in this table, there are some differences in the baseline counts per treatment group. At the beginning of therapy, subjects assigned to treatment B had on average a smaller number of episodes than the other groups. As indicated in Table 18, the treatments A and B were very effective in reducing the number of events. The mean number of events after 2 weeks of treatment was reduced by more than 50% in the groups assigned to treatments A and B and this reduction continued over time. After 6 weeks of treatment, the mean percentage of reduction from baseline to endpoint was close to 90% for treatments A and B versus almost 60% for the placebo group. Assuming that the observed reductions were only due to the treatments would be incorrect, since there is a natural trend as indicated by the reduction observed in the placebo group.

Table 18: Descriptive Statistics: Dependent variable. Means and (Standard Deviations).

Descriptive Statistics per Treatment Group	Counts at week 0	Counts at 2 weeks	Counts at 4 weeks	Counts at 6 weeks
Total Sample	y_0	y_1	y_2	y_3
Mean	19.98	9.89	5.67	4.12
(Std. Deviation)	(12.59)	(9.67)	(7.29)	(7.15)
Treatment A	y_0	y_1	y_2	y_3
Mean	22.26	9.91	4.69	3.37
(Std. Deviation)	(12.99)	(9.56)	(5.77)	(5.40)
Treatment B	y_0	y_1	y_2	y_3
Mean	16.09	6.56	2.59	1.47
(Std. Deviation)	(8.95)	(5.81)	(4.39)	(3.15)
Placebo	y_0	y_1	y_2	y_3
Mean	20.03	13.40	10.73	8.30
(Std. Deviation)	(14.42)	(11.96)	(9.53)	(10.58)

An analysis of the variance-mean ratios of the counts indicates that the distribution of the counts displays signs of overdispersion. The variance is 5 to 13 times greater than the mean depending on the treatment group and the time period, as indicated in Table 18. The data is too overdispersed to consider a Poisson distribution that requires the equality of the mean and the variance. The analysis should also try to determine if the excessive zeros observed in the data are caused by effective treatments or by unobserved heterogeneity that predisposes patients to be cured, for example.

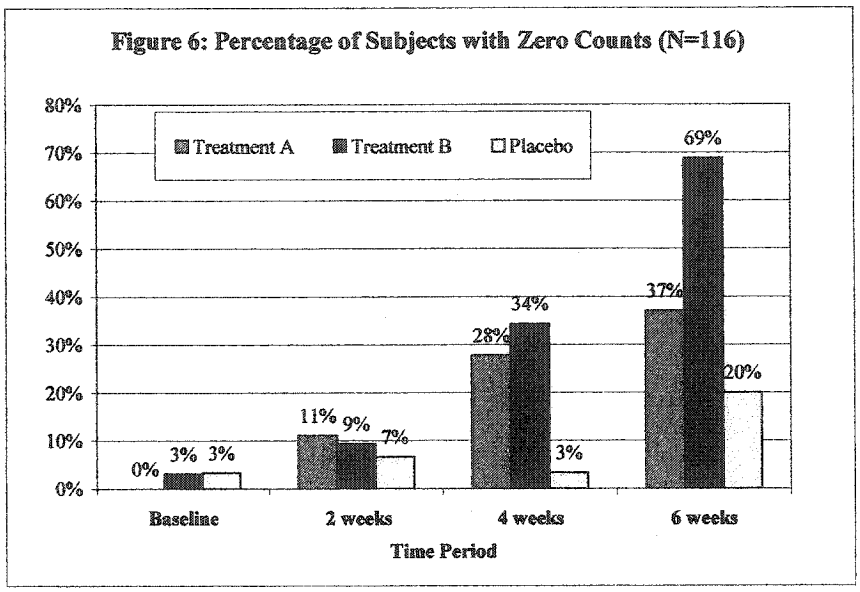
Figure 5: Histogram of the Counts



The analysis of the shape of the probability distribution of the counts indicates that the data is strongly skewed to the right (lack of symmetry) and presents thin-tails for the last two observations. Clearly, the shape of the distribution of the counts changes over time as illustrated in Figure 5, the histogram of the counts per time period. With time, more subjects are cured (i.e.: 0 episodes) due to the treatments. The number of zeros becomes more significant, and as a consequence, the functional form of the distribution changes. In this particular data set, 9 % of all participants had zero episodes 2 weeks following initiation of the treatments versus 41% at six weeks. This is depicted in Figure 6 which represents the percentage of subjects with zero counts per treatment group and time period. Almost 20% of the subjects assigned to the placebo group had zero episodes at 6 weeks which again indicates a trend effect.

Table 19: Correlation Matrix.

	y_0	y_1	y_2	y_3
y_0	1	0.63	0.27	0.26
y_1	0.63	1	0.58	0.63
y_2	0.27	0.58	1	0.83
y_3	0.26	0.63	0.83	1



Finally, due to the repeated nature of the data, the correlation matrix of the counts at each time period given in Table 19 indicates a high positive correlation in the responses reflecting a possible positive contagion effect, especially between 2 consecutive responses.

4.3 Standard Estimations

4.3.1 Univariate and Generalized Estimating Equations Models

Univariate Model Standard analyses include the fitting of the data by the Univariate Negative Binomial defined in equation (24) applied to the last count ob-

served at the end of the trial. The parameterization used for subject i at period t=3 is:

$$\lambda_i = \exp\left[\begin{array}{c}\beta_0 + \beta_1 A_i + \beta_2 W_i + \beta_3 H_i + \beta_4 D_i + \\ (\tau_i TrtA) + (\delta_i TrtB) + \theta_i B\end{array}\right] \tag{109}$$

In the parametrization, A_i, W_i, H_i and D_i are individual covariates for subject i, respectively age, weight, height and duration of illness. $TrtA$ and $TrtB$ are treatment dummy variables taking the value 1 for those patients assigned to treatment A and to treatment B respectively. In this formulation, the value of baseline count B is introduced in the mean as commonly done.

Generalized Estimating Equations Model

A Generalized Estimating Equations model was also applied to the data. In order to demonstrate the importance of the initial value, two Generalized Estimating Equations models were applied to the data. The first model follows the specification of Diggle (1994) and Albert (1999) in which the baseline value is introduced in the regression mean. In order to capture the difference between each treatment and the placebo groups, two additional variables were created by multiplying the baseline value (B) by the dummy treatment ($TrtA$ or $TrtB$). The mean was expressed as follows:

$$\lambda_{it} = \exp\left[\begin{array}{c}\beta_0 + \beta_1 A_i + \beta_2 W_i + \beta_3 H_i + \beta_4 D_i + \\ (\tau_{it} TrtA) + (\delta_{it} TrtB) + \theta_i B + \omega_1 (B * TrtA) + \omega_2 (B * TrtB)\end{array}\right]. \tag{110}$$

The second model is a Generalized Estimating Equations model in which the baseline value is considered endogenous. In this case, the mean function is written as in equation (111) with the exception that the baseline value is now endogenous.

$$\lambda_{it} = \exp\left[\begin{array}{c} \beta_0 + \beta_1 A_i + \beta_2 W_i + \beta_3 H_i + \beta_4 D_i + \\ (\tau_{it} Trt A) + (\delta_{it} Trt B) \end{array}\right] \tag{111}$$

4.3.2 Results

Univariate Negative Binomial Model The results of the Univariate Negative Binomial model which concentrates on the endpoint and considers the baseline value as a regressor are reported in Table 20. The coefficient of overdispersion is significant suggesting unobserved heterogeneity in the data. Surprisingly, as the duration of symptoms increases, the number of events is significantly reduced. While the baseline value is not significant at the 5% level, it is significant at the 10% level in increasing the number of counts. Results indicate that the two treatments are significant in reducing the number of episodes. However, a Likelihood Ratio Test failed to reject the equality of the treatments. The value of the likelihood function of the univariate model was -259.10 when the treatment parameters at the endpoint were set equal ($\tau_3 = \delta_3$) versus -257.60 for the unrestricted model. Twice the difference is 3.01 which is less than 3.84, the critical value of χ^2_1.

Generalized Estimating Equations Model Generalized Estimating Equations estimations were performed using the GenMod procedure of the SAS Institute (version

Table 20: Maximum Likelihood Parameter Estimates: UNB Model.

Maximum Likelihood Estimation		Univariate Negative Binomial Model	
Number of observations:		116	
Convergence achieved after:		9 iterations	
Log of Likelihood:		-257.599	
Variable	Parameter	Estimate	Std. Error
Constant	β_o	-1.176	3.370
Age	β_1	-0.001	0.012
Height	β_2	0.029	0.020
Weight	β_3	-0.013	0.008
Duration	β_4	-0.005*	0.003
Baseline Count	β_5	0.024**	0.012
Efficacy Trt A	τ	-1.021*	0.347
Efficacy Trt B	δ	-1.686*	0.423
Overdisp. Coeff.	α	0.725*	0.068
* indicates significant at the 5 % level.			
** indicates significant at the 10 % level.			

6.2). Column 2 of Table 21 presents the results of a Generalized Estimating Equations (GEE) estimation as conducted by Diggle (1993) and Albert (1999) in their analysis of the number of epilepsy seizures. In this formulation, the baseline value is treated as exogenous. Results indicate that the baseline value is significant and has a positive impact on the number of counts at the end of the trial. The two treatments were statistically significant in reducing the number of episodes. This result is, however, attenuated in this model because for the two treatments the change in the number of episodes before and after randomization was not significant at 95% but only at 90%.

Table 21 presents in column 3 the results when a Generalized Estimating Equations model is applied to the data when the baseline value is considered endogenous. As can be seen, the two Generalized Estimating Equations models have different implications. When the baseline value is treated as endogenous, treatment B is statistically significant at 95% level of confidence but treatment A is only significant

at the 90% level. Due to the characteristics of the Generalized Estimating Equations model, it is not possible to investigate further. This example demonstrates the importance of the treatment of the initial value in the analysis.

Another drawback is that the GenMod procedure from the SAS Institute to estimate Generalized Estimation Equations for longitudinal count data does not allow us to have a sense of the time shape of the efficacy measures. While it is theoretically possible in a Generalized Estimating Equations model to integrate different coefficients for the treatment coefficients at different time periods, the GenMod procedure does not allow it. In addition, based in our example, GenMod estimations were unable to predict any zeros at all as reported in Table A1 of the Appendix. McIntosh (2001) already observed that the results of the maximization of the Poisson likelihood function does not correspond to the results given by the GenMod procedure for Poisson data. This is an important result since the software of the SAS Institute is the standard statistical software used by the pharmaceutical industry to analyze clinical trials. For the analysis for longitudinal count data, the SAS Institute refers to the GenMod procedure using the example of Diggle (1993) but without identifying these important limitations associated with this procedure.

4.4 Quadrivariate Negative Binomial Model

Multivariate Negative Binomial models for longitudinal count data were presented in Section 2.1.2. To my knowledge, Multivariate Negative Binomial models have never been applied to analyze longitudinal clinical trial count data. A Quadrivariate Neg-

Table 21: Parameter Estimates: GEE Models.

Maximum Likelihood Estimation		GEE (Baseline=covariate)		GEE	
Number of observations:		116		116	
Log of Likelihood:		NA		NA	
Variable	Parameter	Estimate	Std. Error	Estimate	Std. Error
Constant	β_o	-1.223	1.377	-0.552	2.40
Age	β_1	0.011	0.008	-0.005	0.006
Height	β_2	0.017**	0.009	0.009	0.013
Weight	β_3	-0.010*	0.004	-0.004	0.004
Duration	β_4	0.154	0.131	0.135	0.100
Baseline Count	β_5	0.016*	0.006	-	-
Efficacy Trt A	τ	-1.142*	0.323	-0.278**	0.144
Efficacy Trt B	δ	-1.649*	0.382	-0.670*	0.160
Baseline*Trt A	$\beta_5\tau$	0.021**	0.010	-	-
Baseline*Trt B	$\beta_5\delta$	0.034**	0.004	-	-
* indicates significant at the 5 % level.					
** indicates significant at the 10 % level.					

ative Binomial model is therefore presented to analyze the clinical trial count data presented previously The model addresses four characteristics that are commonly found in clinical trials: repeated observations on a count used as an efficacy variable, the presence of covariates, unobserved heterogeneity and a count dependence structure.

In this sense, the methodology is particularly relevant for analyzing longitudinal clinical trial count data. The repeated aspect of the data and the correlation between the counts are implicitly modelled by Multivariate Negative Binomial distributions. Unobserved heterogeneity is integrated through the Negative Binomial distribution. The contagion effect is modelled by conditioning on previous events. By conditioning successive responses to previous responses including the initial value, Multivariate Negative Binomial distributions differentiate individual subjects in terms of their specific characteristics, observed or not, and allow us to follow the evolution of the

response of an average or typical subject. Treatment and trend effect can be differentiated by the analysis and the time shape of the responses determined. In addition, the methodology considers the baseline efficacy variable as a random rather than an exogenous variable. This is of relevant importance in the analysis of clinical trial count data due to the generally limited number of repeated observations available for analysis.

4.4.1 Model Specification

In the following, $Y_{it} = (y_{i0}, y_{i1}, y_{i2}, y_{i3})$ is the vector of counts observed for subject i at period t (t = 0, 1, 2 and 3). The length of each period is two weeks. The parameterization used for subject i at time t is:

$$\lambda_{it} = \exp\left[\begin{array}{c} \beta_0 + \beta_1 A_i + \beta_2 W_i + \beta_3 H_i + \beta_4 D_i + \\ (\tau_{it} TrtA) + (\delta_{it} TrtB) + \theta_t T \end{array}\right] \tag{112}$$

with t = 0, 1, 2, 3 ; $\tau_{i0} = \delta_{i0} = 0$. In the parametrization, A_i, W_i, H_i and D_i are individual time-invariant covariates for subject i, respectively age, weight, height and duration. $TrtA$ and $TrtB$ are treatment dummy variables taking the value 1 for those patients assigned to treatment A and to treatment B respectively. Because treatment effects are cumulative over time, different coefficients were attributed to each time period for each treatment allowing the estimation of the trend treatment effects over time. T is the trend variable, which is assumed to be linear. For the initial counts (e.g. t = 0; no treatment), the treatment coefficients γ_{i0} and δ_{i0} were set equal to

zero as well as the coefficient of the trend θ_0.

The likelihood for the Quadrivariate Negative Binomial probability distribution defined in equation (24) has a simple form that is tractable which when applied to our clinical trial involving 116 patients and four repeated measurements is given by:

$$l(\alpha,\beta) = \sum_{i=1}^{N=116} \left[\begin{array}{c} \ln\Gamma(\alpha + \Sigma_{t=0}^{3} y_{it}) - \ln\Gamma(\alpha) - \Sigma_{t=0}^{3}(\ln y_{it}!) + \\ \Sigma_{t=0}^{3}\left\{y_{it}\ln\lambda_{it}\right\} - (\alpha + \Sigma_{t=0}^{3} y_{it})\ln(1 + \Sigma_{t=0}^{3}\lambda_{it}) \end{array} \right] \tag{113}$$

Estimation is performed by maximizing equation (114) with respect to β, the vector of parameters and α, the coefficient of overdispersion, using the Newton-Raphson algorithm.

4.4.2 Results

Parameter Estimates Column 2 of Table 22 presents the maximum likelihood parameter estimates with their corresponding standard errors for the Quadrivariate Negative Binomial model defined by equation (114).

The results of the Quadrivariate Negative Binomial model indicate that the constant and the covariates are not significant in explaining the number of counts. At each time period, both treatments are highly significant in reducing the number of events as well as the trend. As expected, these coefficients are negative and the trend effect is lower than the treatment effects as indicated by the small absolute value of the estimate of the trend. The coefficient of overdispersion (α) is significant reflecting unobserved heterogeneity in the population.

Table 22: Maximum Likelihood Parameter Estimates: QNB Model.

Maximum Likelihood Estimation		Quadrivariate Negative Binomial Model	
Number of observations:		116	
Convergence achieved after:		12 iterations	
Log of Likelihood:		-1,705.924	
Variable	Parameter	Estimate	Std. Error
Constant	β_o	1.903	1.410
Age	β_1	0.003	0.006
Height	β_2	0.003	0.008
Weight	β_3	-0.002	0.003
Duration	β_4	-0.002	0.001
Baseline Count	β_5	-	-
Efficacy Trt A			
at 2 weeks	τ_1	-0.502*	0.056
at 4 weeks	τ_2	-0.954*	0.083
at 6 weeks	τ_3	-0.987*	0.105
Efficacy Trt B			
at 2 weeks	δ_1	-0.623*	0.084
at 4 weeks	δ_2	-1.254*	0.126
at 6 weeks	δ_3	-1.526*	0.167
Trend	θ	-0.148*	0.012
Overdisp. Coeff.	α	1.415*	0.092
* indicates significant at the 5 % level.			
** indicates significant at the 10 % level.			

Treatment B is superior to treatment A at each time period in reducing the number of events as indicated by larger estimates in absolute value than treatment A. Based on a likelihood test ratio, this difference is significant. The value of the log of the likelihood is -1,713.12 when both treatments are set equal at each time period ($\tau_t = \delta_t$ and t =1, 2 and 3) versus -1,705.92 for the unrestricted model. Twice the difference of the likelihood functions is 14.39 which is superior to 7.82, the 95% critical value for the χ^2 with 3 degrees of freedom. The null hypothesis of equality of the two treatments is therefore rejected since twice the difference exceeds the critical value of the χ^2_3.

Additionally, the values of the estimates indicate that while the efficacy of treatment B increases between 4 and 6 weeks (from -1.25 to -1.53), the efficacy of treatment A is stable during the same period (-0.95 and -0.99, respectively). This assumption was accepted based on a likelihood ratio rest. The value of the log-likelihood of the restricted model was -1,705.98 when the efficacy of treatment A at 4 and 6 weeks was set equal ($\tau_2 = \tau_3$), all other coefficients being different, versus -1,705.92 for the unrestricted model. Twice the difference is less than 3.84, the 95% critical value for the χ^2_1. Therefore, the reduction in counts observed between 4 and 6 weeks among patients assigned to treatment A seems to be caused by a natural trend effect rather than by treatment A. In fact, a Quadrivariate Negative Binomial model without a trend variable in the mean function is unable to detect that treatment A reaches its maximum efficacy after 4 weeks of treatment. When the data is estimated by a Quadrivariate Negative Binomial model with no trend as regressor, the value of the

coefficients associated with treatment A increases between 4 and 6 weeks from -1.53 to -1.86 and this difference is significant according to a Likelihood Ratio Test. Excluding a trend in the analysis presumes that the reduction in the number of events is only due to the treatment, which may not be true as seen in our example.

In summary, based on these Likelihood Ratio Tests, both treatments are significant in reducing the number of events. Treatment B is statistically superior to treatment A at each time period and its therapeutic action lasts longer than treatment A. Maximum efficacy of treatment A is reached after 4 weeks. None of the covariates are significant in explaining the reduction in the number of events.

This was not the case when the analysis concentrated on the endpoint and considers the baseline value as regressor as indicated by the results given in Section 4.3.2. In the univariate analysis, a Likelihood Ratio Test failed to reject the equality of the treatments. Finally, while there are no significant differences in the parameter estimates for treatment effects at 6 weeks between the two Negative Binomial models, the standard errors were larger in the univariate model indicating less accuracy with this model.

In conclusion, based on our example, caution should be taken when generalizing results from univariate models when the baseline value is treated as exogenous. Because univariate models do not exploit the richness of the data, they cannot provide accurate information on the treatment efficacy and time shape efficacy parameters. For example, the univariate model was also unable to detect that treatment A reached its maximum efficacy at 4 weeks as indicated by the results of the Quadrivariate Negative

Binomial model. Similarly, the GenMod procedure given in SAS for the Generalized Estimating Equations model does not allow us to estimate the time shape of the efficacy of the treatments and results are unclear.

Predictions Once the parameters of the Quadrivariate Negative Binomial model were estimated, the predicted mean counts at each time period t, were derived from equations (38). As can be seen in columns 3 and 4 of Table 23, the Quadrivariate Negative Binomial model performs well in predicting the mean counts and their evolution over time.

For each treatment group, the Quadrivariate Negative Binomial model predicts exactly the mean percentage reduction of episodes from baseline to endpoint (85% for treatment A, 91% for treatment B and 59% for placebo). The absolute difference in the predicted mean counts and the actual means is less than 5% for the full sample and ranges from 11% to 12% for treatment A, from 22% to 25% for treatment B and from 2% to 9% for the placebo group, depending on the time period. Table 24 presents the observed and predicted coefficients of correlation as defined by equation (31). The predicted coefficients of correlation between two consecutive observations are similar to the observed correlation. However, the model tends to overpredict the correlation for non-successive observations. For example, the observed coefficient of correlation of the efficacy variable between 0 and 4 weeks is 0.28 while the model predicts a value of 0.80.

Finally, it is important to determine the predicted number of zeros by the Quadri-

Table 23: Observed and Predicted Mean Counts. QNB Model.

Treatment Group Time Period	Actual Data	Quadrivariate Negative Binomial Model	% Difference
	Means	Predicted Means	(Predicted versus
Treatment A			
Baseline	22.26	19.67	-12%
2 weeks	9.91	8.85	-11%
4 weeks	4.69	4.19	-11%
6 weeks	3.37	3.01	-11%
Treatment B			
Baseline	16.09	20.10	25%
2 weeks	6.56	8.02	22%
4 weeks	2.59	3.17	22%
6 weeks	1.47	1.79	23%
Placebo			
Baseline	20.03	19.70	-2%
2 weeks	13.40	14.65	9%
4 weeks	10.73	10.89	1%
6 weeks	8.30	8.09	-3%

Table 24: Actual and Predicted Correlations

	Actual	QNB Predicted
	Correlation	Correlation
$\rho(0,1)$	0.63	0.87
$\rho(0,2)$	0.27	0.80
$\rho(0,3)$	0.26	0.74
$\rho(1,2)$	0.58	0.76
$\rho(1,3)$	0.63	0.71
$\rho(2,3)$	0.83	0.66

variate Negative Binomial model in order to have an appreciation of the model in determining how many patients have zero episodes. In this calculation, the predicted number of zeros at t was determined jointly when the count at time t was set equal to 0. The results indicate that the Quadrivariate Negative Binomial model, as the Generalized Estimating Equations model, is unable to predict any zeros at all, offering a very poor fit of the data. In this sense the Univariate Negative Binomial model presented in Table 20 is superior. For the full sample, the Univariate Negative Binomial model predicts that 38 percent of the subjects will be cured at 6 weeks versus an observed percentage of 41%. The predicted mean count by the univariate model is 4.33 at 6 weeks which compares with an observed 4.12. However, as shown previously, the Univariate Negative Binomial model was unable to detect any difference between treatments A and B or that treatment A reaches its maximum efficacy at 4 weeks. Furthermore, the correlation arising from the repeated nature of the clinical trial is not taken into account in univariate models. Therefore other alternatives have to be investigated to obtain a better fit of the data and to draw final conclusions on the assessment of the treatments.

4.5 Quadrivariate Negative Binomial Zero-Inflated (QNB^Z) Model

A Quadrivariate Negative Binomial Zero-Inflated (QNB^Z) model is presented to analyze the data presented in section 4.2, which is characterized by a high proportion of zeros. In Zero-Inflated models, zeros are generated in two regimes. In regime 1

the probability of generating a zero is always one. In regime 2, the 0s but also the positive counts, are generated according to a different probability function.

This Multivariate Zero-Inflated specification allows for a systematic difference in the statistical process governing two regimes of zeros at each time period following initiation of treatment while accounting for the repeated aspect of the longitudinal count data. The idea is simple and is analogous to the definition of the multivariate negative binomial distribution.

4.5.1 Model Specification

The Quadrivariate Negative Binomial Zero-Inflated (QNB^Z) distribution was defined as the product of three Zero-Inflated conditional Negative Binomial distributions (t = 1, 2 and 3) and the Univariate Negative Binomial distribution for the initial value (t = 0). Since some patients with zero counts at t have positive counts at t +1 (t =1 and 2), it was assumed that the allocation of the subjects in each regime was independent across the three time periods. Where unknown parameters and strictly exogenous variables are suppressed, the joint density function of the four repeated counts $Y_{it} = (y_{i0}, y_{i1}, y_{i2}, y_{i3})$ was therefore defined as:

$$QNB^z(Y_{it}) = NB_0(y_{i0}) \prod_{t=1}^{3} NB_t^z(y_{it} \mid Y_{it-1}). \tag{114}$$

In this equation NB_t^z (t = 1, 2 and 3) represents Negative Binomial Zero-Inflated distribution at period t for the counts y_{it} conditional on the vector of previous responses Y_{it-1}. The term NB_0 is the Univariate Negative Binomial density function

for the initial count, which is considered endogenous. A Zero-Inflated specification was not deemed appropriate for the initial value since the subjects are treatment naive at initiation of the trial. The following equation gives in its abbreviated form the associated log-likelihood function when applied to our data set of four repeated observations on 116 patients:

$$l(\phi_t, \beta_t) = \sum_{i=1}^{N=116} (\ln \mathrm{L}NB_0 + \ln \mathrm{L}NB_1^z + \ln \mathrm{L}NB_2^z + \ln \mathrm{L}NB_3^z). \qquad (115)$$

Estimation is conducted by maximizing this log-likelihood function with respect to the vector of parameters ϕ_t associated with the distribution governing the zeros in Regime 1 and with respect to β_t the vector of parameters associated with the distribution governing the zeros and the positive counts in Regime 2.

In this formulation, $\ln \mathrm{L}NB_0$ represents the addition to the likelihood for the Univariate Negative Binomial density for the baseline count. Each $\ln \mathrm{L}NB_t^z$ (t =1, 2 and 3) is the likelihood function of the conditional Negative Binomial Zero-Inflated distribution at time t defined as:

$$\ln \mathrm{L}NB_t^z = \left[\begin{array}{c} 1_{(y_{it}=0)} \ln(\Phi_t + (1 - \Phi_t(y_{it}))NB_t(y_{it} = 0 \mid Y_{it-1})) \\ + \, 1_{(y_{it}>0)} \ln((1 - \Phi_t(y_{it}))NB_t(y_{it} \mid Y_{it-1})) \end{array} \right]. \qquad (116)$$

In this equation $1_{(y_{it}=0)} = 1$ if $y_{it} = 0$, 0 otherwise; and t = 1, 2 and 3. The term Y_{it-1} is the vector of counts observed in the preceding time periods.

The distribution, Φ_t is the cumulative univariate normal distribution governing the zeros in regime 1 (cured with probability 1) at time t. Here Φ_t was assumed to

depend only on treatment dummies and a given constant c. Because treatment effects are cumulative, different treatment coefficients were attributed to each time period as shown below.

$$\Phi_t(y_{it}) = \Phi_t(\varpi_{it} * TrtA + \eta_{it} * TrtB \text{ - c}) \quad \text{t = 1, 2 and 3} \tag{117}$$

In regime 2, the probability of a particular subject being cured or not depends on the number of counts that have already been observed. In particular, in equation (117), NB_t (t=1, 2 and 3), was defined by the $(t+1)$-variate Negative Binomial distribution of y_{it} conditional on the vector of previous responses Y_{it-1}. For example, for t=3, $NB_3(y_{i3} \mid y_{i2}, y_{i1}, y_{i0})$ was defined as the ratio of the Quadrivariate Negative Binomial density function of the vector (y_{i0}, y_{i1}, y_{i2}, y_{i3}) and the trivariate Negative Binomial density of (y_{i0}, y_{i1}, y_{i2}). In this regime, the mean is a function of individual covariates, treatment dummies and a trend effect as defined by equation (113). After this parametrization, estimation is performed by maximizing the log-likelihood function given in equation (116) with respect to β, the vector of parameters associated with the distribution of regime 2, α the coefficient of overdispersion, and to the treatment parameters of regime 1 (the ϖ's and η's). It should be noted that in this Zero-Inflated model, the probability of being in a particular regime depends only on the treatment group to which the subject was assigned and not any other observable subject characteristic. Therefore, an additional source of unobservable heterogeneity is introduced in the Zero-Inflated specification.

4.5.2 Results

The likelihood functions of the Quadrivariate Negative Binomial and Quadrivariate Negative Binomial Zero-Inflated distributions were used to test nested hypotheses since the two distributions are equal if $\Phi_1 = \Phi_2 = \Phi_3 = 0$. Differences among treatments were assessed by carrying out a series of Likelihood Ratio Tests.

Parameter Estimates The results of the maximization of the natural logarithm of the likelihood of the Quadrivariate Negative Binomial Zero-Inflated distribution defined in equation (116) are displayed in Table 25. When compared to the Quadrivariate Negative Binomial model, there is a significant increase in the likelihood function (from -1,705.92 to -1,593.99) since twice the difference in the log-likelihood function is 223.68, which is greater than 12.59, the 95% critical value for the χ^2_6. Additionally all coefficients associated with the treatment effects in regime 1, that is to be cured with probability 1, (the ϖ's and η's) are highly significant. Clearly the data presents a significant excess of zeros caused by treatments. Ignoring this feature by estimating the mean independently of the structure governing the zeros would be incorrect and will result in inconsistent estimates in our case. In particular, in the Quadrivariate Negative Binomial Zero-Inflated model, an additional source of zeros is generated which is parametrized by the ϖ's and η's, all statistically different from 0 in our example. Consequently, the treatment coefficients associated with regime 2 (the τ's and the δ's) have changed with respect to the non-inflated model.

As found in the Quadrivariate Negative Binomial model, the results indicate that the constant and the covariates are not significant. The coefficients of the covariates

in the Quadrivariate Negative Binomial Zero-Inflated model differ slightly but are of the same magnitude and sign as in the Quadrivariate Negative Binomial model except for the constant and the covariate height. The value of the constant is smaller in the Zero-Inflated specification and the covariate "height" now has a positive effect. But again, they are not significant. The trend and the coefficient of overdispersion are significant with similar values. When the excess of zeros is accounted for, both treatments are highly significant regardless of the regimes defining the Quadrivariate Negative Binomial Zero-Inflated model and treatment B is still superior to treatment A as found in the previous model.

The values of the estimates of the parameters given in column 2 of Table 25 suggest that for the process governing regime 1, (i.e. cured with probability 1), treatment B generates almost the same number of zeros as does treatment A at week 2 and week 4, with more zeros than treatment A after 6 weeks following treatment. In this regime, the treatment effects increase over time for treatment B and are almost similar between 4 and 6 weeks for treatment A. In contrast, in regime 2, treatment B is superior to treatment A for the first two time periods and almost similar to treatment A at 6 weeks. Under this regime, both treatments reached their maximum efficacy after four weeks following treatment and the efficacy of treatment A was almost similar at 4 and 6 weeks after initiation of treatment.

Likelihood ratio tests were carried out to test these differences among treatment. The first restricted model was built by imposing on regime 1 the equality of each treatment for the first two periods ($\eta_1 = \varpi_1$ and $\eta_2 = \varpi_2$) and a similar effect of

Table 25: Maximum Likelihood Parameter Estimates - QNBZ Model.

Maximum Likelihood Estimation		Quadrivariate Negative Binomial Zero-Inflated Model	
Number of observations:		116	
Convergence achieved after:		14 iterations	
Log of Likelihood:		-1,593.988	
Variable	Parameter	Estimate	Std. Error
Constant	β_0	2.518	1.321
Age	β_1	0.003	0.006
Height	β_2	-0.002	0.008
Weight	β_3	-0.002	0.003
Duration	β_4	-0.001	0.001
Regime 1			
Efficacy Trt A			
at 2 weeks	η_1	3.761*	0.231
at 4 weeks	η_2	4.346*	0.194
at 6 weeks	η_3	4.584*	0.187
Efficacy Trt B			
at 2 weeks	ϖ_1	3.616*	0.339
at 4 weeks	ϖ_2	4.403*	0.270
at 6 weeks	ϖ_3	5.376*	0.248
Regime 2			
Efficacy Trt A			
at 2 weeks	τ_1	-0.449*	0.057
at 4 weeks	τ_2	-0.746*	0.084
at 6 weeks	τ_3	-0.732*	0.104
Efficacy Trt B			
at 2 weeks	δ_1	-0.597*	0.085
at 4 weeks	δ_2	-1.040*	0.128
at 6 weeks	δ_3	-0.820*	0.169
Trend	θ	-0.135*	0.012
Overdisp. Coeff.	α	1.480*	0.100
* indicates significant at the 5 % level.			

treatment A at 4 and 6 weeks in regime 2 ($\tau_2 = \tau_3$). The value of the likelihood of the restricted model was -1,594.08 versus -1,593.99 for the unrestricted quadrivariate Zero-Inflated model. Twice the difference in the log-likelihood function is less than 7.82, the 95% critical value of χ^2_3, therefore these assumptions were not rejected. Additionally, another Likelihood Ratio Test could not reject the two additional hypotheses that treatment A and B had similar efficacy at 6 weeks in regime 2 ($\tau_3 = \delta_3$) and that the efficacy of treatment A was similar between 4 and 6 weeks in regime 1 ($\eta_2 = \eta_3$). The value of the likelihood of this new restricted model was 1,594.47 and twice the difference was less than 11.07, the 95% critical value of χ^2_5.

In conclusion, as found in the Quadrivariate Negative Binomial model, both treatments were significant in reducing the counts and in curing the patients. Treatment B is statistically superior to treatment A at each time period. This is true at week 2 and 4 in which treatment B is superior to treatment A in regime 2 while in regime 1, both treatments have a similar efficacy. At week 6, treatment B is superior to treatment A in regime 1 and equal to B in regime 2. In addition, treatment A reached its maximum efficacy at week 4 in the two regimes, meaning that the reduction observed in this treatment group between 4 and 6 weeks was due to a natural trend and/or unobserved heterogeneity that predisposed patients to be cured following one month of treatment. By allowing for two regimes, the Zero-Inflated specification defines a more refined mechanism than the Quadrivariate Negative Binomial model and provides additional information on how treatments are effective.

Table 26: Observed and Predicted Means. QNBZ Model.)

Treatment Group Time Period	Actual Data	Quadrivariate Zero-Inflated Negative Binomial Model	% Difference
	Means	Predicted Means	Actual versus
Treatment A			
Baseline	22.26	19.70	-12%
2 weeks	9.91	8.57	-14%
4 weeks	4.69	4.05	-14%
6 weeks	3.37	2.79	-17%
Treatment B			
Baseline	16.09	20.15	25%
2 weeks	6.56	7.77	18%
4 weeks	2.59	3.01	16%
6 weeks	1.47	1.40	-4%
Placebo			
Baseline	20.03	19.46	-3%
2 weeks	13.40	14.86	11%
4 weeks	10.73	11.35	6%
6 weeks	8.30	8.66	4%

Predictions Once the parameters of the Quadrivariate Negative Binomial Zero-Inflated model were estimated, the predicted mean counts at each time period t were derived. As seen in Table 26, which presents these predictions for the three treatment groups, the Quadrivariate Negative Binomial Zero-Inflated model gives a good fit of the data. The mean percentage reduction in the number of episodes from baseline to endpoint model is 86% for treatment A, 93% for treatment B and 55% for the placebo group versus actual reductions of 85%, 91% and 59%, respectively. The difference in the predicted and the actual means was less than 4% for the full sample and ranged from 12% to 17% for treatment A, from 4% to 25% for treatment B and from 3% to 11% for the placebo group, depending on the time period.

The predicted mean number of zeros generated by the Quadrivariate Negative

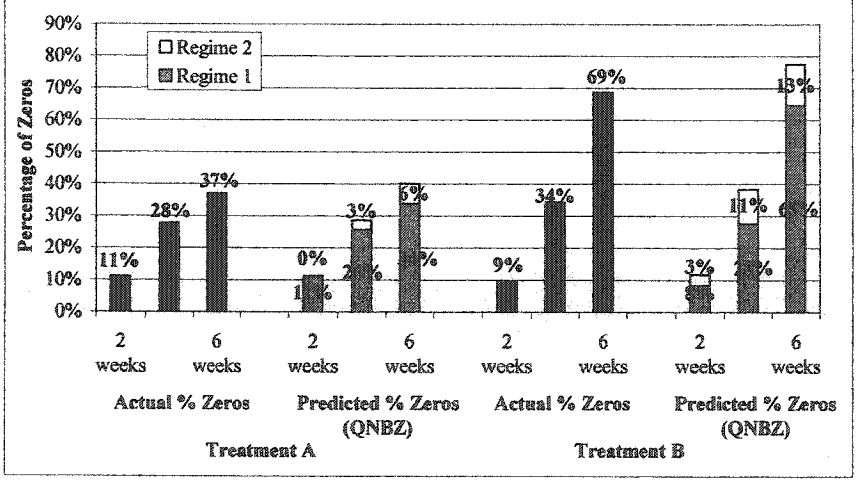
Binomial Zero-Inflated model was computed as the sum of the two probabilities of generating zeros associated with each of the two regimes. In regime 1, the predicted number of zeros at period t was defined by the cumulative normal distribution as $\Phi_t(\varpi_t^p * TrtA + \eta_t^p * TrtB - c)$, the subscript p referring to the QNB^z parameter estimates given in Table 26. In regime 2, the number of zeros at each time period t was defined conditionally on previous responses as $exp(\ -\alpha^p \ln(q_t^p \mid q_{t-1}^p) - \Sigma_1^t y_{t-1}(\ln(q_t^p \mid q_{t-1}^p))\ )$ in which $q_t^p = 1 + \Sigma_0^t \lambda_t^p$.

By generating two ways of zeros, the Quadrivariate Negative Binomial Zero-Inflated model offers a good description of the data. Not only were the mean counts correctly predicted, but also the number of patients who had zero episodes at each time period. For the full sample, the predicted percentage of cured patients (i.e. 0 episodes) was 8.8% at week 2, 22.9% at week 4 and 39.5% at week 6, comparing favorably to the actual percentages of 9.5%, 23.2% and 41.3%, respectively. An analysis per treatment group indicates that the model is very accurate in predicting the percentage of patients cured due to treatment A but less accurate for treatment B and placebo. For example, it was predicted that after 6 weeks of therapy with treatment B, 78% of the patients would be cured versus an observed percentage of 69%. In the placebo group, the number of zeros at week 6 was underpredicted by the model (4.1% versus 20%). Figure 7 presents the actual and predicted percentage of zeros per time period and per regime for treatment A and treatment B.

As expected, this graph illustrates that in a Zero-Inflated specification the zeros are mainly generated in regime 1, governed by cumulative univariate normal distri-

bution, in a proportion superior to 70%. An analysis per treatment group indicates that the proportion of zeros generated in regime 1 was higher for treatment A than treatment B within one month following treatment and equal thereafter. In particular, 96% and 90% of the zeros due to treatment A were generated in regime 1 at 2 weeks and 4 weeks respectively, versus 71% and 72% for treatment B. After 6 weeks of treatment, a similar proportion of 83% and 84% of the zeros were generated in regime 1 for treatment A and B, respectively. However, it is impossible to know if these differences in the allocation of the subjects in each regime are due to the treatments, or if they reflect unobserved heterogeneity that pre-disposes sub-groups of patients to be cured.

Figure 7: Observed and Predicted Percentage of Subjects with Zero Counts. QNBZ Model.



4.6 Conclusions

As also shown in Chapter 3, the analysis of longitudinal (clinical trial) count data characterized by an excess of zeros and correlation over time should be carefully undertaken. The contributions of this chapter are:

1. To have identified important weaknesses in the literature on the analysis of longitudinal clinical trials when the health outcome is a count and the treatments are very effective in reducing the number of episodes.

2. To have illustrated through a fictive cost-effectiveness example why economists should be more concerned about the method used to determine treatment efficacy.

3. To have proposed a new alternative to deal with an excess of zeros in the analysis of longitudinal clinical trial count data by developing a Multivariate Zero-Inflated Negative Binomial model nested with the Multivariate Negative Binomial model.

4. To have provided evidence that a) important differences result from the estimation of the data by a Univariate Negative Binomial model, a Quadrivariate Negative Binomial model and a Generalized Estimating Equations model; b) the presence of excess zeros in the longitudinal context can be tested and accounted for by Multivariate Negative Binomial Zero-Inflated models and c) analyses relying in Generalized Estimating Equations models as recommended by the SAS Institute and several authors (Diggle, 1993; Albert, 1999) could be misleading in the case of an excess of zeros in the data. In our example, the Generalized Estimating Equations model predicted no zero counts at all. In comparison, the Quadrivariate Negative Binomial

Zero-Inflated model was able to predict the mean number of zero counts at each time period as well as the mean number of counts.

In this chapter, a Multivariate Zero-Inflated framework was presented to analyze longitudinal count data in the presence of covariates, overdispersion, correlated counts and an excess of zeros. In accordance with time series literature, the baseline value was considered as a random variable rather than as a regressor.

This methodology is preferable to univariate methods for analyzing longitudinal clinical trial count data because it exploits all the information available yielding more efficient and reliable estimates. In our particular example, the Univariate Negative Binomial model was unable to detect any treatment differences when in fact treatments were different. Moreover, analysis concentrating on the endpoint of the clinical trial could not determine the time shape of the treatment effects nor treat the correlation among the efficacy variables yielding inconsistent estimates.

The methodology presented in this chapter offers a perspective that is complementary to Generalized Estimating Equations models by abandoning the structure used thus far consisting of a given mean function and a variance function defining the dispersion of the data. Because the probability mass function is never specified in a Generalized Estimating Equations model, Generalized Estimating Equations models cannot be inflated to deal with the problem of excessive zeros due to effective medications. If the excess of zeros is significant, the mean function in Generalized Estimating Equations models is not correctly specified resulting in inconsistent estimates as well as a poor fit of the data. Results from a Generalized Estimating

Equations model based on a specification of Diggle (1993) and Albert (1999) indicated that the treatments were significant in reducing the number of outcomes but these results were attenuated by the fact that the change from baseline to endpoint was not significant. In addition, the time shape of the efficacy was not determined by Generalized Estimating Equations models (as estimated by the GenMod procedure, SAS Institute) and the model failed to predict the proportion of subjects cured.

In contrast, by specifying a Multivariate Negative Binomial distribution for all the efficacy variables, the proposed methodology allows to treat for excessive zeros in the multivariate case throughout the development of Multivariate Negative Binomial Zero-Inflated distributions.

The Multivariate Negative Binomial model is especially appropriate for longitudinal count data in which there is a small proportion of zeros. For example, this is the case for clinical trials with stage IV cancer subjects where a few percentage of patients go into remission (e.g., 0 tumors). The Multivariate Negative Binomial model incorporates the longitudinal aspect of the discrete data and the presence of covariates and unobserved random effects. It accounts for correlation among efficacy variables and treats the baseline efficacy variable as a random variable. By identifying trend and treatment effects, Multivariate Negative Binomial models allows us to evaluate the shape of the efficacy of the treatments over time rather than relying on a single endpoint measure. However, this model assumes that only one process generates the data which may not be true especially if the treatments are very effective in reducing the events.

In our data set, an excess of zeros was found to be significant and the Quadrivariate Negative Binomial Zero-Inflated specification defined a more refined mechanism to assess how and whether treatments are effective than the non-inflated model. In addition, to correctly predict the mean number of counts, the Multivariate Zero-Inflated model was able to forecast the number of zeros, offering a good fit of the data. In accordance with previous empirical work in the univariate case, the Zero-Inflated multivariate model for count data is preferred to its parental model in the presence of excessive zeros due to effective treatments.

The main implication of this chapter is to have demonstrated that great care should be taken when analyzing drug treatment effects in longitudinal count clinical trials characterized by very effective treatments. Health economists should be aware of the limitations associated with standard procedures to analyze longitudinal clinical trial count data. If treatments are very effective, a large proportion of zeros will be present in the data and ignoring this feature will result in biased estimates. As a consequence, the economic evaluation of this trial will be erroneous. Failing to detect treatment differences comes at a high price from a societal perspective as indicated by our example in section 4.1.3. This corresponds to the scenario where a drug marketed at a premium versus standard treatment would not be accepted for reimbursement because the statistical analysis wasn't able to prove its superior clinical profile versus usual care. The financial implications of this scenario for the developer of the drug would be catastrophic.

The methodology presented in this chapter has several practical implications. Lon-

gitudinal count data models should require a smaller sample size to detect a treatment difference than univariate models using the baseline and the endpoint value. In our example, the univariate model didn't detect any differences between treatments A and B. This is an interesting area for future research because the sample size in clinical trials is traditionally determined based on mean endpoint differences using two-pair side tests. Enrolling fewer subjects in clinical trials will represent considerable savings to pharmaceutical companies and overall less subjects at risk. In addition, being able to evaluate the efficacy of the treatments at each of the time periods is an important feature for two reasons. Economic results will be more precise if the reported efficacy is time-related rather than represented by a single efficacy endpoint. This will also help the physician to schedule a follow-up visit to assess his patient when the efficacy is at its maximum.

The findings and implications of this research support the development of international or national guidelines for the conduct analysis of clinical trial data. Health economists and econometricians should be involved in this process due to their growing involvement in health care policies.

5 CONCLUSIONS

The main contribution of this thesis is to have identified and addressed the issues associated with an excess of zeros in longitudinal count data. A review of the literature has outlined the need for general models to analyze longitudinal count data characterized by a high proportion of zeros. Because the distribution is not specified in non-parametric models, non-parametric models cannot be inflated or hurdled to test and treat an excess of zeros. Among parametric models, the Multivariate Negative Binomial distribution is an ideal candidate to be inflated due to its simple form.

New methods to inflate or hurdle Multivariate Negative Binomials in the longitudinal framework have been proposed using a conditional approach which allows to model correlation over time of the dependent count variable, unobserved heterogeneity in the data and an excess of zeros. Since the estimation is based on a likelihood approach, it is possible to compare different models if they are nested. This is the case with the Multivariate Negative Binomial Zero-Inflated and the Multivariate Negative Binomial Hurdle models which are each nested to the Multivariate Negative Binomial model. Two applications in health economics suggest that Multivariate Negative Binomial Zero-Inflated or Hurdle models are preferred against traditional parametric or non-parametric models.

The analysis of the number of physician visits generated by a panel of German households provided new evidence in this type of analysis. The first important result is that conclusions based on models applying a Univariate Negative Binomial distribution to the pooled data could be misleading because the longitudinal aspect of the

data is ignored. Pooling the data was done in the preliminary work of Winkelman (2001) in the analysis of doctor visits and by Geil et al. (1997) in modelling the number of hospital visits in Germany. Contrary to Geil et al.'s results (1997) in their analysis of the number of hospital visits, univariate and longitudinal count data models have different interpretations because the number of doctor visits is correlated over time which is not the case for the number of hospital visits. Not taking into account this correlation by using cross-section data or by pooling the data may lead to erroneous conclusions as shown in our example by the different interpretations of univariate and longitudinal models for count data when the analysis is conducted by gender.

However, standard multivariate distributions for longitudinal count data are generally unable to predict correctly the number of zeros at each time period which is a weakness of the generalization of these models in the presence of an excess of zeros. Instead, the method of analysis should take into account the problem of excessive zeros in longitudinal count data while addressing the issues of unobserved heterogeneity and correlation over time due to the repeated nature of the data. A Multivariate Negative Binomial Hurdle model was developed and preferred over the Quadrivariate Negative Binomial model based on specification tests. This result has important practical implications in measuring the impact of an insurance scheme reform, for example, since the longitudinal Hurdle and Non-Hurdle approaches yielded opposite findings with respect to the impact of having private insurance on the number of visits. to General Practitioners. These results confirmed in the longitudinal

context that a different view of the demand for medical care should be considered as in Pohlmeier and Ulrich (1995). It has also been shown that men and women differ in their behaviors to consult physician as a suggested by Geil et al. (1997) in the analysis of the determinants for hospitalization. Consequently, models which ignore this feature and report that gender is a significant variable miscapture the information as shown by our results. Instead, a separate analysis by gender is preferable and the demand for general practitioners and specialists should also be modelled separately as demonstrated in our analyses.

The Multivariate Negative Binomial Zero-Inflated model presented in Chapter 4 to analyze longitudinal clinical trial count data offers several advantages over the current models use to analyze this type of data. The data was characterized by an excess of zeros making incorrect the estimation of the data by any model based on the linear exponential family distribution such as Pseudo Maximum Likelihood or Generalized Estimating Equations models. Instead, a Multivariate Zero-Inflated model allows the zeros to be generated by two different processes at each time period. The results have also shown that important differences exist when the baseline value is treated as exogenous or endogenous. The problem of the treatment of the initial value is also solved by the Quadrivariate Negative Binomial Zero-Inflated and Quadrivariate Negative Binomial Hurdle models because a distribution is assigned to the initial value. Other methodological contributions of this chapter are to have presented a comprehensible way of dealing with trends and treatment effect that differ over time. This last issue is very important for two reasons. From a practitioner point of view,

knowing the shape of the treatment effect over time allows the doctor to know when to assess the treatment. Its also allows the economist to design a most accurate cost effectiveness model. Assuming in a decision tree analysis that the efficacy of a specific treatment is constant over time is a strong assumption. As economic evaluations of drug treatments are now mandatory in various health care settings, economists should be careful in using efficacy data derived from Generalized Estimating Equations estimates or univariate models. Special issues arise when the health outcome is a count observed at several point of times as reported in Chapter 4.

This research has identified several areas of research in the analysis of longitudinal count data. The treatment of the initial value in longitudinal count data is an important area of future research. The Quadrivariate Negative Binomial Hurdle and Quadrivariate Negative Binomial Zero-Inflated models are conditional models and treat the baseline value as an endogenous variable by assigning a distribution to the initial value. Another approach to treat the correlation is the Generalized Auto-Regressive model developed by Cameron and Trivedi (1998). It was argued in section 2.4.5 that the treatment of the initial value was not clear in Auto-Regressive 1 count data models when a limited number of repeated measurements were observed for the dependent count variable.

In order to appreciate this statement, the data was analyzed based on the following formulation of the mean function: $E(y_{it}) = \exp(x_{it}\beta + \rho y_{it-1})$. While Cameron and Trivedi (1998) denounced that this specification was explosive for $\rho > 0$, there is no reason why a-priori this model should not converge for values of $|\rho| < 1$. In Cameron

and Trivedi's model, the initial value (here t=1984) is treated as exogenous. If the conditional density is Univariate Negative Binomial distributed as UNB($y_t \mid x_t, y_{t-1}$), estimation can be done by maximizing the following likelihood function given by:

$$l(\alpha, \beta_t) = \sum_{t=1}^{3} UNB(y_t \mid x_t, y_{t-1}) \text{ with t=1985, 1986 and 1987.} \qquad (118)$$

This model is referred to an Auto-Regressive (1) Univariate Negative Binomial model (Cameron and Trivedi, 1998). Likelihood ratio tests indicated that estimations should be conducted by gender. The estimations of the this model are given in Table A2 in Appendix. According to the results, men and women differ with respect to having children, being married and income. Being married significantly increases the number of GP visits for men but marital status is not significant for women. For them, having children is associated with a significant increase in the number of visits which is not true for men. All the other variables are significant with the expected signs and are similar for both genders. The coefficients of correlation are positive and significant in all cases confirming that the correlation present in the data has to be treated.

However, results based on this model may be erroneous if the model supports an excess of zeros as discussed previously. In order to develop an Auto-Regressive (1) Univariate Negative Binomial Hurdle model, each Univariate Negative Binomial density function of equation (119) was hurdled. In each stage, the count observed at the precedent period is introduced in the mean function. According to Likelihood Ratio Tests, the excess of zeros is significant. Results are presented in Table A3 of

the Appendix and indicated again that men and women differ in terms of having children, being married and income.

If we compare the Auto-Regressive (1) Univariate Negative Binomial Hurdle and the conditional Quadrivariate Negative Binomial Hurdle models, the results indicate different interpretations for several variables and at different stages of the process (contact and frequency). For women, private insurance is associated with a significant reduction in the probability of contacting a doctor in the Auto-Regressive Hurdle model which was not observed in the Quadrivariate Negative Binomial Hurdle model. It is the contrary for education which is significant in the Quadrivariate Negative Binomial Hurdle model for the contact decision but not in the Auto-Regressive(1) Hurdle model. Working is positively associated with an increase in the number of GP visits for women in both specifications. While this is also true for men according to the Auto-Regressive Hurdle model, the working status was not significant according to the Quadrivariate Negative Binomial Hurdle model. It is interesting to note that the two models outline the non importance of income which is only significant in explaining the frequency of men in the Auto-Regressive (1) model. Finally, in both cases, health and chronic conditions are important variables in explaining the demand for doctor visits.

In light of all these results, correlation must be taken into account by either a conditional or an auto-regressive approach. The Quadrivariate Negative Binomial Hurdle and Quadrivariate Negative Binomial Zero-Inflated models take a conditional approach in which a distribution is assigned to the initial count value consistently with

the definition of a multivariate distribution. It is not currently known how to treat the initial value in Generalized Auto-Regressive models when T is small and a trend effect is present. Treating the initial value as exogenous may be problematic since the initial count value may contain a lot of information for analysis. This represents a promising area for future research. In the same avenue of research, the coefficient of correlation of the multivariate Hurdle or Zero-Inflated models are not known at the present time. Attempts to derive them were not successful. This is left for future research.

The review of the literature has also revealed a need for longitudinal count data tests for non-nested longitudinal count data models allowing to test for an excess of zeros. The development of longitudinal non-parametric Hurdle models is a promising avenue for future research if tests are available to compare them with their parental models to test for an excess of zeros. This would also allow us to extend Winkelman's Probit model (2001) to the longitudinal case.

Nonetheless, this doctoral thesis has contributed to the advancement of the analysis of longitudinal count data in the presence of correlation and an excess of zeros by offering simple solutions to important issues generally ignored in the econometric literature on the analysis of longitudinal count data.

REFERENCES

Al-Osh M. A. and Alzaid A. A. (1987). First-order integer valued autoregressive (INAR(1)) process. Journal of Time Series Analysis, Vol 8, No 3, 261-275.

Anderson, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. Journal of the Royal Statistical Society, 32, 283-301.

Albert, P.S. (1999). Tutorial in Biostatistics-Longitudinal data analysis (repeated measures) in clinical trials. Statistics in Medicine, 18, 1707-1732.

Blundel R., Griffith R. and Van Reenen (1995). Dynamic count data models of technological innovation. The Economic Journal, 105 (March), 1933-344.

Cameron A. Colin and Trivedi Pravin K (1998), Regression analysis of count data, Econometric Society Monographs No 30, Cambridge University Press.

Cameron A. C. and Trivedi P. K. (1986). Econometric models based on count data: comparisons and applications of some estimators and tests. Journal of Applied Econometrics, Vol. 1, 29-53.

Cameron, C., Trivedi, P. K. and Piggott, J.(1988). A microeconomic model of the demand for health care and health insurance in Australia. Review of Economic Studies, vol. 55 (4), 85-106.

Chamberlain G. (1992a). Comment: sequential moment restriction in panel data. Journal of Business, Economics and Statistics, 10.

Chamberlain G. (1992b). Efficiency bounds for semi parametric regression. Econometrica, 60, 597-96.

Chiappori P. A., Durand F. and Geoffard P-Y. (1998). Moral hazard and the de-

mand for physician services: first lessons from a French natural experiment. European Economic Review 42 (1998) 499-511.

Cincera M. (1997). Patents, R&D, and technological spillovers at the firm level: some evident from econometric count models for panel data. Journal of Applied Econometrics, Vol 12, 265-280.

Crepon B. and Duguet E. (1997). Research and development, competition and innovation. Pseudo-maximum likelihood and simulated maximum likelihood methods applied to count data models with heterogeneity. Journal of Applied Econometrics 79, 355-378.

Crepon B. and Duguet E.(1997). Estimating the innovation function from patent numbers: GMM n count panel data. Journal of Applied Econometrics, Vol. 12, 243-263.

Dean C.B. and Balshaw R. (1997). Efficiency lost by analyzing counts rather than event times in Poisson and overdispersed Poisson regression models. Journal of the American Statistical Association, 92, 1387-1399.

Deb, P. and Trivedi, P. K. (1997). Demand for medical care by the elderly: a finite mixture approach. Journal of Applied Econometrics, Vol 12, 313-336.

Diggle P. J. Liang K. Y and Zeger S. L. (1993). Analysis of Longitudinal Data. Oxford University Press, Oxford.

Drummond M. F., O'Brien B., Stoddart G. L. and Torrance G. W. (1997) Methods for the economic evaluation of health care programs. Second Edition. Oxford Medical Publications.

Dunson D. B. and Haseman J. K. (1999). Modeling tumor onset and multiplicity using transition models with latent variables. Biometrics, 55, 965-070.

Geil P., Million A., Rotte R. and Zimmermann K. F. (1997). Economic incentives and hospitalization in Germany. Journal of Applied Econometrics, 12, 295-311.

Gerdtham, U. G. (1996). Equity in health care utilisation: furhter tests based on Hurdle models and Swedish micro data. Center for Health Economics, Stockholm School of Economics.

Greene W. H. (1998). Econometric Analysis, 4th Ed. Prentice Hall, Upper Saddle River NJ, USA.

Greene W. H. (2000). Econometric Analysis, 5th Ed. Prentice Hall, Upper Saddle River NJ, USA.

Gourrieroux C., Monfort A. and Trognon A. (1984). Pseudo maximum likelihood methods: applications to Poisson models. Econometrica 52 (3), 701-720.

Gourrieroux C., Monfort A. and Trognon A. (1984). Pseudo maximum likelihood methods: theory. Econometrica 52 (3), 681-700.

Grossman, M. (1972). The demand for health: a theoretical and empirical investigation, NBER, New York.

Gurmu S. and Trivedi P.K. Excess of zeros in count models for recreational trips. Journal of Business& Economic Statistics, 14, 469-77.

Gurmu, S. (1997). Semi-parametric estimation of Hurdle regression models with an application to medicaid utilization. Journal of Applied Econometrics, Vol. 12, 225-242.

Hansen L.P.(1982). Large sample properties of Generalized Method of Moments estimators. Econometrica, 50, 1029-1054.

Haussman J., Hall B., Griliches Z. (1984). Econometric models for count data with an application to the patents-R&D relationship. Econometrica 52 (4), 903-938.

Jaffe, A. (1986). Technological opportunity and spillovers of R&D: Evidence from firms' patents, profit and market value. American Economic Review 76 (5), 984-1001.

Johnston N., Kotz S. and Balakrishnan N. (1997). Discrete multivariate distributions. John Willey, New York.

Kanifiard F. and Gallo PP. (1995). Poisson regression analysis in clinical research. J Biopharm Stat Mar; 5(1): 115-129.

Kocherlakota S. and Kocherlakota K.(1993). Multivariate distributions. New York, Marcel Dreker.

Lahiri K. and Xing G. (2002) An econometric analysis of Veteran's health care utilization using two part-models. Department of Economics, University of Akbany.

Lambert D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, February, Vol. 34, No 1., p 1 -14.

Le N. D., Leroux B. G. and Puterman M. L. (1992). Exact likelihood evaluation in a Markov mixture model for time series of seizure counts. Biometrics 48, 317-323.

Li C-S., Lu J-C, Park J., Kim K., Brinkley P. A. and Peterson J. P. (1999). Multivariate Zero-Inflated Poisson models and their applications. Technometrics, February, Vol. 41, No 1.

Liang K. Y and Zeger S. L.(1986). Longitudinal data analysis using generalized

linear models. Biometrika, 73, 12-22.

Lindsey J. K. (1999). A review of some extensions to generalized linear models. Statistics in Medicine, 18, 2223-2236.

Lindsey J.K. and Lambert P. (1998). On the appropriateness of marginal models for repeated measurements in clinical trials. Statistics in medicine, vol. 17, 447-469.

Lopez Nicolas A. Unobserved Heterogeneity and censoring in the demand for health care. Working Paper. Research Center on Health and Economics, Department of Economics and Business, Universitat Pompeu Fabra.

Marshall A. W. and Olkin I. (1990). Multivariate distributions generated from mixtures of convolution and product families. In H.W. Block, A.R. Simpson and T.H Davis eds, Topics in Statistical Dependance, IMS Lectures Notes Monograph Series, volume 16, 371-393.

Mayberg H.S., Brannan S.K, Tekell J.L., Silva J.A., Mahurin R.K., McGinnis S, and Jerabek P.A.(2000). Regional metabolic effects of fluoxetine in major depression: serial changes and relationship to clinical response. Biol Psychiatry 2000 Oct 15;48(8):830-43.

McIntosh J. (2001). Analyzing counts, duration, and recurrences in clinical trials. Journal of Biopharmaceutical Statistics, 11 (1&2), 65-74.

Montalvo J.G. (1997). GMM estimation of count panel data models with fixed effects and predetermined instruments. Journal of Business and Economic Statistics,15, 82-89.

Mullahy J. (1986). Specification and testing of some modified count data models.

Journal of Econometrics 33, 341-365.

Mullahy J. (1997). Heterogeneity, excess zeros, and the structure of count data. Journal of Applied Econometrics, Vol. 12, 337-350.

Page M. (1995). Racial and ethnic discrimination in urban housing markets: evidence from a recent audit survey. Journal of Urban Economics, 38, 183-206.

Pakes A. and Griliches Z. (1984). Patents and R&D at the firm level: a first look. In Z. Griliches (d.), R&D, Patents and Productivity, National Bureau of Economic Research, University of Chicago Press.

Pinquet J. (1997). Experience rating through heterogeneous models. Working Paper, No 9725, THEMA, Universite Paris X Nanterre

Pohlmeier W. and Ulrich V. (1995). An econometric model of the two part decision making process in the demand for health care. The Journal of Human Resources, XXX, 2339-361.

Ruser J.W. (1991). Workers compensation and occupational injuries and illnesses. Journal of Labor Economics, 9, 325-350.

Rochon S. (2003). Age and presence of chronic conditions, education and the health system reform: impact on utilization of health care services by the Canadian elderly. Thesis submitted for the requirement of the degree of M.A. in Economics, Mc Gill University.

Santos Silva J. (1997). Endogeneity in count data models: an application to demand for health care. Journal of Applied Econometrics, Vol. 12, issue 3, 281-294.

Thall P. F. and Vail S. C. (1990). Some covariance models for longitudinal count

data with overdispersion. Biometrics 46, 657-671.

Vuong Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 57, 307-333.

Wagstaff, A. (986). The demand for health. Some new empirical evidence. Journal of Health Economics, 5, 195-233.

Windmeijer, F. A. G. and Santos Silva, J. M. C. (1997). Endogeneity in count data models: an application to demand for health care. Journal of Applied Econometrics, Vol 12, 281-194

Winkelmann, R. (2000). Econometric analysis of count data, 3rd ed. Springer: Heidelberg, New York

Winkelmann R. (2001). Health Care Reform and the number of doctor visits - An econometric analysis. CEPR, London and the Institute for the Study of Labor (IZA), Discussion paper No 317, June 2001

Winkelmann R. (1994). Count Data Models. Economic theory and application to labor mobility. Lecture Notes in Economic and Mathematical Systems, No 410, Springer-Verlag.

Winkelmann R. and Zimmermann K. F. (1991). A new approach for modelling economic count data. Economic Letters 37, 139-143.

Zeger S. L. and Liang K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes, Biometrics 42, 121-130.

Zeger S. C. (1988). A regression model for time series of counts. Biometrika, 75, 4, 621-9

Zimmermann K.F. (1993). Labor responses to taxes and benefits in Germany. In A. B. Atkinson and G. V. Morgensen (eds), Welfare and Work Incentives. A North European Perspective, Clarendon Press, Oxford, 192-240.

Appendix

Table A1: Predictions Generalized Estimating Equations Model.

The SAS System

The GENMOD Procedure

GEE Observation Statistics

Observation	y	Pred	Xbeta	Std
15	3	2.2785681	0.8235472	0.8016495
16	0	0.0090567	-4.704248	1.0538473
17	6	0.0669206	-2.704248	1.0538473
18	5	0.4944801	-0.704248	1.0538473
19	1	0.0445364	-3.111448	0.460567
20	3	0.329082	-1.111448	0.460567
21	0	2.4316056	0.8885518	0.460567
22	4	0.0076503	-4.873016	1.0684428
23	2	0.0565282	-2.873016	1.0684428
24	3	0.4176898	-0.873016	1.0684428
25	10	0.1213038	-2.109457	0.218789
26	5	0.8963205	-0.109457	0.218789
27	4	6.6229624	1.8905428	0.218789
28	1	0.0285762	-3.55518	0.4464474
29	2	0.2111513	-1.55518	0.4464474
30	1	1.5602092	0.4448199	0.4464474
31	1	0.0810129	-2.513147	0.511328
32	0	0.5986088	-0.513147	0.511328
33	1	4.4231541	1.486853	0.511328
34	1	0.0718868	-2.632662	0.4796825
35	0	0.5311759	-0.632662	0.4796825
36	1	3.9248882	1.3673379	0.4796825
37	6	0.0271929	-3.604801	1.0080783
38	2	0.2009296	-1.604801	1.0080783
39	0	1.4846798	0.3951992	1.0080783
40	18	0.0332496	-3.403714	0.463362
41	8	0.2456828	-1.403714	0.463362
42	2	1.8153644	0.5962862	0.463362
43	19	0.0474921	-3.047192	0.8815957
44	16	0.3509217	-1.047192	0.8815957
45	22	2.5929801	0.9528078	0.8815957
46	20	0.0505878	-2.984044	0.3502257
47	7	0.3737962	-0.984044	0.3502257

Appendix:

Table A2: Maximum Likelihood Estimates.
Auto-Regressive (1) Negative Binomial Model. Gender Analysis

Maximum Likelihood Estimation		AR(1)NB Female	AR(1)NB Male
Number of observations:		2,376	4342
Log of Likelihood:		-11.203.35	-10,265.310
Variable	Parameter		
Constant	β_0	-0.622*	-0.577*
		(0.121)	(0.135)
Age*10^{-1}	$\beta 1$	0.182*	0.140*
		(0.026)	(0.029)
Private Insurance	β_4	-0.431*	-0.490*
		(0.075)	(0.074)
Education	β_5	0.056*	-0.049*
		(0.007)	(0.007)
Working	β_6	0.170*	0.383*
		(0.040)	(0.062)
Children	β_7	0.110*	-0.028
		(0.043)	(0.048)
Income	β_8	-0.146	-0.438*
		(0.110)	(0.136)
Married	β_9	0.202	0.425*
		(0.048)	(0.059)
Health	β_{10}	-0.103*	-0.118*
		(0.008)	(0.009)
Chronic conditions	β_{11}	0.546*	0.808*
		(0.418)	(0.047)
Non West German	β_{12}	0.338*	0.430*
		(0.044)	(0.046)
Overdispersion	α	-0.784*	0.708*
		(0.012)	(0.011)
Auto-correlation	ρ	0.128*	0.100*
		(0.007)	(0.07)
* indicates significant at the 5 % level.			

Appendix:

Table A3: Maximum Likelihood Estimates.

Auto-Regressive (1) Negative Binomial Hurdle Model. Gender Analysis

Maximum Likelihood		AR(1)UNBH		AR(1)UNBH	
Estimation		1984 exogenous		1984 exogenus	
GPs visits		Female		Male	
Number of observations:		2376		2159	
Log of Likelihood:		-11,081.524		-10,162.339	
Variable	Parameter	1st step	2nd step	1st step	2nd step
Constant	β_0	-0.436	0.654*	-0.500*	-0.609*
		(0.232)	(0.176)	(0.228)	(0.183)
Age*10^{-1}	β_1	0.150*	0.171*	-0.131*	0.131*
		(0.054)	(0.037)	(0.050)	(0.038)
Private Insurance	β_4	-0.336*	-0.384*	-0.345*	-0.529*
		(0.107)	(0.107)	(0.109)	(0.111)
Education	β_5	-0.002	0.081	-0.008	0.095
		(0.015)	(0.009)	(0.014)	(0.011)
Working	β_6	0.148*	0.135*	0.611*	0.211*
		(0.066)	(0.053)	(0.133)	(0.081)
Children	β_7	0.028	0.147*	-0.047	-0.022
		(0.072)	(0.058)	(0.072)	(0.065)
Income*10^{-4}	β_8	-0.142	-0.131	-0.548*	-0.240
		(0.47)	(0.190)	(0.187)	(0.203)
Married	β_9	0.143	0.192*	0.499*	-0.273*
		(0.083)	(0.062)	(0.101)	(0.080)
Health	β_{10}	-0.083*	-0.113*	-0.108*	-0.123*
		(0.012)	(0.010)	(0.014)	(0.012)
Chronic conditions	β_{11}	0.543*	0.453*	0.747*	0.754*
		(0.085)	(0.052)	(0.110)	(0.064)
Non West German	β_{12}	-0.134	0.396*	-0.581*	0.571*
		(0.074)	(0.055)	(0.063)	(0.063)
Overdispersion	α	0.876*	0.810*	0.750*	0.631*
		(0.084)	(0.037)	(0.068)	(0.041)
Auto-regression	ρ	0.290*	0.085*	0.273*	0.061*
		(0.007)	(0.007)	(0.050)	(0.007)
* indicates significant at the 5 % level.					