Modeling Usage of Medical Care Services: The Medical Expenditure Panel Survey Data, 1996-2000

Michael Creel, Montserrat Farell

Modeling Usage of Medical Care Services: The Medical Expenditure Panel Survey Data, 1996-2000

Journal:	*Applied Economics*
Manuscript ID:	APE-07-0175.R1
Journal Selection:	Applied Economics
Date Submitted by the Author:	19-Nov-2008
Complete List of Authors:	Creel, Michael; Universitat Autònoma de Barcelona, Dept. of Economics and Economic History Farell, Montserrat; Universitat Autònoma de Barcelona, Dept. of Economics and Economic History
JEL Code:	C25 - Discrete Regression and Qualitative Choice Models < C2 - Econometric Methods: Single Equation Models < C - Mathematical and Quantitative Methods, I10 - General < I1 - Health < I - Health, Education, and Welfare
Keywords:	medical care , count data, maximum likelihood

Modeling Usage of Medical Care Services: The Medical Expenditure Panel Survey Data, 1996-2000

Michael Creel and Montserrat Farell*

November 2008

Abstract

We explore the determinants of usage of six different types of health care services, using the Medical Expenditure Panel Survey data, years 1996-2000. We apply a number of models for univariate count data, including semiparametric, semi-nonparametric and finite mixture models. We find that the complexity of the model that is required to fit the data well depends upon the way in which the data is pooled across sexes and over time, and upon the characteristics of the usage measure. Pooling across time and sexes is almost always favored, but when more heterogeneous data is pooled it is often the case that a more complex statistical model is required.

JEL classifications: C25, I10

Keywords medical care; count data; maximum likelihood

*Both authors are at the Department of Economics and Economic History, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona) Spain. Send email to michael.creel@uab.es. This research was supported by grants SEJ 2006-00712/ECON and CONSOLIDER-INGENIO 2010 (CSD2006-00016).

1 Introduction

The demand for health care services may often be measured as the number of times that some event, for example, a doctor visit, occurs in a given time period. Such variables, defined on the natural numbers, are referred to as count data. There have been many recent advances in the econometric analysis of count data, especially the development of flexible density functions for univariate count data. In many cases, these papers include an empirical analysis of data on demand for health care. Examples of such contributions are Deb and Trivedi (1997), who investigate finite mixture models; Cameron and Johansson (1997), who adapt the polynomial reshaping technique of Gallant and Nychka (1987) to count data, and Gurmu (1997), who uses a flexible density to model latent heterogeneity. All of these approaches define densities that allow for modeling frequently observed features of the data, such as excess zeros and overdispersion, as well as more complicated departures from the behavior implied by standard models such as the Poisson and negative binomial.

At the same time, new sources of data have become available. One of these is the Medical Expenditure Panel Survey (MEPS). The MEPS data is a rich source of recent data on demand for health care, insurance coverage, and related topics. This paper applies many of the recently developed statistical models for univariate count data to the MEPS data, years 1996-2000. This allows comparison of models using a uniform, high quality data set. From this we will be able to determine which models are most successful in capturing the features of six different measures of demand for health care services[1]. Since the six measures exhibit substantially different character-

[1]These are office-based doctor visits (OBDV), outpatient doctor visits (OPV), inpatient visits (IPV), emergency room visits (ERV), dentist visits (DV), and number of prescription drugs taken (RX), all measures on an annual basis.

istics, they form an interesting test bed for the statistical models, at least within the general area of demand for health care. We seek to learn which of the available models seem most useful for analysis of data similar to the usage measures in the MEPS data. We also provide information about the usage measures in the MEPS data, upon which further research can build. Guo and Trivedi (2002) provide a similar, though somewhat less extensive comparison of models, using two data sets on counts of patents. Within the literature on demand for health care we are not aware of any papers that provide a similar comparison of models. Beyond the comparison of the statistical models, we also investigate the stability of parameters over time and across sexes, and we present brief estimation results for the most favored models.

Our criterion for comparison is the consistent Akaike information criterion (CAIC). The CAIC is a penalized goodness of fit criterion that is decreasing in the value of the likelihood function and increasing in the number of parameters. Lower CAIC values correspond to models that offer a good fit to the data without using an excessive number of parameters. While one might be interested in other criteria such as out-of-sample fit, or marginal mean effects, we limit our attention to the CAIC. This is for two reasons. The first is that we believe that the CAIC provides more information, in a paper of limited length, than does any other criterion. Inclusion of other information such as out-of-sample fits would multiply the already large amount of tabular information contained in this paper. The likelihood criterion is probably the most parsimonious summary information about how well a model fits. A model that dominates another in terms of likelihood must on average fit the observed counts (0,1, etc.) better than the other model, in spite of the fact that the second model may better fit a par-

ticular count, such as the observed number of zeros. If too many extra parameters are needed to improve fits to certain counts, the CAIC penalizes the likelihood value to control for this, thus avoiding overparameterized models. Features of interest such as conditional moments and their derivatives are functions of the estimated probabilities of counts. The CAIC picks the model that best assigns probabilities to the counts observed in the data, in a certain sense. The second reason that we focus on the CAIC and not on features such as fitted marginal effects is that we do not know the true marginal effects, so comparisons across models are difficult to evaluate. If we were doing Monte Carlo work, it would certainly be of interest to look at criteria of this sort. However, with real sample data as is used here, we believe that the fitted marginal effects of a model that is strongly dominated by another in terms of the CAIC are simply not of much interest, since the model is almost certainly very poorly specified. When a model is poorly specified, the fitted marginal effects will in general be biased and inconsistent. Our focus on the CAIC is admittedly narrow, but it is an attempt to provided as much useful information as is possible in a limited number of pages. Future work could explore other features of interest in a Monte Carlo context, focusing on the models that this paper identifies as the most promising according to the penalized goodness of fit criterion.

To summarize the main results, we find that some of the newer models are useful additions to the toolbox for analysis of health care usage, but others are almost always dominated. The complexity of the model that is favored depends upon the type of data that is analyzed. For variables that have relatively high means, significant overdispersion, and relatively few zeros, relatively complex models are needed to fit the data well. For other variables such as the number of inpatient hospitalizations, the simple neg-

ative binomial density is perfectly adequate. Another result is that pooling data across time and sexes leads to a parsimonious model that still fits the data as well as separate models that allow all parameters to vary. Pooling should be done when the data allow it. When more heterogeneous data is pooled, it is more likely that a relatively complex statistical model will be required. With relatively homogeneous data, the simple negative binomial statistical model often fits well.

2 Statistical models for count data dependent variables

Data on health care demand often exhibits overdispersion, which means that the ratio of the conditional variance to the conditional mean is greater than one (Cameron and Trivedi, 1986; Pohlmeier and Uhlrich, 1995). Another common characteristic is that many zeros are observed, possibly more than can be accounted for by simple count densities (Pohlmeier and Uhlrich, 1995; Gerdtham, 1997). Factors such as latent variables or latent population groups could induce more marked departures from standard densities, leading to bimodality or especially fat right tails, for example. In this section we briefly survey some of the newer univariate count data models that can allow for such departures. Before surveying the recent models, we briefly discuss the more standard models upon which the newer approaches build.

Poisson (POISSON)

The Poisson density for a count random variable Y is

$$f_Y(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}.$$

To allow for covariates, λ is usually parameterized as $\lambda = e^{\mathbf{x}\beta}$. The Poisson density implies that the conditional mean and the conditional variance of y are both equal to λ. Since data on health care demand usually exhibit overdispersion and possibly excess zeros, the basic Poisson model will usually not be suitable for analyzing demand for health care.

Negative binomial (NB)

If the Poisson mean contains a latent component, marginalization, under some assumptions, will lead to a negative binomial density (see for example Cameron and Trivedi, 1998, pp. 100-102) . The negative binomial density may be written as

$$f_Y(y|\phi) = \frac{\Gamma(y+\psi)}{\Gamma(y+1)\Gamma(\psi)} \left(\frac{\psi}{\psi+\lambda}\right)^{\psi} \left(\frac{\lambda}{\psi+\lambda}\right)^{y} \tag{1}$$

where $\phi = \{\lambda, \psi\}$, $\lambda > 0$ and $\psi > 0$.[2] When $\psi = \lambda/\alpha$ we have the negative binomial-I model (NB-I), and $\psi = 1/\alpha$ gives the negative binomial-II (NB-II) model. Though other versions exist, we limit attention to these in this paper. The moment generating function of the NB density, which is needed below, is

$$M_Y(t) = \psi^{\psi} \left(\lambda - e^t\lambda + \psi\right)^{-\psi}. \tag{2}$$

For the NB-I density, $V(Y) = \lambda + \alpha\lambda$. In the case of the NB-II model, we have $V(Y) = \lambda + \alpha\lambda^2$. For both forms, $E(Y) = \lambda$. Thus, both forms capture overdispersion, with the NB-II model allowing for a more extreme form. As with the Poisson models, the usual means of incorporating conditioning variables is the parameterization $\lambda = e^{\mathbf{x}\beta}$. When this is done, the previous

[2] Among the numerous examples of application of the NB model to health care demand are Cameron *et al.* (1988), Pohlmeier and Ulrich (1995), Geil *et al.* (1997) and Dismuke and Guimares (2002).

formulae for moment will give the conditional moments.

Hurdle negative binomial (HNB)

As noted by Pohlmeier and Uhlrich (1995) and Gerdtham (1997), health care demand may exhibit excess zeros with respect to what a NB model can accommodate. This leads us to consider the hurdle version of the NB model. The hurdle[3] negative binomial model first models the zero *vs.* positive outcome using a probit or similar model. Then, conditional on positive visits, the count follows a zero-truncated negative binomial density. Different parameter vectors are associated with the binary and truncated densities. Hurdle count models were introduced by Cragg (1971) and Mullahy (1986), who also presented "with-zeros" (also known as "zero-inflated") models. Here we present only the hurdle model, since it seems to have been used more widely than the "with-zeros" model for analysis of data on on usage of health care services.[4] We follow Deb and Trivedi (1997), who use a NB model to parameterize the Bernoulli trial. For a NB random variable,

$$\begin{aligned} \Pr(Y=0) &= f_Y(0,\phi_h) = \left(\frac{\psi_h}{\psi_h+\lambda_h}\right)^{\psi_h} \\ \Pr(Y>0) &= 1-\Pr(Y=0), \end{aligned}$$

where the parameter of the hurdle process is $\phi_h = \{\lambda_h, \psi_h\}$. To achieve identification one can set $\alpha_h = 1$ when parameterizing ψ_h, which may be done as with the NB model. The above probabilities are used to estimate the binary $0/1$ hurdle process. Then, for the observations where visits are

[3] Hurdle models are also known as "two-part" models.

[4] Examples of applications of the HNB model to health care demand include Pohlmeier and Ulrich (1995), Gerdtham (1997), Deb and Trivedi (1997) and Yoshida and Kim (2008).

positive, a truncated NB density, with a different parameter $\phi = \{\lambda, \psi\}$ is estimated. This density is

$$\begin{aligned} f_Y(y, \phi | y > 0) &= \frac{f_Y(y, \phi)}{1 - \left(\frac{\psi}{\psi+\lambda}\right)^{\psi}} \\ &= \frac{\Gamma(y+\psi)}{\Gamma(y+1)\Gamma(\psi)} \left[\left(\frac{\psi}{\psi+\lambda}\right)^{\psi} - 1\right]^{-1} \left(\frac{\lambda}{\psi+\lambda}\right)^{y} \end{aligned}$$

Since the hurdle and truncated components of the overall density for Y share no parameters, they may be estimated separately, which is computationally less burdensome than estimating the overall model. The expectation of Y is

$$E(Y) = \left[1 - \left(\frac{\psi_h}{\psi_h + \lambda_h}\right)^{\psi_h}\right] \left[1 - \left(\frac{\psi}{\psi+\lambda}\right)^{\psi}\right]^{-1} \lambda.$$

NB-I and NB-II versions that allow for conditioning variables follow from the appropriate parameterizations of ψ_h, ψ, λ_h and λ.

The HNB model could possibly be considered the most sophisticated attempt to deal with the issues of excess zeros and overdispersion in the modeling of health care demand count data, up until 1996. Shortly after, the following models were introduced. All of these models can account for excess zeros and overdispersion, so they can deal with the issues the HNB model was designed to address. Some of the models are also more flexible than the HNB model, even though they may be more parsimonious.

2.1 A semiparametric approach (PSP, HPSP)

A semiparametric approach to modeling count data has been developed by Gurmu and Trivedi (1996), Gurmu (1997) and Gurmu *et al.* (1999). This approach introduces unobserved heterogeneity in a Poisson model, and al-

lows the unobserved heterogeneity to follow a semi-nonparametric density. This is conceptually similar to the way that a negative binomial model is obtained as a Poisson-gamma mixture density, but is more flexible in that the latent variable is not restricted to follow a one parameter gamma density. The semi-nonparametric density of the latent variable is closely related to that proposed by Gallant and Nychka (1987). The difference is that Laguerre polynomials are used instead of Hermite polynomials. Gurmu *et al.* (1999) show that, under weak assumptions, the Laguerre expansion density can consistently estimate densities of unknown form. As such, the mixture density is semiparametric, since the Poisson specification is parametric but the modelization of the heterogeneity is not.

Gurmu and Trivedi (1996) found that the basic semiparametric approach of Gurmu *et al.* (1999)[5] did not fit data well - specifically, excess zeros were a problem. To overcome this problem, Gurmu (1997) proposed a hurdle version of the semiparametric model.

The original semiparametric model is based upon an infinite mixture of a Poisson random variable and an independent random variable V which captures unobserved heterogeneity. The assumption is that the Poisson mean is random, so that $E(Y|V = v) = \lambda v$. Integrating out the heterogeneity, one obtains the marginal density:

$$f_Y(y, \lambda, \phi) = \int \frac{e^{-\lambda v}(\lambda v)^y}{y!} g_V(v, \phi) dv \quad (3)$$

$$= \frac{\lambda^y}{y!} M_V^y(-\lambda) \quad (4)$$

where $M_V^y(-\lambda)$ is the yth order derivative of the moment generating function of V, evaluated at $-\lambda$. $M_V^0(-\lambda) = M_V(-\lambda)$, is the moment generating

[5]The 1999 paper is based upon a 1996 working paper, which explains the dates of these references.

function itself.

To model the density $g_V(v,\phi)$ flexibly, Gurmu *et al.* (1999) use a normalized Laguerre polynomial expansion around a gamma baseline density. The gamma baseline density is

$$f(v,\phi) = \left(\frac{v^{\alpha-1}\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta v} \right)$$

where $\phi = (\alpha, \beta)$. The semi-nonparametric density for v is

$$g_V(v|\phi,\gamma) = \frac{\left[h_p\left(y,\gamma\right)\right]^2 f(v|\phi)}{\eta_p(\phi,\gamma)}$$

where

$$h_p\left(y,\gamma\right) = \sum_{k=0}^{p} \gamma_k P_k(v), \tag{5}$$

$\gamma = (1, \gamma_1, \gamma_2, ..., \gamma_p)$, and $P_k(v)$ is the kth order Laguerre polynomial. The term $\eta_p(\phi,\gamma) = \gamma'\gamma$ is the normalization factor that makes the density sum to one. The restriction $\gamma_0 = 1$ is used to achieve identification, since the density is homogeneous in γ. This density is semi-nonparametric in the sense that, under weak assumptions, there exist (ϕ,γ) such that a density of unknown form can be approximated arbitrarily well as p goes to infinity. Gurmu *et al.* (1999) provide the consistency proof, which is similar to that of Gallant and Nychka (1987).

Next, they are able to obtain a closed form for $M_V^y(-\lambda)$, which upon substitution into equation 4 yields the semiparametric density for the count random variable Y. In estimation, a restriction is imposed such that $E(V) = 1$, which leads to $E(Y) = \lambda$. In the course of the empirical work reported below, we have found that the model is poorly identified without this restriction, and that it is very difficult to obtain convergence if it is not im-

posed. The results we report always impose the restriction. We will refer to this model as the Poisson semiparametric model (PSP). To incorporate conditioning variables, the Poisson-style parameterization $\lambda = e^{\mathbf{x}\beta}$ is used, so that $E(Y|\mathbf{x}) = \exp(\mathbf{x}\beta)$.

To extend this to the hurdle case, Gurmu (1997) allows a first PSP model to determine whether the zero/positive hurdle is crossed, and a second PSP model is used to model the positives. For the hurdle crossing process, the relevant probabilities are

$$\begin{aligned} \Pr(Y = 0) &= M_V(-\lambda_h) \\ \Pr(Y > 0) &= 1 - \Pr(Y = 0). \end{aligned}$$

The truncated version of the PSP density is

$$f_Y(y|y > 0, \lambda, \phi) = \frac{\frac{\lambda^y}{y!} M_V^y(-\lambda)}{1 - M_V(-\lambda)}.$$

Just as in the case of the HNB model, the binary and truncated components of the hurdle Poisson semiparametric (HPSP) model may be estimated separately. Notationally, we will let PSP(k) or HPSP(k) refer to a model that uses a k-order expansion.

2.2 Semi-nonparametric approaches (PSNP, NBSNP)

Cameron and Johansson (1997) directly adapt Gallant and Nychka's (1987) semi-nonparametric density to the count data case. They reshape a Poisson baseline density using a squared polynomial, and then normalize the result to sum to one. We shall refer to this as the Poisson semi-nonparametric (PSNP) approach, though there has been no formal proof of the condi-

tions under which the density has nonparametric properties.[6] The PSP and HPSP models embed the semi-nonparametric density in a parametric density to obtain a semiparametric model, after marginalization of the latent variable. As such, one expects that the approach of Cameron and Johansson should be able to capture more extreme departures from the baseline model, though perhaps at the cost of needing to estimate many parameters. For example, the PSNP model can accommodate bimodal densities, while the PSP density cannot.

The PSNP density is

$$f_Y(y|\lambda,\gamma) = \frac{\left[h_p\left(y|\gamma\right)\right]^2}{\eta_p(\phi,\gamma)}\frac{e^{-\lambda}\lambda^y}{y!}, \tag{6}$$

where

$$h_p\left(y|\gamma\right) = \sum_{k=0}^{p}\gamma_k y^k, \tag{7}$$

and $\eta_p(\phi,\gamma)$ is a normalizing factor to make the density sum to one. The normalizing factor is

$$\eta_p(\lambda,\gamma) = \sum_{y=0}^{\infty}\left[h_p\left(y|\gamma\right)\right]^2\frac{e^{-\lambda}\lambda^y}{y!}.$$

Cameron and Johansson show that this has the closed form

$$\eta_p(\lambda,\gamma) = \sum_{k=0}^{p}\sum_{l=0}^{p}\gamma_k\gamma_l m_{k+l} \tag{8}$$

where $m_r(\lambda)$ is the rth noncentral moment of the Poisson density. Because

[6]The consistency proofs of Gallant and Nychka (1987) and Gurmu *et al.* (1999) are for continuous random variables. While it seems reasonable to expect that the proofs could be adapted to discrete random variables, this has not yet been done, to our knowledge.

the term

$$\frac{\left[h_p\left(y|\gamma\right)\right]^2}{\eta_p(\lambda,\gamma)}$$

that reshapes the baseline density in equation 6 is a homogeneous function of γ, it is necessary to impose a normalization to achieve identification: γ_0 is set to 1. The moments of Y may be calculated using the closed form expression in Cameron and Johansson's equation 4. The typical Poisson-style parameterization of the mean is used to incorporate conditioning variables.

Since the NB model usually fits health care data dramatically better than does the Poisson model, using only one more parameter, one might suspect that changing the baseline model to the NB might allow the model to fit well using fewer parameters. What we shall refer to as the negative binomial semi-nonparametric (NBSNP) model is obtained by making this change. The density is

$$f_Y(y|\phi,\gamma) = \frac{\left[h_p\left(y|\gamma\right)\right]^2}{\eta_p(\phi,\gamma)} \frac{\Gamma(y+\psi)}{\Gamma(y+1)\Gamma(\psi)} \left(\frac{\psi}{\psi+\lambda}\right)^{\psi} \left(\frac{\lambda}{\psi+\lambda}\right)^{y},$$

where $h_p\left(y|\gamma\right)$ and $\eta_p(\phi,\gamma)$ are defined as in equations 7 and 8, respectively, and the raw moments $m_r(\lambda,\psi)$ are obtained from equation 2. The moments of Y are again obtained from Cameron and Johannson's equation 4, substituting the NB raw moments for those of the Poisson density.[7] The model can use either the NB-I or the NB-II as the baseline model. We investigate both versions in what follows. Notationally, let NBSNP-I(3), for example, indicate the NBSNP model using a NB-I baseline density, and a 3rd order polynomial expansion. To our knowledge, this is the first paper that applies this model to data on health care demand. Guo and Trivedi (2002) apply a version of this model based upon an NB-II base density to

[7] We used MuPAD version 2.51 to perform these calculations.

patent data.

2.3 A finite mixture approach (MNB, CMNB)

The finite mixture approach to fitting data on health care demand was introduced by Deb and Trivedi (1997). The mixture approach can be interpreted as allowing for latent groups in the population. The data for each group may be characterized by a parameter vector. Since the group to which an individual belongs is not observed, a mixing probability is used to classify individuals probabilistically. There may be two or more latent groups. The mixture approach has been also applied by Gerdtham and Trivedi (2000), who find that it performs better than the HNB approach.

The mixture negative binomial (MNB) model has the virtue of being conceptually simple. The density is

$$f_Y(y, \phi_1, ..., \phi_p, \pi_1, ..., \pi_{p-1}) = \sum_{i=1}^{p-1} \pi_i f_Y^{(i)}(y, \phi_i) + \pi_p f_Y^p(y, \phi_p),$$

where $\pi_i > 0, i = 1, 2, ..., p$, $\pi_p = 1 - \sum_{i=1}^{p-1} \pi_i$, and $\sum_{i=1}^{p} \pi_i = 1$. The $f_Y^{(i)}(y, \phi_i)$, $\phi_i = \{\lambda_i, \psi_i\}$ are p separate NB-I or NB-II densities, as in equation 1. Identification requires that the π_i be ordered in some way. We follow Deb and Trivedi (1997) by imposing $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_p$ and $\phi_i \neq \phi_j, i \neq j$. This is simple to accomplish post-estimation by rearrangement of the component densities. Another issue is how to consistently estimate the number of component densities, supposing that the true density is in fact a mixture density (see James, *et al.* 2001, for example). We skirt this issue by considering only the possibility of 2 component densities.

The properties of the mixture density follow in a straightforward way from those of the components. In particular, the moment generating func-

tion is the same mixture of the moment generating functions of the component densities, whence $E(Y) = \sum_{i=1}^{p} \pi_i \lambda_i$.

The MNB density may suffer from overparameterization, since the total number of parameters grows rapidly with the number of component densities. To address this problem, Deb and Trivedi propose a constrained mixture negative binomial model (CMNB) which restricts all the slope parameters in $\lambda_j = e^{x\beta_j}$ to be the same across all component densities. The constants and the overdispersion parameters α_j are allowed to differ.

3 The MEPS data

3.1 Data Sources

The Medical Expenditure Panel Survey composed of four surveys of individuals, nursing homes, health care providers, and employers in the United States. We use only the Household Component, which is a survey of a nationally representative sample of households. The Household Component uses an overlapping panel design where individuals are interviewed five times over the course of 2.5 years, such that complete data for two calendar years is collected. Each year, a new series of contacts is initiated. Thus, data for a given individual will appear in the data files for two consecutive years, and the samples for consecutive years are not independent. The raw MEPS data files are available at the site `http://www.meps.ahrq.gov`. The data files used are the Household Component Full-Year files for years 1996-2000, which are files HC-012, HC-020, HC-028, HC-038 and HC-050, respectively. These data files collect responses to many questions related to health care usage, health, insurance coverage, income, *etc.*

3.2 Variables

From these files we use six different measures of annual health care usage, for each of the five years. These are office-based doctor visits (OBDV), outpatient doctor visits (OPV), emergency room visits (ERV), inpatient hospital visits (IPV), dental visits (DV), and number of prescription drugs taken (RX).

The explanatory variables used are months of public insurance coverage during the year, divided by 12 (PUB), sex (SEX - coded as 0 for men and 1 for women), age (AGE), years of schooling (EDUC), and family income in thousands of dollars (INC). Since health care issues change considerably with age, we limit the sample to individuals between the ages of 40 and 65, inclusive. Work not reported here revealed that models that pool data for broader age groups often do not pass specification tests. Also, extremely few younger people have publicly provided insurance coverage. This fact causes problems in obtaining convergence of models that use data limited to that for younger people. We also suspect that women's and men's health issues are different enough to warrant the consideration of models that allow the form of the model and all parameters to differ by sex. We investigate the possibility of pooling the form of the model or some of the parameters across sexes.

We limit the sample to people who have private health insurance coverage during the entire year. Originally we used months of private coverage as an explanatory variable. This variable is very likely to be endogenous in a model for health care usage, since latent health status will likely simultaneously affect choices regarding health care usage and purchase of health care insurance.[8] The econometric problem is to find convincing in-

[8]Exploratory work with Hausman-type tests suggested that endogeneity of private in-

strumental variables for private coverage, that can reasonably be excluded from the equation that explains health care usage. Since we were not able to find such variables in the survey data, we prefer to simply estimate models conditional on full private insurance coverage, and avoid the issue of possible endogeneity entirely. The analysis is more limited, but the results are sharper and more reliable. Depending upon the year and the value of SEX, we loose between 20% and 35% of the available sample due to this decision. We include the measure of publicly provided insurance, PUB, to investigate the effects of double coverage. We believe that PUB and the other explanatory variables may be safely considered as exogenous, *a priori*.

All the variables with the exception of PUB and INC are directly available from the survey data. PUB is simply the sum of the monthly indicators of public health care coverage, divided by 12. Thus, it runs from 0 to 1, with 1 indicating that a person enjoys publicly provided insurance coverage during the full year. INC was constructed by summing the incomes of all members of the family. In the MEPS data, total personal income is the sum of many different sources of income, which may or may not be directly reported. Observations for which any source of any family member's income was "hot decked" were dropped, since hot decking introduces measurement error which leads to inconsistent estimation in the context of regression analysis.[9,10]

We do not use any information on health status, and instead treat health

surance coverage is in fact a problem, especially for the OBDV and RX use variables. These results are not entirely reliable, however, due to the problem of poor instruments, discussed in the body of the paper, and thus we do not present them in detail.

[9] "Hot decking" is a term used in the MEPS documentation to describe a method of replacing missing data with conditional or unconditional sample means of the variable. See the documentation at `www.meps.ahrq/Pubdoc/H12DOC.PDF` for more details.

[10] The programs used to process the raw data, as well as the resulting data files are available upon request from the authors.

status as entirely latent. This is in contrast to many studies that have incorporated objective and/or subjective indicators of health status. The health status information in the MEPS data include measures of perceived health status as well as objective measures of limitations to activities. The recorded data is based upon one family member's assessment of all family members' healths. We have the problem that individual A and individual B may evaluate individual B's health very differently, which at a minimum implies that this data will be subject to measurement errors which can lead to inconsistency if not properly addressed (Windmeijer and Santos Silva, 1997). In the case of limitations to activities, many variables are recorded in the data sets. These include, for example, indicators of whether or not individuals have difficulty standing 20 minutes, or difficulty in reaching over the head, and a number of similar variables, and again, one family member reports for the entire family. These variables are likely to be highly collinear, and none of them seems suitable as a single measure of overall health status. Furthermore, it is not clear that results that are conditional on such measures of health status are directly useful for many sorts of economic analysis. Since an economic analysis would likely need to marginalize results that are conditional on these variables, and since the only means of marginalizing them is using the sample information itself, we prefer simply not to condition on them from the outset. Thus, we treat health status as a purely latent source of heterogeneity, and we model it as such. The primary concern in treating health status in this way is the possibility that latent health status might be correlated with conditioning variables such as private insurance coverage, which would induce problems of endogeneity. Our solution, as noted above, is to condition on full private insurance coverage, so that its level disappears as a regressor. We think that the other

conditioning variables may safely be assumed to be exogenous.

3.3 Descriptive Statistics

As noted above, we limit the data used in this paper to people between 40 and 65 years of age, and initially we estimate separate models for women and men. The sample sizes by year and sex are found in Table 1.

To obtain a first idea of the characteristics of the six measures of use, Tables 2 through 5 give descriptive statistics for women's and men's health care usage, for the the years 1996 and 2000[11]. Studying these tables, we can make a few observations:

- Women, on average, use all six forms of health care more frequently than do men. This result is very uniform and is stable over time. This suggests that models that pool across sexes will require a dummy variable for sex.

- Men are more likely than women never to use forms of care. The difference is especially notable in the cases of OBDV, DV, and RX, which are probably more elective forms of care than are the other three measures.

- There is considerable temporal stability of the statistics, for all six measures of usage. However, there are some temporal variations that are notable. For example, the mean of ERV for women is 50% higher in 1996 than in 2000. This suggests the use of dummy variables for years in models that pool across time.

- Most measures of health care use exhibit considerable unconditional overdispersion. The IPV and ERV measures are in some cases rea-

[11] Results for the other years are very similar and are omitted to save space.

sonably close to unconditional equidispersion. For these measures it is possible that conditional equidispersion might hold, and that a Poisson model might be adequate. In the other cases the models that allow for overdispersion will likely be preferred.

- The percentage of zeros for the OPV, IPV and ERV measures is usually around 90% or higher. The OBDV, DV, and RX measures have positive usage by a much larger proportion of the sample.

Next, to obtain an idea of the characteristics of the explanatory variables, Tables 6 and 7 present descriptive statistics for the four conditioning variables, for women and men. We present these statistics only for the year 2000 data, since the other years are substantially similar. Highlights include:

- The means of AGE and EDUC are quite similar across sexes.
- There is a notable difference in the mean of INC, which presumably is due to a sex differential in the incomes of single people (recall that INC is defined as family income). The fact that maximum values of INC are the same is because INC was top-coded during the execution of the survey.
- Only a small part of the population has access to publicly-provided health care insurance.

4 Model selection

We have under consideration 6 measures of health care usage, 2 sex groups, and 5 years of data. For each of these 60 data sets we wish to determine which of a number of statistical models is most appropriate. Some of the statistical models require determination of the specific parameterization

(*e.g.*, whether to use an NB-I or NB-II base model, or the degree of the polynomial expansion for the PSNP, PSP, HPSP and NBSNP models). In the face of so many comparisons to make we use an information criterion approach, concretely the consistent Akaike information criterion (CAIC). The CAIC is defined as $CAIC = -2\ln L + p(\ln n + 1)$, where $\ln L$ is the log-likelihood value, p is the number of parameters of the model, and n is the sample size. The CAIC is a penalized goodness of fit criterion. Additional parameters usually allow for better fit, in terms of the log-likelihood value, but the penalty term prevents selection of overparameterized models. The CAIC is a consistent model selector, in the sense that the correct model in a set of models will have the lowest CAIC value, as the sample size tends to infinity (Sin and White, 1996)[12]. The simple Akaike information criterion (AIC), which has been used in some of the related literature, is not consistent, in that it can favor overparameterized models. The Bayes (or Schwartz) information criterion (BIC) that also appears in the literature can be expressed as $BIC = CAIC - p$. This criterion is also consistent. It may favor a somewhat more highly parameterized model than the CAIC. The BIC can be calculated using the information we provide in our results, but we do not report it here, to save space.

We report CAIC values starting with the models that allow parameters to vary by both sex and year, then we report results where parameters are constant across sexes (except for the coefficient of a dummy variable for sex) but vary by year, and finally we report results that pool both across sexes and time. The pooling across time is only for the years 1996, 1998 and 2000, so that no individual enters the sample in more than one year. In this

[12] If none of the models is the correct model, then the model that is closest to the correct model in the sense of the Kullback-Leibler information criterion will have the smallest CAIC value, asymptotically.

way, the observations are independent of one another. MLE estimation of models with dependent observations would require specifying the nature of the dependence, which is a step we prefer to avoid in this work.

It may be shown that, for two models that share no parameters and use disjoint data sets, the overall CAIC value is simply the sum of the CAIC values of the two models. Thus, one can compare the sum of the CAIC values for separate models for men and women in a given year with the CAIC value of a model that pools across sexes using the data of the same year. If the CAIC of the pooled model is lower, pooling is supported, otherwise, separate models are favored. Likewise, we can compare models that pool across time with analogous models that allow coefficients to vary by time. In this way we can determine what level of pooling is supported by the data, for each of the 6 measures of use.

With regard to estimation details, some of the models lead to a log-likelihood function that may have local maxima. For the models that do not have a globally concave log-likelihood function, we used simulated annealing to find a rough maximizer which satisfied convergence of the log-likelihood function out to 2 decimal places, then iterated to convergence using a BFGS maximizer. For the other models we used the BFGS maximizer directly. All estimation routines were programmed using GNU Octave (`www.octave.org`) and are available from the authors.

Separate models by sex and year We begin with the CAIC values of the various models, for the year 2000 data. For the other years we only report (below) the results for the favored models, to avoid overwhelming the reader with details. Tables 8 and 9 report the relative CAIC values, for women and men, respectively, for the statistical models that were discussed

in Section 2. For the model with the minimum CAIC value, the tables report this value. For the other models the tables report the CAIC value relative to that of the favored model, to facilitate comparisons. For the OBDV, DV and RX measures of use (which are those with higher sample means), we can see that there are a number of models that reach a CAIC within 1% of that of the favored model, while for the OPV, IPV and ERV measures of use, the distances between the favored model and the other models are often larger.

Tables 10 and 11 report which are the CAIC-favored models for each of the five years, for women and men, respectively. Some points to note:

- The numbers of times models are favored are: NB - 44 times; PSP - 8 times; NBSNP - 6 times; and CMNB and HPSP, one time each. The other models are never favored.
- Some use measures exhibit considerable variation over time in the models that are favored (*e.g.*, OBDV and DV). The CAIC-favored model for these use measures has "close competitors" in Tables 8 and 9. The favored model is very stable over time for the IPV and RX use measures.
- One result that stands out is that the simple Poisson-style specification of the conditional mean, $E(y|x) = e^{x\beta}$, is used by the favored model in 52 of 60 cases (86.6%).

With relatively homogeneous data that are for single sex groups and single years, simple models work well in the great majority of cases.

Pooling across sexes Table 12 reports the CAIC values for models that pool the coefficients across sexes, and add a dummy variable that allows the constant to vary by sex, for the year 2000. In the last row we present

the sum of the CAIC values of the models that allow all parameters to vary by sex, relative to the CAIC value of the favored model. We do not present such detailed results for the other years, but instead only report the favored models for this level of pooling, for each of the five years, in Table 13. In this table we can observe that:

- Pooling across sexes is favored in all cases except for ERV in 1998. In all other cases, use of a dummy variable and a common model and slope coefficients is favored.
- Only four models (apart from parameterization details) are ever favored: they are NB (14 times), NBSNP (8 times), PSP (6 times), and CMNB (3 times).
- The simple Poisson-style specification of the conditional mean $E(y|x) = e^{\mathbf{x}\beta}$ is implied by the favored model in 20 of 31 cases (the NB and PSP models) which is 64.5% of the cases, down from the 86.6% for separate models by sexes. Pooling is supported in all cases but 1 out of 30 (ERV, 1998), but pooling seems to require more flexible densities to capture the greater heterogeneity of the data.
- There is considerable stability over time. For example, the NBSNP model is favored in 4 of 5 years for the DV use measure, and the PSP model is favored 4 of 5 times for the OPV use measure.
- The Poisson model and the more highly parameterized hurdle and mixture models (HPSP, MNB) are never favored.
- When the NBSNP model is favored, it is always the version that uses a NB-I base density.

Pooling across years We have seen that pooling across sexes is almost always favored. Next we present CAIC results for models that pool across the years 1996, 1998 and 2000, adding dummy variables that allow the constant to vary by both sex and year. We do not use the data from 1997 and 1999 so that a given individual appears only once in the sample, and thus the data consists of independent observations.[13] Table 14 presents the results. We note that

- Pooled models are always favored. Time-wise heterogeneity seems to be adequately captured by a dummy variable.
- Relatively complicated, newer models (PSP, NBSNP, CMNB) are favored in 5 of 6 cases. However, for the ERV data where the PSP model is favored, the NB-I and NB-II models have only slightly higher CAIC values.
- The simple Poisson-style mean function $E(y|x) = e^{\mathbf{x}\beta}$ is favored in only 2 of 6 cases (for IPV and ERV). Again, as we pool more heterogeneous data, more complicated densities are required to fit the data well. These more complicated densities imply more complicated conditional moments. Note that the cases where the simple mean function is accepted are those where the unconditional mean of the dependent variable is lowest, and the percentage of zeros is highest, and the mean/variance ratio is closest to 1 (see Tables 2-5).

The overall conclusion is that pooling by age and sex is almost always favored, when data is available to make it possible. Simpler models often work well when the data is relatively homogeneous (for example, separate

[13] If we were to include the data from 1997 and 1999, but still treat the observations as independent, the estimators would not truly be maximum likelihood estimators, and thus the use of the CAIC to compare models would not be valid.

models by sexes, for a single year) and more complex models are often required when more heterogeneous data is pooled. Of the statistical models compared in this paper, some (the Poisson, PSNP, HNB and MNB) are always dominated, and the HPSP model is likely too highly parameterized for all but exceptional cases. Of the more complicated newer models, the NBSNP, PSP and CMNB models are found to be useful contributions for analysis of this sort of data and probably deserve consideration in future work.

5 Estimation results

Though a detailed economic analysis of estimation results for the favored models is beyond the scope of this paper, Table 15 presents estimation results for the CAIC-favored models for the pooled 1996-1998-2000 data, with pooling across sexes. The Table contains results for all six use measures. The models are the favored models that appear in Table 14.

Examining the estimation results, we can make several notes:

- With respect to time trends, DV usage has declined significantly over the 1996-2000 period. Consumption of prescription drugs (RX) has increased significantly. No other trends are clear.
- For all usage measures except DV, holding publicly-provided health care insurance (PUB) has a positive and strongly significant effect on usage levels.
- The dummy variable that indicates that the individual is a woman is positive in all cases, and is highly significant except for the IPV and ERV measures. The IPV and ERV measures are often associated

with events such as serious illness or accidents that are in a large part beyond the control of the individual.

- Age always has a positive coefficient, and is highly significant in all cases except the ERV usage measure.
- Income is negative and significant for the IPV, ERV and RX use measures. It is positive and significant for the DV measure.
- Education has a positive and significant effect upon the OBDV and DV measures, and it has a negative and significant impact upon the IPV and ERV measures.
- There is evidence that low-income, low-education individuals use IPV and ERV services more than the average individual. They make less use of dental care visits than average. Other effects are not so clear.
- The CMNB model used for the OBDV, OPV and RX use measures is characterized by mixing two NB densities, both of which are overdispersed, and at least one of which is highly overdispersed. The mix (π) parameter is estimated with poor precision in all three cases. The constant shifter for the second NB density is highly significant.
- The α and γ parameters of the PSP(1) density for the ERV usage measure are estimated imprecisely. It appears that they are not well identified separately for this data set, but that their joint impact is important (since the model had the best CAIC score).
- The α and one of theγ parameters are significant for the NBSNP model used for the DV use measure. There appears to be a problem of poor separate identification similar to that of the PSP(1) model for the ERV

data. This problem was noted by Cameron and Johansson (1997) for the PSNP model.

Of the forms of health care under consideration, OBDV and DV are those that are most likely to include preventive visits such as checkups. We can see that more educated, and in the case of DV, higher income individuals, use these two forms of care more frequently than average. Likewise, IPV and ERV may be used more than average by people who have not taken care of their health through preventive care, or who are seeking to use emergency room visits in place of ordinary doctor visits in an attempt to avoid insurance copayments. The fact that poorer, less educated people use these forms of care more frequently than average might be explained by such factors.

6 Conclusions

This paper has surveyed a number of statistical models for univariate count data and has applied them data on health care usage from the Medical Expenditure Panel Survey, years 1996-2000. The objective of the paper has been to attempt to determine which models are most appropriate for this sort of data. A secondary objective has been to determine which level of pooling across time and sexes is supported by the data.

We have found that some of the newer models are quite useful and warrant serious consideration when undertaking empirical work with this sort of data. In particular, depending upon the usage measures and the level of pooling, the NBSNP, PSP, and CMNB models are found to fit the data better than more traditional models such as the NB and especially the HNB. Other newer models such as the MNB and HPSP are found to be

excessively parameterized for the usage measures in the MEPS data used here, according to the CAIC criterion.

Another result is that pooling the data, both across time and across sexes, is almost always favored. There is enough parameter stability so that dummy variables can be used to capture the important variations in a simple and parsimonious way, without imposing overly strong restrictions on the model. As more heterogeneous data is pooled, more complex statistical models become necessary so that the assumption of parameter constancy (except changes in the constant captured by dummy variables) can be maintained. The basic finding of the paper is that it is more parsimonious to use a relatively complex statistical model with parameter constancy than to use simple statistical models with parameters that vary across data groups. The degree of complexity of the statistical mode required for adequate fit to the data depends upon the usage measure under consideration. Factors that lead to more complicated models being needed are a high mean, low proportion of zeros, and overdispersion.

This paper has not focused upon estimation results or economic analysis of the such results. Nevertheless, we have presented some limited results using the pooled by time and sex data, which is the favored approach in all cases. We have seen that the coefficients of the variables have signs that can be given a plausible economic interpretation. However, the discussion has not been deep, since this sort of analysis is not the focus of this paper.

Some directions for further work are quite clear. Given that pooling across time has been found to be desirable, it would be useful to develop models that allow for dependent observations, so that the entire data set for all years could be used. This will require explicit modeling of the de-

pendency of use measures over time, which will lead to the consideration of multivariate count data densities and issues of estimation of such non-linear models with panel data. Another direction for work would to be to try to tackle the endogeneity of private health care insurance in a convincing way. This may not be possible using the MEPS data due to lack of good instruments, but with other data sets it could be undertaken.

Table 1: Sample Sizes

	1996	1997	1998	1999	2000
Women	737	1205	477	830	817
Men	680	1104	478	802	800

Table 2: Descriptive Statistics, Use Variables, Women, 1996

	OBDV	OPV	IPV	ERV	DV	RX
mean	4.970	0.288	0.117	0.156	1.640	13.323
st. dev.	6.965	1.597	0.423	0.474	2.432	20.377
mean/var	0.102	0.113	0.651	0.696	0.277	0.032
min	0.000	0.000	0.000	0.000	0.000	0.000
max	93.000	37.000	4.000	5.000	32.000	142.000
% zero	0.195	0.870	0.913	0.877	0.392	0.187

Table 3: Descriptive Statistics, Use Variables, Men, 1996

	OBDV	OPV	IPV	ERV	DV	RX
mean	3.240	0.253	0.099	0.106	1.481	7.684
st. dev.	5.665	1.395	0.410	0.388	2.234	19.270
mean/var	0.101	0.130	0.585	0.702	0.297	0.021
min	0.000	0.000	0.000	0.000	0.000	0.000
max	54.000	19.000	4.000	4.000	17.000	296.000
% zero	0.309	0.893	0.929	0.915	0.488	0.349

Table 4: Descriptive Statistics, Use Variables, Women, 2000

	OBDV	OPV	IPV	ERV	DV	RX
mean	4.721	0.279	0.072	0.109	1.515	14.034
st. dev.	6.680	2.741	0.307	0.366	1.887	18.379
mean/var	0.106	0.037	0.768	0.813	0.425	0.042
min	0.000	0.000	0.000	0.000	0.000	0.000
max	89.000	77.000	4.000	3.000	11.000	124.000
% zero	0.170	0.864	0.936	0.906	0.424	0.176

Table 5: Descriptive Statistics, Use Variables, Men, 2000

	OBDV	OPV	IPV	ERV	DV	RX
mean	3.027	0.161	0.083	0.130	1.100	8.810
st. dev.	4.532	0.734	0.344	0.588	1.639	16.566
mean/var	0.147	0.299	0.697	0.376	0.410	0.032
min	0.000	0.000	0.000	0.000	0.000	0.000
max	50.000	13.000	3.000	11.000	12.000	160.000
% zero	0.314	0.909	0.936	0.911	0.515	0.354

Table 6: Descriptive Statistics, Explanatory Variables, Women, 2000

	mean	st. dev.	min	max
PUB	0.045	0.197	0.000	1.000
AGE	50.770	7.140	40.000	65.000
INC	69.682	44.486	0.000	323.033
EDUC	13.542	2.521	0.000	17.000

Table 7: Descriptive Statistics, Explanatory Variables, Men, 2000

	mean	st. dev.	min	max
PUB	0.056	0.223	0.000	1.000
AGE	50.300	7.247	40.000	65.000
INC	71.963	43.813	0.000	323.033
EDUC	13.504	2.926	0.000	17.000

Table 8: CAIC Values, Women, 2000

	OBDV	OPV	IPV	ERV	DV	RX
POISSON	1.6320	1.8007	1.0103	1.0094	1.1131	2.8494
PSNP(1)	1.5060	1.7249	1.0227	1.0189	1.1162	2.5116
PSNP(2)	1.3932	1.5585	1.0066	1.0126	1.0222	2.2547
PSNP(3)	1.2782	1.5037	1.0230	1.0250	1.0250	2.2560
PSP(1)	1.0019	1.0234	1.0119	1.0129	1.0032	1.0017
PSP(2)	1.0080	1.0582	1.0774	1.0927	1.0097	1.0069
PSP(3)	4332.1958	1.0126	1.0262	1.0253	1.0060	1.0030
HPSP(1)	1.0103	1.0638	1.1030	1.1175	1.0152	1.0092
HPSP(2)	1.0005	1.0096	1.0415	1.0372	1.0085	1.0028
HPSP(3)	1.0139	1.0732	1.1290	1.1407	1.0197	1.0118
NB-I	1.0031	1.0530	1.0008	622.1461	2752.9661	5844.7338
NB-II	1.0027	1.0514	471.1805	1.0021	1.0051	1.0012
HNB-I	1.0047	1.0864	1.0598	1.0551	1.0042	1.0032
HNB-II	1.0040	1.0616	1.0539	1.0574	1.0034	1.0032
MNB-I	1.0115	1.0289	1.0990	1.0729	1.0158	1.0048
MNB-II	1.0125	1.0234	1.1045	1.0841	1.0169	1.0069
CMNB-I	1.0014	1.0007	1.0400	1.0357	1.0053	1.0008
CMNB-II	1.0015	911.6805	1.0400	1.0367	1.0090	1.0033
NBSNP-I(1)	1.0057	1.0128	1.0200	1.0139	1.0027	1.0023
NBSNP-I(2)	1.0014	1.0212	1.0236	1.0258	1.0006	1.0036
NBSNP-I(3)	1.0031	1.0010	1.0456	1.0367	1.0032	1.0038
NBSNP-II(1)	1.0056	1.0139	1.0175	1.0155	1.0028	1.0062
NBSNP-II(2)	1.0031	1.0040	1.0236	1.0277	1.0045	1.0039
NBSNP-II(3)	1.0049	1.0125	1.0482	1.0367	1.0067	1.0033

Table 9: CAIC Values, Men, 2000

	OBDV	OPV	IPV	ERV	DV	RX
POISSON	1.4931	1.2675	1.0317	1.1393	1.1277	3.1952
PSNP(1)	1.4693	1.2760	1.0439	1.1486	1.1307	2.7935
PSNP(2)	1.2857	1.0421	1.0143	1.0342	1.0254	2.4597
PSNP(3)	1.1948	1.0532	1.0309	1.0463	1.0287	2.4614
PSP(1)	1.0071	1.0054	1.0218	1.0005	1.0130	1.0065
PSP(2)	1.0118	1.0621	1.1195	1.0710	1.0242	1.0122
PSP(3)	1.0051	1.0164	1.0372	1.0099	1.0153	1.0071
HPSP(1)	1.0140	1.0795	1.1526	1.0908	1.0308	1.0136
HPSP(2)	1.0033	1.0274	1.0528	1.0221	1.0170	1.0074
HPSP(3)	1.0153	1.1010	1.1856	1.1144	1.0364	1.0168
NB-I	3564.8823	693.6498	463.0455	1.0027	2298.0098	4589.4350
NB-II	1.0051	1.0025	1.0064	634.1129	1.0097	1.0057
HNB-I	1.0056	1.0439	1.0748	1.0512	1.0129	1.0066
HNB-II	1.0083	1.0490	1.0706	1.0532	1.0141	1.0077
MNB-I	1.0099	1.0710	1.1045	1.0677	1.0170	1.0102
MNB-II	1.0138	1.0678	1.1128	1.0695	1.0313	1.0161
CMNB-I	1.0018	1.0261	1.0477	1.0244	1.0096	1.0032
CMNB-II	1.0032	1.0275	1.0545	1.0230	1.0185	1.0083
NBSNP-I(1)	1.0018	1.0125	1.0167	1.0167	1.0036	1.0022
NBSNP-I(2)	1.0005	1.0186	1.0324	1.0140	1.0066	1.0036
NBSNP-I(3)	1.0027	1.0293	1.0489	1.0260	1.0099	1.0049
NBSNP-II(1)	1.0069	1.0140	1.0233	1.0106	1.0127	1.0066
NBSNP-II(2)	1.0079	1.0239	1.0398	1.0195	1.0159	1.0081
NBSNP-II(3)	1.0100	1.0340	1.0553	1.0300	1.0192	1.0097

Table 10: CAIC-Favored Models, Women, 1996-2000

	OBDV	OPV	IPV	ERV	DV	RX
1996	PSP(3)	NB-II	NB-II	NB-I	NBSNP-I(3)	NB-I
1997	NBSNP-I(2)	NB-II	NB-I	NB-I	NBSNP-I(3)	NB-II
1998	NB-I	NB-I	NB-II	NB-II	PSP(1)	NB-II
1999	HPSP(2)	NB-II	NB-I	NB-II	NB-I	NB-II
2000	PSP(3)	CMNB-II	NB-II	NB-I	NB-I	NB-I

Table 11: CAIC-Favored Models, Men, 1996-2000

	OBDV	OPV	IPV	ERV	DV	RX
1996	NBSNP-II(2)	PSP(1)	NB-I	NB-I	NB-I	NB-I
1997	PSP(3)	PSP(1)	NB-II	PSP(2)	NBSNP-I(3)	NB-I
1998	NB-I	NB-II	NB-I	NBSNP-I(3)	NB-I	NB-I
1999	NB-I	PSP(1)	NB-I	NB-I	NB-I	NB-II
2000	NB-I	NB-I	NB-I	NB-II	NB-I	NB-I

Table 12: CAIC Values, Pooling Across Sexes, 2000

	OBDV	OPV	IPV	ERV	DV	RX
POISSON	1.5736	1.5941	1.0273	1.0767	1.1222	3.0133
PSNP(1)	1.5727	1.5688	1.0346	1.0821	1.1240	2.6808
PSNP(2)	1.3665	1.3916	1.0085	1.0143	1.0231	2.6816
PSNP(3)	1.2445	1.3309	1.0177	1.0211	1.0247	2.3243
PSP(1)	1.0036	1.0100	1.0087	1233.0104	1.0094	1.0083
PSP(2)	1.0053	1.0329	1.0729	1.0452	1.0096	1.0064
PSP(3)	1.0001	1.0052	1.0179	1.0048	1.0100	1.0091
HPSP(1)	1.0048	1.0376	1.0905	1.0588	1.0129	1.0069
HPSP(2)	1.0012	1.0041	1.0270	1.0116	1.0087	1.0099
HPSP(3)	1.0062	1.0478	1.1089	1.0650	1.0150	1.0085
NB-I	1.0026	1.0274	912.9064	1.0025	5019.8251	10405.3000
NB-II	1.0046	1.0297	1.0004	1.0009	1.0078	1.0084
HNB-I	1.0034	1.0503	1.0469	1.0369	1.0044	1.0024
HNB-II	1.0028	1.0387	1.0452	1.0308	1.0052	1.0040
MNB-I	1.0096	1.0276	1.0613	1.0403	1.0103	1.0035
MNB-II	1.0091	1.0233	1.0646	1.0393	1.0157	1.0130
CMNB-I	7865.6354	1581.2384	1.0270	1.0121	1.0038	1.0002
CMNB-II	1.0009	1.0013	1.0268	1.0121	1.0110	1.0085
NBSNP-I(1)	1.0042	1.0331	1.0102	1.0104	1.0009	1.0014
NBSNP-I(2)	1.0003	1.0061	1.0191	1.0061	1.0025	1.0022
NBSNP-I(3)	1.0013	1.0027	1.0270	1.0129	1.0033	1.0023
NBSNP-II(1)	1.0062	1.0091	1.0101	1.0047	1.0093	1.0104
NBSNP-II(2)	1.0056	1.0144	1.0193	1.0113	1.0070	1.0110
NBSNP-II(3)	1.0066	1.0087	1.0284	1.0181	1.0085	1.0117
Separate Models	1.0040	1.0152	1.0234	1.0189	1.0062	1.0028

Table 13: CAIC-Favored Models, Pooling Across Sexes, 1996-2000

	OBDV	OPV	IPV	ERV	DV	RX
1996	NBSNP-I(2)	PSP(1)	NB-I	NB-I	NBSNP-I(3)	NBSNP-I(2)
1997	NBSNP-I(2)	PSP(1)	NB-I	NB-I	NBSNP-I(3)	NB-I
1998	PSP(3)	PSP(1)	NB-I	Women: NB-II Men: NBSNP-I(3)	NBSNP-I(3)	NB-I
1999	CMNB-I	PSP(3)	NB-I	NB-I	NBSNP-I(3)	NB-I
2000	CMNB-I	CMNB-I	NB-I	PSP(1)	NP-I	NB-I

Table 14: CAIC Values, Pooling across 1996-1998-2000 and Sexes

	OBDV	OPV	IPV	ERV	DV	RX
POISSON	1.6126	1.5517	1.0515	1.0583	1.1692	2.9997
PSNP(1)	1.5255	1.5465	1.0547	1.0608	1.1699	2.6795
PSNP(2)	1.4060	1.3059	1.0033	1.0076	1.0537	2.3916
PSNP(3)	1.2867	1.1573	1.0073	1.0106	1.0544	2.3920
PSP(1)	1.0047	1.0042	1.0050	3051.3255	1.0074	1.0057
PSP(2)	1.0055	1.0158	1.0390	1.0283	1.0008	1.0029
PSP(3)	1.0045	1.0005	1.0076	1.0019	1.0076	1.0051
HPSP(1)	1.0020	1.0203	1.0469	1.0344	1.0021	1.0024
HPSP(2)	1.0037	1.0000	1.0115	1.0050	1.0083	1.0055
HPSP(3)	1.0030	1.0241	1.0548	1.0399	1.0033	1.0032
NB-I	1.0043	1.0217	2338.3952	1.0003	1.0010	1.0004
NB-II	1.0061	1.0231	1.0003	1.0001	1.0081	1.0053
HNB-I	1.0040	1.0285	1.0272	1.0212	1.0031	1.0014
HNB-II	1.0042	1.0247	1.0274	1.0209	1.0030	1.0019
MNB-I	1.0077	1.0183	1.0350	1.0272	1.0046	1.0022
MNB-II	1.0044	1.0211	1.0362	1.0240	1.0097	1.0070
CMNB-I	19541.3171	4029.5548	1.0099	1.0052	1.0005	25190.7329
CMNB-II	1.0004	1.0002	1.0115	1.0052	1.0059	1.0050
NBSNP-I(1)	1.0050	1.0242	1.0043	1.0037	1.0016	1.0010
NBSNP-I(2)	1.0000	1.0112	1.0083	1.0027	13100.3787	1.0010
NBSNP-I(3)	1.0005	1.0054	1.0121	1.0057	1.0005	1.0011
NBSNP-II(1)	1.0068	1.0221	1.0042	1.0033	1.0088	1.0059
NBSNP-II(2)	1.0054	1.0117	1.0080	1.0044	1.0085	1.0060
NBSNP-II(3)	1.0058	1.0140	1.0117	1.0072	1.0092	1.0064
Separate Models	1.0057	1.0192	1.0354	1.0125	1.0039	1.0029

Table 15: Estimation Results, Overall CAIC-Favored Models, 1996-1998-2000

	OBDV	OPV	IPV	ERV	DV	RX
Const.	0.2098	-3.8315	-2.6857	-1.7703	-1.6691	-0.1809
	(0.210)	(0.673)	(0.549)	(0.561)	(0.200)	(0.172)
	[0.318]	[0.000]	[0.000]	[0.002]	[0.000]	[0.294]
1998	0.0215	0.0038	-0.2639	-0.0865	-0.0112	0.0498
	(0.044)	(0.127)	(0.168)	(0.218)	(0.049)	(0.049)
	[0.627]	[0.976]	[0.115]	[0.692]	[0.818]	[0.306]
2000	-0.0163	-0.0603	-0.1818	-0.0635	-0.1732	0.1048
	(0.038)	(0.110)	(0.138)	(0.133)	(0.044)	(0.041)
	[0.672]	[0.582]	[0.187]	[0.633]	[0.000]	[0.011]
Pub	0.2804	0.4658	1.0211	0.6232	-0.1300	0.5118
	(0.076)	(0.208)	(0.202)	(0.333)	(0.102)	(0.104)
	[0.000]	[0.025]	[0.000]	[0.061]	[0.205]	[0.000]
Woman	0.3794	0.3271	0.1028	0.1323	0.2581	0.5490
	(0.034)	(0.097)	(0.125)	(0.114)	(0.038)	(0.037)
	[0.000]	[0.001]	[0.410]	[0.244]	[0.000]	[0.000]
Age	0.0279	0.0394	0.0260	0.0132	0.0185	0.0412
	(0.002)	(0.007)	(0.009)	(0.010)	(0.003)	(0.002)
	[0.000]	[0.000]	[0.004]	[0.178]	[0.000]	[0.000]
Income	-0.0004	-0.0007	-0.0042	-0.0037	0.0017	-0.0012
	(0.000)	(0.001)	(0.002)	(0.002)	(0.000)	(0.000)
	[0.315]	[0.594]	[0.026]	[0.044]	[0.000]	[0.009]
Education	0.0175	0.0200	-0.0660	-0.0643	0.1115	0.0078
	(0.007)	(0.020)	(0.022)	(0.024)	(0.009)	(0.008)
	[0.012]	[0.318]	[0.003]	[0.006]	[0.000]	[0.305]
alpha	9.8375	0.4879	0.4579	0.4146	2.8892	12.7021
	(0.224)	(0.590)	(0.173)	(1.151)	(0.078)	(0.313)
	[0.000]	[0.408]	[0.008]	[0.719]	[0.000]	[0.000]
gam1/Const2	1.0000	-1.5483	na	0.2949	-0.0948	2.3131
	(0.096)	(0.732)		(0.323)	(0.055)	(0.131)
	[0.000]	[0.034]		[0.361]	[0.084]	[0.000]
gam2/alpha2	2.5659	17.5843	na	na	0.0030	56.6040
	(0.166)	(0.546)			(0.036)	(0.400)
	[0.000]	[0.000]			[0.934]	[0.000]
mix	0.2278	0.5515	na	na	na	0.6577
	(0.461)	(1.354)				(0.933)
	[0.621]	[0.684]				[0.481]

( ) = standard errors; [ ] = p-values

References

[1] Cameron, A.C. and P.K. Trivedi (1986a), Econometric models based on count data: comparisons and applications of some estimators and tests, *Journal of Applied Econometrics*, **1**, 29-54.

[2] Cameron, A.C. and P.K. Trivedi (1998), *Regression analysis of count data*, Econometric Society Monographs, Cambridge University Press.

[3] Cameron, A.C. *et al.* (1988), A microeconometric model of the demand for health insurance and health care in Australia, *Review of Economic Studies*, **55**, 85-106.

[4] Cameron, A.C. and P. Johansson (1997), Count data regression using series expansions: with applications, *Journal of Applied Econometrics*, **12**, 203-23.

[5] Cragg, J.G. (1971), Some statistical models for limited dependent variables with applications to the demand for durable goods, *Econometrics*, **39**, 829-44.

[6] Deb, P. and P.K. Trivedi (1997), Demand for medical care by the elderly: a finite mixture approach, *Journal of Applied Econometrics*, **12**, 313-36.

[7] Dismuke, C.E. and P. Guimares (2002), Has the caveat of case-mix based payment in¯uenced the quality of inpatient hospital care in Portugal?, *Applied Economics*, **34**, 1301-1307.

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

[9] Gallant, A.R. and D.W. Nychka (1987), Seminonparametric maximum likelihood estimation, *Econometrica*, **55**, 363-90.

[10] Gerdtham, U-G. (1997), Equity in health care utilisation: further evidence based on hurdle models and Swedish macro data, *Health Economics*, **6**, 303-19.

[11] Gerdtham, U-G. and P.K. Trivedi (2000), Equity in Swedish health care reconsidered: new results based on the finite mixture model, mimeo, http://swopec.hhs.se/hastef/abs/hastef0365.htm

[12] Guo, J.Q. and P.K. Trivedi (2002), Flexible parametric models for long-tailed patent count distributions, *Oxford Bulletin of Economics and Statistics*, **64**, 63-82.

[13] Gurmu, S. (1997), Semi-parametric estimation of hurdle regression models with an application to medicaid utilization, *Journal of Applied Econometrics*, **12**, 225-42.

[14] Gurmu, S. and P.K. Trivedi (1996), Excess zeros in count models for recreational trips, *Journal of Business and Economic Statistics*, **14**, 469-77.

[15] Gurmu, S., P. Rilstone and S. Stern (1999), Semiparametric estimation of count regression models, *Journal of Econometrics*, **88**, 123-50.

[16] James, L., C. Priebe and D. Marchette (2001), Consistent estimation of mixture complexity, *Annals of Statistics*, **29**, 1281-1296.

[17] Mullahy, J. (1986), Specification and testing of some modified count data models, *Journal of Econometrics*, **33**, 341-65.

[18] Pohlmeier, W. and V. Ulrich (1995), An econometric model of the two-part decision-making process in the demand for medical care, *Journal of Human Resources*, **30**, 339-61.

[19] Sin, Chor-Yiu, and H. White (1996), Information criteria for selecting possibly misspecified parametric models, *Journal of Econometrics*, **71,** 207-225.

[20] Windmeijer, F. and J. Santos Silva (1997), Endogeneity in count data models: an application to demand for health care, *Journal of Applied Econometrics,* **12**, 281-94.

[21] Yoshida, A. and Y.-S. Kim (2008), Sharing health risk and income risk within households: evidence from Japanese data, *Applied Economics*, **40**, pp. 1723 - 1735