14 Empirical models of health care use

Partha Deb and Pravin K. Trivedi

1. Background

In many studies of the use of medical care services, utilization is measured as a count variable taking only non-negative integer values. In such cases, for example, the number of visits to a doctor, number of hospital stays or the number of prescription medicines, the sample is concentrated on a few small values, the distribution is substantially skewed and the distribution of counts is intrinsically integer-valued and non-negative. In addition, these counts include both non-users and users with the former group usually contributing a nontrivial fraction to the total sample size. This “zeros problem”, which generates a nonlinearity in the response to determinants of use, is a significant econometric complication that has motivated a variety of specification and estimation strategies.

Typical analyses of health care use analyze the role of its potential determinants and are focused on the effects of these determinants on the expected number of events. There are, however, practical and policy-related applications in which predictions of the distributions of event-counts or probabilistic statements regarding the distributions, and not only the expected number of events, are of interest. In many applications, the determinants of health care use are treated as if they are exogenous variables. However, this is not always the case; two leading examples of potentially endogenous variables in
such models, central to the economics of the utilization of health care services, are price and insurance effects. When individuals choose health insurance plans, insurance status and service-level out-of-pocket prices are likely to be endogenous. This problem is ubiquitous in the US system where most individuals purchase insurance via employer-sponsored health care plans and where others can choose to purchase private insurance coverage independently; it also exists in certain countries in Europe and elsewhere in mixed public-private systems where individuals can purchase private insurance supplemental to that provided in the universal public plan. In either case, the plan is not exogenously assigned, which may lead to self-selection. Self-selection, which may take the form of favourable or adverse selection, arises because optimizing individuals, possessing knowledge of their own health attributes, proclivities, and economic constraints, select plans accordingly. Favourable selection refers to the tendency for even the more healthy individuals to purchase insurance, while adverse selection means that the less healthy purchase more insurance. Therefore, these attributes, which partly determine the individual's choice of health plans, also affect their expected utilization of services. Consequently, insurance status and prices of services cannot be treated as exogenous. Instead, identification of the causal impact of insurance on health care use requires joint modeling of insurance choice and use of care.

The objective of successful identification and consistent estimation of these key parameters faces several challenges, many of which are the subject of current research. In this article we summarize the leading issues in econometric analyses of the use of health care services based on individual-level cross section data. We consider first the case of exogenous regressors, and then the more general model with endogenous regressors.
2. Methods 1: exogenous covariates

The Poisson regression model, derived from the Poisson distribution for the number of occurrences of the event provides a natural starting point for analyses of count data. The first two moments of the distribution are \( E[y] = \mu \), and \( V[y] = \mu \). The regression model is derived by parameterizing the relation between the mean \( \mu \) and predictors \( x \).

The standard assumption uses the exponential mean parameterization with \( K \) linearly independent covariates,

\[
\mu_i = \exp(x_i \beta), i = 1,...,N.
\]

Because \( V[y_i | x_i] = \exp(x_i \beta) \) the regression is intrinsically heteroskedastic.

The Poisson is an attractive starting point as it respects the integer-valued property of outcomes. However, it implies the equality of mean and variance, i.e. the equidispersion property, is usually too restrictive for counts of health care use. Most measures of health care use display overdispersion, with conditional variance exceeding the conditional mean. In many applications a Poisson density predicts the probability of a zero count to be considerably less than is actually observed in the sample, which reflects the failure of the equidispersion. This is termed the "excess zeros problem", as there are more zeros in the data than the Poisson predicts.

The phenomenon of overdispersion has several explanations. In count data it may be due to unobserved heterogeneity. Counts are viewed as being generated by a Poisson process, but the researcher is unable to correctly specify the rate parameter of this process. Hence the rate parameter is itself modeled as a random variable whose
variation is partly induced by unobserved factors that are assumed to be uncorrelated with observed covariates. The resulting model is a mixture model. If the unobserved heterogeneity is multiplicative (insert expression here) and has an independent gamma distribution, then the marginal distribution of event counts is the widely-used negative binomial (NB) model. Unlike the Poisson, the NB model allows for overdispersion as its conditional variance exceeds the conditional mean. In applications, however, the NB often underpredicts the number of zeros, especially for measures such as number of doctor visits, with a substantial fraction of zeros and a long right tail. In response to these concerns, a number of modified Poisson alternatives have been developed in the literature, which we describe below. In practice, it is desirable to consider and compare its adequacy with these modified Poisson alternatives.

**Hurdle or Two-part Models**

Overdispersion may also arise because the process generating the first event may differ from the process determining later events. In modeling the usage of medical services, the two-part model (TPM) has served as a methodological cornerstone of empirical analysis. The first part of the TPM is a binary outcome model that describes the determinants of use versus nonuse. The second part describes the distribution of use conditional on some use. Although in health economics the term "two-part model" has been used predominantly to refer to models of health expenditures, the structure of the TPM is equally applicable for discrete or continuous outcomes. The TPM for count data has, on the other hand, often been referred to as a hurdle model (Mullahy, 1986). This leads to a modified model for count data. Suppose the zeros are determined by the
density $f_i(.)$, so that $\Pr[y = 0] = f_i(0)$. The positive counts come from the truncated density $f_2(y | y > 0) = f_2(y)/(1 - f_i(0))$, which is multiplied by $\Pr[y > 0] = 1 - f_i(0)$ to ensure that probabilities sum to unity. Then

$$g(y) = \begin{cases} f_i(0) & \text{if } y = 0 \\ f_2(y)(1 - f_i(0)) & \text{if } y > 0. \end{cases}$$

This reduces to the standard model only if $f_i(.) = f_2(.)$. Thus, in the modified model, the two processes generating the zeros and the positives are not constrained to be the same. Additional flexibility results from being able to choose suitable component densities that improve on the Poisson; for example, one could use a NB density for the positive counts, thus allowing for two sources of overdispersion. Note that while the motivation for this model is to handle excess zeros, it is also capable of modeling too few zeros.

The appeal of the TPM is partly driven by an important feature of the demand for medical care, which is the high incidence of zero usage. For example, approximately 30% of typical cross-sectional samples of non-institutionalized individuals in the United States report no outpatient visits in the survey year. The TPM is also well supported empirically, with explanatory variables often playing different roles in the two parts of the model. The appeal of the TPM in health economics is also based on its connection to a principal-agent model (see, for example, Zweifel, 1981). In this framework, it is argued, the decision to seek care is largely the patient's (the principal) whereas the physician (the agent) determines utilization on behalf of the patient once initial contact is made (this argument is well motivated in Manning et al., 1987; Pohlmeier and Ulrich, 1995). Finally, because maximum likelihood estimation of the hurdle model involves separate maximization of the two terms in the likelihood, one corresponding to the zeros
and the other to the positives, the TPM is easy to estimate and typically does not suffer from computational instabilities.

Pohlmeier and Ulrich (1995) develop a TPM based on the NB distribution and estimate its parameters for counts of GP and specialist visits using data from the German Socioeconomic Panel. They find strong evidence favouring the TPM over Poisson and NB models and conclude that contact and frequency decisions are governed by different stochastic processes, hence should be modeled using a TPM.

**Zero-inflated models**

A second modified count model is the zero-inflated model. This supplements a count density \( f_2(.) \) with a binary process with density \( f_1(.) \). If the binary process takes value 0, with probability \( f_1(0) \), then \( y = 0 \). If the binary process takes value 1, with probability \( f_1(.) \), then \( y \) takes count values 0, 1, 2, ... from the count density \( f_2(.) \). This lets zero counts occur in two ways: as a realization of the binary process and as a realization of the count process when the binary random variable takes value 1. In the context of health care use, suppose the binary process determines whether or not an individual participates in the market for health care services, i.e., distinguishing between those who have demand for medical care from those who do not. Suppose, in addition that the count process determines the use of health care services among those who have demand. Then the appropriate statistical model is the zero-inflated model. The density is

\[
g(y) = f_1(0) + (1 - f_1(0))f_2(0) \quad \text{if } y = 0 \\
= (1 - f_1(0))f_2(y) \quad \text{if } y > 0.
\]

Regression models typically let \( f_1(.) \) be a logit model and \( f_2(.) \) be a Poisson or negative binomial density. Like the hurdle or TPM, the zero-inflated model is capable of
modeling too many or too few zeros, which can be verified by checking the first-order condition for likelihood maximization.

Although the zero-inflated model provides a natural way to introduce extra zeros, it is considerably less used as compared to the TPM. We speculate this is because the computational task is considerably more complex than the TPM as the binary and count parts of the model cannot be estimated separately. If the regression specifications for the binary and count have the same covariates, then identification of parameters of the binary and count parts separately can be a difficult problem in small to medium sized samples.

Chang and Trivedi (2003) find that the zero-inflated Poisson and negative binomial models used to analyze Vietnamese health care utilization data considerably improve the fit of the Poisson regression. Wang (2003) estimates a zero-inflated negative binomial regression model using data from the Australian Health Survey and it provides a plausible explanation for the relatively large fraction "permanent" non-users of health-care services.

Latent class models

Overdispersed counts can also be modeled using a discrete representation of unobserved heterogeneity. This generates the finite mixture class of models, and a particular subclass of this is the latent class models (LCM).

Within the LCM class the density of $y$ is a linear combination of $m, m \geq 2$, different densities, where the $j^{th}$ density is $f_j(y | \theta_j), j = 1, \ldots, m$ Thus an $m$-component finite mixture is
\[ g(y | \theta) = \sum_{j=1}^{m} \pi_j f_j(y | \theta), \quad 0 \leq \pi_j \leq 1, \quad \sum_{j=1}^{m} \pi_j = 1. \]

Here the components of the mixture are assumed, for generality, to differ in all their parameters. Less general formulations assume that only some parameters differ across the components. Zero-inflated class is a special case; additional flexibility comes from parameterizing \( \pi_i \).

Plausibly, different classes of users of health care services exist if the population is split by the latent health status of individuals. Note that, although surveys typically have some measures of health status and administrative data considerably richer diagnosis-based measures, few would argue that these data contain adequate information on the severity of the illnesses, for example, among latent dimensions of health status. The healthy subpopulation, perhaps the majority, might account for low average demand, whereas those who are ill may account for high average demand. When the observed health status is imperfectly observed, the finite mixture model may do a good job of separating subpopulations.

The LCM is an attractive candidate for modeling health care use for a number of reasons. First, it provides a flexible and parsimonious method of modeling the data. Each mixture component provides a local approximation to some part of the true distribution. Second, the LCM is semiparametric in the sense that it does not require any distributional assumptions for the mixing variable. Third, in many cases the latent classes may correspond to specific subpopulations of relevance from the viewpoint of public policy; e.g. healthy and sick, or high users and low-users in which case the finite mixture characterization has a natural interpretation. However, this is not essential. A caveat is that the LCM may fit the data better simply because outliers, influential observations, or contaminated observations, present in the data, are captured by the
additional components of the LCM model. Hence it is desirable that the hypothesis of LCM should be supported both by a priori reasoning and by meaningful a posteriori differences in the behaviour of latent classes. Finally, the approach can flexibly incorporate departures from random sampling, e.g. as in on-site and truncated samples, by substituting component densities that capture such departures (Lourenço and Ferreira, 2004).

A preference for the LCM over TPM in the context of models of the use of health care services may also be because the sharp dichotomy between users and non-users may not be tenable in the case of typical cross-sectional data-sets. In these data, health care events are recorded over a fixed time period and not over an episode of illness. Note that the TPM is motivated by the principal-agent model over an episode of illness. The LCM provides a more tenable distinction for typical cross-sectional data, distinguishing between an "infrequent user" and a "frequent user" of medical care, the difference being determined by health status, attitudes to health risk, and choice of lifestyle.

Practical limitations of LCM include: lack of theoretical guidance on specifying the number of components; inability to reliably distinguish between some of insufficiently-different components; the computational difficulties due to the multiple maxima in the likelihood function. The usual practice is to start with a few components and then add additional components if the fit of the model is significantly improved by doing so. Information criteria are often used to select between models with different number of components.

Deb and Trivedi (1997) estimated NB, TPM and LCM models of use for a number of types of services in a study of medical care demand by the elderly in the US. Using a
variety of tests, they conclude that the LCM is superior to the TPM for most types of services and speculate that the source of latent heterogeneity might be unobserved health status. Deb and Trivedi (2002) compare TPM and LCM models in a reexamination of data from the RAND Health Insurance Experiment. They also find that LCM performs better than TPM in most cases. Because heterogeneity due to insurance status can be ruled out in the RAND study, the authors argue that the case for health status as the source of latent heterogeneity is strong. Jimenez-Martin, et al. (2002), using data from the European Community Household Panel, estimate demand for physician services equations for 12 European countries. They find that LCM is preferred to TPM for visits to GPs while the opposite is true for visits to specialists.

3. Methods 2 - endogenous covariates

We now consider complications that arise from the presence of endogenous regressors in a model of counts. Neglecting endogeneity means that estimates will not be consistent. A leading example of an endogenous regressor in the context of the use of health care services is insurance status. In the simplest case, this might be a binary variable that takes the value 1 if the subject is insured, and zero otherwise. The full model then consists of two equations, a binary choice model for insurance and a count model for health care use. More generally, we may have a multinomial choice model for insurance (if the choice is between three or more insurance plans) and a count model for utilization (see Deb and Trivedi, 2006). If self-selection or adverse selection is an important feature of individual behaviour then modeling interdependence between use ("outcome") and insurance choice ("treatment") is very important. The econometric
challenge in these cases is to specify the precise manner in which the interdependence between the use and choice equations arises, and to control for the interdependence in estimation.

As is well-known, in the linear instrumental variable (LIV) approach the outcome equation is estimated after specifying instruments, denoted $W$, defined as variables that are correlated with treatments, but uncorrelated with the outcomes, conditional on other exogenous variables (see the chapter by Auld in this Companion). That is, valid instruments are those that impact on outcomes solely through the treatment variables. Important advantages of LIV are well documented. Under appropriate conditions they include consistent estimation, computational simplicity, and an absence of strong distributional assumptions. Such models have been used in models of health care use when insurance is binary. Linearization of the models for use and insurance can be justified (Angrist, 2001). However, extensions of such models to the case in which the treatment is multinomial do not exist. In addition, ignoring the discreteness of the count and the substantial mass at zero, often leads to poorly fitting models.

Among nonlinear models, for consistent estimation, the literature suggests two broad estimation approaches. The first is a limited information ("semiparametric") approach based on nonlinear instrumental variables (NLIV) or generalized method of moments (GMM); see Mullahy (1997). The second is a fully parametric approach that makes distributional assumptions about all endogenous variables (see Terza, 1998; Deb and Trivedi, 2006). Terza also provides a two-step approach that mimics the linear two-stage least squares estimator in which the endogenous variable is replaced by a fitted value from a reduced form.
The first approach begins with a moment condition. Then, assuming that there are enough moment conditions available, the GMM or NLIV estimation is feasible. But for nonlinear and limited dependent variable models, the advantages of NLIV methods are less well understood (Angrist, 2001). First, efficient estimation is computationally awkward because optimal nonlinear instruments are hard to find (Amemiya, 1985). Second, in implementing this approach the discrete nature of the variable is ignored and the model is treated like any other nonlinear model with an exponential mean. This is likely to lead to a substantial loss of goodness of fit.

Yet another approach for allowing for endogenous regressors is the control function approach, of which residual inclusion is a special case. The essential idea of this approach is to add an auxiliary variable (also known as a generated regressor) in the equation, called the control function, such that conditional on including this variable, the endogeneity problem can be ignored; see Navarro (2008) for a discussion of the properties of such an estimator. An example of this approach is to include a residual from an auxiliary regression of the endogenous variable on other regressors; see Terza et al. (2008).

In contrast to GMM and two-step approaches, Munkin and Trivedi (2003) and Deb and Trivedi (2006) develop a joint parametric model of counts with insurance plan variables as regressors and a choice model for the insurance plans. Endogeneity arises from the presence of correlated unobserved heterogeneity in the outcome (count) equation and the binary choice equation. Their model has the following structure:

\[ \Pr(Y_i = y_i \mid x_i, d_i, I_i) = f(x_i' \beta + \gamma d_i + \lambda I_i) \]

\[ \Pr(d_i = 1 \mid z_i, I_i) = g(z_i' \alpha + \delta I_i). \]
Here \( \mathbf{I}_i \) is a vector of latent factors reflecting unobserved heterogeneity and \( \lambda \) and \( \delta \) are associated vectors of factor loadings. The joint distribution of selection and outcome variables, conditional on the common latent factors, can be written as

\[
\Pr(Y_i = y_i \mid \mathbf{x}_i, d_i, \mathbf{I}_i) = f(x_i' \beta + y d_i + \lambda \mathbf{I}_i) \\
\times g(z_i' \alpha + \delta \mathbf{I}_i)
\]

because \((y, D)\) are conditionally independent.

The problem in estimation arises because the \( \mathbf{I}_i \) are unknown. Although the \( \mathbf{I}_i \) are unknown, assume that the distribution of \( \mathbf{I}_i, h, \) is known and can therefore be integrated out of the joint density, i.e.,

\[
\Pr(Y_i = y_i, d_y = 1 \mid \mathbf{x}_i, z_i) = \int [f(x_i' \beta + y d_i + \lambda \mathbf{I}_i) \\
g(z_i' \alpha + \delta \mathbf{I}_i)] h(\mathbf{I}_i) d\mathbf{I}_i
\]

The unknown parameters of the model may be estimated by maximum likelihood. The main problem of estimation, given suitable specifications for \( f, g \) and \( h \), is the fact that the integral does not have, in general, a closed form solution. A simulated log-likelihood function for the data can then be defined and maximized as in Deb and Trivedi (2006).

The above approach is equally applicable to multiple treatments and multiple outcomes, discrete or continuous. The limitation comes from the burden of estimation which is very heavy compared with an IV type estimator.

It is also possible to specify the error structure of our model using discrete distributions, often called the discrete factor model, as described by Mroz (1999). There are two advantages to such an approach. First, because such models are latent class models, they are semiparametric and the discrete distributions can, in principle,
approximate any continuous distributions. Second, because such models replace the integration with summation, their likelihood functions are considerably simpler to compute. There is one substantial drawback of such a specification. Because the discrete factor model is a finite mixture model, its likelihood function is known to have multiple maxima (Lindsay, 1995), a feature more likely to be encountered in applications involving multiple treatments and nonlinear outcomes.

4. Which Parameters?

In many nonlinear models including the models for count data described here, direct interpretation of coefficients may be difficult. In such cases, marginal effects can be calculated as functions of estimated parameters and data. The effect of covariates such as insurance status can be measured using a parameter like the average treatment effect (ATE) within the framework of the potential outcome model (POM).

Let $D_i$ be a (1,0) binary indicator of insurance status, and let $y_{i1}$ denote use when insured, and $y_{i0}$ denote use when uninsured. In the framework POM, we interpret $D_i$ as an indicator of treatment. Assume that every element of the target population is potentially exposed to the treatment. Then the triple $\{(y_{i1}, y_{i0}, D_i, i = 1, \ldots, N)\}$ forms the basis of treatment evaluation, with $y_{i1}$ measuring the response if receiving treatment, and $y_{i0}$ when not receiving treatment. Since the receipt and nonreceipt of treatment are mutually exclusive states for individual $i$, only one of the two measures is available for any given $i$, and the unavailable measure is the counterfactual. The effect of the cause $D$ on the outcome of individual $i$ is measured by $(y_{i1} - y_{i0})$. The average causal effect of
$D_i = 1$, relative to $D_i = 0$, is measured by ATE:

$$ATE = E[y \mid D = 1] - E[y \mid D = 0]$$

where expectations are with respect to the probability distribution over the target population (see chapters by Auld and Polsky and Basu in this Companion for further discussion of ATEs).

The POM framework is simpler to apply if treatment is randomly assigned. But in observational data, random assignment of treatment is generally not a reasonable assumption. Then for consistent estimation of the ATE it is necessary to control for possible correlation between the outcomes and treatment. However, if we do so, then ATEs, or any other interesting quantile of the distribution of treatment effects, can be estimated, although one may need to resort to simulation in order to do so (see Deb and Trivedi, 2006).

The POM framework puts emphasis on the ATE as the key parameter of interest in policy analysis. This is reasonable if the treatment effect can be assumed to be constant. However, the more realistic assumption of heterogeneity in response to a policy change encourages a search for an alternative more flexible framework in which the effects of a covariate on the outcome may vary across expenditure quantiles. For example, raising the level of insurance deductible may impact lower quantiles much more than upper quantiles of expenditure. Quantile regression can be applied in such cases (Machado and Silva, 2005; Cameron and Trivedi, 2010); this methodology, originally developed for continuous outcomes can also be extended to counted data (Winkelmann, 2006). The presence of heterogeneity in response to treatment has also led to a large literature on other measures of policy impact, e.g. local average treatment effect (LATE).

5. Conclusion
In this chapter we have summarized the leading issues in econometric analyses of the use of health care services based on individual-level cross section data, both for the case of exogenous regressors and for the case with endogenous regressors. We have drawn from studies on the demand for medical care using counts of service use by Cameron et al. (1988), Chang and Trivedi (2003), Deb and Trivedi (1997, 2002), Kenkel and Terza (2001), Jimenez-Martín, Labeaga and Martinez-Granado (2002), Mullahy (1997), Munkin and Trivedi (1999, 2003), Pohlmeier and Ulrich (1995), Schellhorn (2001), Windmeijer and Santos-Silva (1997), and Winkelmann (2004).

The literature on count data models with endogenous regressors is still nascent and an active area of research. For example, while there are numerous applications of modified count-data models in the context of exogenous regressors, there are few applications when one or more regressor is endogenous. In addition, we have not described extensions of these models to panel data, another area in which the literature is growing.

References


