CEMA WORKING PAPER SERIES

Simulation-Based Estimation of the Structural Errors-in-Variables Negative Binomial Regression Model with an Application

Jie Q. Guo
Wylie Hall 105 Department of Economics
Indiana University
Bloomington, IN 47405

Tong Li
Wylie Hall 105 Department of Economics
Indiana University
Bloomington, IN 47405

Working Paper 46

May 2000

Working Paper 46, 1–22

Simulation-Based Estimation of the Structural Errors-in-Variables Negative Binomial Regression Model with an Application[*]

Jie Q. Guo

Wylie Hall 105 Department of Economics
Indiana University
Bloomington, IN 47405
E-mail: jiguo@indiana.edu

and

Tong Li[†]

Wylie Hall 105 Department of Economics
Indiana University
Bloomington, IN 47405
E-mail: toli@indiana.edu

This paper studies the effects and estimation of errors-in-variables negative binomial regression model. We prove that in the presence of measurement errors, in general, maximum likelihood estimator of the overdispersion using the observed data is biased upward. We adopt a structural approach assuming that the distribution of the latent variables is known and propose a simulation-based corrected maximum likelihood estimator and a simulation-based corrected score estimator to estimate the errors-in-variables negative binomial model. Though having similar asymptotic properties to the simulation-based corrected maximum likelihood estimator, the simulation-based corrected score estimator has a better finite sample performance as evidenced by the Monte Carlo studies. An application to the elderly demand for medical care using Medical Expenditure Panel Study is illustrated.

Key Words: Count Data; Measurement Errors; Overdispersion; Simulation-based Corrected Score Estimator; Health Care Demand.

JEL Classification Numbers: C13, C15, C51.

[*] We are grateful to Partha Deb, Pravin K. Trivedi, and the seminar participants at the Indiana University-Purdue University at Indianapolis for their comments and suggestions. Financial support to Li from the National Science Foundation (SES-0001663) is gratefully acknowledged.

[†] Corresponding author.

1. INTRODUCTION

There is an impressive body of work in biostatistics and econometrics using count regression models to explain the frequency of events, in which case, the dependent variable takes non-negative integer values. See, e.g. Cameron and Trivedi (1998) for a survey on count models in the regression context. Many count models that have been proposed and studied are variants of the Poisson model. As is well known, the Poisson model has limited applicability in practice because of its implication of equidispersion, that is, the variance of the dependent variable equals its mean both conditional on the explanatory variables. Among the variants of the Poisson regression model, the negative binomial model turns out to be a widely used one for its flexibility and parsimony. For the applications of negative binomial model, see, e.g., Hausman, Hall and Griliches (1984) on patents and R & D relationship, Cameron, Trivedi, Milne and Piggott (1988) on the determination of health service utilization and health insurance service, and Haab and McConnell (1996) on recreation demand analysis, among others.

This paper studies the effects of measurement errors in the negative binomial model and the estimation of a negative binomial model when the measurement errors are present. The negative binomial model has been widely applied in biostatistics and microeconometrics, where the presence of measurement errors is a common problem. See, e.g., Griliches (1986) for a thorough discussion of the errors-in-variables problems in micro data. As is well known, unobserved heterogeneity is the source of the overdispersion in the negative binomial model. An interesting question is: when the measurement errors are present in the explanatory variables, what their effect on overdispersion is conditional on the observed data. This question is important because if the dispersion of the data is affected by the measurement errors, then the standard negative binomial model is no longer valid to treat overdispersion as arising from unobserved heterogeneities. This paper proves that measurement errors in covariates will, in general, increase the extent of the "observed" overdispersion, compared with the negative binomial model without measurement errors.

Another issue of interest is how to estimate a negative binomial model with errors-in-variables. The estimation of errors-in-variables negative binomial model is a special case of the general nonlinear errors-in-variables models, which have been of great interest to statisticians and econometricians since 1980s. Carroll, Ruppert and Stefanski (1995) provide an extensive survey of the literature on the nonlinear errors-in-variables models. To the best of our knowledge, however, the estimation of errors-in-variables negative binomial model has not been studied yet. This paper follows a structural approach by assuming that the distribution of the true but unobserved variables is known and proposes a simulation-based corrected maximum likelihood estimator and a simulation-based corrected score estimator. Our estimators provide another application of the recently developed simulation based methods in latent variables models.[1] While sharing a similar asymptotic property to that of a simulation-based corrected maximum likelihood estimator, the simulation-based corrected score estimator proposed in the paper has better finite sample performance as found in our Monte Carlo studies.

We apply our methodology to study the elderly demand for medical care using data from the 1996 Medical Expenditure Panel Study (MEPS). We model counts of visits to physicians with income as one of the covariates. There has been a long-standing suspicion, however, about the accuracy of reported income in survey data. For example, Hausman, Newey and Powell (1995) apply the Hausman specification test to some Engel curves and reject the accuracy of reported income in the 1982 Consumer Expenditure Survey data. Using German Socio-Economic Panel survey data, Rendtel and Langeheine (1998) try to identify the potential effects of measurement error in income on poverty dynamics. Deb and Trivedi (1997) examine the elderly demand for medical care using the 1987 National Medical Expenditure Survey (NMES). It is noted that both NMES and MEPS, the latter of which is used in this paper, have similar sampling designs and are implemented by the same agency - Agency for Health Care Policy and Research (AHRQ). One of models they use is the negative binomial regression. However, they do not discuss the possible consequence of using potentially error contaminated reported income. We estimate the demand model by both the “naive” negative binomial model and the corrected negative binomial model assuming that there may be measurement errors in the reported income. The comparison of the results from these two models suggest that income reported in MEPS is accurately measured.

This paper is organized as follows. In section 2, we prove that measurement errors in covariates will increase the extent of overdispersion on the observed data if the true data generating process is the negative binomial. Section 3 proposes a simulation-based corrected maximum likelihood estimator and a simulation-based corrected score estimator for the errors-in-variables negative binomial models. Section 4 is devoted to the Monte Carlo studies, which demonstrate the good finite sample performance of the simulation-based corrected score estimator. Section 5 presents an application to the elderly demand for medical care using the 1996 MEPS data. Section 6 concludes.

[1] For a survey on the simulation based methods developed by McFadden (1989) and Pakes and Pollard (1989), see Gourieroux and Monfort (1996). Also, Li (2000) proposes a simulated minimum distance estimator in the estimation of the structrual errors-in-variables models.

2. OVERDISPERSION CAUSED BY ERRORS-IN-VARIABLES

The standard negative binomial model can be regarded as a mixture of the Poisson model with a random variable that follows a gamma distribution. Such a gamma distributed random variable is used to control for the unobserved heterogeneity that gives rise to overdispersion. In particular, a negative binomial model has a probability mass function

$$f(y_i|\lambda_i) = \frac{\Gamma(y_i+v)}{\Gamma(y_i+1)\Gamma(v)}\left(\frac{v}{v+\lambda_i}\right)^{v}\left(\frac{\lambda_i}{v+\lambda_i}\right)^{y_i}, \tag{1}$$

where $v^{-1} = \alpha$ $(\alpha > 0)$ is a scalar (overdispersion) parameter. In the regression framework where some explanatory variables $\mathbf{x}$ are included, λ_i is usually specified as $\lambda_i = \exp(\beta'\mathbf{x}_i)$ leading to a negative binomial regression model. The conditional mean and variance are given by

$$\begin{aligned} \mathsf{E}(y_i|\mathbf{x}_i) &= \exp(\beta'\mathbf{x}_i), & (2)\\ \mathsf{var}(y_i|\mathbf{x}_i) &= \left[1+v^{-1}\mathsf{E}(y_i|\mathbf{x}_i)\right]\mathsf{E}(y_i|\mathbf{x}_i), & (3) \end{aligned}$$

respectively. It is clear from (3) that the ratio of the conditional variance to the mean is a linear function of the conditional mean. (3) also indicates that $\mathsf{var}(y_i|\mathbf{x}_i) > \mathsf{E}(y_i|\mathbf{x}_i)$. As a result, the negative binomial model explicitly takes into account the overdispersion. The Poisson model is obtained as a limiting case of the negative binomial model as $v \to \infty$.

This paper considers the case where covariates are measured with errors. Specifically, we assume that $\mathbf{x}$, a vector of K covariates, is unobserved; instead, we observe $\mathbf{z}$ such that $\mathbf{z} = \mathbf{x} + \epsilon$ where ϵ are independent of $\mathbf{x}$ with mean 0 and variance-covariance matrix Ω.[2] While the conditional mean and variance of y given $\mathbf{x}$ have the relationship given by (2) and (3), it is interesting to investigate the relationship between the conditional mean and variance of y given $\mathbf{z}$ when the measurement errors are present and only $\mathbf{z}$ are observed. The next proposition establishes such a relationship.

PROPOSITION 2.1. *Consider a negative binomial model that has the feature given by (2) and (3) with* $\mathbf{z} = \mathbf{x} + \epsilon$ *where* ϵ *are independent of* $\mathbf{x}$ *with mean 0 and variance-covariance matrix* Ω*, then with measurement errors in covariates, we have*

$$\mathsf{var}[y|\mathbf{z}] \geq \left[1+v^{-1}\mathsf{E}[y|\mathbf{z}]\right]\mathsf{E}[y|\mathbf{z}],$$

[2] For ease of exposition, we assume that all the explanatory variables are measured with errors. Nonetheless, it is general enough to include the case where only some of the explanatory variables are measured with errors, which will be the case in our application.

where the equality holds only when $\mathsf{E}[y|\mathbf{x}] = 1$ *or the conditional density of* y *given* $\mathbf{x}$ *is a.e. 0.*

Proposition 2.1 indicates that given that the true data generating process is the negative binomial regression model and the overdispersion satisfies (3) for $(y,\ \mathbf{x})$, the extent of "observed" overdispersion will generally be larger for $(y,\ \mathbf{z})$. This is an interesting result which implies that if the true model is the negative binomial model and some explanatory variables are measured with errors, direct estimation of the negative binomial model using the observed data will yield upward biased estimates of the overdispersion parameter. Also, Proposition 2.1 can be extended to more general cases such that conditional mean and variance of y given the latent variables $\mathbf{x}$ satisfy

$$\mathsf{var}(y_i|\mathbf{x}_i) = \left[1 + v^{-1}\mathsf{E}^k(y_i|\mathbf{x}_i)\right]\mathsf{E}(y_i|\mathbf{x}_i),$$

where k is a positive integer. The variation of k allows different rates of increment in the conditional variance. In particular, k equals 0 for a negative binomial-1, equals 1 for a negative binomial-2 (Cameron and Trivedi (1986)), equals 2 for a Poisson inverse Gaussian regression (Dean, Lawless, and Willmot (1989)), equals 1 and v equals 1 for a geometric model. A larger k can be used to accommodate highly overdispersed data.

3. ESTIMATION OF THE ERRORS-IN-VARIABLES NEGATIVE BINOMIAL REGRESSION MODEL

As indicated by Proposition 2.1, measurement errors increase the extent of the "observed" overdispersion if the true data generating process is the negative binomial. One implication of this result, as mentioned earlier, is that if one ignores measurement errors and uses a "naive" negative binomial model, which treats contaminated $\mathbf{z}$ as a true variable, upward biased estimate of the overdispersion parameter will occur. On the other hand, as a folklore in econometrics, measurement errors may cause downward biases in the estimation of coefficients of latent variables in a nonlinear model (see, e.g. Hausman et al. (1995) and Chesher (1991) for some discussion of the so-called "attenuation" caused by errors-in-variables). Given the popularity of the negative binomial model in count data analysis and the importance of measurement errors problem, it is surprising that little work has been done in the estimation of errors-in-variables negative binomial model.[3] This section is devoted to proposing a simulation-based corrected

[3] For the estimation of errors-in-variables Poisson regression model, see Nakamura (1990) and Guo and Li (2000).

maximum likelihood estimator and a simulation-based corrected score estimator for estimating errors-in-variables negative binomial models.

As assumed in Section 2, while we observe y_i, $i = 1, \ldots, n$, we do not observe $\mathbf{x}_i$ but instead we observe $\mathbf{z}_i = \mathbf{x}_i + \varepsilon_i, .i = 1, 2, ..., n$. Suppose that the true data generating process of the dependent variable is the negative binomial regression model with the density function

$$\Pr(Y_i = y_i|\mathbf{x}_i) = \frac{\Gamma(y_i + v)}{\Gamma(y_i + 1)\Gamma(v)} \left(\frac{\lambda_i(\beta_0)}{v + \lambda_i(\beta_0)}\right)^{y_i} \left(\frac{v}{v + \lambda_i(\beta_0)}\right)^{v},$$

where $\lambda_i(\beta_0) = \exp(\beta_0'\mathbf{x}_i)$. Then the maximum likelihood estimate (MLE) $\hat{\beta}_{MLE}$ based on $(y_i, \mathbf{x}_i)$, $i = 1, 2, ..., n$ is consistent. The (average) log-likelihood function using $(y_i, \mathbf{x}_i)$, $i = 1, 2, ..., n$ is

$$L_0 = \frac{1}{n}\sum_{i=1}^{n}\left[\ln\frac{\Gamma(y_i + v)}{\Gamma(y_i + 1)\Gamma(v)} + v\ln v + y_i\beta'\mathbf{x}_i - (y_i + v)\ln(v + \exp(\beta'\mathbf{x}_i)\right],$$

which converges to

$$q = \mathsf{E}_{\mathbf{x},y}\left[\ln\frac{\Gamma(y + v)}{\Gamma(y + 1)\Gamma(v)} + v\ln v + y\beta'\mathbf{x} - (y + v)\ln(v + \exp(\beta'\mathbf{x}))\right]. \tag{4}$$

The "naive" (average) log-likelihood function is

$$\begin{aligned} L_1 &= \frac{1}{n}\sum_{i=1}^{n}\left[\ln\frac{\Gamma(y_i + v)}{\Gamma(y_i + 1)\Gamma(v)} + v\ln v + y_i\beta'\mathbf{z}_i - (y_i + v)\ln(v + \exp(\beta'\mathbf{z}_i))\right] \\ &= L_0 + \frac{1}{n}\sum_{i=1}^{n}\left[y_i\beta'\varepsilon_i + (y_i + v)(\ln(v + \exp(\beta'\mathbf{x}_i)) - \ln(v + \exp(\beta'\mathbf{z}_i)))\right], \end{aligned}$$

which no longer converges to q because the second term in the last line does not converge to 0. Therefore, using a "naive" negative binomial model will result in biased estimates in general.[4] However,

$$L_1 - \frac{1}{n}\sum_{i=1}^{n}\left[y_i\beta'\varepsilon_i + (y_i + v)(\ln(v + \lambda_i(\mathbf{x}_i)) - \ln(v + \lambda_i(\mathbf{z}_i)))\right],$$

or equivalently,

$$\frac{1}{n}\sum_{i=1}^{n}\left[\ln\frac{\Gamma(y_i + v)}{\Gamma(y_i + 1)\Gamma(v)} + v\ln v + y_i\beta'\mathbf{z}_i - (y_i + v)\ln(v + \exp(\beta'\mathbf{x}_i))\right], \tag{5}$$

[4] As discussed in Chesher (1991), in general, $\mathsf{E}(y|\mathbf{z}) \neq \mathsf{E}(y|\mathbf{x})$. As a result, the pseudo maximum likelihood estimation does not yield consistent estimates either.

still converges to (4). We refer to (5) as a corrected log-likelihood function of the errors-in-variables negative binomial model. The corresponding corrected score equations are

$$\beta \ : \ \frac{1}{n}\sum_{i=1}^{n}\left[y_i\mathbf{z}_i - \frac{v+y_i}{v+\exp(\beta'\mathbf{x}_i)}\exp(\beta'\mathbf{x}_i)\mathbf{x}_i\right] = 0, \tag{6}$$

$$\begin{aligned} v \ &: \ \frac{1}{n}\sum_{i=1}^{n}\left[(\ln\frac{\Gamma(y_i+v)}{\Gamma(y_i+1)\Gamma(v)})' + 1 + \ln v \right.\\ &- \left.\ln(v+\exp(\beta'\mathbf{x}_i)) - \frac{v+y_i}{v+\exp(\beta'\mathbf{x}_i)}\right] = 0. \end{aligned} \tag{7}$$

In principle, (5) can be maximized to obtain consistent estimates for β and v. Such a maximization problem, however, is infeasible to implement in practice due to the fact that the last term in (5) involves the unobservables $\mathbf{x}$. To resolve this problem, we adopt a structural approach by assuming the distribution of the latent variables $\mathbf{x}$ is known.[5] If this is the case, then the maximization of (5) can be implemented using the simulation based methods.

To proceed, note that the last quantity $\frac{1}{n}\sum_{i=1}^{n}(y_i+v)\ln(v+\exp(\beta'\mathbf{x}_i))$ of (5) converges to

$$\begin{aligned} &\mathsf{E}_{\mathbf{x},y}\left[(y+v)\ln(v+\exp(\beta'\mathbf{x}))\right] \\ = &\int\left[\int(y+v)g(y|\mathbf{x})dy\right]\ln(v+\exp(\beta'\mathbf{x}))f(\mathbf{x})d\mathbf{x}. \end{aligned} \tag{8}$$

As a result, if the former is replaced by the latter in (3.2), we have that

$$\frac{1}{n}\sum_{i=1}^{n}\left[\ln\frac{\Gamma(y_i+v)}{\Gamma(y_i+1)\Gamma(v)} + v\ln v + y_i\beta'\mathbf{z}_i\right] - \mathsf{E}_{\mathbf{x},y}\left[(y+v)\ln(v+\exp(\beta'\mathbf{x}))\right] \tag{9}$$

converges to (4) as well. Therefore, maximization of (9) also yields consistent estimates. Although maximization of (9) is straightforward in some cases when $\mathbf{x}$ is a scalar, in general it is much more involved if $\mathbf{x}$ is a multivariate vector. This is due to the multiple integrals involved in the expectation with respect to $\mathbf{x}$.[6] This difficulty highlights the benefit of

[5] This assumption is not as strong as it might seem. In fact, it has been a common practice in the estimation of structural nonlinear errors-in-variables models to assume that the functional forms of latent distributions are known. Such an assumption can be justified in those situations where there are replications for the latent variables or the validation data are available. See, e.g., Hsiao (1989, 1992) and Lee and Sepanski (1995).

[6] In some cases even with scalar $\mathbf{x}$, the expectation involving an integral is also difficult to compute.

using the simulation based methods. Specifically, from (8),

$$\mathsf{E}_{\mathbf{x}}\left[(v+\exp(\beta'\mathbf{x}))\ln(v+\exp(\beta'\mathbf{x}))\right],$$

can be approximated by its unbiased estimator

$$\frac{1}{S}\sum_{s=1}^{S}\left[\ln(v+\exp(\beta'\tilde{\mathbf{x}}_s))(v+\exp(\beta'\tilde{\mathbf{x}}_s))\right],$$

where $\tilde{\mathbf{x}}_s$, $s=1,\ldots,S$, are drawn from the density function of $\mathbf{x}$. Consequently, one can now maximize

$$\begin{aligned}\tilde{q}_n &= \frac{1}{n}\sum_{i=1}^{n}\left[\ln\frac{\Gamma(y_i+v)}{\Gamma(y_i+1)\Gamma(v)}+v\ln v+y_i\beta'\mathbf{z}_i\right.\\ &- \left.\frac{1}{S}\sum_{s=1}^{S}\ln(v+\exp(\beta'\tilde{\mathbf{x}}_s))(v+\exp(\beta'\tilde{\mathbf{x}}_s))\right].\end{aligned} \tag{10}$$

Since the estimator obtained by maximizing (10) is a (partially) simulation-based corrected likelihood estimator, it has the same properties as simulated maximum likelihood (SML) estimators that have been extensively studied recently (see, e.g., Hajivassiliou (1997)). In particular, the estimator is consistent as S and n both go to infinity.

As indicated in Hajivassiliou (1997), sometimes, SML estimators do not perform well in finite samples. This turns out to be true for our (partially) simulation-based corrected likelihood estimator as well.[7] Poor finite sample performance of the SML estimators is mainly due to the fact that $\mathsf{E}\tilde{q}_n = q$ does not necessarily imply $\mathsf{E}\arg\max\tilde{q}_n = \arg\max q$ (Hajivassiliou (1997)). To alleviate the instability of the simulation-based (corrected) maximum likelihood estimation, we propose to work on the score equations (6) and (7). This is motivated by the fact that the instability of the simulation-based corercted maximum likelihood estimator is caused by the quite different presentation of score equation of (10) from the original one. As the comparision reveals, simulation in the log-likelihood function introduces additional variation that cannot be ignored.

Note that the terms involving the observed $\mathbf{x}$ in the corrected score equations (6) and (7) satisfy

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{v+y_i}{v+\exp(\beta'\mathbf{x}_i)}\exp(\beta'\mathbf{x}_i)\mathbf{x}_i\right] \rightarrow \mathsf{E}\left[\exp(\beta'\mathbf{x})\mathbf{x}\right], \tag{11}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[1-\ln(v+\exp(\beta'\mathbf{x}_i))-\frac{v+y_i}{v+\exp(\beta'\mathbf{x}_i)}\right] \rightarrow \mathsf{E}\left[-\ln(v+\exp(\beta'\mathbf{x}))\right] \tag{12}$$

[7] The Monte Carlo results are available from the authors upon request.

As a result, these two terms in (6) and (7) can be approximated by their corresponding probability limits given in (11) and (12), respectively, which in turn can be approximated by simulations if the expectations are difficult to calculate. Specifically, we can draw $\tilde{\mathbf{x}}_s$, $s = 1, \ldots, S$, from the density of $\mathbf{x}$ and obtain simulation-based corrected score estimators $\hat{\beta}$, $\hat{v}$ by solving the following equations

$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i\mathbf{z}_i - \frac{1}{S}\sum_{s=1}^{S}\exp(\hat{\beta}'\tilde{\mathbf{x}}_s)\tilde{\mathbf{x}}_s\right] = 0, \quad (13)$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[(\ln\frac{\Gamma(y_i+\hat{v})}{\Gamma(y_i+1)\Gamma(\hat{v})})' + \ln\hat{v} - \frac{1}{S}\sum_{s=1}^{S}\ln(\hat{v}+\exp(\hat{\beta}'\tilde{\mathbf{x}}_s))\right] = 0. \quad (14)$$

As for the simulation-based maximum likelihood estimator, the consistency of the simulation-based corrected score estimator requires that both n and S go to infinity.[8] The following result gives its asymptotic distribution.

Proposition 3.2. *Consider a negative binomial model, with assumption that both* n *and* $S \to \infty$, *and* $n/S \to 0$, $\sqrt{n}(\tilde{\Theta}-\Theta_0)$ *converges in distribution to* $N(0, I^{-1}(J_1 + J_2 + J_3)I^{-1})$, *where* I, J_1, J_2, *and* J_3 *are defined in the proof of this proposition in the Appendix.*

Although it has similar asymptotic properties to a simulation-based corrected maximum likelihood estimator, the simulation-based corrected score estimator performs quite well in a Monte Carlo study reported in the next section.

4. MONTE CARLO EXPERIMENTS

To assess the finite sample properties of our simulation-based corrected score estimator, we conduct Monte Carlo experiments using a regression model with only one regressor.

The experiments are designed as follows. The latent variable x is normally distributed with mean $\mu = 1.1$ and standard deviation $\sigma_x = 0.4$. The dependent variable y is generated from a negative binomial model with the

[8] Because of special feature of our problem, it is difficult to interpret corrected score equations (6) and (7) as simple analogs to some moment conditions constructed from residuals that are differences between the observed y and conditional expectations. As a result, we are unable to construct a method of simulated score estimator in the sense of Hajivassiliou and McFadden (1998). Our simulation-based corrected score estimator, on the other hand, can be viewed essentially as a simulation-based corrected maximum likelihood estimator, whose consistency requires that both the sample size and number of simulations go to infinity.

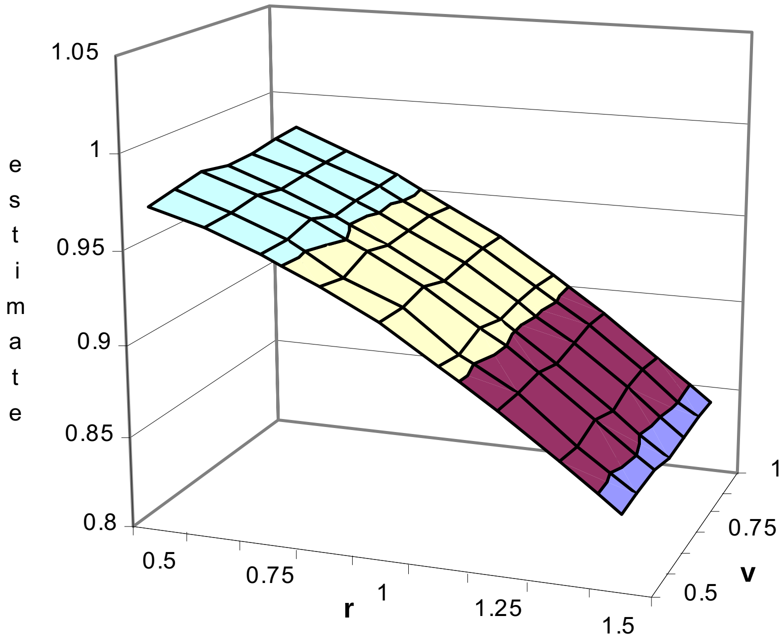
covariate x and true value of the parameter $\beta_0 = 1$. We vary overdispersion parameter v_0 from 0.5, 2/3, to 1 in experiments. Smaller value of v_0 means greater overdispersion of the data. The measurement error ϵ is distributed as $N(0, \sigma_\epsilon^2)$, where σ_ϵ varies in experiments from 0.6, 0.5, 0.3 to 0.2. In this setting, we define r to be the ratio of σ_ϵ and σ_x ($r = \sigma_\epsilon/\sigma_x$), which is used as a measure for the relative magnitude of measurement error. Intuitively, when r is relative small, the bias from using the "naive" negative binomial model should also be small because the magnitude of variation of regressor dominates that of the variation of the error. The observed variable z is generated from $z = x + \epsilon$, where x and ϵ are independently simulated. Each experiment is replicated 500 times with 500 observations in each replication. For the choice of S, we choose S to be 1000, twice of the sample size. Also, for comparability, we keep the same random number generating seeds for each experiment. Of course, the seed for data generation within an experiment changes. To summarize, we vary two values in experiments: the standard deviation of the measurement error σ_ϵ and the overdispersion parameter v.

The results of Monte Carlo experiments are given in Table 1, where $\hat{\beta}_T$ is the estimate from true negative binomial model that uses observations on true x, $\hat{\beta}_C$ is the simulation-based corrected score estimate and $\hat{\beta}_N$ is the estimate from the "naive" model using the observations on z. We define R $(= (\hat{\beta} - \beta_0)/\beta_0$ or $(\hat{v} - v_0)/v_0))$ as the relative bias for estimate. The simulations show that the "naive" negative binomial model always gives upward biased estimate of the overdispersion as established in Proposition 2.1. It is interesting to see that there is a downward biased estimate of the slope, which is consistent with the attenuation effect as described in Chesher (1991) and Hausman et al. (1995). Also, as anticipated, the bias arising from the "naive" model is small when r is small. For instance, when v_0 is fxed at 0.5 and r declines from 1.5 to 0.5, the relative biases for $\hat{\beta}_N$ reduces from -15.9% to only -2.2% while the relative biases for $\hat{v}_N$ reduces from -16.0% to -1.8%. This dramatical change pattern, however, cannot be observed by changing v and fixed r. The simulation-based corrected score estimates perform consistently well regardless of the size of v and r. The negative binomial model using true x as explanatory variable gives nice estimates. It is reasonable because the model uses the accurate data and correctly specifies the data generating process.

These results can be seen more clearly from Figures 1 and 2, which show the estimates of β from corrected and naive negative binomial models, respectively, where the true $\beta = \beta_0 = 1$. The X axis is the relative magnitude of measurement error r (it also indirectly demonstrates the impact of σ_ϵ, because $r = \sigma_\epsilon/\sigma_x$ and σ_x is fixed as 0.4 in all experiments), the Y axis is the overdispersion parameter v, and the Z axis is the estimate of β . Figure 1 reveals that the ratio r has more effects on "naive" negative binomial estimate than overdispersion, while corrected score estimates are not affected significantly by either ratio or overdispersion. Also because we keep the same seeds in all experiments, it is not surprising to see in Table 1 that the estimates $\hat{\beta}_T$ and $\hat{v}_T$ for the true model are the same in all experiments. The simulation is programed using SAS and run on a 160MHz POWER2SuperChip processor in an IBM RS/6000 Scalable POWER parallel System. It takes about 75 minutes CPU time to obtain results for Table 1.

The following figures show the estimates of slope β from naive negative binomial and corrected negative binomial models. The true value $\beta_0 = 1$. The X axis is the relative magnitude of measurement error r, the Y axis is the overdispersion v, and the Z axis is the estimate of β.

FIG. 1. parameter estimate of $\hat{\beta}_N$



5. APPLICATION

The data (HC-003) are obtained from the 1996 Medical Expenditure Panel Survey (MEPS) Household Component (HC), which are collected by the Agency for Health Care Policy and Research (AHRQ) (1998). MEPS provides information about financing and use of medical care in the United

TABLE 1.

	$v_0 = 0.5$			$v_0 = 2/3$			$v_0 = 1$		
	Mean	Std Err	$R(\%)$	Mean	Std Err	$R(\%)$	Mean	Std Err	$R(\%)$
$\hat{\beta}_T$	0.995	0.057	-0.5	0.998	0.051	-0.2	0.998	0.045	-0.2
$\hat{v}_T$	0.504	0.042	0.8	0.675	0.062	1.2	1.005	0.094	0.5
$\hat{\beta}_C$	0.995	0.072	-0.5	0.998	0.068	-0.2	0.999	0.060	-0.1
$\hat{v}_C$	0.510	0.065	2.0	0.685	0.101	2.7	1.039	0.244	3.9
$\hat{\beta}_N$	0.841	0.060	-15.9	0.842	0.053	-15.8	0.839	0.046	-16.1
$\hat{v}_N$	0.420	0.038	-16.0	0.539	0.050	-19.2	0.751	0.072	-24.9

$\sigma_\epsilon = 0.5, r = 1.25.$

	$v_0 = 0.5$			$v_0 = 2/3$			$v_0 = 1$		
	Mean	Std Err	$R(\%)$	Mean	Std Err	$R(\%)$	Mean	Std Err	$R(\%)$
$\hat{\beta}_T$	0.995	0.057	-0.5	0.998	0.051	-0.2	0.998	0.045	-0.2
$\hat{v}_T$	0.504	0.042	0.8	0.675	0.062	1.2	1.005	0.094	0.5
$\hat{\beta}_C$	0.995	0.070	-0.5	0.998	0.066	-0.2	0.999	0.059	-0.1
$\hat{v}_C$	0.510	0.062	2.0	0.683	0.097	2.4	1.032	0.199	3.2
$\hat{\beta}_N$	0.888	0.059	-11.0	0.889	0.053	-11.8	0.885	0.046	-11.5
$\hat{v}_N$	0.440	0.039	-12.0	0.567	0.052	-14.7	0.803	0.076	-19.7

$\sigma_\epsilon = 0.3, r = 0.75.$

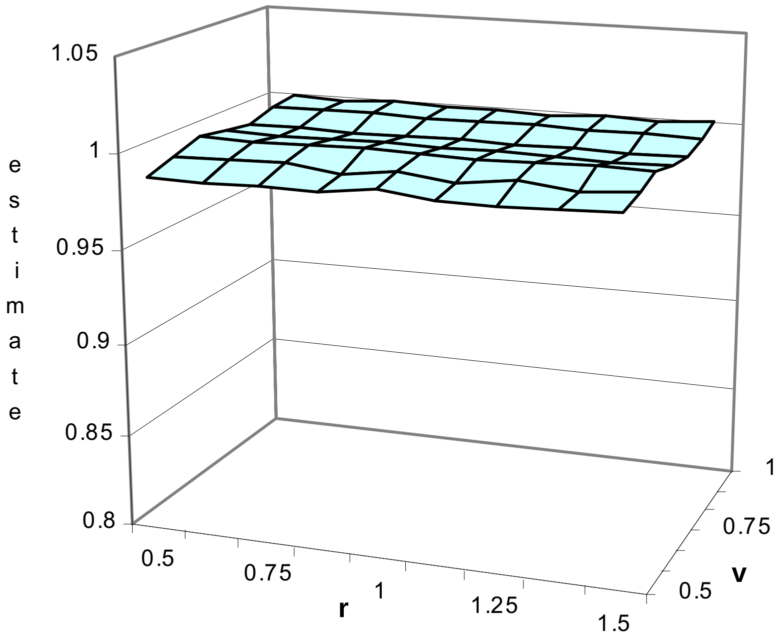
	$v_0 = 0.5$			$v_0 = 2/3$			$v_0 = 1$		
	Mean	Std Err	$R(\%)$	Mean	Std Err	$R(\%)$	Mean	Std Err	$R(\%)$
$\hat{\beta}_T$	0.995	0.057	-0.5	0.998	0.051	-0.2	0.998	0.045	-0.2
$\hat{v}_T$	0.504	0.042	0.8	0.675	0.062	1.2	1.005	0.094	0.5
$\hat{\beta}_C$	0.995	0.068	-0.5	0.998	0.064	-0.2	0.999	0.057	-0.1
$\hat{v}_C$	0.509	0.059	1.8	0.681	0.090	2.1	1.027	0.173	2.7
$\hat{\beta}_N$	0.956	0.059	-4.4	0.959	0.052	-4.1	0.956	0.046	-4.4
$\hat{v}_N$	0.477	0.041	-4.6	0.629	0.057	-5.7	0.913	0.085	-8.7

$\sigma_\epsilon = 0.2, r = 0.50.$

	$v_0 = 0.5$			$v_0 = 2/3$			$v_0 = 1$		
	Mean	Std Err	$R(\%)$	Mean	Std Err	$R(\%)$	Mean	Std Err	$R(\%)$
$\hat{\beta}_T$	0.995	0.057	-0.5	0.998	0.051	-0.2	0.998	0.045	-0.2
$\hat{v}_T$	0.504	0.042	0.8	0.675	0.062	1.2	1.005	0.094	0.5
$\hat{\beta}_C$	0.995	0.067	-0.5	0.999	0.063	-0.1	0.998	0.056	-0.2
$\hat{v}_C$	0.509	0.058	1.8	0.680	0.088	1.9	1.026	0.168	2.6
$\hat{\beta}_N$	0.978	0.058	-2.2	0.980	0.051	-2.0	0.979	0.045	-2.1
$\hat{v}_N$	0.491	0.042	-1.8	0.653	0.059	-2.5	0.960	0.090	-4.0

There are 500 replications with 500 observations in each replication. The latent variable x is distributed as $N(1.1, 0.4)$. The dependent variable y is generated from a negative binomial model with the covariate x and true value of the parameter $\beta_0 = 1$. The observed variable z is generated from $z = x + \epsilon$, where x and ϵ are independently generated. The measurement error ϵ is distributed as $N(0, \sigma_\epsilon)$, where σ_ϵ varies from 0.6, 0.5, 0.3 to 0.2. $r = \sigma_\epsilon \sigma_x$ is the relative magnitude of the measurement error. The overdispersion parameter v varies from 0.5 2/3, to 1. In the table, $\hat{\beta}_T$ is the estimate from the true negative binomial model based on true x, both $\hat{\beta}_C$, the simulated corrected score estimate, and $\hat{\beta}_N$, the estimate from the "naive" model, are based on observations on z. $S = 1000$. R is the percentage of relative bias to the true value. $\sigma_\epsilon = 0.6, r = 1.50.$

FIG. 2. parameter estimate of $\hat{\beta}_C$



States. The HC-003 is designed for three interviews covering calendar year 1996 to provide a nationally representative sample of the U.S. civilian non-institutionalized population on health care utilization. The interview interval is fairly close. The public use dataset contains 23,230 persons.

It has been documented in the literature that the demand behaviors for medical services between the elderly and young people are significantly different (see e.g. Deb and Trivedi, (1997)). Thus, we focus on those people who are 65 or over. After reorganizing the data, a sample of 2218 observations is obtained.

The HC-003 file we use contains two rounds survey (information for some variables from round three is also available). The main variables in MEPS are classified into six groups, namely, survey administration, demographic, employment, health status, health insurance, and utilization. Except administration information, we select related variables from each of groups. The demographic variables include AGE, MALE, BLACK, MARRIED (marital status), DEGREE (education level), NOREAST, MIDWEST, and

WEST (used to control the regional difference), and INCOME.[9] The health variables contain POORHLTH, EXCLHLTH (both are self-perceived measures of health status), ADLHELP (a measure of disability status), and COGLIMT (a measure of cognitive ability). The employment and health insurance variables contain EMPLOYED (employment status), MEDICAID, PRIVINS (supplementary private insurance).[10] The number of physician hospital outpatient visits (OPDRV) is treated as the dependent variable. The variable definitions and summary statistics are given in Table 2.

We pay special attention to the accuracy of income. To examine whether there is a measurement error in the income variable, we apply the method discussed in previous sections. Adopting a similar approach used in Hausman et al. (1995), we utilize income from first two rounds (INCOME1, INCOME2). Because the income sources for the elderly are more stable than for the youth and the two rounds are interviewed at close time periods, which greatly eliminates the income variation due to job change and other unobserved factors, we view INCOME1 and INCOME2 as repeated measures of the true income variable. Following Lewbel (1996), we assume that the true income variable is distributed as a log-normal. Assuming that the recorded income from the survey is equal to the sum of the true income and a measurement error that is independent of the true income, the mean of the recorded income therefore equals the mean of the true but unobserved income.

With the mean and variance calculated as above, the simulated income can be easily generated. The mean and standard deviation of simulated income are listed at the bottom of Table 2. As can be seen, they are fairly close to but smaller than the sample mean and standard deviation, which implies that the sample data contain some extreme values of income. Actually, we find that the largest weekly income is \$1538.40 (or 7.338 after log-transformation), while 99% sample has weekly income less than \$590.80 (or 6.369 after log-transformation).

Both naive and corrected negative binomial regression models are estimated. The results are given in Table 3. Since the asymptotic variance-covariance matrix of the simulation-based corrected score estimator as given in Proposition 3.1 is complicated, we calculate the standard errors of this estimator using bootstrap. To make it comparable, we also use bootstrap to calculate the standard errors of the estimates from the "naive"

[9] Since the original dataset does not contain the income variable, we calculate it by using the product of weekly work hours and hourly wage rate. Actually, the weekly work hours and hourly wage rate reported in the survey are imputed from a person's income based on the time period on which the income was based and the number of hours worked per time period. So our calculation is essentially a reverse step.

[10] Cartwright, Hu, and Huang (1992) and Deb and Trivedi(1997) have shown that the supplementary private insurance in a utilization regression for people 65 and over is exogeneous.

TABLE 1.

Variable definitions and summary statistics

Variable	Definition	Mean	Std. dev.
	utilization		
OPDRV	number of physician hospital outpatient visits	0.431	1.808
	health insurance		
MEDICAID	= 1 if the person is a recipient of Medcaid	0.119	0.324
PRIVINS	= 1 if the person is covered by private insurance	0.578	0.494
	health employment		
EMPLOYED	= 1 if the person is employed	0.094	0.292
	health status		
POORHLTH	= 1 if perceived health status of the person is poor	0.092	0.288
EXCLHLTH	= 1 if perceived health status of the person is excellent	0.171	0.376
ADLHELP	= 1 if activities of daily living need help	0.087	0.283
COGLIMT	= 1 if the person has cognitive limitation	0.127	0.333
	demongraphic		
NOREAST	= 1 if the person lives in northeastern US	0.223	0.416
MIDWEST	= 1 if the person lives in midwestern US	0.239	0.427
WEST	= 1 if the person lives in western US	0.203	0.403
MALE	= 1 if the person is male	0.390	0.488
AGE	log-transferred age	4.300	0.087
BLACK	= 1 if the person is African American	0.120	0.325
MARRIED	= 1 if the person is married	0.959	0.199
DEGREE	highest degree (1-no degree;2-GED;3-high school;4-bachelor;5-master;6-ph.d;0-other)	2.204	1.263
INCOME1	log-transferred weekly income in round 1	0.451	1.494
INCOME2	log-transferred weekly income in round 2	0.459	1.478
INCOME	simulated log-transferred weekly income	0.424	1.404

TABLE 2.

Parameter estimates of “naive” and corrected negative binomial regression models

	Negative binomial		Corrected Negative binomial	
MEDICAID	0.725**	(0.34)	0.722**	(0.35)
PRIVINS	0.114	(0.16)	0.114	(0.18)
EMPLOYED	0.072	(0.44)	0.070	(0.34)
POORHLTH	0.613**	(0.29)	0.611**	(0.30)
EXCLHLTH	−0.141	(0.21)	−0.140	(0.21)
ADLHELP	−0.620*	(0.34)	−0.621*	(0.35)
COGLIMT	0.071	(0.32)	0.069	(0.33)
NOREAST	0.452**	(0.23)	0.452**	(0.23)
MIDWEST	0.140	(0.18)	0.140	(0.18)
WEST	−0.357*	(0.22)	−0.357*	(0.22)
MALE	0.384**	(0.14)	0.379**	(0.15)
AGE	−2.540**	(0.85)	−2.561**	(0.85)
BLACK	0.121	(0.28)	0.119	(0.30)
MARRIED	−0.071	(0.37)	−0.076	(0.38)
DEGREE	0.009	(0.06)	0.045	(0.09)
INCOME	−0.093	(0.09)	−0.040	(0.04)
Constant	9.667**	(3.69)	9.662**	(3.69)
v	0.171**	(0.02)	0.170**	(0.02)
HAUSMAN’S TEST			0.146	

Notes: standard errors are in parentheses.
* statistically significant at the 10% level.
** statistically significant at the 5% level.

negative binomial model. The estimates from both models are quite close, which makes the existence of the measurement errors suspicious. To further investigate, we apply Hausman's specification test. The idea of this test is that under the null hypothesis of no measurement error, both "naive" negative binomial and corrected negative binomial give consistent estimators, while the former is more efficient. If the null is false, however, then only corrected negative binomial gives consistent estimates. The Hausman's test statistics is 0.146, which is not high enough to reject the null that there is no measurement error. Table 3 also shows that most variables have smaller standard errors (more efficient) in "naive" than corrected negative binomial model.

A closer look of the results indicates the following. First, Medicaid insurance coverage has strong positive effects on the outpatient visits, while this is not shown for private insurance status (PRIVINS). Considering Medicaid as an insurance plan for the poor, it confirms, at least partially, the importance of this policy for vulnerable population in society. Second, the income effect for the elderly is not significant, which has been also found by Cartwright et al. (1992) and Deb and Trivedi (1997). This result can be explained by the dominance of insurance coverage on the income effect. The insensitivity of utilization to marginal changes in income may be due to the generosity of Medicare that is irrespective of family income. The employment status is not significant either. This finding may be understandable, as when a certain level of income (e.g. Social Security Income) for the elderly is guaranteed, whether to hold a job is no longer as important as is for young people. On the other hand, even if a senior citizen is employed, the capability of his or her physical condition still limits the amount of income he or she could receive from the job. Therefore, the marginal utility of employment is negligible. Third, the health status affects the utilization of medical care, especially for those who are with poor self-perceived health status. The limitation of daily living activities, however, decreases the number of visits, which may reflect the inconvenience for those with difficulty in daily living activities to access the services. Similarly to the mobility issue for those who have difficulty in daily living activities, people tend to decrease the number of outpatient visits with age going up. This is illustrated in Deb and Trivedi (1997) as well. Also men seek outpatient visits more often than women, which might contradict the conventional view. But the race, marital status, and education level have no significant impact on the utilization, which are consistent with previous findings.

6. CONCLUSION

This paper considers the effects of measurement errors in the negative binomial regression models.We show that in general the errors in covariates increase the degree of the "observed" overdispersion and also result in biased estimates for mean parameters if the model is estimated with the "naive" method. This also justifies the upward bias in the overdispersion parameter if the "naive" model is used ignoring the measurement errors. We also propose a simulation-based corrected maximum likelihood estimator and a simulation-based corrected score estimator to consistently estimate the errors-in-variables negative binomial model assuming that the distribution of the latent variables is known. The Monte Carlo study shows that the simulation-based corrected score estimator is preferred to the simulation-based corrected maximum likelihood estimator in finite samples although they share similar asymptotic properties. The application to the elderly demand for medical care using MEPS shows that the corrected model can be used to detect the potential problem of measurement error and provide consistent estimates of the regression model.

APPENDIX A

Proof of Proposition 2.1 As assumed, we have $z = x + \epsilon$ where ϵ are assumed to be independent of $\mathbf{x}$ with mean 0. Let $g(\mathbf{x}|\mathbf{z})$ denote the conditional density of $\mathbf{x}$ given $\mathbf{z}$. Since the true density distribution is negative binomial, we have

$$\mathsf{var}[y|\mathbf{x}] = \left[1 + v^{-1}\mathsf{E}[y|\mathbf{x}]\right] \mathsf{E}[y|\mathbf{x}],$$

or

$$\mathsf{E}[y^2|\mathbf{x}] = \mathsf{E}[y|\mathbf{x}] + (1 + v^{-1})(\mathsf{E}[y|\mathbf{x}])^2.$$

Using Cauchy-Schwartz inequality, we have

$$\begin{aligned}\left[\int \mathsf{E}[y|\mathbf{x}] g(\mathbf{x}|\mathbf{z}) dx\right]^2 &= \left[\int \mathsf{E}[y|\mathbf{x}]\sqrt{g(\mathbf{x}|\mathbf{z})}\sqrt{g(\mathbf{x}|\mathbf{z})} dx\right]^2 \\ &\leq \int (\mathsf{E}[y|\mathbf{x}]\sqrt{g(\mathbf{x}|\mathbf{z})})^2 dx \int \left(\sqrt{g(\mathbf{x}|\mathbf{z})}\right)^2 dx \\ &= \int (\mathsf{E}[y|\mathbf{x}])^2 g(\mathbf{x}|\mathbf{z}) dx. \qquad \text{(A.1)}\end{aligned}$$

It is straightforward to show that

$$\begin{aligned}&v^{-1}\left[\int \mathsf{E}[y|\mathbf{x}] g(\mathbf{x}|\mathbf{z}) dx\right]^2 + \int \mathsf{E}[y|\mathbf{x}] g(\mathbf{x}|\mathbf{z}) dx \\ \leq\ &\int \left[\mathsf{E}[y|\mathbf{x}] + (1 + v^{-1})(\mathsf{E}[y|\mathbf{x}])^2\right] g(\mathbf{x}|\mathbf{z}) dx - \left[\int \mathsf{E}[y|\mathbf{x}] g(\mathbf{x}|\mathbf{z}) dx\right]^2,\end{aligned}$$

or

$$
\begin{aligned}
&\left[1+v^{-1}\int \mathsf{E}[y|\mathbf{x}]g(\mathbf{x}|\mathbf{z})dx\right]\int \mathsf{E}[y|\mathbf{x}]g(\mathbf{x}|\mathbf{z})dx \\
\leq\ &\int \mathsf{E}[y^2|\mathbf{x}]g(\mathbf{x}|\mathbf{z})dx-\left[\int \mathsf{E}[y|\mathbf{x}]g(\mathbf{x}|\mathbf{z})dx\right]^2.
\end{aligned}
$$

That is

$$\left[1+v^{-1}\mathsf{E}[y|\mathbf{z}]\right]\mathsf{E}[y|\mathbf{z}]\leq \mathsf{var}[y|\mathbf{z}].$$

The equality holds only when $\mathsf{E}[y|\mathbf{z}]=1$ or $g(\mathbf{x}|\mathbf{z})=0$ almost everywhere.

Proof of Proposition 3.1 Denote $\Theta=(\beta v)'$. The corrected score functions for Θ are

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}S(\Theta)\equiv\begin{pmatrix}\frac{\partial L_C}{\partial\beta}\\ \frac{\partial L_C}{\partial v}\end{pmatrix}=\begin{pmatrix}\frac{1}{n}\sum\limits_{i=1}^{n}s_i(\beta)\\ \frac{1}{n}\sum\limits_{i=1}^{n}s_i(v)\end{pmatrix}\\
=\begin{pmatrix}\frac{1}{n}\sum\limits_{i=1}^{n}\left[y_i\mathbf{z}_i-\frac{v+y_i}{v+\exp(\beta'\mathbf{x}_i)}\exp(\beta'\mathbf{x}_i)\mathbf{x}_i\right]\\ \frac{1}{n}\sum\limits_{i=1}^{n}\left[(\ln\frac{\Gamma(y_i+v)}{\Gamma(y_i+1)\Gamma(v)})'+1+\ln v-\ln(v+\exp(\beta'\mathbf{x}_i))-\frac{v+y_i}{v+\exp(\beta'\mathbf{x}_i)}\right]\end{pmatrix}
\end{aligned}
$$

which converge to $\begin{pmatrix}\mathsf{E}y\mathbf{z}-\mathsf{E}_x\exp(\beta'\mathbf{x})\mathbf{x}\\ \mathsf{E}\left[(\ln\frac{\Gamma(y_i+v)}{\Gamma(y_i+1)\Gamma(v)})'+\ln v\right]-\mathsf{E}_x\ln(v+\exp(\beta'\mathbf{x}))\end{pmatrix}$ when $n\to\infty$.

The simulation-based corrected score equations are

$$
\begin{aligned}
0\ &=\ \frac{1}{n}\sum_{i=1}^{n}\tilde{S}(\tilde{\Theta})\equiv\begin{pmatrix}\frac{1}{n}\sum\limits_{i=1}^{n}\tilde{s}_i(\tilde{\beta})\\ \frac{1}{n}\sum\limits_{i=1}^{n}\tilde{s}_i(\tilde{v})\end{pmatrix}\\
&=\begin{pmatrix}\frac{1}{n}\sum\limits_{i=1}^{n}\left[y_i\mathbf{z}_i-\frac{1}{S}\sum_{s=1}^{S}\exp(\tilde{\beta}'\tilde{\mathbf{x}}_s)\tilde{\mathbf{x}}_s\right]\\ \frac{1}{n}\sum\limits_{i=1}^{n}\left[(\ln\frac{\Gamma(y_i+\tilde{v})}{\Gamma(y_i+1)\Gamma(\tilde{v})})'+\ln\tilde{v}-\frac{1}{S}\sum_{s=1}^{S}\ln(\tilde{v}+\exp(\tilde{\beta}'\tilde{\mathbf{x}}_s))\right]\end{pmatrix},
\end{aligned}
\tag{A.2}
$$

where $\tilde{\mathbf{x}}_s$ is the simulated covariates from the distribution of $\mathbf{x}$.

The simulation-based corrected score function can be rewritten as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{S}_i(\tilde{\Theta})=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}S_i(\tilde{\Theta})+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_n+\frac{\sqrt{n}}{S}\sum_{s=1}^{S}B_s,$$

where

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} A_n \\
= \ &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix} \left[\frac{\tilde{v}+y_i}{\tilde{v}+\exp(\tilde{\beta}'\mathbf{x}_i)}\exp(\tilde{\beta}'\mathbf{x}_i)\mathbf{x}_i - \mathsf{E}_x\exp(\tilde{\beta}'\mathbf{x})\mathbf{x}\right] \\ \left[\frac{\tilde{v}+y_i}{\tilde{v}+\exp(\tilde{\beta}'\mathbf{x}_i)} - 1 + \ln(\tilde{v}+\exp(\tilde{\beta}'\mathbf{x}_i)) - \mathsf{E}_x\ln(\tilde{v}+\exp(\tilde{\beta}'\mathbf{x})\right] \end{pmatrix},
\end{aligned}$$

and

$$\frac{\sqrt{n}}{S}\sum_{s=1}^{S} B_s = \frac{\sqrt{n}}{S}\sum_{s=1}^{S}\begin{pmatrix} \left[\mathsf{E}_x\exp(\tilde{\beta}'\mathbf{x})\mathbf{x} - \exp(\tilde{\beta}'\tilde{\mathbf{x}}_s)\tilde{\mathbf{x}}_s\right] \\ \left[\mathsf{E}_x\ln(\tilde{v}+\exp(\tilde{\beta}'\mathbf{x}) - \ln(\tilde{v}+\exp(\tilde{\beta}'\tilde{\mathbf{x}}_s))\right] \end{pmatrix}.$$

By Taylor's expansion,

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_i(\tilde{\Theta}) &\approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(S_i(\Theta_0)+S_{i\Theta}(\Theta_0)(\tilde{\Theta}-\Theta_0)) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_i(\Theta_0) + \sqrt{n}(\tilde{\Theta}-\Theta_0)\frac{1}{n}\sum_{i=1}^{n} S_{i\Theta}(\Theta_0),
\end{aligned}$$

with $\mathsf{E}S_{i\Theta}(\Theta_0) = -I$.

Therefore,

$$\sqrt{n}(\tilde{\Theta}-\Theta_0) \approx I^{-1}[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_i(\Theta_0)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n} A_n + \frac{\sqrt{n}}{S}\sum_{s=1}^{S} B_s].$$

With assumption that B_s has finite second order moment, we have

$$\frac{\sum\limits_{s=1}^{S} B_s}{\sqrt{S}} \sim N(0,\Omega_0),$$

when $S \to \infty$. As a result, when $n \to \infty$, $s \to \infty$, and $\frac{n}{S} \to 0$, we have $\frac{\sqrt{n}}{S}\sum\limits_{s=1}^{S} B_s = \mathrm{o}_p(1)$.

It is clear that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} A_n \sim N(0,J_2),$$

where

$$J_2 = \mathsf{E}m_i m_i',$$

$$m_i = \begin{pmatrix} \frac{\tilde{v}+y_i}{\tilde{v}+\exp(\mathbf{x}_i\tilde{\beta}')} \exp(\tilde{\beta}'\mathbf{x}_i)\mathbf{x}_i - \mathsf{E}_x \exp(\tilde{\beta}'\mathbf{x})\mathbf{x} \\ \frac{\tilde{v}+y_i}{\tilde{v}+\exp(\tilde{\beta}'\mathbf{x}_i)} - 1 + \ln(\tilde{v} + \exp(\tilde{\beta}'\mathbf{x}_i)) - \mathsf{E}_x \ln(\tilde{v} + \exp(\tilde{\beta}'\mathbf{x})) \end{pmatrix}.$$

And

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_i(\Theta_0) \sim N(0, J_1),$$

where

$$J_1 = \mathsf{E} S_i(\Theta_0) S_i'(\Theta_0).$$

Overall, $\sqrt{n}(\tilde{\Theta} - \Theta_0)$ converges in distribution to $N(0,\ I^{-1}(J_1 + J_2 + J_3)I^{-1})$, where

$$J_3 = \mathsf{E}[m_i S_i(\Theta_0)' + S_i(\Theta_0)' m_i].$$

REFERENCES

Rockville, MD., 1998, Agency for health care policy and research. *MEPS HC-003: 1996 panel opulation characteristics and utilization data for 1996.*

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

Cameron, A. C. and P. K. Trivedi, 1998 *Regression analysis of count data.* Cambridge, UK.

Cameron, A. C., P. K. Trivedi, F. Milne, and J. Piggott, 1988, A microeconometric model of the demand for health care and health insurance in Australia. *Review of Economic Studies* **55**, 85-106.

Carroll, R. J., D. Ruppert, and L. A. Stefanski, 1995, *Measurement error in nonlinear models.* Chapman and Hall, London.

Cartwright, W. S., T. W. Hu, and L. F. Huang, 1992, Impact of varying medigap insurance coverage on the use of medical services of the elderly. *Applied Economics* **24**, 529-539.

Chesher, A., 1991, The effect of measurement error. *Biometrika* **78**, 451-462.

Dean, C., J. F. Lawless, and G. E. Willmot, 1989, A mixed poisson-inverse-gaussian regression model. *The Canadian Journal of Statistics* **17,2**, 171-181.

Deb, P. and P. K. Trivedi, 1997, Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics* **12**, 313-326.

Gourieroux, C. and A. Monfort, 1996, *Simulation-based econometric methods.* Oxford University Press, Oxford and New York.

Griliches, Z., 1986, Economic data issues. In: *Handbook of Econometrics*, Edited by Z. Griliches and M.D. Intriligator, Vol III, 1456-1514, North-Holland, Amsterdam.

Guo, J. Q. and T. Li, 2000, Poisson regression models with errors-in-variables: Implication and treatment. Working Paper, Indiana University.

Haab, T. C. and K. E. McConnell, 1996, Count data models and the problem of zeros in recreation demand analysis. *American Journal of Agricultural Economics* **78**, 89-102.

Hajivassiliou, V. A., 1997, Some practical issues in maximum simulated likelihood. In: *Simulation-Based Inference in Econometrics: Methods and Applications,* Edited by R. Mariano, M. Weeks, and T. Schuermann, Cambridge University Press.

Hajivassiliou, V. A. and D. L. McFadden, 1998, The method of simulated scores for the estimation of LDV models. *Econometrica* **66**, 863-896.

Hausman, J. A., B. H. Hall, and Z. Griliches, 1984, Econometric models for count data with an application to the patents-R and D relationship. *Econometrica* **52**, 909-938.

Hausman, J., W. K. Newey, and J. L. Powell, 1995, Nonlinear errors-in-variables estimation of some engel curves. *Journal of Econometrics* **65**, 205-233.

Hsiao, C., 1989, Consistent estimation for some nonlinear errors-in-variables models. *Journal of econometrics* **41**, 159-185.

Hsiao, C., 1992, Nonlinear latent variable models. In: *The Econometrics of Panel Data*, Edited by Matyas, L. and P. Sevestre, Kluwer Academic Publishers, Dordrecht.

Lee, L. F. and J. H. Sepanski, 1995, Estimation of linear and nonlinear errors-in-variables models using validation data. *Journal of the American Statistical Association* **90**, 130-140.

Lewbel, A., 1996, Demand estimation with expenditure measurement errors on the left and right hand side. *Review of Economics and Statistics* 718-725.

Li, T., 2000, Estimation of nonlinear errors-in-variable models: A simulated minimum distance estimator. *Statistics and Probability Letters* **47**, 243-248.

McFadden, D. L., 1989, A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* **57**, 995-1026.

Nakamura, T., 1990, Corrected score function for errors-in-variables models: methodology and application to generalized linear models. *Biometrika* **77**, 127-137.

Pakes, A. and D. Pollard, 1989, Simulation and the asymptotics of optimization estimators. *Econometrica* **57**, 1027-1057.

Rendtel, U. and R. Langeheine, 1998, The estimation of poverty dynamics using different measurements of household income. *Review of Income and Wealth* **44**, 81-98.