

An exact solution to the endogeneous regressor problem in zero-inflated regression models

Sencer Ecer*

December 17, 2003

Abstract

In this paper I present an exact solution for endogeneous regressor problem in zero-inflated regression models.

1 Introduction

Zero-inflated (ZI) regression models are being increasingly used in health economics, especially in health care demand estimations. In these estimations, health insurance dummy is usually among the independent variables. However, due to adverse selection in the choice of health insurance, endogeneity of this dummy variable is an issue that needs to be addressed. Terza (1998)'s solution, which applies to Poisson and Negative Binomial regressions, for endogeneous regressor problem in count data models is not directly applicable to ZI regression models.

2 Econometric methodology

In general, finite mixture models are appropriate when it is possible that the same outcome can arise from more than one processes. A special case of finite mixture models is

*I thank Stephen Donald, Don Fullerton, Li Gan, Çağatay Koç and Paul Wilson for comments and suggestions. Naturally, all errors are solely mine.

the Zero-Inflated (ZI) regression model. Estimating a ZI regression model involves estimating the switching probability between the regimes using the logit regression model,¹ and estimating the second regime by any non-degenerate count-data regression model. Poisson or Negative Binomial Regression models are by far the only count-data models used for the purpose of estimating the second regime, rendering Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) models, respectively.² (Poisson, Negative, Binomial and Hurdle models are frequently used in health care demand estimations, see Cameron and Trivedi (1998) for a survey). Thus, one can analyze the effect of an independent variable in different decision stages, on the assumption that it plays a role in switching between these underlying regimes. Technically, the choice of covariates for the switching probability (Stage 1) is completely independent from the choice of covariates for the ex post demand D_2 (Stage 2). However, using the same covariates in both parametrizations serves the purpose of identifying the possibly different roles of the same determinant in each stage. In this chapter, I use identical sets of determinants both for the switching probability between D_1 and D_2 , and for the demand in D_2 .

The basic distribution underlying the Zero-Inflated regression model is:³

$$y_i = 0 \quad \text{with probability } P(D_1)$$

$$y_i \sim \text{Poisson}[\alpha_i] \quad \text{with probability } P(D_2)$$

¹Using probit or Prentice's F distribution, which is a generalization of the logistic distribution (see Cameron and Trivedi (1998)), are other possibilities but logit is more commonly used, and my results are robust to using Probit.

²Extending ZIP and ZINB to, respectively, ZIP-Hurdle and ZINB-Hurdle models are also possible (provided identification is achieved). Hurdle models are count data models that assume a non-linearity in the process, and hence let the zero outcome come from a different distribution than the distribution of the positive outcomes. Using a Hurdle model, doctor-induced moral hazard can be accounted for, as in Pohlmeijer and Ulrich (1995). Note that in a Hurdle model, zero need not necessarily be the "Hurdle", for example, as suggested by Wilson (1998), a positive hurdle in a Hurdle model is conceivable.

³There are some typos in Greene (1997), on p.943. Following his notation there, $Prob(Z_i = 0) = F(w_i, \gamma)$ and the mean should be $(1 - F)\lambda_i$. For the correct mean see Lambert (1992), Greene (1994) or Cameron and Trivedi (1998).

where $P(D_1)+P(D_2) = 1$, and Lambert (1992) proposes to estimate $P(D_1)$ by using either of the logit or probit models and employing the Poisson (or any other count-data) regression model for estimating the coefficients in D_2 . That is, letting $F(\cdot)$ be the logistic distribution, Zero-Inflated Poisson (ZIP) model is given as:

$$P(D_1) = F(Z, \gamma)$$

$$P(y_i = j|D_2) = \frac{\exp(-\lambda_i)\lambda_i^j}{j!} \quad j = 0, 1, 2, \dots$$

where $\lambda_i = e^{X_i\beta}$ as in all the familiar count models.⁴ Even though $P(D_1)$ is parametrized, the marginal effects on the probability of switch, that is, $P(D_2)$ is straightforward since, $P(D_2) = 1 - P(D_1)$.

In the present application, $F(Z, \gamma) = F(Z\gamma)$ and $F(\cdot)$ are the cumulative probability function of the logistic distribution. Also the same determinants are used for both parametrizations, that is, X and Z will be identical. Now I proceed to the ZINB regression model.

The two models ZIP and ZINB have no technical difference except that the decision in Stage 2 is parametrized by NB distribution in the case of ZINB (Greene, 1994), that is,

$$P(D_1) = F(Z, \gamma) = F(Z\gamma)$$

$$P(y_i = j|D_2) = \frac{\Gamma(j + \alpha^{-1})}{\Gamma(j + 1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu}\right)^j.$$

where, j is a non-negative integer and α is a non-negative real number, and $F(\cdot)$ is the logistic distribution in the present application. The NB distribution reduces to Poisson when $\alpha = 0$, so ZIP is nested in ZINB, which makes the likelihood ratio test applicable. For the gradients of the likelihood function, I refer the reader to Greene (1994).

When ZINB is estimated, one gets a coefficient for each covariate for each decision stage. If $X = Z$ as in my case, then there are two coefficients for each covariate, one

⁴See Cameron and Trivedi (1998) for an account of parametrization of mean using the exponential function in count data models.

for Stage 1 and the other for Stage 2. These two coefficients allow to differentiate between the roles of the same covariate in different decision stages. In the case of health insurance dummies, this enables me to differentiate between the types of moral hazard effects.

Marginal Effects in ZI models are not trivial but relatively straightforward. For the sake of completeness and clarity of exposition they are given here. The mean of ZIP and ZINB are the same and given by:

$$E(Y|X, Z) = (1 - F)\lambda$$

where $\lambda = e^{X\beta}$ and $F = F(Z, \gamma) = F(Z\gamma)$. Thus the marginal effect of X_i is

$$\frac{\partial E(Y|X, Z)}{\partial X_i} = \frac{\partial}{\partial X_i} [(1 - F(Z\gamma))e^{X\beta}]$$

Consider the following two cases. In the first case, the covariate X_i is a regressor in both of the stages. In the second case, it is a regressor in only one stage. These two cases lead to different marginal effects. In the present application Case 1 holds, and therefore Case 2 is not detailed in what follows.

Case 1: Let the variable X_i be a covariate in both regimes, i.e., $\exists i$ such that $X_i = Z_i$. Then

$$\frac{\partial E(Y|X, Z)}{\partial X_i} = -\gamma_i F'(Z\gamma)e^{X\beta} + (1 - F(Z\gamma))\beta_i e^{X\beta}$$

The first part, i.e., $-\gamma_i F'(Z\gamma)e^{X\beta}$ is the marginal effect in Stage 1 and the second part, i.e., $(1 - F(Z\gamma))\beta_i e^{X\beta}$ is the marginal effect in Stage 2. Since F' , $1 - F$, and $e^{X\beta}$ are all positive, the marginal effect in the first stage (the marginal effect on probability of switching to D_2) has the opposite sign with the estimated coefficient γ_i , whereas the marginal effect in the second stage, has the same sign with the estimated coefficient β_i . The sign of the overall marginal effect differs depending on signs and ratios of the coefficients with respect to the hazard function associated with the logistic density. Define this hazard function by

$$\frac{F'(Z\gamma)}{1 - F(Z\gamma)} = h(Z\gamma)$$

Note that since logistic distribution is used, $F(Z\gamma) = \frac{\exp(Z\gamma)}{1+\exp(Z\gamma)}$, the hazard function coincides with the cumulative density function, that is, $h(Z\gamma) = F(Z\gamma)$. The sign of the Overall Marginal Effect (OME) can be given as:

Case 1.1. $(\beta_i, \gamma_i > 0)$. OME > 0 iff $\frac{\beta_i}{\gamma_i} > h(Z\gamma)$.

Case 1.2. $(\beta_i, \gamma_i < 0)$. OME > 0 iff $\frac{\beta_i}{\gamma_i} < h(Z\gamma)$.

Case 1.3. $(\beta_i > 0, \gamma_i < 0)$. OME > 0 .

Case 1.4. $(\beta_i < 0, \gamma_i > 0)$. OME < 0 .

In this case, the percentage change in the dependent variable as a result of a marginal change in a given independent variable X_i is (note that $X = Z$ in the present application)

$$\frac{\frac{\partial E(Y|X)}{\partial X_i}}{E(Y|X, Z)} = \beta_i - h(Z\gamma)\gamma_i = \beta_i - F(Z\gamma)\gamma_i.$$

Case 2: Now consider a variable X_i that is used as a covariate in only Stage 2, that is, $X_i \in (X \setminus Z)$. Then the marginal effects and percentage changes can easily be found by directly following the steps in Case 1.

In ZIP or ZINB, when the same covariates are used for both decision stages, as in the present application, the estimated percentage changes in the dependent variable (evaluated at the sample mean) as a function of a given covariate is given as

$$\hat{\beta}_i - h(\bar{Z}\hat{\gamma})\hat{\gamma}_i \tag{1}$$

where $\hat{\beta}_i$ is the estimated coefficient of Z_i in Stage 2 and $\hat{\gamma}_i$ is the estimated coefficient of Z_i in Stage 1, and \bar{Z} is the vector of means of Z_i 's.

3 Solution for endogeneity

Inconsistent estimates may arise because of the possibility that unobservable characteristics of an individual (e.g. an expectation of a poor health status in the future) may be correlated with the insurance dummy. This problem is called the problem of endogeneity of health insurance, which is a concern in health care demand estimations. The Generalized Method of Moments (GMM) is employed by Windmeijer and Santos-Silva (1997), CDK (2001a,b) and Craig and Koç (2001). This approach can be applied

to the present context, but I pursue an exact explicit solution here.

The problem of endogeneity is also addressed by Greene (1994), who accounts for the closely related problem, sample-selection, in a different context. On the other hand, the solution provided by Greene (1994) closely follows Heckman (1979) and ignores the specific properties of ZI model. The solution proposed by Greene (1994) is practical because it involves values from the cumulative normal distribution. The exact solution that is presented below requires a non-linear least squares estimation, where the mean function contains an integral that does not belong to a known distribution.

Terza (1997) accounts for endogeneity in Poisson and Negative Binomial models by using Heckman's (1979) two-step procedure. Since the means of the Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) regression models are different from the means of Poisson and Negative Binomial, Terza's (1997) solution is not valid for ZIP and ZINB models. Below I account for endogeneity for a class of models including the fully parametric models ZIP and ZINB using a semi-parametric approach.

Let the dependent variable y be formulated as

$$y = f(y|w, d, \eta, \epsilon) \tag{2}$$

where w is the set of all exogeneous variables, η and ϵ are nuisance parameters to account for possible heterogeneity in respectively Stage 1 and Stage 2, and d is the possibly-endogeneous dummy variable where,

$$d = \begin{cases} 1 & \text{if and only if } T\alpha + \nu > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $T \subset W = (X \cup Z) \setminus \{d\}$, α is a vector of coefficients, and ν is the error term. Note that $d \in (X \cup Z)$. Let

$$E[y|w, d, \epsilon, \eta] = \frac{\exp(X\beta + \epsilon)}{1 + \exp(Z\gamma + \eta)}. \tag{3}$$

In particular, this assumption is satisfied by ZIP and ZINB models if the regime switching probability is modeled by Logit. The parameters ϵ and η are unobservable heterogeneities, and they are possibly correlated with d . For example, it may be true

that individuals who buy health insurance are those who are more likely to demand more health care, the phenomenon of adverse selection. The opposite effect is also possible, implying the presence of screening. Similarly, individuals who don't have access to or appropriate educational background inducing a sufficient level of self-control for proper self-preventive care may be more inclined to buy insurance, suggesting a possible correlation between d and η .⁵

This conditional mean motivates the following regression:

$$y = \frac{\exp(X\beta + \epsilon)}{1 + \exp(Z\gamma + \eta)} + u \quad (4)$$

where $E[u|w, d, \epsilon, \eta] = 0$.

Let (ϵ, η, ν) have a joint normal distribution. This can be used to motivate a regression in which the possibly endogeneous dummy is no longer correlated with the error term. For this purpose I need to derive $E[y|w, d]$, which by the Law of Iterated Expectations is equal to $E_{[\epsilon, \eta]}[y|w, d, \epsilon, \eta]$. Then $E[y|w, d]$ can be rewritten as:

$$\begin{aligned} E[y|w, d] &= dE_{[\epsilon, \eta]} \left[\frac{\exp(X\beta + \epsilon)}{1 + \exp(Z\gamma + \eta)} | v > -T\alpha, w \right] + \\ &\quad (1 - d)E_{[\epsilon, \eta]} \left[\frac{\exp(X\beta + \epsilon)}{1 + \exp(Z\gamma + \eta)} | v \leq T\alpha, w \right] \end{aligned} \quad (5)$$

Using

$$\frac{\exp(X\beta + \epsilon)}{1 + \exp(Z\gamma + \eta)} = E \left[\frac{\exp(X\beta + \epsilon)}{1 + \exp(Z\gamma + \eta)} | v, w \right],$$

(5) can be written as

$$\begin{aligned} E[y|w, d] &= dE_{[\epsilon, \eta]} \left[E_{[\epsilon, \eta]} \left[\frac{\exp(X\beta + \epsilon)}{1 + \exp(Z\gamma + \eta)} | v, w \right] | v > -T\alpha, w \right] + \\ &\quad (1 - d)E_{[\epsilon, \eta]} \left[E_{[\epsilon, \eta]} \left[\frac{\exp(X\beta + \epsilon)}{1 + \exp(Z\gamma + \eta)} | v, w \right] | v \leq T\alpha, w \right]. \end{aligned} \quad (6)$$

⁵Ruling out the possibility of individual heterogeneity in Stage 1, i.e., setting $\eta = 0$, or letting the individual heterogeneity be characterized by the same parameter in both regimes, i.e., setting $\eta = \epsilon$, can substantially simplify the following solution. However, for now I proceed with the most general case.

Since

$$f(\epsilon, \eta, \nu) \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & 1 \end{bmatrix} \right)$$

where $\sigma_{ij} = \sigma_{ii}\sigma_{jj}\rho_{ij}$, the conditional density of (ϵ, η) becomes

$$f_{\epsilon, \eta}(\epsilon, \eta | \nu) \sim N \left(\begin{pmatrix} \sigma_{13}\nu \\ \sigma_{23}\nu \end{pmatrix}, \begin{bmatrix} \sigma_{11} - \sigma_{13}^2 & \sigma_{12} - \sigma_{13}\sigma_{23} \\ \sigma_{12} - \sigma_{13}\sigma_{23} & \sigma_{22} - \sigma_{23}^2 \end{bmatrix} \right). \quad (7)$$

From this conditional density we can find

$$E_{[\epsilon, \eta]} \left[\frac{\exp(X\beta + \epsilon)}{1 + \exp(Z\gamma + \eta)} | \nu, w \right],$$

which is an expectation of a non-linear transformation of the joint density given by (7).

By a first-order Taylor approximation, using the (population) mean as the expansion point, and letting

$$g(\epsilon, \eta) = \frac{\exp(X\beta + \epsilon)}{1 + \exp(Z\gamma + \eta)},$$

one can write

$$E[g(\epsilon, \eta) | \nu, w] \approx g(\mu_\epsilon, \mu_\eta) = \frac{\exp(X\beta + \sigma_{13}\nu)}{1 + \exp(Z\gamma + \sigma_{23}\eta)}.$$

Thus, (6) becomes

$$\begin{aligned} E[y|w, d] &\approx dE \left[\frac{\exp(X\beta + \sigma_{13}\nu)}{1 + \exp(Z\gamma + \sigma_{23}\nu)} | v > -T\alpha, w \right] + \\ &(1-d)E \left[\frac{\exp(X\beta + \sigma_{13}\nu)}{1 + \exp(Z\gamma + \sigma_{23}\nu)} | v \leq T\alpha, w \right] \end{aligned} \quad (8)$$

This approximation of $E[y|w, d]$ can be written as follows:

$$\begin{aligned} E[y|w, d] &= d \int_{-T\alpha}^{\infty} \frac{\exp(X\beta + \sigma_{13}\nu)}{1 + \exp(Z\gamma + \sigma_{23}\nu)} \phi(\nu | \nu > -T\alpha, w) d\nu + \\ &(1-d) \int_{-\infty}^{-T\alpha} \frac{\exp(X\beta + \sigma_{13}\nu)}{1 + \exp(Z\gamma + \sigma_{23}\nu)} \phi(\nu | \nu \leq -T\alpha, w) d\nu \end{aligned}$$

which after simplification and defining $h(\cdot)$ as (note that $1 - \Phi(-T\alpha) = \Phi(T\alpha)$):

$$\begin{aligned}
& h(w, d, \beta, \gamma, T, \alpha, \sigma_{13}, \sigma_{23}) = \tag{9} \\
& \exp(X\beta) \left[\frac{d}{\Phi(T\alpha)} \int_{-T\alpha}^{\infty} \frac{\exp(\sigma_{13}\nu)}{1 + \exp(Z\gamma + \sigma_{23}\nu)} \phi(\nu) d\nu + \right. \\
& \left. \frac{1-d}{\Phi(-T\alpha)} \int_{-\infty}^{-T\alpha} \frac{\exp(\sigma_{13}\nu)}{1 + \exp(Z\gamma + \sigma_{23}\nu)} \phi(\nu, w) d\nu \right] \tag{10}
\end{aligned}$$

gives

$$E[y|w, d] \approx h(w, d, \beta, \gamma, T, \alpha, \sigma_{13}, \sigma_{23}).$$

Motivated by (9) one can write, $y = E[Y|w, d] + e$, that is,

$$\begin{aligned}
y = \exp(X\beta) & \left[\frac{d}{\Phi(T\alpha)} \int_{-T\alpha}^{\infty} \frac{\exp(\sigma_{13}\nu)}{1 + \exp(Z\gamma + \sigma_{23}\nu)} \phi(\nu|w) d\nu + \right. \\
& \left. \frac{1-d}{\Phi(-T\alpha)} \int_{-\infty}^{-T\alpha} \frac{\exp(\sigma_{13}\nu)}{1 + \exp(Z\gamma + \sigma_{23}\nu)} \phi(\nu|w) d\nu \right] + e \tag{11}
\end{aligned}$$

where $e \equiv \frac{\exp(X\beta + \epsilon)}{1 + \exp(Z\gamma + \eta)} + u - E[y|w, d]$. Let $\tau = [\beta \ \gamma \ \alpha \ \sigma_{13} \ \sigma_{23}]$. Then it can be checked that $E[\frac{\partial h(\cdot)}{\partial \tau} e] = 0$, so the NLS estimator is consistent. To reduce the computational burden, one can use a two-stage (TS) method as in Heckman (1976, 1979). By Theorem 6.11 in White(1994) $\sqrt{n}(b^{TS} - b^{NLS})$ converges in distribution to a normal distribution with mean zero. For the solution, first $\hat{\alpha}$ is to be found by, say, probit. Then $\beta, \gamma, \sigma_{13}\nu$ and $\sigma_{23}\nu$ are to be estimated by NLS using the following specification of the conditional mean:

$$\begin{aligned}
y = \exp(X\beta) & \left[\frac{d}{\Phi(T\hat{\alpha})} \int_{-T\hat{\alpha}}^{\infty} \frac{\exp(X\beta + \sigma_{13}\nu)}{1 + \exp(Z\gamma + \sigma_{23}\nu)} \phi(\nu) d\nu + \right. \\
& \left. \frac{1-d}{\Phi(-T\hat{\alpha})} \int_{-\infty}^{-T\hat{\alpha}} \frac{\exp(X\beta + \sigma_{13}\nu)}{1 + \exp(Z\gamma + \sigma_{23}\nu)} \phi(\nu) d\nu \right] + e^0 \tag{12}
\end{aligned}$$

where $e^0 \equiv e + E[y|w, d, \alpha] - E[y|d, w, \hat{\alpha}]$.

Equation (12) solves the problem of endogeneity. As can be seen from (12), an integral with an infinite limit has to be computed at every iteration during the estimation process.

References

Arrow, K.J., 1963. Uncertainty and the welfare economics of medical care, *American Economic Review*, 53, 941-973.

Arrow, K.J., 1968, The economics of moral hazard: Further Comment. *American Economic Review*, 58, 537-538.

Cameron, A.C., Trivedi, P.K., 1998. *Regression Analysis of Count Data*, Econometric Society Monographs No. 30, Cambridge University Press.

Cameron, A.C., P.K. Trivedi, F. Milne and J. Piggott (1988), A Microeconomic Model of the Demand for Health Care and Health Insurance in Australia, *Review of Economic Studies*, 55, 85-106.

Craig, B., Dusansky, R. and C. Koç (2001a), Insurance Endogeneity, Moral Hazard and the Demand for Health Care, Department of Economics, University of Texas, photocopy.

Craig, B., Dusansky, R. and C. Koç (2001b), GMM Estimation and the Variation of Moral Hazard across Medical services, Department of Economics, University of Texas, photocopy.

Craig, B., C. Koç (2001), The Moral Hazard Effect of Insurance Across Health Cohorts, Department of Preventive Medicine, University of Wisconsin Medical School at Madison, photocopy.

Crepon, B., Duguet E., 1997. Research and Development, competition and innovation Pseudo-maximum likelihood and simulated maximum likelihood methods applied to count data models with heterogeneity, *Journal of Econometrics*, 79, 355-78.

DeGroot, M.H., 1986. *Probability and Statistics*, Addison-Wesley, Massachusetts.

Dionne, G., St-Michel P., 1991. Workers' Compensation and Moral Hazard, *Review of Economics and Statistics*, 73(2), 236-244.

Greene, W. H. (1994), Accounting for Excess Zeros and Sample Selection in Poisson and negative Binomial Models, Stern School of Business Working Papers, EC-94-10, New York University.

Greene, W.H., 1997. *Econometric Analysis*, Prentice Hall, New Jersey.

Grootendorst, P.V. (1995), A comparison of alternative models of prescription drug utilization, *Health Economics*, 4, 183-198

Heckman, J., 1976. The common Structure of statistical models of truncation sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5, 475-492.

Heckman, J., 1979. Sample selection bias as a specification error. *Econometrica* 47, 153-161.

Lambert, D. (1992), Zero-Inflated Poisson Regression with an application to defects in manufacturing, *Technometrics*, 34(1),1-14

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

Pauly, M. V., 1968, The economics of moral hazard: Comment. *American Economic Review*, 58, 531-537.

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

Terza, J.V. 1998. Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects, *Journal of Econometrics*, 84, 129-154.

Terza, J.V. and Wilson, P.W., 1990. Analyzing frequencies of several types of events: a mixed multinomial-Poisson approach, *Review of Economics and Statistics*, 72, 108-15.

Vuong, Q.,1989. Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses, *Econometrica*, 57, 307-34.

Wilson, P.W., 1992. Count data models without mean variance restrictions, presented at the European Meeting of the Econometric Society, Brussels.

Wilson, P.W., 1998. Estimating counts of related events with endogenous switching using complex survey data, photocopy, Department of Economics, University of Texas.

Windmeijer, F.A.G., Santos-Silva, J.M.C., 1997. Endogeneity in count data models: An application to demand for health care, *Journal of Applied Econometrics*, 12(3), 281-294.