A discusión

COUNTS WITH AN ENDOGENOUS BINARY REGRESSOR: A SERIES EXPANSION APPROACH*

Andrés Romeu and Marcos Vera-Hernández**

WP-AD 2004-36

Corresponding author: A. Romeu, Universidad de Alicante, Departamento de Fundamentos del Análisis Económico, Campus Sant Vicent 03690 Alicante, Spain. E-mail: aromeu@merlin.fae.ua.es

Editor: Instituto Valenciano de Investigaciones Económicas, S.A.

Primera Edición Septiembre 2004.

Depósito Legal: V-3979-2004

IVIE working papers offer in advance the results of economic research under way in order to encourage a discussion process before sending them to scientific journals for their final publication.

---

* We thank Professor Pravin Trivedi and two anonymous referees for their valuable comments. We wish to thank Professor Michael Creel for useful help and advice, and Professor Joseph Terza for kindly providing the data and the TSM and WNLS software code. We thank Professor Richard Blundell and Frank Windmeijer for their interesting comments. We also wish to thank participants at the CEMMAP lunch seminar and ENTER *meeting* at University College of London, the *Royal Economic Society Easter School* in Birmingham, the *6th Computational Economics Conference* in Barcelona, the *CEMAPRE Conference* in Lisbon, the *XXV Symposium of Economic Analysis* in Barcelona, the *ASSET Euroconference* in Lisbon and seminar participants in the Universidad de la Laguna and Universidad de Alicante. All remaining errors are our responsibility. We benefited from financial support of the *Commisionat per a Universitats I Recerca de la Generalitat de Catalunya* grant nº. 1997FI-436, *Universitat Autónoma de Barcelona* AP92-34967274 and from Spanish Ministry of Education DGCYT PB96-1160.

** A. Romeu: Universidad de Alicante. M. Vera: Institute for Fiscal Studies, 7 Ridgmount St., WC1E 7AE, London, UK, E-mail: marcos.vera@ifs.org.uk

COUNTS WITH AN ENDOGENOUS BINARY REGRESSOR: A SERIES EXPANSION APPROACH

Andrés Romeu and Marcos Vera-Hernández

ABSTRACT

We propose an estimator for count data regression models where a binary regressor is endogenously determined. This estimator departs from previous approaches by using a flexible form for the conditional probability function of the counts. Using a Monte Carlo experiment we show that our estimator improves the fit and provides a more reliable estimate of the impact of regressors on the count when compared to alternatives which do restrict the mean to be linear-exponential. In an application to the number of trips by households in the US, we find that the estimate of the treatment effect obtained is considerably different from the one obtained under a linear-exponential mean specification.

Keywords: Count data, Polynominal Poisson Expansions, Flexible Functional Form.

JEL classification: C35, C52

1 Introduction

In the empirical analysis of count data, it is not uncommon to find situations where one or more of the regressors are presumed to be simultaneously determined with the outcome of interest.[1] In this situation, the Poisson model will yield biased estimates of the parameters of interest. Moreover, this model imposes severe restrictions on the shape of the conditional probability function of the counts. The researcher may want to use a flexible model that accommodates data generating processes (DGP) that might exhibit excess of zeros, multi-modalities and/or other non-Poisson characteristics. Partial solutions to the problem of endogeneity bias and flexible estimation may be found elsewhere in the literature. To our knowledge, a treatment of both problems remains unexplored except for Kenkel and Terza (2001) who consider the family of inverse Box-Cox functions for the conditional mean.

In this paper, we propose an estimator that deals simultaneously with a binary endogenous variable and departures of standard assumptions such as a linear-exponential specification and/or Poisson or Negative Binomial distribution of the counts. Mullahy (1997) and Windmeijer and Santos-Silva (1997) use GMM estimation based on a linear-exponential mean (say LEF) specification and a set of instruments. The GMM estimators are robust as far as the true data generating process of the counts shows a LEF for the conditional mean. Alternatively, the Two-Stage Method (TSM) and the Weighted Non-Linear Least Squares (WNLS) estimators proposed in Terza (1998) require some additional distributional assumptions with respect to the joint distribution of the unobserved components of the model. However, these assumptions allow us to incorporate a wider range of functional forms other than a LEF for the mean. In this paper, we exploit the advantage given by the distributional assumptions over the unobserved components of the models to gain flexibility in the modelling of the counts.

Allowing for specifications alternative to the LEF is not irrelevant in practice. Popular alternatives as the Hurdle model (Pohlmeier & Ulrich, 1995) or finite mixture models (Deb & Trivedi, 1997) do not exhibit a LEF. Gurmu and Trivedi (1996) find evidence of misspecification of the LEF in a data set with a large frequency of zero counts. When they tried to solve this problem by adding the squares and the cross-products of the regressors, the fit deteriorated appreciably.

To solve the problem of flexible estimation, some researchers have proposed to use polynomial series expansions over a baseline probability function. These polynomial expansions have been introduced in the context of exogenous regressors in Gurmu (1997), Gurmu and Trivedi (1996), Gurmu, Rilstone, and Stern (1999), Cameron and Johansson (1997), Creel and Farrell (2001) and Guo and Trivedi (2002). In general these estimators have been shown to work better than the standard Poisson or Negative Binomial models in terms of goodness of fit and information criteria under several forms of non-Poissonness. For this reason, we propose to apply these type of polynomial expansions to get a Polynomial Poisson Full-Information

[1] See Coulson, Terza, Neslulan, and Stuart (1995), Mullahy (1997), Windmeijer and Santos-Silva (1997), Vera-Hernandez (1999), Schellhorn (2001), Kenkel and Terza (2001), Deb and Trivedi (2002, 2003), Munkin and Trivedi (2003) in the case of health economics and Terza (1998) in the case of transportation economics.

Maximum Likelihood (PP-FIML) estimator to be used within the framework of a binary endogenous regressor.

Section 2 explains the data generating process of the Terza (1998) model and discusses three alternative estimators: the TSM, the Weighted Non-Linear Least Squares (WNLS) and the Poisson/Negative Binomial FIML. Actually, the work of Terza (1998) concentrates on the TSM and WNLS, while the Poisson/Negative Binomial FIML is only cited but estimation of these models is not carried out, nor their properties are explored. We contribute to the literature with the study of the bias that might arise due to misspecification in the Poisson-FIML and a discussion of the Negative Binomial FIML approach particularly in terms of identification.

As an alternative to these estimators, we present in section 3 the Polynomial Poisson Full Information Maximum Likelihood (PP-FIML) estimator. The PP-FIML is based on a polynomial expansion around a baseline Poisson probability function. Consequently, the conditional expectation of the counts is no longer given by the mean of the Poisson, being instead a non-linear function of it. This allows a departure from the standard LEF specification. Additional considerations on the choice of the degree of the polynomial expansion and on how to compute a measure of the impact of the binary endogenous on the counts are also included in section 3.

In section 4, we report the results of a small Monte Carlo experiment with data generating processes that exhibit over-dispersion, excess of zeros and bi-modality. In general, the PP-FIML estimator is shown to perform better than the alternative LEF-restricted approaches in terms of fit and estimation of the treatment effect. Also, as an example of a field-data application, we use a data set on the demand of trips by households already analyzed in Terza and Wilson (1990) and Terza (1998). We show that a Poisson with unobserved heterogeneity fails to generate predictions of zeros and ones and is rejected at 5% confidence level. Instead, a polynomial expansion of degree 2 is enough to improve the fit significantly, as shown by a battery of information criteria and goodness of fit tests. We find that the impact of the endogenous binary regressor on the counts (Treatment Effect) differs significantly between the PP-FIML and a model which imposes LEF, showing that the inference on the Treatment Effect under a LEF assumption could contain an important bias in practice.

At the end of the paper, the appendix provides technical details about the computational specifics of the PP-FIML model as a help for the reader interested in such issues.

2 Counts with an endogenous binary regressor

Assume we have a sample of size N from a count random variable Y and covariates X, d and Z where d is a binary variable taking values zero or one, X is a vector of regressors and Z is a vector of other covariates possibly containing at least some or all of the regressors in X. For each $i = 1, \ldots, N$, the conditional probability function of observation y_i is given by $f(y_i \mid x_i', d_i, \varepsilon_i)$ where ε is an unobserved random variable with zero mean. We will also assume that the covariance between d and ε is different from zero which implies that d is

endogenously determined.

Mullahy (1997)[2] assumes a linear exponential specification (LEF) for the conditional mean, i.e.,

$$E(y_i \mid x_i', d_i, \varepsilon_i) = \exp(X_i'\beta + \gamma d_i + \varepsilon_i), \tag{1}$$

where β and γ are unknown parameters. He shows that if $E(z_i\varepsilon_i) = 0$ for all $i = 1, \ldots, n$, then

$$E\left[\exp(-x_i'\beta - \gamma d_i)y_i - 1 \mid z_i\right] = 0, \tag{2}$$

where the covariates Z play the role of instrumental variables. Thus, the orthogonality condition in (2) permits us to define an Instrumental Variables estimator through the Generalized Method of Moments (GMM) techniques. This GMM estimator does not require additional distributional assumptions on either ε or Y. Note however that the assumption of a LEF mean for the counts (see equation 1) is necessary for (2) to hold. Alternative specifications of the conditional mean would require that a closed-form expression similar to (2) was available, which is not true in general.

In the context where the only source of endogeneity is given by the binary regressor, an alternative appears in Terza (1998). This author assumes that the binary variable d_i is generated by the following process

$$d_i = \begin{cases} 1 & \text{if } z_i'\alpha + v_i > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

where α is a vector of parameters conformable with the instruments and v is another error term. It is assumed that conditional on the exogenous variables, the ε_i and the v_i have non-zero correlation. Assume that ε can be decomposed into two independent random variables, $\varepsilon = \epsilon + \zeta$ such that (ϵ, v) follows a bivariate normal distribution with zero mean and variance-covariance matrix

$$\Omega = \begin{bmatrix} \sigma_\epsilon^2 & \sigma_\epsilon\rho \\ \sigma_\epsilon\rho & 1 \end{bmatrix}.$$

It must be noted that the specification above generalizes the model in Terza (1998) as we have distinguished between the heterogeneity which is correlated (ϵ) and uncorrelated (ζ) by the binary variable (d). Note also that the distribution of (ε, v) will not be completely characterized until the distribution of the uncorrelated error term ζ is defined. Indeed, Terza (1998) is a particular case which assumes that ζ is normal. For the moment however, we only need to assume that ζ has a moment generating function $M_\zeta(s) \equiv E[\exp(\zeta s)]$ with well-defined derivatives up to the third order in an open interval containing zero. This amounts

[2]Windmeijer and Santos-Silva (1997) follow a similar approach.

to ensuring that there exists $\bar{\zeta} \equiv \left.\frac{\partial \ln M_\zeta(s)}{\partial s}\right|_{s=0} = E(\zeta)$ [3] and $\sigma_\zeta^2 \equiv \left.\frac{\partial^2 \ln M_\zeta(s)}{\partial^2 s}\right|_{s=0} = V(\zeta)$.

Proposition 1 (Second Order Normal Approximation). *: Say $w = (\epsilon, v)$ and let*

$$\widetilde{w} \sim N\left[\begin{pmatrix} \bar{\zeta} \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 + \sigma_\zeta^2 & \rho\sigma_\epsilon \\ \rho\sigma_\epsilon & 1 \end{pmatrix}\right].$$

Let $M_w(s), M_{\widetilde{w}}(s)$ be the moment generating functions of w and $\widetilde{w}$ respectively for any $s = (s_1, s_2) \in \mathbb{R}^2$. Then,

(i) *$M_w(s) = M_{\widetilde{w}}(s) \exp[R(s_1)]$ where $R(s_1)$ is a polynomial in s_1 such that $\lim_{s_1 \to 0} \exp[(R(s_1)] = 1$ and there exists $\lim_{s_1 \to 0} R(s_1)/s_1^3 \neq 0$*

(ii) *In particular, if ζ follows a normal distribution, then $M_w(s) = M_{\widetilde{w}}(s)$.*

Proof. First, note that $w = \begin{pmatrix} \epsilon \\ v \end{pmatrix} + \begin{pmatrix} \zeta \\ 0 \end{pmatrix}$. Therefore,

$$M_w(s) = E\left(e^{s'w}\right) = \exp\left\{\tfrac{1}{2} s'\Omega s + \ln M_\zeta(s_1)\right\}. \tag{4}$$

Now, consider the Taylor expansion of $\ln M_\zeta(\cdot)$ around zero:

$$\ln M_\zeta(s_1) = \bar{\zeta} s_1 + \frac{1}{2}\sigma_\zeta^2 s_1^2 + R(s_1) \tag{5}$$

where $R(s_1)$ is a Taylor polynomial in s_1 which means that $\lim_{s_1 \to 0} \exp[R(s_1)] = 1$ and that there exists $\lim_{s_1 \to 0} R(s_1)/s_1^3 \neq 0$. Substituting (5) in (4) we prove (i). Finally note that because the sum of normals is itself normal, the proof of (ii) is trivial. ■

Proposition 1 defines a second order normal approximation to the (unknown) distribution of the unobserved heterogeneity vector (ε, v). This approximation is exact in the case of a normally distributed ζ which is the case in Terza (1998). Otherwise, it will depend on how far the moments of ζ are from those of a normal, in particular the moments of order higher than or equal to three. This is what the Taylor rest polynomial $R(s_1)$ accounts for. Because the estimators proposed in Terza (1998) are obtained under the assumption of joint normality of the vector (ε, v), proposition 1 allows us to interpret the properties of these estimators through the assumptions on the (unknown) distribution of the error term ζ. We will consider two useful cases:

[3] Note also that we have not assumed here a zero mean for ζ. This is left intentionally, as one of the two cases that we analyze later is one where the expectation of the exponential of the ζ variable has mean equal to one which does not imply in general that $E(\zeta) = 0$.

2.1 Case 1: Normal Error

Terza (1998) uses this assumption jointly with a linear-exponential function for the first order moment of Y (see equation 1) to build a Two-step Heckman-type estimator (TSM). Because the TSM estimator uses only the first order moment of the dependent variable, a natural question is whether TSM could be improved through the use of higher order moments. Thus, a Weighted Nonlinear Least Squares (WNLS) is also proposed. This estimator loses some robustness with respect to the TSM as it requires us to assume a Poisson process for Y.

But also under the assumption of a Poisson for Y, a FIML estimator is readily available. Say $\tilde{\sigma}^2 = \sigma_\epsilon^2 + \sigma_\zeta^2$ and $\tilde{\rho} = \rho\sigma_\epsilon/\tilde{\sigma}$, then it follows that $v = (\tilde{\rho}/\tilde{\sigma})\varepsilon + u$ where $u \sim N(0, 1-\tilde{\rho}^2)$, independent with respect to ε. Say

$$\begin{aligned} f_P(y_i \mid x_i', d_i, \varepsilon_i) &\equiv \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} \\ \text{where } \lambda_i &= \exp(x_i'\beta + d_i\gamma + \varepsilon_i). \end{aligned} \tag{6}$$

Collect all parameters of interest in $\theta = (\beta, \gamma, \alpha, \tilde{\sigma}, \tilde{\rho})$. From (6) and (3) we have that

$$f(y_i, d_i | x_i', z_i, \theta) = \int_{-\infty}^{+\infty} f_P(y_i | X_i, d_i, \varepsilon) \Phi^*(\varepsilon)^{d_i} [1 - \Phi^*(\varepsilon)]^{1-d_i} \, \mathrm{d}\Phi(\varepsilon/\tilde{\sigma}) \tag{7}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal and $\Phi^*(\varepsilon)$ is defined as $\Phi\left([z_i\alpha + (\tilde{\rho}/\tilde{\sigma})\varepsilon]/\sqrt{1-\rho^2}\right)$.

Full Information Maximum Likelihood estimation of the model has an important advantage over its TSM and WNLS alternatives, namely that it is expected to use the information more efficiently. In general, the Poisson-FIML is inconsistent if the count variable does not follow a Poisson distribution. The same can be said of the WNLS as it also requires a Poisson assumption. The TSM, though, retains consistency as long as the first order moment is correct and the error terms follow a normal joint distribution. For this reason, we will propose in section 3 a FIML estimator which is expected to be flexible and robust against data generating processes which include instances where log-linearity of the mean is not fulfilled. From a computational point of view, FIML requires us to use numerical integration in (7), a minor difficulty thanks to the availability of numerical integration software packages. See appendix B on details of computation.

Note that, though consistent estimators of $\tilde{\sigma}$ and $\tilde{\rho}$ can be found, identification of σ_ϵ and ρ is not feasible. The reason is that the unobservability of ε makes it impossible to distinguish between the heterogeneity induced by ϵ from the one induced by ζ. In other words, using data on Y, X, d and Z it is possible to find an estimate of the variance of $\zeta + \epsilon$, but not of the variance of ζ and ϵ separately.

2.2 Case 2: Exp-Gamma Error

Consider now that $\exp(\zeta)$ follows a Gamma distribution $\text{Gamma}(\eta, \eta)$ where $\eta > 0$. It can be shown by integration of ζ (see Cameron & Trivedi, 1998) that if Y follows a Poisson conditional on ϵ then the distribution of Y conditional on ϵ is Negative Binomial with mean λ and variance $\lambda + \eta\lambda^2$. Consequently, the Negative Binomial (NegBin, henceforth) itself accounts for the overdispersion induced by the error term ζ, but not for that of ϵ.

The probability function of the Negative Binomial is defined in this context as,

$$\begin{aligned} f_{NB}(y_i \mid x_i', d_i, \epsilon_i) &\equiv \frac{\Gamma\left(Y_i + \eta^{-1}\right)}{\Gamma\left(\eta^{-1}\right)\Gamma\left(Y_i + 1\right)} \eta^{Y_i} \left(\frac{1}{1+\eta\lambda_i}\right)^{\eta^{-1}} \left(\frac{\lambda_i}{1+\eta\lambda_i}\right)^{Y_i}, \qquad (8) \\ \text{where } \lambda_i &= \exp\left(X_i\beta + d_i\gamma + \epsilon_i\right) \text{and } \eta > 0. \end{aligned}$$

with $\Gamma(\cdot)$ denoting the Gamma function. Thus, we would use (8) instead of (6) in (7). This, defines a Negative Binomial FIML (NB-FIML) estimator.

The Negative Binomial law is a popular choice in count data models. Typically, the Negative Binomial allows for overdispersion without the need of numerical integration. In our context though, this does not represent an advantage as the Poisson-FIML also allows for overdispersion and numerical integration is required in both Poisson and NegBin FIML models anyway. Consequently, the benefits from using a NB-FIML are not expected to increase dramatically from those of the Poisson-FIML, both in terms of fit and/or computational effort. Moreover, there is no a priori assessment of why the NegBin should show better fit than Poisson-FIML for it is not clear why a exp-gamma for the unobserved heterogeneity should be preferred to any other distribution like Normal, as above. And more importantly, the identification of the NB-FIML is challenging. The main reason is that the NB-FIML approach adds a new parameter, η, which determines the variance of ζ. As said above, the variance of ζ is not identified should ζ follow a Normal law since the sample only gives information on the variance of $\zeta + \epsilon$ together and not separately. In the context of FIML based on (8) though, identification of η may be achieved but it should be based on the moments of order higher or equal than three, as the second order moments of the sample only give information about the *joint* variation of $\epsilon + \zeta$. Relying on these high order moment conditions for identification poses a technical challenge to the NB-FIML estimator as the objective function is likely to be almost flat in a region around the optimum if the high order sample moments of ζ do not depart substantially from those of a normal.

Finally, the following corollary shows that the Poisson-FIML retains consistency even when the data have been generated through a NegBin distribution, at least for the most relevant parameters.

Corollary 1. *Assume that* $\exp(\zeta) \sim$ *Gamma*(η, η). *Say* $\Psi_{(k)}(\cdot)$ *the digamma function defined as the* $\Psi_{(k)}(\cdot) = \partial_k \ln\Gamma(\cdot)$. *Then, the Poisson-FIML estimates* $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{dim(X)})$ *and* $\hat{\gamma}$ *are consistent for* $(\beta_0 + \ln\eta + \Psi_{(1)}(\eta), \beta_1, \ldots, \beta_{dim(X)})$ *and* γ *respectively.*

Proof. First, because $\exp(\zeta) \sim \text{Gamma}(\eta, \eta)$, then

$$\ln M_\zeta(s_1) = \ln \Gamma(\eta + s_1) - \ln \Gamma(\eta) + s_1 \ln \eta \tag{9}$$

By proposition 1.i) we have that

$$M_w(s) = \exp\left\{ \frac{1}{2} s' \Omega s + s_1 \left(\ln \eta + \Psi_{(1)}(\eta) \right) + \frac{\Psi_{(2)}(\eta) s_1^2}{2}) \exp\left[R(s_1) \right]. \right\} \tag{10}$$

Arranging terms, we find that w follows approximately

$$N \left[\begin{array}{c} \ln \eta + \Psi_{(1)}(\eta) \\ 0 \end{array} , \left(\begin{array}{cc} \sigma_\epsilon^2 + \sigma_\zeta^2 & \rho\sigma_\epsilon \\ \rho\sigma_\epsilon & 1 \end{array} \right) \right]. \tag{11}$$

Therefore, the Poisson-FIML estimate of the constant term β_0 is shifted by $\ln \eta + \Psi_{(1)}(\eta)$, while the other coefficients β and γ remain unaffected. ■

Thus, a Poisson-FIML with a constant term provides consistent estimators of the parameters affecting covariates even when the data have been generated with a Negative Binomial.

3 Polynomial Poisson FIML (PP-FIML)

In light of the previous discussion, it could be argued that an estimator which did not impose a priori parametric restrictions in the distribution of the ζ should be preferred to restricting to a Normal or Gamma specification. But the choice of a distribution for this error term determines to a great extent the conditional distribution of the dependent variable once integrated out from $f(y_i \mid x_i', d_i, \epsilon + \zeta)$. Indeed, a Poisson count conditional on a Gamma ζ leads to Negative Binomial conditional on ϵ. Then, there are at least two ways in which the parametric restrictions in our model can be relaxed: one is to keep a simple specification of Y conditional on ζ (Poisson for instance) while relaxing the assumption on the distribution of ζ. The other is to keep a simple specification for ζ (Normal for instance) and relax the assumptions on the conditional distribution of Y. We found the second approach much simpler to implement since we can benefit from the work by Cameron and Johansson (1997) [4].

Another reason to opt for a polynomial expansion of $f(y_i \mid x_i', d_i, \epsilon + \zeta)$ is that we do not restrict the mean of the process to be based on a LEF. The estimators proposed in Terza (1998), i.e., the TSM, the WNLS and the Poisson/NegBin-FIML, assume a LEF function for the mean, pretty much as the instrumental variables approach in Mullahy (1997) does. Then, it is reasonable to ask what is the benefit against the "cost" of assuming a normal distribution for (ε, v) instead of using such a GMM approach which does not requires distributional assumptions on the unobserved heterogeneity. Our answer here is that assuming normality

[4] Note that our model differs from Cameron and Johansson (1997) in two ways. First we introduce overdispersion through the normal unobserved heterogeneity. The second is that we take into account endogeneity.

of the unobserved heterogeneity facilitates the task of implementing a "flexible" specification of $f(Y_i|X_i,\ d_i,\ \varepsilon_i)$ and particularly of the mean of the counts.

Assume that, conditional on X,d and ε, Y has a probability function given by a polynomial expansion around a baseline Poisson. Thus, say $a = (a_0, a_1, \ldots, a_K)$ the vector of coefficients of the polynomial and say $\theta = (\beta, \gamma, \alpha, \tilde{\sigma}, \tilde{\rho}, a)$, then

$$f_{PP}(y_i \mid x_i', d_i, \varepsilon_i, \theta) \propto \left(\sum_{k=1}^{K} a_k y_i^k \right)^2 f_P\left(y_i \mid x_i', d_i, \varepsilon_i\right), \tag{12}$$

where f_P is defined as in 6. The PP-FIML estimators would then be obtained by maximization of the FIML objective function

$$(1/N) \sum_{i=1}^{N} \ln \int_{-\infty}^{+\infty} f_P(y_i|X_i, d_i, \varepsilon) \Phi^*\left(\varepsilon\right)^{d_i} \left[1 - \Phi^*\left(\varepsilon\right)\right]^{1-d_i} \mathrm{d}\Phi(\varepsilon/\tilde{\sigma}). \tag{13}$$

Our PP-FIML estimator has several interesting features:

First, by simple algebra on (12), the mean of the count variable conditional on both observable and unobservable variables is given by

$$E\left(y_i|x_i', d_i, \varepsilon\right) = \sum_{j=0}^{K} \sum_{h=0}^{K} a_i a_h m_{j+h}, \tag{14}$$

where m_j stands for the j^{th} non-central moment of the Poisson density with mean $\lambda_i = \exp\left(x_i'\beta + d_i\gamma + \varepsilon_i\right)$. Thus, a polynomial expansion implies a departure from the standard LEF specification, allowing for a more flexible modelling. Note that when $K = 0$, the expression in (14) reduces to λ_i, thus nesting the linear exponential case.

Second, it is expected that a polynomial expansion could approximate any model arbitrarily well as long as we increase the order of the polynomial.[5] However, increasing the polynomial size arbitrarily when the sample size is fixed may lead to overfitting the data, which leads to the question of how to fix the size K in applications.

3.1 Deciding on the polynomial size: Specification Test

We propose two sets of rules for deciding on the polynomial size: information criteria and specification testing. In the literature on polynomial expansions it is common (see Gallant & Tauchen, 1997; Gurmu & Trivedi, 1996; Gurmu, 1997, or Cameron & Johansson, 1997) to use information criteria such as the Bayesian Information Criteria (BIC) and/or Consistent Akaike Criteria (CAIC). If $\mathfrak{L}$ denotes the log-likelihood and p is the number of parameters to be estimated, then BIC $= -2\ln\ \mathfrak{L} + p\ln\ n$ and CAIC $= -2\ln\ \mathfrak{L} + p(\ln\ n + 1)$. Gallant and

[5] Cameron and Johansson (1997) or Creel (1999) do not provide a proof of this claim but show numeric simulations. This has been proven in the continuous case by Gallant and Nychka (1997).

Tauchen (1997) advocate the use of BIC as a parsimonious criterion on the order of the polynomial. The BIC imposes a bigger penalty on the number of parameters than the standard Akaike, but not as big as that imposed by the CAIC. Considering a penalty on the number of parameters is interesting, since one would like to avoid overparameterized models.

A disadvantage of the information criteria is that the unconstrained model needs to be estimated to decide on the polynomial size. Moreover, it is not clear when a difference in information criteria is statistically significant. We believe that specification test can be useful for the applied econometrician to decide on the polynomial size, as well as whether or not is necessary to estimate more complicated models. The FIML approach permits us to define such a test which is based on Andrews (1988b, 1988a) and compares the expected and the observed frequencies of the counts. First, we partition the range of the count variable into J intervals, where $c_1 \geq c_2 \geq ... \geq c_{J-1} > 0$ are the endpoints of the intervals. The observed frequency, p_j, of the interval $j = 1,\ 2,\ ...,\ J$ is given by

$$p_j = \frac{1}{N} \sum_{i=1}^{N} \mathbf{I}_{[c_j \leq y_i \leq c_{j+1}]}, \tag{15}$$

where $\mathbf{I}[.]$ is the indicator function. Now, the expected frequency $\hat{p}_j$ of the $j^{th.}$ interval can be computed using $f(Y, d \mid x_i', z_i, \hat{\theta})$ as an estimate of the true joint distribution of Y and d, then marginalizing the count variable

$$f\left(y_i | x_i', z_i, \hat{\theta}\right) = f\left(y_i, 1 | x_i, z_i, \hat{\theta}\right) + f\left(y_i, 0 | x_i, z_i, \hat{\theta}\right). \tag{16}$$

Thus, an estimate of the expected frequency $\widehat{p}_j$ is given by

$$\widehat{p}_j = \sum_{i=1}^{N} \sum_{Y \in c_j} f\left(Y | \, x_i, z_i, \hat{\theta}\right) \tag{17}$$

Under the null of a correct specification, $\Delta_j \equiv |p_j - \widehat{p}_j|$ converges to zero. The goodness of fit measures used in Gurmu and Trivedi (1996) and Cameron and Johansson (1997) based on the sum of the differences Δ_j can thus be extended to our context of endogenous regressors as a moment conditions test on Δ_j. Numerical integration is needed at some steps of the implementation to evaluate (17). The interested reader may consult the appendix B on computational methods at the end of the paper.

3.2 Estimating the impact of regressors

In a LEF model, the coefficients β and γ have a deep structural meaning. Say x_{ij} the $jth.$ regressor in X_i, then, from (1) it follows that

$$\begin{aligned} \frac{\partial \ln E\left(y_i|\, x_i', d_i, \varepsilon_i\right)}{\partial \ln x_{ij}} &= \beta_j \quad \text{and also} \\ \ln \frac{E\left(y_i|\, x_i', d_i = 1, \varepsilon_i\right)}{E\left(y_i|\, x_i', d_i = 0, \varepsilon_i\right)} &= \gamma. \end{aligned} \tag{18}$$

for all $i = 1, \ldots, N$. Hence, these coefficients are the elasticities of expected value of the count with respect to covariates. However, in a PP-FIML the coefficients β and γ lose such a meaning as the expectation $E(y_i \mid x_i', d, \varepsilon)$ is no longer linear-exponential (see equation 14). Since the researcher is often interested in a measure of the impact of covariates in the counts, the question we want to address is how to recover such a measure with the PP-FIML model. The idea is that once an estimate of the joint distribution of the count and the dummy is available, estimates of any moment of Y and/or d conditional on covariates can be computed. In particular, the quantity $E[y_i|x_i', Z_i, d_i]$ can be estimated by

$$\hat{E}[y_i|x_i', Z_i, d_i] = \sum_{y=1}^{\infty} y \frac{f(y, d_i|X_i, Z_i, \hat{\theta})}{\hat{f}(d_i|Z_i, \hat{\theta})}. \tag{19}$$

The marginal density of d in the previous expression can be computed as follows. Define $\hat{\Phi}^*(\varepsilon)$ as in (7) where the unknown parameters are replaced by their corresponding PP-FIML estimates, then,

$$\hat{f}(d|z_i', \hat{\theta}) = \int_{-\infty}^{+\infty} \left\{ d\hat{\Phi}^*\left(\varepsilon\right) + (1-d)\left[1 - \hat{\Phi}^*\left(\varepsilon\right)\right] \right\} d\Phi(\varepsilon). \tag{20}$$

Define the *treatment* effect as the variation in $\hat{E}[Y_i|X_i, Z_i, d_i]$ induced by the change of treatment from $d = 0$ to $d = 1$. An estimate of this quantity is given by

$$\frac{1}{N} \sum_{i=1}^{N} \frac{\hat{E}(y_i|x_i', z_i', d = 1) - \hat{E}(y_i|x_i', z_i', d = 0)}{\hat{E}(y_i|x_i', z_i', d = 0)}. \tag{21}$$

for each $i = 1, \ldots, N$. Similar measures can be defined for each of the regressors in the X vector. As usual, an estimate of the standard deviation of (21) can be computed with a first order linear approximation of (21) around the true value of the parameters (Delta method).

4 Empirical Results

4.1 A Simulation Exercise

This section provides an illustration of the properties of the PP-FIML estimator using simulated data generated by random sampling from five different specifications or data generating processes (DGP). The number of simulations used in all experiments is set to 100 and the

sample size was fixed to 1000 observations per Monte Carlo iteration. In order to prevent convergence to local optima we used in each run several starting values for a gradient-based optimization algorithm. In this simulation exercise we fix the polynomial size at 2. The results for the PP-FIML reported in this simulation exercise are somewhat a lower bound of what can be get with this technique as they could be improved if we allowed for larger polynomials. This strategy would entail to look for the better specifications at each Monte Carlo. This task would be computationally very intensive and time consuming.

Prior to any simulation we draw samples of two independent N(0,1) regressors. One of them will be excluded from the count equation to ensure that we have enough instruments in the binary equation. The draws of covariates are kept constant across all the 100 simulations. Then, for each simulation and for each DGP, we draw a sample of size 1000 of the unobserved heterogeneity vector (ε, v) from a bivariate normal with mean zero and variance-covariance matrix Σ. Then, we draw a sample from the probability function $f(\cdot\,|x, d, \varepsilon)$. The definitions of this probability function and of Σ for each DGP appear in Table 1 in detail.

The five DGP's are labelled DGP1 trough DGP5. These specifications and their parameters have been selected intentionally to yield data with different degrees of overdispersion and/or excess of zeros. Table 2 shows the ratios between the observed frequencies in the simulations (numerator) and the frequencies predicted under Poisson (denominator)with the correct mean evaluated at the average covariates.[6] DGP1 is not included in this table as it is our benchmark model. DGP1 is generated from a Poisson distribution with LEF mean $\lambda = \exp(x\beta + \gamma d + \varepsilon)$ and normal unobserved heterogeneity. DGP2 keeps the assumption of a LEF but data is generated from a Negative Binomial distribution with overdispersion parameter $1/\eta \equiv 2$. DGP3 simulates a Hurdle Poisson model with a non-LEF mean. Hurdle models (Pohlmeier & Ulrich, 1995) are very popular in many applications. They are a mixture of two processes: one driving the zero and non zero observations and the other one being a count process truncated at zero. Hurdle models typically show an excess of zeros and this is confirmed by comparing the frequency of zero and one counts in Table 2 for the DGP3 column. DGP4 simulates a Hurdle Negative Binomial model. Finally, DGP5 simulates from an equally weighted mixture of Poisson distributions. Mixture models have been also proposed in the literature as a flexible estimator for count processes. We use a mixture model to assess the performance of our flexible PP-FIML estimator when the counts have been generated by other well-known flexible model. Looking at Table 2, it can be noted that parameters of the mixture have been selected in order to have very different shapes of the counts depending on the treatment. Under no treatment (d=0), the mixture shows long tails, while under treatment, it shows a very high proportion of zeros.

The performance of the PP-FIML model will be tested against different alternatives for each of the DGP's. In the case of the DGP1 the TSM, the Poisson-FIML and the PP-FIML yield consistent estimators of the parameters of interest, as the DGP1 maintains the LEF assumption. Thus, the results for this DGP can be used as a benchmark for comparing all

[6]Except for the mixture model where the x_1 is set to 1. Notice that when x_1 equals zero then both parts of the mixture collapse to a Poisson density.

three estimators in term of efficiency because, as expected, the bias is very small in all cases. Table 3 reports the squared errors of the three estimators averaged through simulations. The reader should pay attention to the upper panel of this table which reports the variance, as the size of the bias is almost negligible. The Poisson-FIML shows the lowest error, even lower than the PP-FIML with a polynomial of size 2. Recall that the Poisson-FIML can be interpreted as a PP-FIML estimator which incorporates the restriction that the coefficients of the polynomial are zero. Note that the differences between the TSM and the Poisson-FIML are larger at the parameter (γ) associated with the endogenous binary. Finally, note that the degree 2 PP-FIML performs poorly for the constant term of the count equation (β_0). According to Cameron and Johansson (1997) the same values of the count mean and variance are compatible with different combinations of the λ and the first parameter of the polynomial. This feature does not preclude identification by FIML as higher order moments will differ but it would affect the precision of the estimation of β_0 and the first order parameter of the polynomial. In any case it could be solved by just adding the appropriate restrictions on the first coefficient of the polynomial.

The LEF assumption still holds for the DGP2 but the Poisson-FIML and the PP-FIML are expected to give inconsistent estimates of the parameter for the constant term β_0. Given our corollary 1, they should provide consistent estimates of β_1 and γ. As shown in section 2 the bias is given by the expression in corollary 1. Table 4 reports the squared error for all parameters where the error for the constant term in the FIML approach has been computed with respect to the "shifted" constant term, i.e., $\beta_0 + \ln(\eta) + \Psi_(1)(\eta)$. Note that the bias of β_1 and γ is not bigger for FIML than it is for TSM. Notice that this provides the empirical counterpart to our corollary 1. Comparing with DGP1, the results in terms of bias and MSE of the different estimators are qualitatively similar.

The LEF assumption does not hold for the DGP3, DGP4 and DGP5. In these cases, all estimators are expected to yield inconsistent estimates of the parameters θ. Therefore, the comparison should be done on the basis of the ability of these estimators to yield good inference on the impact of the regressors on the mean of the counts and the treatment effect, as discussed in the previous subsection. Table 5 performs such an exercise. In Hurdle models, the estimators obtained under the LEF assumption, i.e., the TSM and the Poisson-FIML provide a biased estimate of the impact of the binary regressor on the counts, while the bias of the PP-FIML estimator is almost negligible. In the case of DGP5 mixture of Poisson, the LEF assumption induces a significant bias on the estimate of the treatment effect overestimating this quantity by around 19% for the TSM and 8% for the Poisson-FIML. Our PP-FIML estimator overestimates the treatment by just a 3%. Finally, it must be noted that the Polynomial Poisson provides a better BIC and CAIC than the Poisson-FIML.

As a summary, the TSM model performed well in DGP1 and DGP2 where the LEF assumption is verified. We were not expecting TSM to give good estimates of the treatment effect for DGP3 to DGP5 where this assumption does not hold. Manning and Mullahy (2001) have found that moment based estimators based on a LEF are prone to overfitting when there is substantial skewness in the count distribution as it is the case in DGP3 to DGP5 which

show long tails that produce skewness. The PP-FIML provides better estimates than TSM and Poisson-FIML for the three designs that do not show a LEF (DGP3, DGP4 and DGP5).

4.2 An Application to Data on Trip Frequency

Terza (1998) used data on the number of trips by households (Tottrips) to specify a model where vehicle ownership (OwnVeh) is included as a binary regressor. Indeed, it is reasonable to believe that there may exist unobserved variables such as the personal predisposition (or aversion) to travel which may be positively (or negatively) correlated with the decision of purchasing a vehicle. For instance, an individual may like to travel but may detest traffic jams, and such an aversion will be negatively correlated with the ownership of a vehicle. Thus, the one would wish to isolate the effect of vehicle ownership accounting for the correlation with the unobservables while being confident that its estimate will be robust against misspecification on the first order moment of the distribution of the counts.

Table 6 describes the variables in the data set. Some variables have been scaled with respect to the original source and they have been divided in two groups attending to their status: endogenous (number of total trips and vehicle ownership) and exogenous regressors. The Tottrips variable has a sample variance which is almost five times greater than the sample mean. Additionally, it also shows a relatively big frequency of zeros. The sample contains a frequency of zeros which is 17 times greater than would be expected from a Poisson with mean equal to sample mean.

The choice of regressors for the count and the binary equations must take into account issues of identification regarding exclusion restrictions of the coefficients in the count equations. The question is then to decide which variables to exclude. This dilemma is not exclusive to our framework and it is encountered in previous approaches such as, among others, Mullahy (1997) and Windmeijer and Santos-Silva (1997). Because this application is intended as a comparative exercise for illustrative purposes, we will adopt the specification in Terza (1998) and exclude the Adults regressor from the count equation.

The estimation of Nonlinear Least Squares (NLS), the TSM, and WNLS models in Tables 7 and 8 give a first impression of the consequences of endogeneity of the OwnVeh variable. As mentioned in Section 2, TSM and WNLS correct for endogeneity using a estimator similar to the one proposed by Heckman (1978) but adapted to this particular count data framework. Column 1 of table 7 contains the estimates of a PROBIT model of OwnVeh on a set of regressors. The estimates of the PROBIT part are used in a second stage process (Table 8) to compute the corrected moment conditions, which define the TSM and WNLS estimators. The value of the OwnVeh coefficient estimated with TSM and WNLS increases from between 30% to 75% with respect to NLS. This indicates that the sign of the correlation between the unobserved heterogeneity and the endogenous dummy is negative. The WNLS pursues a more efficient estimation than TSM at the price of restricting the parametric family of the conditional counts to be a Poisson. For instance, a test of the significance of some variables like FullTime may lead to different conclusions under TSM or WNLS.

The fourth column of table 8 contains the estimates of the count equation for the Poisson

FIML model. A Poisson FIML approach (K=0)[7] as shown in section 2 is based on the same assumptions as the WNLS, namely a Poisson conditional probability function for the counts in addition to joint bivariate normality and a linear exponential specification for the mean. The estimates of the coefficients are more similar to the WNLS than to TSM, particularly for the DistoCbd variable.

The FIML approach allows further diagnostics to be made on both the Poisson assumption and LEF, by means of comparing the expected and observed frequencies. Table 9 shows the empirical and expected frequencies for several counts intervals. The K=0 model underpredicts the frequency of zeros and overpredicts the frequency of counts one and two, as usually happens when the empirical distribution puts an excess of mass in the zero counts. In fact, the Andrews' test rejects the null of a correct specification at 1% for the K=0 model. Using an informal test, Terza (1998) also found evidence of misspecification for the Poisson assumption. This makes it clear that a model that specifies a LEF and accounts simultaneously for overdispersion may not adequately fit a sample with an excess of zeros such as the one at hand (for another example of this problem, see Gurmu and Trivedi, 1996). Consequently, this provides motivation for a more flexible specification introducing a polynomial series expansion over a Poisson baseline density as proposed in section 3. We started with the K=1 specification and sequentially increased the size of the polynomial. The parameter estimates for the count and binary equations are shown in the last columns of Table 7 and 8 respectively, while expected frequencies appear in table 9. In term of goodness of fit, a considerable gain is obtained by the model with K=2 with respect to K=0 and K=1. As Table 9 shows, the measure of the distance between observed and predicted frequency decreases considerably and the test does not reject the null for a size of the polynomial of two or higher. This suggest that the polynomial terms considerably improve the fit of the model.

This leads to the problem of having to take a decision on where to stop adding new terms to the polynomial expansion. We used the information criteria in section 3, and the results are shown in the last rows of Table 8. The BIC favors K=2 with respect to any other model which is not rejected by the Andrews test. It must be noted, though that the CAIC for K=0 almost matches that of the K=2.

Table 8 also shows that the OwnVeh coefficient differs across models, being 2.796 for TSM and 2.369 for K=2. It is important to recall that once a LEF specification is not accepted, these estimates have no direct structural interpretation. Following the discussion in section 3 the researcher should not be interested in the coefficients themselves, rather than on the way they can affect (cause) the characteristics of the count variable (for instance, its mean). In order to make comparisons of these mean effects, we used the formula in section 3.1. Table 10 shows the estimates of the mean effect computed for the OwnVeh variable and other regressors. The point estimates for the mean effect of the OwnVeh variable are 0.5798 for TSM and 1.3718 for K=2 in per-unit points. A 95% confidence interval for the latter is approximately (0.5326,2.211). Although the confidence interval is too wide, the difference

[7]In the tables, we chose this nomenclature to emphasize that a Poisson distribution is a particular case of our polynomial expansion with a polynomial size of zero.

between both estimates is considered important enough to serve as an illustration that a researcher should consider using a flexible alternative as our PP-FIML model when there is evidence that a LEF specification is not appropriate.

5 Conclusions

Terza (1998), Mullahy (1997) and Windmeijer and Santos-Silva (1997) provide consistent estimators for count data in presence of a dummy endogenous variable provided a LEF assumption holds. These estimators can be useful in many applications. However, the literature with exogenous regressors has shown that the LEF assumption does not always hold. Our paper builds on Terza (1998) and Cameron and Johansson (1997) to provide an estimator based on a polynomial expansion of a baseline Poisson process. The literature with exogenous regressors has already shown that flexible models that depart from the linear exponential specification fit the data better. We extend this idea to the case where a potentially endogenous binary variable is included as a regressor. Our estimator is expected to improve the fit with respect to LEF based alternatives in those cases where such an assumption does not hold. Deb, Ming, and Trivedi (2001) argued that some distribution characteristics such as an excess of zeros or overdispersion are not likely to be captured by estimators which use only low order moment restrictions. For this reason we base our estimation strategy on Maximum Likelihood estimation using a flexible form.

A small Monte Carlo experiment shows that, at least for our specific coefficient values, the Polynomial Poisson does a good job in improving the fit and getting good estimates of the Treatment Effect parameter even if the true DGP is generated from non-LEF data generating processes, like a finite mixtures of Poisson or Hurdle models. As an illustration, we use a data set on the number of trips by households already analyzed in the literature. The results show that flexible estimation of the conditional probability function of the count helps to significantly improve the fit with respect to LEF alternatives. In particular, we find that a model with a polynomial expansion of size two can not be rejected by the data, while a model based on a LEF specification is. We also find that the estimates of the impact of regressors on the counts differ, so stressing the need of using a flexible form like the PP-FIML when the data is suspected of showing non-LEF.

Finally, at the beginning of section 3 we claimed that our approach is far from being unique. Semiparametric alternatives based on a flexible specification of the distribution of the error term can be devised. [8]

[8]Newey, Powell, and Walker (1990), page 328, claim that "*specification of the regression function and set of instrumental variables appears to be more important than specification of the error distribution for these data*". This evidence is further supported in Vella (1995).

References

Andrews, D. (1988a). Chi-square diagnostic tests for econometric models: Introduction and applications. *Journal of Econometrics*, *37*, 135-156.

Andrews, D. (1988b). Chi-square diagnostic tests for econometric models: Theory. Econometrica, 56, 1419-1453.

Cameron, A. C., and Johansson, P. (1997). Count data regression using series expansions with applications. *Journal of Applied Econometrics, 12,* 203-233.

Cameron, A. C., and Trivedi, P. K. (1998). *Regression analysis of count data* (Vol. 30). New York: Cambridge University Press.

Coulson, N. E., Terza, J. V., Neslulan, C., and Stuart, C. (1995). Estimating the moral hazard effect of suplemental medical insurance in the demand for prescription drugs by the elderly. *AER papers and proceedings, 85*, 122-126.

Creel, M. (1999). *A flexible and parsimonious density for count data.* (Universitat Autonoma de Barcelona)

Creel, M., and Farrell, M. (2001). *Likelihood demand approaches to modelling demand for medical care*. Universitat Autonoma de Barcelona, working paper 498-01.

Deb, P., Ming, X., and Trivedi, P. K. (2001). *Finite mixture count regression: Maximum likelihood versus extended moment-based estimation*. mimeo. (Indiana University)

Deb, P., and Trivedi, P. (2002). *Specification and simulated likelihood estimation of a non-normal outcome model with selection: Application to health care utilization".*

Deb, P., and Trivedi, P. (2003). *Gatekeeping, self-selection, and utilization of curative and preventive health care services.* mimeo.

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

Gallant, A. R., and Nychka, D. W. (1997). Estimation of continuous time models for stock return and interest rates. *Econometrica*, *55*, 363-390.

Gallant, A. R., and Tauchen, G. (1997). Estimation of continous-time models for stock return and interest rates. *Macroeconomic Dynamics*, *1* (1), 135-168.

Goffe, W. L., Ferrier, G. D., and Rogers, J. (1994). Optimization of statistical functions with simulated annealing. *Journal of Econometrics*, *60*, 65-99.

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

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

Gurmu, S., Rilstone, P., and Stern, S. (1999). Semi-parametric estimation of count regression models. *Journal of Econometrics*, *88*, 123-150.

Gurmu, S., and Trivedi, P. K. (1996). Excess zeros in count models for recreational trips. *Journal of Business and Economics Statistics*, 469-477.

Kenkel, D., and Terza, J. (2001). The effect of physician advice on alcohol consumption: Count regression with an endogenous treatment effect. *Journal of Applied Econometrics*, 165-184.

Manning, W., and Mullahy, J. (2001). Estimating log models: To transform or not to transform. *Journal of Health Economics*, *20*, 461-494.

Mullahy, J. (1997). Instrumental-variable estimation of count-data models: Application to models of cigarette smoking behavior. *Review of Economics and Statistics*, *79*, 586-593.

Munkin, M., and Trivedi, P. (2003). Bayesian analysis of a self-selection model with mul-tiple outcome using simulation-based estimation: An application to the demand for healthcare. *Journal of Econometrics*, *114* (2), 197-220.

Newey, W. K., Powell, J. L., andWalker, J. R. (1990). Semi-parametric estimation of selection models: Some empirical results. *American Economic Review*, *80* (2), 324-328.

Pohlmeier, W., and Ulrich, V. (1995). An econometric model of the two-part decisionmaking process in the demand for health care. *Journal of Human Resources*, *30*, 338-361.

Schellhorn, M. (2001). The effect of variable health insurance deductibles on the demand for physician visits. *Health Economics*, 441-456.

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*, 108-115.

Vella, F. (1995). Estimating models with sample selection bias: A survey. *The Journal of Human resources*, *33* (1), 127-169.

Vera-Hernandez, A. M. (1999). Duplicate coverage and demand for health care: The case of catalonia. *Health Economics*, *8*, 579-598.

Windmeijer, F., and Santos-Silva, J. (1997). Endogeneity in count-data models: An applica-tion to demand for health care. *Journal of Applied Econometrics*, *12*, 281-294.

Appendix. Details on Computation

The numerical routine for integration of unobserved heterogeneity in (7) is based on the Gauss-Hermite quadrature. This is a popular choice when integrating normal variables across the whole real line. (Judd 1999). The procedure requires to specify the number of points for quadrature evaluation. We used 26 points of quadrature. Choosing a larger number of quadrature points did not influence the results.

The objective function of the Poisson-FIML was optimized using the Broyden-Fletchell-Golden-Shannon (BFGS) algorithm. We never found problems of local optima. Convergence time is quite low (between 2 and 4 minutes) depending on the initial conditions. Using TSM starting values improves convergence time substantially. The TSM estimator converges extremely fast. We have noted that the TSM might exhibit computational instability if the count takes very large values (i.e. 150). However, we have not pursued an in-depth analysis of this issue.

The objective function of the PP-FIML was optimized using the BFGS algorithm. We tried several initial conditions as local optima is a problem often encountered when using series expansion. In order to choose the initial conditions for the polynomial coefficients, it is advisable not to choose large numbers. We normally found that large values of the polynomial coefficients rarely provide an appropriate convergence. In all our results, the coefficients are not larger than one in absolute values. We found that local optima are an issue for the PP-FIML. This was particularly true for the DGP generated as mixtures of Poisson. Each run with the BFGS took approximately between 2 and 10 minutes in a Pentium III, depending on the polynomial degree and the initial condition.

For the application with real data, we wanted to ensure that the global maxima was reached, so we decided to implement, as a final step, several runs with a local-robust optimization algorithm like the simulated annealing (SA), which is a search method specifically designed to deal with the problem of multiple local optima (see Goffe, Ferrier, and Rogers, 1994). We benefitted from the code written by E.G. Tsionas. For the application, the SA algorithm matched the best result obtained using BFGS. All code is available from the authors on request.

TABLES

Name	Prob. Funct.	Parameter Values
DGP1	$y \sim P(\lambda)$ $\lambda = \exp(x\beta + \gamma d + \varepsilon)$	$\beta = (0.5, 0.5);\ \gamma = 1$ $\alpha = (0, 0.5, 0.5)$ $\Sigma = \begin{pmatrix} 0.25 & 0.25 \\ 0.25 & 1 \end{pmatrix}$
DGP2	$y \sim NB(\lambda, \eta)$ $\lambda = \exp(x\beta + \gamma d + \varepsilon)$	$\beta = (0.5, 0.5)\ ; \gamma = 1$ $\alpha = (0, 0.5, 0.5); \eta = 0.5$ $\Sigma = \begin{pmatrix} 0.25 & 0.25 \\ 0.25 & 1 \end{pmatrix}$
DGP3	$y_1^* \sim P(\lambda_1)$ $y_2^* \sim P(\lambda_2 \mid y_2^* > 0)$ $y = \begin{cases} 0 & \text{if } y_1^* = 0 \\ y_2^* & \text{if } y_1^* > 0 \end{cases}$ $\lambda_1 = \exp(x\beta_1 + \gamma_1 d + \varepsilon)$ $\lambda_2 = \exp(x\beta_2 + \gamma_2 d + \varepsilon)$	$\beta_1 = (-1, 0.5)\ ; \gamma_1 = 1$ $\beta_2 = (0.75, 0.5)\ ; \gamma_2 = 1$ $\alpha = (0, 0.5, 0.5)$ $\Sigma = \begin{pmatrix} 0.25 & 0.25 \\ 0.25 & 1 \end{pmatrix}$
DGP4	$y_1^* \sim NB(\lambda_1, \eta_1)$ $y_2^* \sim NB(\lambda_2, \eta_2)$ $y = \begin{cases} 0 & \text{if } y_1^* = 0 \\ y_2^* & \text{if } y_1^* > 0 \end{cases}$ $\lambda_1 = \exp(x\beta_1 + \gamma_1 d + \varepsilon)$ $\lambda_2 = \exp(x\beta_2 + \gamma_2 d + \varepsilon)$	$\beta_1 = (-1, 0.5)\ ; \gamma_1 = 1$ $\beta_2 = (0.75, 0.5)\ ; \gamma_2 = 1$ $\alpha = (0, 0.5, 0.5); \eta_1 = \eta_2 = 0.5$ $\Sigma = \begin{pmatrix} 0.25 & 0.25 \\ 0.25 & 1 \end{pmatrix}$
DGP5	$y_1^* \sim P(\lambda_1)$ $y_2^* \sim P(\lambda_2)$ $y = \begin{cases} y_1^* & \text{with prob. } 0.5 \\ y_2^* & \text{with prob. } 0.5 \end{cases}$ $\lambda_1 = \exp(x\beta_1 + \gamma_1 d + \varepsilon)$ $\lambda_2 = \exp(x\beta_2 + \gamma_2 d + \varepsilon)$	$\beta_1 = (-1.5, -2)\ ; \gamma_1 = 1$ $\beta_2 = (-0.5, 2);\ \gamma_2 = 1$ $\alpha = (0, 0.5, 0.5)$ $\Sigma = \begin{pmatrix} 0.25 & 0.25 \\ 0.25 & 1 \end{pmatrix}$

TABLE 1: Specifications of the experiments. Each column shows the name of the DGP in the text, the specification of the conditional probability function $\mathbf{f}(\mathbf{y} \mid \mathbf{x}, \mathbf{d}, \varepsilon)$ and the values given to the parameters in simulations.

Count	d=0				d=1			
	DGP2	DGP3	DGP4	DGP5	DGP2	DGP3	DGP4	DGP5
0	1.03	1.45	1.28	4.68	1.20	14.14	4.15	212.1
1	0.87	0.25	0.40	0.16	0.80	0.12	0.45	2.84
2	1.10	0.72	0.81	0.21	0.80	0.19	0.38	0.04
3	1.86	2.06	2.17	0.42	1.07	0.30	0.43	0.01
4	3.78	5.89	7.26	0.84	1.78	0.47	0.62	0.02
5	9.95	16.84	29.13	1.67	3.57	0.75	1.06	0.03

TABLE 2: Each entry shows the ratio of the frequency of each count predicted by a Poisson vs. the observed across all simulations (total sample size 100000).

	Parameter						
	β_0	β_1	γ	α_0	α_1	α_2	$\rho\sigma$
	Variance						
TSM	0.01863	0.00287	0.05179	0.00156	0.00176	0.00221	0.02247
Poisson-FIML	0.00574	0.00097	0.01683	0.00157	0.00180	0.00221	0.00589
PP-FIML	0.02240	0.00179	0.01900	0.00160	0.00178	0.00220	0.00677
	Squared Bias ($\times 10^{-3}$)						
TSM	0.02462	0.01129	0.00210	0.00248	0.01353	0.01871	0.00342
Poisson-FIML	0.00001	0.00029	0.00276	0.00197	0.01608	0.01441	0.01180
PP-FIML	1.02802	0.02378	0.21219	0.00088	0.01654	0.01250	0.01279

TABLE 3: Average Squared Error of Parameter Estimates in DGP1computed over 100 replications. The squared error has been decomposed on variance (upper panel) and squared bias (lower panel). The squared bias is measured in 1×10^{-3} units. In TSM, the error for β_0 has been computed discounting the shift of $\sigma^2/2$ (see Terza, 1998).

	Parameter						
	β_0	β_1	γ	α_0	α_1	α_2	$\rho\sigma$
Variance							
TSM	0.07844	0.00945	0.19823	0.02282	0.00251	0.00232	0.07840
Poisson-FIML	0.02906	0.00510	0.09587	0.00228	0.00251	0.00234	0.03761
PP-FIML	0.17373	0.00646	0.09822	0.00231	0.00252	0.00233	0.03767
Squared Bias ($\times 10^{-3}$)							
TSM	0.1838	0.04132	1.38812	0.00090	0.01671	0.04607	0.12462
Poisson-FIML	0.3653	0.02165	1.76942	0.00087	0.01761	0.04571	0.50431
PP-FIML	2.5227	0.06332	0.47548	0.00119	0.02020	0.03670	0.31429

TABLE 4: Average Squared Error of Parameter Estimates in DGP2 computed over 100 replications. The squared error has been decomposed on variance (upper panel) and squared bias (lower panel portion). The squared bias is measured in 1×10^{-3} units. In Poisson-FIML and PP-FIML, the average squared error for the β_0 parameter has been computed discounting the shift due to Negative Binomial (see corollary 1). In TSM, we discount the usual (see caption of table 3) $\sigma^2/2$ for this parameter.

	In Sample		Out of Sample			
	Average	Sq. Err.	Average	Sq. Err.	**BIC**	**CAIC**
DGP3 - Poisson Hurdle						
True: 7.3525						
TSM	6.8385	1.6835	6.8506	1.6708		
Poisson-FIML	6.8436	2.0174	6.8477	2.0336	5463.15	5471.95
PP-FIML	7.3604	1.0627	7.3743	1.0818	5181.63	5191.63
DGP4 - Negative Binomial Hurdle						
True: 5.3997						
TSM	6.1144	2.9345	6.1090	3.0391		
Poisson-FIML	6.1454	1.6687	6.1352	1.7433	4582.51	4606.32
PP-FIML	5.5140	0.8653	5.5230	0.8955	4535.49	4545.49
DGP5 - Poisson Mixture						
True: 2.8178						
TSM	3.3717	0.7985	3.3717	0.7988		
Poisson-FIML	3.0477	0.7031	3.4615	0.7037	5074.88	5082.88
PPFIML	2.7362	0.2111	2.7368	0.2157	5016.77	5026.77

TABLE 5: Average Estimates of the Treatment Effect and Information Criteria. Number of Monte-Carlo replications, 100. Sample size *in sample*= Sample size *out of sample*= 1000. The treatment effect is computed according to equation (21) over sample size. Estimates computed across the 100 Monte-Carlo replications.

Variable	Mean	Std.	Description
		Endogenous	
Tottrips	4.5511	4.9351	Number of trips by members of the household in 24 hrs.
OwnVeh	0.8492	0.3581	1 if household owns at least one motorized vehicle.
		Exogenous	
WorkSchl	0.2622	0.3278	% of total trips for work vs. personal.
Hhmem	2.9289	1.6127	Number of individuals in the household.
DistoCbd$^{(a)}$	0.2887	0.4932	Distance to the central business district in kilometers.
AreaSize	0.3761	0.4848	1 if area is bigger than 2,5 million population.
FullTime	0.9792	0.8475	Number of full time workers in the household.
DistoNod$^{(b)}$	2.0272	3.1378	Distance from home to the nearest transit node in blocks.
RealInc$^{(c)}$	0.8042	0.9197	Household income divided by median income of census tract.
Weekend	0.2236	0.4170	1 if 24 hours survey period is Saturday or Sunday.
Adults	2.0797	0.8978	Number of adults in the household 16 years or older.

TABLE 6: Descriptive Statistics of the variables in the data set on household trips.
NOTES: (a) In 1/30 of original units. (b) In 1/5 of original units. (c) In 1/3 of original units.

Instrument$^{(a)}$ (Z)	**Probit**$^{(b)}$ $^{NLS,TSM}_{WNLS}$	**Polynomial Degree**				
		K=0	K=1	K=2	K=3	K=4
Constant	$^{**}-0.633$	-0.533	-0.532	-0.496	-0.495	-0.487
	0.237	0.354	0.347	0.353	0.354	0.353
WorkSchl	0.152	0.326	0.344	0.325	0.324	0.331
	0.265	0.328	0.325	0.320	0.320	0.318
Hhmem	0.003	0.047	0.056	0.053	0.053	0.054
	0.068	0.072	0.072	0.071	0.071	0.071
DistoCbd	0.629	*0.676	*0.672	*0.666	*0.666	*0.667
	0.399	0.379	0.375	0.374	0.374	0.372
AreaSize	-0.206	-0.247	-0.234	-0.246	-0.245	-0.240
	1.242	0.157	0.155	0.156	0.156	0.155
FullTime	**0.871	**1.014	**1.003	**1.009	**1.009	**1.001
	0.155	0.181	0.175	0.176	0.176	0.174
Adults	**0.381	0.252	0.244	0.225	0.225	0.219
	0.145	0.178	0.174	0.176	0.176	0.175
DistoNod	0.048	0.050	*0.052	*0.050	*0.050	*0.049
	0.033	0.031	0.031	0.030	0.030	0.030
RealInc	**0.472	0.353	0.324	*0.346	*0.346	*0.339
	0.177	0.216	0.214	0.206	0.205	0.201

TABLE 7: Estimates of the binary equation for the data set on household trips. *NOTES*: (a) Upper row shows point estimate, lower row in small type shows standard deviation. The superscripts *, ** preceding an entry denote significance at 5% and 1% respectively. (b) The NLS, TSM and WNLS estimates for the binary equation are equal to a Probit model.

Regressor$^{(a)}$	**NLS**	**TSM**	**WNLS**	**PP-FIML**				
(X)				K=0	K=1	K=2	K=3	K=4
Constant	$^{**}-0.600$	$^{**}-1.445$	$^{**}-1.005$	$^{**}-1.404$	$^{**}-1.383$	$^{**}-2.313$	$^{**}-2.497$	$^{**}-2.930$
	0.225	0.258	0.181	0.239	0.234	0.382	0.415	0.545
WorkSchl	$^{**}-0.527$	$^{**}-0.554$	$^{**}-0.363$	$^{**}-0.339$	$^{**}-0.362$	$^{**}-0.392$	$^{**}-0.397$	$^{**}-0.420$
	0.143	0.147	0.128	0.137	0.134	0.150	0.152	0.158
Hhmem	$^{**}0.166$	$^{**}0.148$	$^{**}0.134$	$^{**}0.154$	$^{**}0.145$	$^{**}0.180$	$^{**}0.184$	$^{**}0.191$
	0.027	0.031	0.028	0.023	0.021	0.028	0.029	0.031
DistoCbd	-0.149	-0.268	$^{**}-0.057$	$^{**}-0.062$	$^{**}-0.068$	$^{**}-0.078$	$^{**}-0.080$	$^{**}-0.092$
	0.136	0.172	0.024	0.040	0.042	0.047	0.048	0.053
AreaSize	-0.034	-0.008	0.038	0.040	0.019	0.048	0.050	0.045
	0.097	0.100	0.085	0.086	0.087	0.105	0.108	0.109
FullTime	$^{**}0.189$	$^{**}0.205$	$^{**}0.220$	$^{**}0.226$	$^{**}0.234$	$^{**}0.249$	$^{**}0.255$	$^{**}0.267$
	0.048	0.101	0.073	0.065	0.052	0.070	0.072	0.076
DistoNod	$^{**}0.002$	$^{*}0.021$	0.019	$^{*}0.022$	0.018	$^{*}0.024$	$^{*}0.025$	$^{**}0.027$
	0.010	0.012	0.013	0.013	0.011	0.014	0.014	0.014
RealIinc	0.041	0.020	0.007	0.025	0.026	0.034	0.036	0.043
	0.048	0.052	0.051	0.026	0.023	0.028	0.028	0.030
Weekend	-0.155	-0.165	-0.029	-0.098	-0.099	-0.122	-0.125	-0.135
	0.112	0.115	0.080	0.093	0.095	0.108	0.111	0.117
OwnVeh	$^{**}1.607$	$^{**}2.796$	$^{**}2.079$	$^{**}2.157$	$^{**}2.221$	$^{**}2.376$	$^{**}2.427$	$^{**}2.535$
	0.185	0.613	0.312	0.304	0.280	0.311	0.320	0.360
Polynomial Coefficients								
a_1					-0.027	0.032	0.110	-0.157
					0.001	0.091	0.103	0.302
a_2						$^{*}0.174$	$^{*}0.179$	0.742
						0.104	0.104	0.485
a_3							0.012	-0.117
							0.011	0.125
a_4								0.020
								0.022
Variance-Covariance Matrix								
ρ				$^{**}-0.762$	$^{**}-0.781$	$^{**}-0.764$	$^{**}-0.761$	$^{**}-0.767$
				0.036	0.041	0.082	0.089	0.117
$\sigma^{(b)}$				$^{+}0.726$	$^{+}0.761$	$^{++}0.895$	$^{++}0.921$	$^{++}0.975$
				0.153	0.131	0.130	0.129	0.125

TABLE 8: Estimates of the count equation for the data set on household trips.
NOTES: (a) Upper row shows point estimate, lower row in small type shows standard deviation. The superscripts * ** preceding an entry denote significance at 5% and 1% respectively. (b) The superscripts $^{+}$, $^{++}$ preceding an entry in the last row denote significance of the test $\sigma = 1$ at 5% and 1% respectively.

Count	**Sample**	**PP-FIML**				
Interval	**Average**	K=0	K=1	K=2	K=3	K=4
0	0.1854	0.1528	0.1558	0.1840	0.1854	0.1880
1	0.1196	0.1572	0.1572	0.1285	0.1246	0.1170
2	0.1092	0.1343	0.1327	0.1122	0.1133	0.1227
3	0.1248	0.1100	0.1080	0.1038	0.1060	0.1096
4	0.0919	0.0881	0.0864	0.0921	0.0931	0.0891
5-6	0.1161	0.1246	0.1226	0.1398	0.1381	0.1283
7-9	0.1231	0.1048	0.1046	0.1173	0.1139	0.1148
10-14	0.0780	0.0748	0.0771	0.0760	0.0754	0.0828
>14	0.0520	0.0533	0.0556	0.0463	0.0501	0.0476
Abs.diffs.[a]		0.2514	0.2468	0.1240	0.1122	0.1148
Andrews test		20.917	18.665	6.4407	4.6528	4.2915
P-Value		0.007	0.0167	0.5979	0.7939	0.8299
-$(1/N)$ $*$Log-lik		2.6695	2.6687	2.6573	2.6577	2.6546
Number parameters		21	22	23	24	25
BIC		3214.1	3219.6	3212.8	3219.6	3222.4
CAIC		3235.1	3241.6	3235.8	3243.6	3247.4

TABLE 9: Andrews Test and Fit Criteria of the models for the data set on household trips.
NOTES: (a) The absolute differences are computed as $\frac{1}{n}\sum_{j=1}^{J} \mid p_j - \hat{p}_j \mid$.

Regressor (X)	**TSM**	**K=0**	**K=2**
OwnVeh	0.5798	1.8032	1.3718
WorkSchl	-0.1453	-0.0872	-0.0865
Hhmem	0.4357	0.4583	0.4467
DistoCbd	-0.0776	-0.0179	-0.0187
AreaSize	-0.0033	0.0149	0.0152
FullTime	0.1037	0.2224	0.2057
DistoNod	0.0438	0.0479	0.0425
RealInc	0.0161	0.0145	0.0237
Weekend	-0.0369	-0.0223	-0.0228

TABLE 10: Estimates of the Mean Effects of the regressors for the data on trip frequency.