ISSN 1755-5361

University of Essex

Department of Economics

Discussion Paper Series

No. 657 July 2008

Identification issues in models for underreported counts

Georgios Papadopoulos and J. M. C. Santos Silva

Identification issues in models for underreported counts*

Georgios Papadopoulos
Department of Economics, University of Essex

J. M. C. Santos Silva
Department of Economics, University of Essex and CEMAPRE

July 25, 2008

Abstract

In this note we study the conditions under which leading models for underreported counts are identified. In particular, we highlight a peculiar identification problem that afflicts two of the most popular models in this class.

JEL classification code: C13, C25.

Key words: Poisson-logit; Restrictions, Stopped-sum distributions.

*We are grateful to Bill Greene, Pravin Trivedi and Rainer Winkelmann for helpful comments and discussions. The usual disclaimer applies. The data used in the illustration in Section 3 were kindly made available by Rainer Winkelmann and are from the public use version of the German Socio-Economic Panel Study. We thank the Deutsches Institut für Wirtschaftsforschung for authorizing its use. Santos Silva gratefully acknowledges the partial financial support from Fundação para a Ciência e Tecnologia, program POCTI, partially funded by FEDER.

Address for correspondence: Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK. E-mail: gpapad@essex.ac.uk (Papadopoulos), jmcss@essex.ac.uk (Santos Silva).

1. INTRODUCTION

Underreporting is likely to be pervasive in survey data. Therefore, models that account for this type of measurement error are possibly useful in many applications. For the particular case of count data, models accounting for underreporting are described in the monographs of Cameron and Trivedi (1998) and Winkelmann (2008), and are implemented in popular statistical packages (e.g., Econometric Software, Inc., 2007). In this note we study the conditions under which leading models for underreported counts are identified, and highlight a peculiar identification problem that afflicts two of the most popular specifications.

The reminder of this note is organized as follows. Section 2 presents the identification results, Section 3 discusses the practical consequences of the main results and, finally, Section 4 contains brief concluding remarks.

2. RESULTS

Models for underreported counts are based on the assumption that the number of occurrences reported in a given period by individual i is given by

$$y_i = \sum_{j=1}^{y_i^*} b_{ij}, \tag{1}$$

where y_i^* is the total (unobserved) number of occurrences and b_{ij} is a Bernoulli random variable that takes the value 1 when the jth occurrence is reported. Throughout, we assume a regression framework in which the object of interest is the conditional distribution of y_i^*, given the set of regressors x_i. For convenience, x_i is written as $x_i = (x_{1i}, x_{2i})$, where x_{1i} and x_{2i} may be identical, overlapping or disjoint.

Perhaps the most popular specification for underreported counts is the Poisson-logit model introduced by Winkelmann and Zimmermann (1993),[1] which is obtained

[1] See also Mukhopadhyay and Trivedi (1995).

as a special case of (1) when it is assumed that, conditionally on x_i: y_i^* follows a Poisson distribution with parameter $\lambda_i = \exp(x_{1i}'\beta)$, $\Pr(b_{ij} = 1|x_i) = \Lambda_i = \exp(x_{2i}'\gamma) / (1 + \exp(x_{2i}'\gamma))$, and y_i^* and b_{ij} are independent.[2] Under these assumptions, it is possible to show that, conditionally on x_i, y_i follows a Poisson law with parameter $\mu_i = \lambda_i\Lambda_i$.[3]

Although Poisson regression is generally well behaved, the Poisson-logit is a double-index model whose likelihod function may have multiple maxima. Moreover, identification of $\theta = (\beta', \gamma')'$ is problematic, even under the maintained strong parametric assumptions. For example, θ is not identified when x_{2i} is a subset of x_{1i} and Λ_i is constant (see, Cameron and Trivedi, 1998, or Winkelmann, 2008). Even in less extreme situations, identification of the Poisson-logit model is afflicted by a subtle problem. Indeed, it is trivial to show that

$$\mu_i \equiv \exp(x_{1i}'\beta) \frac{\exp(x_{2i}'\gamma)}{1 + \exp(x_{2i}'\gamma)} = \exp(x_{1i}'\beta + x_{2i}'\gamma) \frac{\exp(-x_{2i}'\gamma)}{1 + \exp(-x_{2i}'\gamma)} \equiv \mu_i^a. \quad (2)$$

Since the likelihood function of a Poisson regression model depends on x_i only through the conditional mean, (2) implies that there are two Poisson-logit regression models with conditional means $\mu_i \equiv \lambda_i\Lambda_i$ and $\mu_i^a \equiv \lambda_i \exp(x_{2i}'\gamma)(1 - \Lambda_i)$, which will lead exactly to the same value of the likelihood function. Therefore, unless appropriate restrictions are imposed on θ, these two models with very different specifications of $E[y_i|x_i]$ are observationally equivalent.[4]

[2] If y_i^* and b_{ij} are conditionally independent, (1) implies that y_i has a stopped-sum distribution (see Johnson, Kemp and Kotz, 2005, Ch. 9).

[3] This result can be traced back to Catcheside (1948).

[4] Zero inflated models of the type introduced by Mullahy (1986) and Lambert (1992) are often specified with a conditional mean of the form $\mu_i \equiv \lambda_i\Lambda_i$, where Λ_i represents the probability of zero-inflation. The likelihood function for zero-inflated models, however, depends separately on μ_i and Λ_i, and therefore the identification problem discussed here does not arise if the model is estimated by maximum-likelihood.

To explore the consequences of (2), it is convenient to consider the case in which $x_i = x_{1i} = x_{2i}$, which leads to

$$\exp(x_i'\beta)\frac{\exp(x_i'\gamma)}{1+\exp(x_i'\gamma)} = \exp(x_i'(\beta+\gamma))\frac{\exp(-x_i'\gamma)}{1+\exp(-x_i'\gamma)}. \quad (3)$$

Now, identification can be studied by analyzing the non-sample information needed to distinguish $\theta = (\beta', \gamma')'$ from $\theta^a = (\beta' + \gamma', -\gamma')'$.

Starting with the case in which the researcher has information on the logit part of the model, it is obvious that θ is identified when the sign of at least one (non-zero) element of γ is known *a priori*. Alternatively, when some elements of γ are known to be zero, although there are still two observationally equivalent models, it is possible to identify the elements of β corresponding to the zeros in γ because in this case the relevant elements of β and $\beta^a = \beta + \gamma$ are identical. Turning now to the possible non-sample information on β, identification of θ requires the knowledge of at least one element of this vector, for example as a result of an exclusion restriction. When that is the case, μ_i and μ_i^a can be distinguished because μ_i^a is not consistent with the non-sample information. The practical consequences of these results will be explored below.

Although consistency of the Poisson-logit estimator only requires $E[y_i|x_i] = \mu_i$, it is possible to generalize this model to account for possible overdispersion. The standard way of doing this is to assume that, conditionally on x_i and on an unobservable individual effect ε_i, y_i^* follows a Poisson distribution with parameter $\lambda_i = \exp(x_{1i}'\beta + \varepsilon_i)$, $\Pr(b_{ij} = 1|x_i, \varepsilon_i) = \exp(x_{2i}'\gamma)/(1+\exp(x_{2i}'\gamma))$, and y_i^* and b_{ij} are independent. Now, conditionally on x_i and ε_i, y_i has a Poisson distribution with parameter $\lambda_i\Lambda_i\exp(\varepsilon_i)$, where λ_i and Λ_i are defined as before.

Under the usual assumption that $\exp(\varepsilon_i)$ has a gamma distribution with unit mean and variance α_i, the distribution of y_i, conditionally on x_i only, is negative-binomial with mean $\mu_i = \lambda_i\Lambda_i$ and variance $\omega_i = \mu_i + \alpha_i\mu_i^2$.

The leading member of this family of models is the NegBin2-logit, for which α_i is constant. As in the Poisson-logit, the likelihood function of the NegBin2-logit depends on x_i only through μ_i (see Cameron and Trivedi, 1998). Therefore identification of θ requires exactly the same conditions established for the Poisson-logit model.

When α_i is allowed to depend on the regressors, identification may be easier. In particular, identification is possible when α_i is a function of λ_i. For example, assuming that $\alpha_i = \delta/\lambda_i$, we obtain a NegBin1-logit model with mean $\mu_i = \lambda_i \Lambda_i$ and variance $\omega_i = \mu_i + \delta \lambda_i \Lambda^2$. In this case, due to additional structure imposed on the variance, the likelihood function does not depend on x_i only through μ_i and therefore the result in (2) does not imply the existence of an identification problem, even when $x_{1i} = x_{2i}$.

It should be noted, however, that in this case identification of the conditional mean of y_i is achieved by assuming a parametric specification for the conditional variance. This has obvious consequences for the robustness of the estimator.

An alternative way of accounting for unobserved heterogeneity in count data is to use of finite-mixture models, which in some cases have a natural and attractive interpretation (see, e.g., Deb and Trivedi, 1997 and 2002). Although we are not aware of any model based on finite-mixtures that also accounts for underreporting, it is easy to see that, if the probability of underreporting is allowed to vary across classes, the conditions for the identification of the mixture model will be exactly the same that are required for the identification of the models for each class. Therefore, the results presented above can easily be extended to this class of models.

Of course, strictly speaking, the identification problems of the Poisson-logit and NegBin2-logit do not extend to models where $\Pr(b_{ij} = 1|x_i)$ is not of the logit form, like in a Poisson-probit model or in the specification suggested by Winkelmann (1998). In spite of this, the findings in the next section suggest that the identification results presented here are likely to be useful at least for some of these models.

3. CONSEQUENCES FOR PRACTITIONERS

As noted before, the likelihood function of a Poisson-logit model may have multiple maxima. Moreover, the results in Section 2 suggest that very different sets of parameters may lead to very similar values of the likelihood function. This will certainly be the case when identification hinges on an exclusion restriction that is "weak" in the sense that either the regressors excluded from x_1 are highly colinear with remaining elements of this vector, or the elements of γ corresponding to the regressors excluded from x_1 are small. Naturally, the same result is true for the NegBin2-logit model.

If the practitioner is unaware of this, he may be puzzled to find that his estimated model fits the data quite well, despite having estimated parameters with implausible values or "wrong signs". To exemplify this situation, we reconsider a well known empirical illustration.

Winkelmann (2008) uses a Poisson-logit to model the number of job offers when only data on voluntary job changes is available. The partial observability resulting from the fact that only the accepted job offers are counted makes the use of the Poisson-logit potentially appropriate.[5] The data used in the illustration is a subsample of the German Socio-Economic Panel consisting of 1962 males workers aged between 25 and 50 in 1974. Table 1 presents two sets of results from the estimation of a Poisson-logit model using the sample and specification considered in Winkelmann (2008).[6] The names of regressors in Table 1 are self-explanatory, but a complete description of the regressors and further details on the data can be found in Winkelmann (2008).

The first two columns in Table 1, labeled $\tilde{\beta}$ and $\tilde{\gamma}$, coincide with the estimates given in Winkelmann (2008). The final two columns, labeled $\hat{\beta}$ and $\hat{\gamma}$, correspond to

[5] It should be made clear that the results presented in Winkelmann (2008) are merely illustrative and that the same data set is used to exemplify the estimation of many different count data models.

[6] Standard errors computed from the observed information matrix are given in parenthesis.

the estimates obtained when using as starting values $\beta^0 = \tilde{\beta} + \tilde{\gamma}$ and $\gamma^0 = -\tilde{\gamma}$, and imposing the that the element of β^0 corresponding to SINGLE is zero.

Table 1: Poisson-logit regressions for number of job changes

	$\tilde{\beta}$	$\tilde{\gamma}$	$\hat{\beta}$	$\hat{\gamma}$
INTERCEPT	0.81208	0	0.84395	0
	(0.21841)	—	(0.22583)	—
EDUCATION	−0.32237	3.73211	3.25160	−3.58186
	(0.15870)	(2.05257)	(2.04117)	(1.98077)
EXPERIENCE	−0.66889	−6.04420	−6.40329	5.72807
	(0.13372)	(3.89619)	(3.70472)	(3.67749)
EXPERIENCE^2	0.07179	3.32178	3.19228	−3.12084
	(0.04796)	(2.24479)	(2.11346)	(2.09315)
UNION	−0.29149	0	−0.29352	0
	(0.06498)	—	(0.06496)	—
SINGLE	0	0.37970	0	−0.08352
	—	(1.37139)	—	(0.11589)
GERMAN	−0.39741	0	−0.39614	0
	(0.07614)	—	(0.07622)	—
QUALIFIED WHITE COLLAR	0.06936	0	0.06771	0
	(0.13112)	—	(0.13117)	—
ORDINARY WHITE COLLAR	0.17865	0	0.18113	0
	(0.14752)	—	(0.14770)	—
QUALIFIED BLUE COLLAR	0.13240	0	0.13270	0
	(0.08245)	—	(0.08252)	—
Log-likelihood	−2039.3549		−2039.1312	

In this particular sample, the model is identified by the exclusion of SINGLE from x_{1i}. However, SINGLE has a relatively small, and statistically insignificant, effect on the logit part of the model. Therefore, although $\tilde{\theta} = \left(\tilde{\beta}', \tilde{\gamma}'\right)'$ and $\theta^0 = \left(\beta^{0\prime}, \gamma^{0\prime}\right)'$ do not lead to the same value for the likelihood function, the results of the two last columns reveal that there is a second maximum of the likelihood function for θ very close to θ^0. Indeed, we note that $\hat{\gamma} \simeq -\tilde{\gamma}$ and $\hat{\beta} \simeq \tilde{\beta} + \tilde{\gamma}$. Moreover, we find that $\hat{\theta} = \left(\hat{\beta}', \hat{\gamma}'\right)'$ leads to a value of the log-likelihood function which is extremely close to, and actually slightly better than, the one obtained with $\tilde{\theta}$.

These results are not limited to the Poisson-logit specification. For this particular data, comparable results can be found for the Poisson-probit and for the NegBin2-logit models (these results are available from the authors upon request). Therefore, even in cases where identification is guaranteed by appropriate restrictions or by the structure of the model, the results of the previous section may be helpful in that they can be used to guide the researcher in the search for the global maximum of the likelihood function.

4. CONCLUDING REMARKS

This note reviews the conditions needed for the identification of the Poisson-logit and other leading models for underreported counts. In general, identification is easier when at least one of the regressors affects the probability of reporting without affecting the underlying count process. In case such regressor is not available, identification may still be achieved by using additional information. This information can take the form of restrictions on the signs of some coefficients in the probability of reporting, or it can be used to specify a skedastic function that disentangles the effect of the regressors on the mean of the count process and on the probability of reporting.

Even when the model is identified, the likelihood function is likely to have multiple maxima. Therefore, when estimating this sort of model, it is important not to accept

a maximum of the likelihood function as global without having performed a thorough search for other maxima. In particular, having obtained one set of estimates, our results can be used to obtain starting values that are often close to a set of parameter values that leads to an alternative maximum of the likelihood function.

REFERENCES

Cameron, A.C. and P.K. Trivedi (1998). *Regression Analysis of Count Data*, Cambridge, MA: Cambridge University Press.

Catcheside, D.G. (1948). "Genetic Effect of Radiation," *Advances in Genetics*, 2, 271-358.

Deb, P. and P.K. Trivedi (1997). "Demand for Medical Care by the Elderly: A Finite Mixture Approach," *Journal of Applied Econometrics*, 12, 313-336.

Deb, P. and P.K. Trivedi (2002). "The Structure of Demand for Health Care: Latent Class Versus Two-Part Models," *Journal of Health Economics*, 21, 601-625.

Econometric Software, Inc. (2007). *LIMDEP 9.0*, New York: Plainview.

Johnson, N.L.; A.W. Kemp, and S. Kotz, (2005). *Univariate Discrete Distributions*, 3rd ed., New York: John Wiley & Sons.

Lambert, D. (1992). "Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing," *Technometrics*, 34, 1-14.

Mukhopadhyay, K. and P.K. Trivedi (1995). *Regression Models for Under-recorded Count Data*, Paper presented at the Econometric Society 7th World Congress, Tokyo.

Mullahy, J. (1986). "Specification and Testing in Some Modified Count Data Models," *Journal of Econometrics*, 33, 341-365.

Winkelmann, R. (1998). "Count Data Models with Selectivity," *Econometric Reviews*, 17, 339-359.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*, 5th ed., Berlin: Springer-Verlag.

Winkelmann, R and K.F. Zimmermann (1993). *Poisson Logistic Regression*, Department of Economics, University of Munich, Working Paper No 93-18.