Endogeneity in a Binomial Count Data Model*

Kurt Brännäs[1] and Maria Melkersson[2]
[1] Department of Economics, Umeå University
[2] SOFI, Stockholm University

Abstract

A characterization of endogeneity in a binomial regression model is given. The relationship to a model subject to selection is indicated. Maximum likelihood and nonlinear instrumental variable estimators are obtained and evaluated in a Monte Carlo study. A new LM test and a Hausman test against endogeneity are also evaluated. Maximum likelihood estimation based on an instrumental variable and the LM test are found to perform well. An illustration based on choice set sizes for Swedish unemployed is included.

1. Introduction

The purposes of this paper are threefold: First, it aims at obtaining a Lagrange multiplier (LM) or score test statistic against endogeneity[1] for a binomial regression model. Second, it introduces a nonlinear instrumental variable (NLIV) estimator based on an instrumental variable correction for endogeneity. As a by-product a Hausman-type test is obtained. Third, the paper also aims at comparing and evaluating the proposed estimators and tests in finite samples.

Section 2 demonstrates how endogeneity can arise in the binomial model framework. While there may be other ways of modelling, the presented one is based on a binary endogeneous variable and is the reasonable choice with respect to our empirical interests. The introduced characterization of the problem appears a new and coherent one, carrying over to selection model formulations as well as to the Poisson count data framework. Amemiya (1978) and Even (1988) consider the related case of a continuous endogeneous variable in a probit model framework.

Section 3 considers maximum likelihood (ML) estimation and derives the LM test against endogeneity. By some quite small changes to the expressions, the obtained results include the essentials also for estimating selection models and testing against selection bias (cf. Appendix B). Selection models for nonbinomial count data have recently been studied by Terza (1998) and others. In the same section, instrumental variable correction is considered and a Hausman-type (Hausman, 1978) test against endogeneity is obtained.

---

*The financial support from The Expert Group on Labour Market Policy Evaluation Studies, Ministry of Labour, Stockholm, is acknowledged. William H Greene and Xavier de Luna are thanked for helpful comments on an earlier version of the paper.

[1]See Engle, Hendry and Richard (1984) for a formal definition of (weak) exogeneity. A variable that is not exogeneous is endogeneous.

Endogeneity correction using instrumental variables for nonlinear models was first considered by Amemiya (1974). A recent account for Poisson count data is given by Windmeijer and Santos Silva (1998).

The proposed test statistics and estimators are evaluated and compared by finite sample Monte Carlo experimentation in Section 4. The section also provides evidence of the consequences of neglected endogeneity and of the use of an instrumental variable for ML estimation. An empirical illustration based on choice set sizes among Swedish unemployed (see Melkersson, 1998) is given in Section 5. The final section contains a few concluding remarks and summarizes the main findings.

2. Model

To construct a binomial model with a potential endogeneity problem we proceed in two steps. First, we think of a binary endogeneous variable w, that arises from a latent variable model such that

$$w = \begin{cases} 1, & w^* = \mathbf{z}\boldsymbol{\gamma} + \epsilon \geq 0 \\ 0, & \text{otherwise} \ , \end{cases} \tag{1}$$

where $\mathbf{z}$ is a vector of fixed explanatory variables, $\boldsymbol{\gamma}$ is a vector of unknown parameters, and ϵ is an error term. Second, the count or frequency arises as

$$y = \sum_{j=1}^{m} d_j, \tag{2}$$

where the $d_j, j = 1, \ldots, m$, arise from latent variable models

$$d_j^* = \mathbf{x}\boldsymbol{\beta} + w\alpha + \xi_j, \qquad j = 1, \ldots, m, \tag{3}$$

by the indicator variable $d_j = 1(d_j^* \geq 0)$. Note that neither the $1 \times k$-vector $\mathbf{x}$, w, nor the parameters vary with j, and that m is fixed.

We make the following assumptions about the random disturbance terms: all disturbance terms are normally distributed, and in addition $E(\epsilon) = E(\xi_j) = 0$, $V(\xi_j) = 1$, $V(\epsilon) = \sigma_\epsilon^2 = 1 + m\theta^2$, $E(\xi_i \xi_j) = 0, i \neq j$, and $E(\epsilon\xi_j) = \theta$, $i, j = 1, \ldots, m$.

It follows that the correlation between ϵ and $\xi_j, j = 1, \ldots, m$, is constant and equal to $\rho = \theta/\sigma_\epsilon$. This constancy appears reasonable for the binomial setting. A correlation that varies with j would be further away from the binomial model and make for larger technical problems later on. The correlation coefficient ρ only varies in a subinterval $(-c, c)$ of the $(-1, 1)$ interval; for $m = 10$ we get $c \approx 0.32$. The constant c is decreasing in m. Note also that as θ varies so does σ_ϵ^2.

For $\theta = 0$, the right hand side variable w of the latent d_j^* expression (3) is exogenous, i.e. uncorrelated with ξ_j. In this case and with m fixed, y has, conditional on $\mathbf{x}$ and w, a binomial distribution with parameters m and $\pi = \Pr(\xi_j \geq -\mathbf{x}\boldsymbol{\beta} - w\alpha) = \Phi(\mathbf{x}\boldsymbol{\beta} + w\alpha)$, where $\Phi(\cdot)$ is the standard normal distribution function. When m in (2) is random, other distributions than the binomial arise. For instance, when m is Poisson distributed with mean λ, the distribution of y is Poisson with mean $\lambda\pi$.

For an interpretation of (1)-(3) as a model subject to selection, we set $\alpha = 0$ and observe y only when either of $w = 1$ or $w = 0$ is observed.

3. Estimation and Testing

We first consider the construction of the likelihood function and obtain an LM test against endogeneity. Later we consider instrumental variables, NLIV estimation and give a Hausman-type test against endogeneity.

3.1 ML Estimation

When $\theta = 0$, ML estimation based on the binomial model is straightforward, as it will be for the probit model for w, where $\Pr(w = 1) = \Phi(\mathbf{z}\boldsymbol{\gamma}^*)$, with $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}/\sqrt{1 + m\theta^2}$.

For $\theta \neq 0$, so that w is endogenous in the d_j^* expression (3), we have for given w and a fixed m, that y is binomially distributed with parameters m and $\pi(w)$. The probabilities are obtained as

$$\begin{aligned}
\pi(1) &= \Pr(d_j^* \geq 0 | w = 1) = \Pr(\xi_j \geq -\mathbf{x}\boldsymbol{\beta} - \alpha, \epsilon \geq -\mathbf{z}\boldsymbol{\gamma})/\Pr(w = 1) \\
&= \int_{-\mathbf{z}\boldsymbol{\gamma}}^{\infty} \int_{-\mathbf{x}\boldsymbol{\beta}-\alpha}^{\infty} f(\xi_j, \epsilon)\ d\xi_j d\epsilon/\Phi(\mathbf{z}\boldsymbol{\gamma}^*) \\
&= \int_{-\mathbf{z}\boldsymbol{\gamma}}^{\infty} f(\epsilon)\ [\int_{-\mathbf{x}\boldsymbol{\beta}-\alpha}^{\infty} f(\xi_j|\epsilon)\ d\xi_j]\ d\epsilon/\Phi(\mathbf{z}\boldsymbol{\gamma}^*)
\end{aligned}$$

$$\begin{aligned}
\pi(0) &= \Pr(d_j^* \geq 0 | w = 0) = \Pr(\xi_j \geq -\mathbf{x}\boldsymbol{\beta}, \epsilon < -\mathbf{z}\boldsymbol{\gamma})/\Pr(w = 0) \\
&= \int_{-\infty}^{-\mathbf{z}\boldsymbol{\gamma}} \int_{-\mathbf{x}\boldsymbol{\beta}}^{\infty} f(\xi_j, \epsilon)\ d\xi_j d\epsilon/\Phi(-\mathbf{z}\boldsymbol{\gamma}^*) \\
&= \int_{-\infty}^{-\mathbf{z}\boldsymbol{\gamma}} f(\epsilon)\ [\int_{-\mathbf{x}\boldsymbol{\beta}}^{\infty} f(\xi_j|\epsilon)\ d\xi_j]\ d\epsilon/\Phi(-\mathbf{z}\boldsymbol{\gamma}^*).
\end{aligned}$$

Making the transformation $\epsilon' = \sigma_\epsilon^{-1}\epsilon$, the conditional distribution of ξ_j given ϵ' is univariate normal with mean $\theta\epsilon'/\sigma_\epsilon$ and variance $1 - \rho^2 = 1 - \theta^2/\sigma_\epsilon^2$. It follows that, for $w = 0$ or 1,

$$\int_{-\mathbf{x}\boldsymbol{\beta}-\alpha w}^{\infty} f(\xi_j|\epsilon')\ d\xi_j = \Phi[(\mathbf{x}\boldsymbol{\beta} + \alpha w - \theta\epsilon'/\sigma_\epsilon)/\sqrt{1-\rho^2}].$$

Hence,

$$\begin{aligned}
\pi(1) &= \int_{-\mathbf{z}\boldsymbol{\gamma}^*}^{\infty} \phi(\epsilon')\Phi[(\mathbf{x}\boldsymbol{\beta} + \alpha - \theta\epsilon'/\sigma_\epsilon)/\sqrt{1-\rho^2}]\ d\epsilon'/\Phi(\mathbf{z}\boldsymbol{\gamma}^*) \\
\pi(0) &= \int_{-\infty}^{-\mathbf{z}\boldsymbol{\gamma}^*} \phi(\epsilon')\Phi[(\mathbf{x}\boldsymbol{\beta} - \theta\epsilon'/\sigma_\epsilon)/\sqrt{1-\rho^2}]\ d\epsilon'/\Phi(-\mathbf{z}\boldsymbol{\gamma}^*),
\end{aligned}$$

where $\phi(\cdot)$ is the standard normal density function.

The likelihood function for the observations can then be written on the form

$$\begin{aligned}
\mathcal{L} &= \prod_{w=0} \Pr(y_i, w_i = 0) \prod_{w=1} \Pr(y_i, w_i = 1) \\
&= \prod_{w=0} \Pr(y_i | w_i = 0) \Pr(w_i = 0) \prod_{w=1} \Pr(y_i | w_i = 1) \Pr(w_i = 1) \\
&= \prod_{i=1}^{n} \pi(w_i)^{y_i}\ (1 - \pi(w_i))^{m-y_i}\ \ \Phi(\mathbf{z}_i\boldsymbol{\gamma}^*)^{w_i}\Phi(-\mathbf{z}_i\boldsymbol{\gamma}^*)^{1-w_i}.
\end{aligned}$$

The likelihood function is related to the Type 5 model of the Amemiya (1984) classification.

Under the assumption $\theta = 0$, the likelihood function factors as required for the weak exogeneity of the w variable for the parameters of interest $\boldsymbol{\beta}$ and α, since $\pi(1|\theta = 0) = \Phi(\mathbf{x}\boldsymbol{\beta} + \alpha)$, $\pi(0|\theta = 0) = \Phi(\mathbf{x}\boldsymbol{\beta})$, and $\boldsymbol{\gamma}$ is neither related to $\boldsymbol{\beta}$ nor to α. In the general case of $\theta \neq 0$, ML estimation will numerically be demanding as numerical integration will be required.

While it is well-known that endogeneity among the explanatory variables in linear models causes an asymptotic bias, much less is known about the finer details in nonlinear problems of the type considered here. By a standard Taylor expansion of the likelihood equation for the binomial model to the first order around the true parameter value $\boldsymbol{\psi}_0 = (\boldsymbol{\beta}_0', \alpha_0)'$ it follows that the ML estimator is consistent under exogeneity. When $\theta \neq 0$, the estimator is asymptotically biased, due to the fact that $\pi(w_i) \neq \Phi(\mathbf{u}_i\boldsymbol{\psi}_0)$, and where $\mathbf{u}_i = (\mathbf{x}_i, w_i)$.

Note also that for the closely related selection problem, with $\alpha = 0$, the likelihood function simplifies to $\mathcal{L} = \prod_{w=0} \Pr(w_i = 0) \prod_{w=1} \Pr(y_i|w_i = 1) \Pr(w_i = 1)$.

3.2 LM Test

Even if the likelihood function for ML estimation is complicated, obtaining an LM test statistic against endogeneity, i.e. of $\mathrm{H}_0 : \theta = 0$ against $\mathrm{H}_1 : \theta \neq 0$, is quite straightforward. Its practical evaluation will also be simple using standard functions contained in many software packages. The resulting LM statistic may be expressed in various ways, of which we prefer the following

$$\mathrm{LM} = S \sum_{i=1}^{n} a_i \tilde{y}_i \tilde{w}_i, \tag{4}$$

where

$$\begin{aligned}
\tilde{y}_i &= \frac{y_i - m\Phi(\mathbf{x}_i\boldsymbol{\beta} + \alpha w_i)}{m\Phi(\mathbf{x}_i\boldsymbol{\beta} + \alpha w_i)(1 - \Phi(\mathbf{x}_i\boldsymbol{\beta} + \alpha w_i))} \\
\tilde{w}_i &= \frac{w_i - \Phi(\mathbf{z}_i\boldsymbol{\gamma})}{\Phi(\mathbf{z}_i\boldsymbol{\gamma})(1 - \Phi(\mathbf{z}_i\boldsymbol{\gamma}))} \\
a_i &= -m\phi(\mathbf{z}_i\boldsymbol{\gamma})\phi(\mathbf{x}_i\boldsymbol{\beta} + \alpha w_i).
\end{aligned}$$

Obviously, $\tilde{y}_i$ is the residual for the binomial model divided by the variance of the residual, and analogously for $\tilde{w}_i$ from the probit model. When $\theta = 0$ these residuals are uncorrelated. The S is obtained from the appropriate element of the inverse of the outer product of the gradient vector. The elements of the gradient vector are given in Appendix A.

The asymptotic distribution of the LM statistic in (4) is $N(0, 1)$ and the test is two-sided. To evaluate the statistic, one needs estimates separately from the binomial model and the binary 'endogeneity' model. The form of (4) may also be used to generate simpler conditional moment tests or informal plots. The form of the test also suggests that the sign of the statistic can be expected to be opposite to the sign of θ.

An LM test against selection bias can be obtained by simplifying (4), see Appendix B.

3.3 Instrumental Variable Estimation

Consider the use of instrumental variables in an attempt to protect against the negative effects of endogeneity. A straightforward use of an instrument such as $\hat{p}_i = \Phi(\mathbf{z}_i\hat{\boldsymbol{\gamma}})$ for w_i from a first step probit ML estimator appears tempting. Note that its use in place of

w_i for binomial model, ML estimation does not yield a consistent estimator. Nonlinear instrumental variable (NLIV, see, e.g., Davidson and MacKinnon, 1993, ch. 7) or generalized method of moments (GMM, see, e.g., Hansen, 1982) estimation are alternative approaches that yield consistent estimators.

The residual for the binomial model is of the form

$$\eta_i = y_i - m\Phi(\mathbf{u}_i\boldsymbol{\psi}).$$

Windmeijer and Santos Silva (1998) demonstrate in a Poisson count data setting that $\mathbf{x}_i$ and $\hat{p}_i$ may be taken as instruments for $\mathbf{x}_i$ and w_i, respectively. Collecting variables we let $\mathbf{V} = (\mathbf{X}, \hat{\mathbf{p}})$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)'$. The NLIV estimator minimizes $\boldsymbol{\eta}'\mathbf{V}(\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\boldsymbol{\eta}$. Let $Cov(\boldsymbol{\eta}) = \boldsymbol{\Omega}$, then an NLIV estimator that accounts for heteroskedasticity is given as the argument that minimizes the GMM criterion:

$$\tilde{\boldsymbol{\psi}} = \arg\min_{\psi} \ \boldsymbol{\eta}'\mathbf{V}(\mathbf{V}'\boldsymbol{\Omega}\mathbf{V})^{-1}\mathbf{V}'\boldsymbol{\eta}.$$

An estimator of the asymptotic covariance matrix is given by

$$\mathbf{S}_{NLIV} = Cov(\tilde{\boldsymbol{\psi}}) = (\mathbf{U}'\mathbf{V}(\mathbf{V}'\boldsymbol{\Omega}\mathbf{V})^{-1}\mathbf{V}'\mathbf{U})^{-1},$$

where $\mathbf{U} = -\partial\boldsymbol{\eta}/\partial\boldsymbol{\psi}$ and all matrices are evaluated at estimates. Note that nothing prevents the inclusion of additional instrumental variables beyond $\mathbf{X}$ and $\hat{\mathbf{p}}$. In practice, a first step with $\boldsymbol{\Omega}$ set equal to the identity matrix is required to obtain an estimated $\hat{\boldsymbol{\Omega}}$ matrix for use in a second step.

3.4 Hausman Test

The NLIV estimator is consistent under both exogeneity and endogeneity, but is not efficient. The ML estimator for the binomial model is consistent and efficient under exogeneity, while inconsistent under endogeneity. Therefore, we may form a Hausman-type test based on the difference between the parameter estimators, i.e. on $\mathbf{d} = \tilde{\boldsymbol{\psi}}_{NLIV} - \hat{\boldsymbol{\psi}}_{ML}$, as

$$\mathrm{H} = \mathbf{d}'(\mathbf{S}_{NLIV} - \mathbf{S}_{ML})^{-1}\mathbf{d}.$$

Here, $\mathbf{S}_{NLIV}$ and $\mathbf{S}_{ML}$ are the covariance matrices of the NLIV and ML estimators, respectively. The H-test statistic is asymptotically $\chi^2(k+1)$ distributed under H_0, with $k+1$ the number of parameters.

4. Finite Sample Performance

The finite sample study is based on Monte Carlo experimentation. The following questions are focused: (i) Are the properties of the LM and Hausman type test statistics comparable?, and (ii) Which if any are the advantages in terms of estimator performance of using instrumental variables in small samples?

To provide, at least, a tentative answer we use the model framework of Section 2 to generate data. For the endogeneity model (2), we set $\mathbf{z} = (1, z_1, z_2, z_3)$, with independently $z_1 \sim N(4,4)$, $z_2 \sim N(2,1)$, $z_3 \sim N(8,9)$, and with $\boldsymbol{\gamma} = (0.5, -0.6, -0.1, 0.2)'$. For the binomial model we set $m = 10$, $\mathbf{x} = (1, x_1)$, with $x_1 \sim N(3, 9/4)$, and $\boldsymbol{\beta} = (0.1, -0.3)'$. The parameters to be varied are $\alpha = -0.4, 0$, and 0.4, and $\theta = -0.7$ (0.1) 0.7. Two sample sizes, $n = 100, 500$, are employed. The explanatory variables $\mathbf{x}$ and $\mathbf{z}$ are fixed over the 1000 replications in each cell of the design.

The following estimators are applied: Binomial model ML using w (BIN_w) and $\hat{p}$ (BIN_p), and NLIV. The covariance matrices for the binomial models are obtained from the information matrix, while for NLIV it is based on an Eicker-White estimator, i.e. on $\hat{\mathbf{\Omega}}$ =diag$(\hat{\eta}_i^2, i = 1, \ldots, n)$. For each parameter estimator bias and mean square error (MSE) results are obtained. Besides reporting the power functions of the LM and H test statistics of $\theta = 0$ against $\theta \neq 0$, we report on sizes for $\alpha = 0$ versus $\alpha \neq 0$. To evaluate power functions we throughout use a nominal significance level of 0.05.

The biases and MSEs of the estimators of α are for selected θ values given in Table 1. It is quite obvious that neglecting endogeneity (BIN_w) has severe biasing and MSE enhancing effects. The size of the bias is almost linear in θ, while the MSE is quadratic. The bias has the sign of θ. There is no apparent sample size effect. While the ML estimator based on the idea of replacing w with $\hat{p}$ is known to be biased (essentially due to the difference between $\pi(w)$ and $\Phi(\mathbf{x}\boldsymbol{\beta} + \alpha w)$), the results suggest that the bias can be small and moreover decreasing with an increasing sample size. Note that in some cases for large $|\theta|$ (less than 10 out 1000 for $|\theta| \approx 0.7$) the BIN_p estimator diverges. On comparison the consistent NLIV estimator has a bias that frequently, but not throughout, is marginally smaller than that of BIN_p. The MSE of NLIV is larger for $n = 100$ and more or less identical for $n = 500$.

When it comes to the sizes of the conventional t-test for $\alpha = 0$ against $\alpha \neq 0$, we find that the sizes of BIN_w are not reliable. For BIN_p we find that for $n = 100$ the size varies (over $\theta \in [-0.7, 0.7]$) between 0.048 and 0.064, and for $n = 500$ the range is $(0.038, 0.064)$. By the use of the Murphy and Topel (1985) covariance matrix estimator[2] the sizes are, for $n = 100$, $(0.035, 0.053)$, while for $n = 500$, $(0.036, 0.060)$, i.e. for the smaller sample, sizes are sometimes too small. The corresponding sizes for NLIV are $(0.044, 0.071)$ and $(0.049, 0.062)$, respectively. Hence, the size properties are not very different for the tests of $\alpha = 0$.

Figure 1 gives the power functions for the LM test of $\theta = 0$. As expected the power is higher for the larger sample size. For $n = 500$ the sizes are in no case significantly larger than the nominal 0.05 level. For $n = 100$, however, sizes are throughout too large. The dependence on α becomes weak for a larger sample size. Note the unusual feature of decreasing power for larger $|\theta|$. The explanation to this feature is that the variance of the endogeneity model is a function of θ, such that for $\theta = 0$ the $R^2 = 0.64$ in the latent model, for $\theta = 0.5$ $R^2 = 0.34$, and for $\theta = 1$ $R^2 = 0.14$. Therefore, the power functions also highlight the importance of the explanatory power of the endogeneity model (cf. Brännäs and Karlsson, 1996).

In Figure 2 the power functions of the H test are shown. The H statistic is negative in up to 25 percent of the replications for $\alpha = -0.4$ (up to 15 percent for $\alpha = 0.4$). The reported powers are based on only those replications that have a positive H test. Given this, the powers of the H test are much lower than those of the LM test and particularly so for the smaller sample size. For $n = 100$ sizes are not significantly different from the nominal size, whereas for $n = 500$ they are too small. Among these two test statistics against endogeneity the LM test appears the more promising one.

[2] Let ℓ_1 and ℓ_2 be the log-likelihood functions for the probit and binomial models. Murphy and Topel (1986) show that $\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \overset{a}{\sim} N(\mathbf{0}, \mathbf{\Sigma})$, with $\mathbf{\Sigma}$ a function of $R_1 = -E(\partial^2\ell_1/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}')$, $R_2 = -E(\partial^2\ell_2/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}')$, $R_3 = -E(\partial^2\ell_2/\partial\boldsymbol{\gamma}\partial\boldsymbol{\psi}')$ and $R_4 = E(\partial\ell_1/\partial\boldsymbol{\gamma})(\partial\ell_1/\partial\boldsymbol{\gamma})'$, cf. eq. (34).

5. Empirical Illustration

In Melkersson (1998) choice set sizes of 781 unemployed was modelled by binomial models. Those with many contacts with officials at employment offices were found to have significantly larger choice sets. However, the contact frequency variable was suspected to be endogeneous. By a Hausman test exogeneity could not be rejected.

Assume the probability for individual i to discuss one alternative is $\Phi(\mathbf{x}_i\boldsymbol{\beta}+\alpha w_i)$, where $\mathbf{x}_i$ is a vector of individual characteristics, while $w_i = 1$ if the individual has 7 contacts or more with officials during the previous year (cf. Table III in Melkersson, 1997). Of the observations 53 percent have $w_i = 1$. The observed correlation between y and w is 0.32. In a first step, we estimate a model for the contact frequency variable by probit ML and in a second step the binomial model with the maximum choice set size $m = 15$.

For the contact frequency we control for age, sex, profession, labour market history, etc. Altogether 29 explanatory variables are used and $R^2 = 0.12$ (explained divided by total variation). Only four variables have significant effects.

In Table 2 we present estimates for BIN_w, where w_i is treated as exogenous, NLIV with $\hat{p}_i$ used as an instrument, and BIN_p in which w_i is replaced by $\hat{p}_i$. For BIN_p we present t-values based on the information matrix (t_I) and on a Murphy and Topel (1985) covariance matrix.

The LM-statistic takes the value 1.999, so that the exogeneity of the contact frequency variable can be rejected ($p = 0.046$). The Hausman-test can not reject exogeneity. The statistic takes the value 15.043 ($p = 0.448$) when the covariance matrix for the binomial model is based on the Hessian matrix, while the value 5.124 ($p = 0.991$) is obtained when the information matrix is used.

The parameter α is estimated to 0.22 with BIN_w using w_i, while $\hat{\alpha} = 0.53$ for BIN_p and $\hat{\alpha} = 0.54$ for NLIV. This relationship between parameter estimates can be expected for a negative θ, cf. Table 1. Using an approximation (cf. Melkersson, 1998) for the effect of a dummy variable, we find that in the first model (BIN_w) unemployed with seven contacts or more would have 19 percent larger choice sets, than the reference category of fewer contacts. In the other models the difference in choice set size would be 50 percent. Among the significant coefficients we see that unemployed with a handicap, have larger choice sets while choice set size decreases with age of the unemployed.

6. Conclusion

The paper has developed LM and Hausman test statistics against endogeneity in a binomial model framework and studied the performance of these as well as of some associated estimators.

From the Monte Carlo experiments we find that the consistent NLIV estimator does not perform uniformly best. What is surprising is that the binomial model based on the instrumental variable $\hat{p}$ instead of the original endogeneous w variable appears to do quite well. Both of these estimators are markedly better than the estimator based on a complete neglect of the endogeneity of w. Related conclusions hold for the testing of the parameter in front of w. The use of a Murphy and Topel covariance matrix estimator has only a small effect. The LM test statistic of exogeneity has too large sizes for the very small samples but is doing well in this respect for more conventional micro-sample sizes. The sizes of the Hausman test are frequently too low and so is the power. In addition, the statistic is frequently negative. Both the suggested LM and Hausman test statistics as well as the NLIV estimator are easy to apply using standard software.

References

Amemiya, T. (1974) The Non-Linear Two-Stage Least Squares Estimator. *Journal of Econometrics*, 2, pp. 105-110.

Amemiya, T. (1978) The Estimation of a Simultaneous Equation Generalized Probit Model. *Econometrica*, 46, pp. 1193-1205.

Amemiya, T. (1984) Tobit Models: A Survey. *Journal of Econometrics*, 24, pp. 3-61.

Brännäs, K. and Karlsson, N. (1996) Endogeneity Testing in Micro Econometric Models. *Proceedings of the Third Umeå - Würzburg Conference in Statistic*s. Umeå University.

Davidson, R. and MacKinnon, J.G. (1993) *Estimation and Inference in Econometrics.* (Oxford, Oxford University Press).

Engle, R.F., Hendry, D.F. and Richard, J-F. (1984) Exogeneity. *Econometrica*, 51, pp. 277-304.

Even, W.E. (1988) Testing Exogeneity in a Probit Model. *Economics Letters*, 26, pp. 125-128.

Hansen, L.P. (1982) Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50, pp. 1029-1054.

Hausman, J.A. (1978) Specification Tests in Econometrics. *Econometrica*, 46, pp. 1251-1271.

Melkersson, M. (1998) Explaining Choice Set Size for Unemployed. To appear in *Applied Economics.*

Murphy, K.M. and Topel, R.H. (1985) Estimation and Inference in Two-Step Econometric Models. *Journal of Business & Economic Statistics*, 3, pp. 370-379.

Terza, J.V. (1998) Estimating Count Data Models with Endogeneous Switching: Sample Selection and Endogeneous Treatment Effects. *Journal of Econometrics*, 84, pp. 129-154.

Windmeijer, F.A.G. and Santos Silva, J.M.C. (1998) Endogeneity in Count Data Models: An Application to Demand for Health Care. *Journal of Applied Econometrics*, 12, pp. 281-294.

Table 1: Bias and MSE results ($\times 10$) for BIN_w, BIN_p and NLIV estimators of α.

		$n = 100$						$n = 500$					
		BIN_w		BIN_p		NLIV		BIN_w		BIN_p		NLIV	
α	θ	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
	-0.5	-3.67	1.47	0.22	0.53	0.07	0.61	-3.66	1.36	0.37	0.11	0.23	0.11
-0.4	0	-0.02	0.12	0.07	0.28	-0.10	0.33	-0.01	0.02	0.11	0.05	0.00	0.06
	0.5	3.55	1.35	0.14	0.54	-0.24	1.06	3.47	1.22	-0.11	0.10	-0.29	0.13
	-0.5	-3.60	1.39	-0.10	0.47	-0.09	0.50	-3.59	1.30	-0.04	0.08	-0.04	0.08
0	0	0.00	0.10	0.02	0.24	0.00	0.25	-0.01	0.02	0.00	0.04	0.00	0.04
	0.5	3.64	1.40	0.27	0.46	0.22	0.53	3.53	1.26	0.06	0.08	0.05	0.08
	-0.5	-3.52	1.31	-0.08	0.44	-0.01	0.49	-3.51	1.25	-0.04	0.08	-0.03	0.08
0.4	0	0.02	0.08	0.03	0.21	0.04	0.22	0.00	0.02	0.01	0.04	0.00	0.04
	0.5	3.73	1.46	0.30	0.39	0.32	0.41	3.59	1.30	0.18	0.08	0.16	0.08

Table 2: Estimation results.

	BIN_w		NLIV		BIN_p		
Variable	$\hat{\psi}$	t	$\hat{\psi}$	t	$\hat{\psi}$	t_I	t_{MT}
Constant	-0.94	-15.55	-1.10	-7.95	-1.08	-16.12	-7.75
Age/100	-0.71	-9.55	-0.72	-4.65	-0.60	-7.93	-3.68
Education, years/100	0.80	2.23	0.85	1.11	0.78	2.17	1.07
Male (=1)	0.03	1.95	0.04	1.20	0.04	2.58	1.20
Swedish (=1)	-0.06	-2.30	-0.06	-1.05	-0.05	-1.98	-0.94
Handicap (=1)	0.30	10.18	0.31	5.26	0.29	9.62	5.11
Number of periods unemployed	0.01	1.03	0.01	0.68	0.01	1.62	0.86
Number of states	0.00	0.96	-0.00	-0.27	-0.00	-0.67	-0.30
Days/100 of unemployment, total	0.00	1.24	0.00	0.37	0.00	0.82	0.40
Days/100 in present period	0.02	6.31	0.02	2.43	0.02	4.63	2.39
Days/100 in present state	0.00	0.04	-0.01	-0.95	-0.01	-1.96	-1.02
Some experience (=1)	-0.04	-1.54	-0.04	-0.83	-0.04	-1.83	-1.00
Long experience (=1)	-0.10	-4.26	-0.08	-1.71	-0.12	-5.28	-2.72
Education in profession (=1)	0.02	1.02	0.02	0.56	0.02	1.34	0.67
Contact with official $\geq$7 times (=1)	0.22	14.78	0.54	14.37	0.53	7.70	3.54

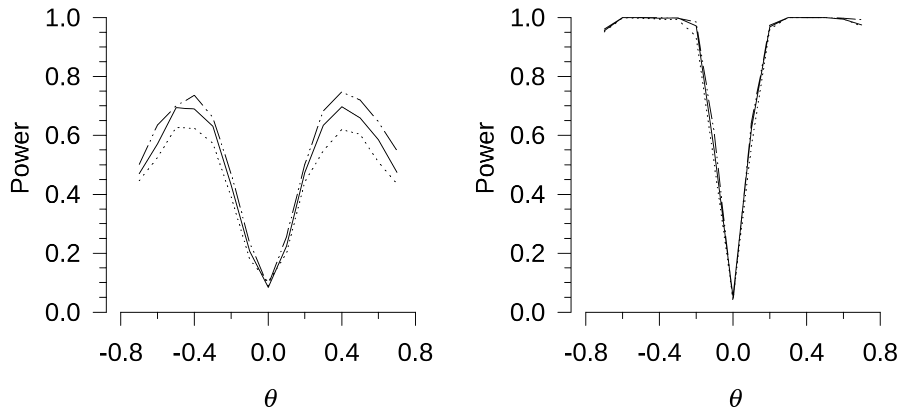

Figure 1: Power functions for the LM test statistic of $\theta = 0$ against $\theta \neq 0$ for $n = 100$ (left) and $n = 500$ (right), with $\alpha = 0.4$ (dot-dash line), $\alpha = 0$ (solid line), and $\alpha = -0.4$ (dotted line).

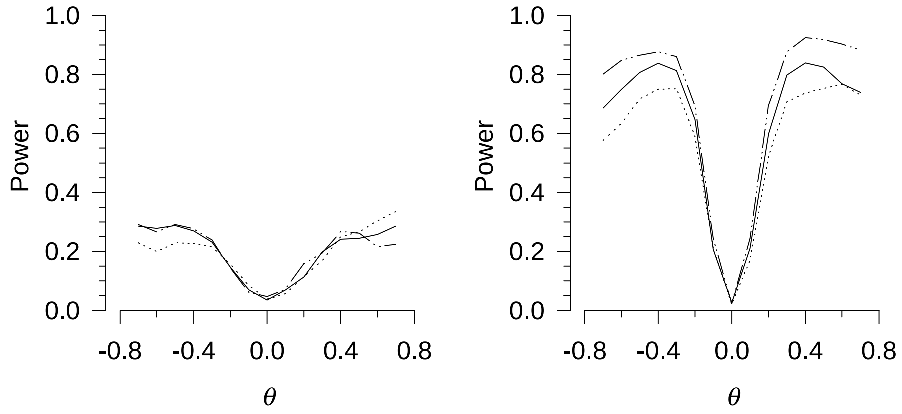

Figure 2: Power functions for the H test statistic of $\theta = 0$ against $\theta \neq 0$ for $n = 100$ (left) and $n = 500$ (right), with $\alpha = 0.4$ (dot-dash line), $\alpha = 0$ (solid line), and $\alpha = -0.4$ (dotted line).