Communications in Statistics—Simulation and Computation®, 38: 1834–1855, 2009

ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610910903147789

Taylor & Francis
Taylor & Francis Group

Simple Tests for Exogeneity of a Binary Explanatory Variable in Count Data Regression Models

KEVIN E. STAUB

Statistics and Empirical Economic Research, Socioeconomic Institute, University of Zurich, Zurich, Switzerland

This article investigates power and size of some tests for exogeneity of a binary explanatory variable in count models by conducting extensive Monte Carlo simulations. The tests under consideration are Hausman contrast tests as well as univariate Wald tests, including a new test of notably easy implementation. Performance of the tests is explored under misspecification of the underlying model and under different conditions regarding the instruments. The results indicate that often the tests that are simpler to estimate outperform tests that are more demanding. This is especially the case for the new test.

Keywords Dummy variable; Endogeneity; Poisson; Testing.

Mathematics Subject Classification 62H15; 62F03.

1. Introduction

This article is concerned with inference about endogeneity caused by a binary variable in count data models. Unlike the case with a continuous endogenous regressor, such models cannot be consistently estimated by two-stage residual-inclusion procedures, making it necessary to use other estimation techniques. For instance, nonlinear instrumental variables estimation, as introduced by Mullahy (1997), is general enough to be applicable irrespective of the binary nature of the endogenous regressor, and can therefore be used to conduct Hausman tests of endogeneity. If the focus is solely on testing exogeneity, however, easily implementable two-stage residual-inclusion also provides a valid test which was first proposed by Wooldridge (1997). Furthermore, if the researcher is willing to introduce parametric assumptions about the error structure of the model (Terza, 1998), significant efficiency gains might be exploited and alternative tests for exogeneity can be implemented.

Despite its rather specific nature, estimation of count data models with a potentially endogenous dummy variable is very common in the empirical economics

Received February 16, 2009; Accepted June 26, 2009
Address correspondence to Kevin E. Staub, Statistics and Empirical Economic Research, Socioeconomic Institute, University of Zurich, Zuerichbergstr. 14, Zurich CH-8032, Switzerland; E-mail: staub@sts.uzh.ch

literature, and with estimation routines for this models becoming available in statistical software packages[1] the number of applications is bound to increase further. Earlier examples of count data models with an endogenous dummy variable include Windmeijer and Santos Silva (1997), who studied the effect of a binary measure of self-reported health on the number of physician consultations; Terza (1998) who investigated the impact of vehicle ownership on the number of recreational trips; and Kenkel and Terza (2001) who analyzed how physician advice affects the consumption of alcoholic drinks. To cite just a few, more recent work studies whether educational attainment decreased women's fertility (Miranda, 2004), or if U.S. residence of Mexican women influenced their relationship power as measured by the number of less egalitarian responses to a questionnaire (Parrado et al., 2005). The model has also been used to test for possible endogeneity of the mechanism to price initial public offerings (bookbuilding or auction) in a regression on the number of buy recommendations for a company (Degeorge et al., 2007). Quintana-Garcia and Benavides-Velasco (2008) investigated if an increase of diversification in firm technology leads to a higher number of patents.

The model has also been the subject of more theoretically-oriented work, which developed semiparametric procedures to estimate the model under less stringent assumptions (e.g., Masuhara, 2008; Romeu and Vera-Hernandez, 2005); a Bayesian version of the model is analyzed in Kozumi (2002). However, since the impact of these developments on applied work is more modest, and given that the focus of this article is on tests for exogeneity that are relevant for applied empirical practice, the analysis will be limited to exogeneity tests obtained under more widespread—if more restrictive—model assumptions.

Below, various tests for exogeneity in a count data model with a binary endogenous regressor are presented and their performance is compared in small and moderately-sized samples through Monte Carlo simulation. This article is restricted to the just-identified case with one instrument. As a benchmark, the Hausman test that contrasts efficient and consistent estimates is evaluated against various univariate Wald tests based on an estimated parameter that captures the degree of endogeneity. Among them, a new test of particularly easy implementation is presented. The tests are assessed with regards to sensitivity to instrument strength and to mild and moderate model misspecification of the data generating process. A key result of interest to practitioners is that, overall, the two most easy-to-implement tests, including the new test, displayed very acceptable empirical size and power properties among the presented tests, often outperforming the other tests.

Frequently endogeneity tests are conceived as pretests to decide whether a model estimated with an estimator that is consistent under endogeneity can be re-estimated with a more efficient estimator that is only consistent under exogeneity. However, recent work by Guggenberger (2008) in a linear IV model context demonstrates that using a Hausman pretest can be devastating for inference on second stage tests. Thus, further simulations are performed to address the question of how exogeneity pretests affect inference about the effect of the potentially endogenous binary variable in count data models. Here, the results turn out to be less encouraging, as severe size distortions suggest that researchers should refrain from using these exogeneity tests as pretests.

[1]For example, there are routines for both Mullahy's (1997) NLIV/GMM estimator and Terza's (1998) full information maximum likelihood estimator in STATA. See Nichols (2007) and Miranda (2004), respectively.

The rest of the article is organized as follows. Section 2 presents the model under consideration. The tests for exogeneity are introduced in the next section. The design of the Monte Carlo experiment and its results are discussed in Sec. 4, while Sec. 5 contains some conclusions.

2. Count Data Regression Models with a Potentially Endogenous Binary Variable

The model considered here will be a model for a count dependent variable, y, whose mean, conditional on a vector of observed explanatory variables x, a binary variable d and an unobserved error component ε, is an exponential function of a linear index of (x, d, ε):

$$\mathrm{E}(y \mid x, d, \varepsilon) = \exp(x'\beta + \beta_d d + \varepsilon). \tag{1}$$

Concentrating the analysis to this class of models means that the conclusions of this article are relevant to a wide range of applied work, since both Poisson and Negative Binomial regression, the two most extensively used count model estimators, fall by default into the class defined in (1).[2] Note that including the error term ε in the exponential function as opposed to additively outside the function corresponds to the interpretation of ε as further variables that affect the expectation of y (but that are unobservable to the econometrician) and should be treated symmetrically to the observed variables.[3]

If the regressors x and the dummy variable d are statistically independent from ε, the conditional expectation function (1) marginal of ε is

$$\mathrm{E}(y \mid x, d) = \exp(x'\beta + \beta_d d)\,\mathrm{E}[\exp(\varepsilon \mid x, d)] = \exp(x'\beta^* + \beta_d d), \tag{2}$$

assuming that the mean of $\exp(\varepsilon)$ is constant and that x includes a constant first element, as then β^* is equal to β but with first element shifted by $\ln \mathrm{E}[\exp(\varepsilon)]$ (cf. Windmeijer and Santos Silva, 1997). Note that assuming zero correlation between regressors and errors as in the linear case is not sufficient for (2) to hold, as this does not warrant that $\mathrm{E}[\exp(\varepsilon) \mid x, d] = \mathrm{E}[\exp(\varepsilon)]$.

Equation (2) represents the case of exogeneity, and efficient estimation of the model depends on the distribution of ε and of $y \mid x, d, \varepsilon$. For instance, with the latter being Poisson-distributed, if ε is distributed as normal or exp-gamma, then the resulting models marginal of ε are the Poisson-log-normal and the negative binomial regression model, respectively. However, because of its robustness to distributional misspecification and easy implementation, it is very common to give up full efficiency and estimate models satisfying (2) by Poisson pseudo maximum likelihood (cf. Wooldridge, 1997), which yields consistent estimates of (β^*, β_d) irrespective of the distribution of ε. Nonlinear least squares estimation is also consistent, but is less frequently encountered in the count data context as it neglects the count nature of the dependent variable. Consistency up to the first element

[2]Evidently, exponential conditional mean functions are not limited to count data, and many of the procedures and results discussed here are in principle applicable to continuous data as well.

[3]An alternative justification for this representation is by means of the interpretability of the model in terms of ceteris paribus marginal effects (cf. Winkelmann, 2008, p. 160).

does not hold in general for nonlinear models but is a specific consequence of the multiplicative separability of linear combinations in the exponential function.

For continuous elements of x, the parameters β have the interpretation of (semi-)elasticities with respect to the conditional expectation function (CEF), i.e., for the kth regressor

$$\frac{\partial \mathrm{E}(y \mid x, d)/\mathrm{E}(y \mid x, d)}{\partial x_k} = \beta_k$$

while for discrete regressors, as for instance the binary variable of interest here, direct interpretation of the coefficients is only suitable as an approximation to the discrete partial effect $\exp(\beta_d) - 1$. Note that for both marginal and discrete partial effects as well as for predictions of CEF, inconsistent estimation of the first element of β is inconsequential.[4]

The binary variable d is endogenous in model (1) whenever it is not statistically independent from ε and, thus, the second equality in (2) does not hold. Estimation of the model neglecting endogeneity yields inconsistent estimates of all parameters, even when the regressors are orthogonal. To pin down the source of this dependence one can recur to modelling d as

$$d = \begin{cases} 1 & \text{if } z'\gamma \geq v \\ 0 & \text{if } z'\gamma < v \end{cases}, \tag{3}$$

where z is a vector of observable variables, possibly including at least some elements from x, and the unobserved error component v follows some joint distribution with ε from (1). Terza (1998) proposed to specify the distribution of $(\varepsilon, v)'$ conditional on the exogenous variables (x, z) as bivariate normal according to

$$\begin{pmatrix} \varepsilon \\ v \end{pmatrix} \,\Big|\, x, z \sim \text{Normal}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}\right] \tag{4}$$

which defines a probit model for (3). Also, statistical dependence is captured entirely by the correlation parameter $\rho \in [-1, 1]$ which yields independence whenever $\rho = 0$. Thus, the hypothesis of exogeneity can be stated as $H_0{:}\rho = 0$ with alternative $H_1{:}\rho \neq 0$ corresponding to endogeneity.

3. Tests for Exogeneity

The most widely used test for exogeneity is probably the Hausman test, since it is applicable in a vast number of situations. In the context of the model discussed here, it has the advantage that it does not require assumption (4). After shortly discussing Hausman tests, the exposition will turn to univariate Wald tests, first presenting two tests based on Terza's (1998) full information maximum likelihood estimator and a more general two-stage method of moments estimator. Finally, two tests of particularly easy implementation are discussed, which also rely on estimation in two stages: a new test based on a first-order approximation to the method of moments estimator and a residual inclusion estimator.

[4]While the partial effects do not depend on the first element of β, predictions of CEF are consistent because $x'\hat{\beta}^*$ is consistent for $x'\beta + \ln \mathrm{E}[\exp(\varepsilon)]$.

3.1. *Hausman Contrast Tests*

The Hausman test (Hausman, 1978) in its most general form contrasts two estimates obtained from different estimators. In the case of endogeneity, one of the estimators is consistent under both the null hypothesis (exogeneity) and the alternative (endogeneity) while the second estimator is inconsistent under the alternative but efficient (relative to any linear combination of the two estimators) under the null hypothesis. Then, denoting by $\hat{\beta}_C$ the consistent estimate and by $\hat{\beta}_E$ the efficient one, the Hausman test statistic is

$$h = (\hat{\beta}_E - \hat{\beta}_C)'\left[\mathrm{Var}(\hat{\beta}_C) - \mathrm{Var}(\hat{\beta}_E)\right]^{-1}(\hat{\beta}_E - \hat{\beta}_C) \sim \chi^2_j$$

with the degrees of freedom of the χ^2 distribution, j, being equal to the dimension of the β-vectors involved in h.

An early application of a Hausman test to count data models with endogeneity is provided by Grogger (1990), who suggested calculating the corresponding test statistic with estimates from Poisson ML and a nonlinear instrumental variables (NLIV) estimator based on an additive error to the CEF. However, this estimator is inconsistent under a multiplicative error defined implicitly as $\exp(\varepsilon)$ in (1) (Dagenais, 1999; Terza, 2006), and Mullahy's (1997) GMM estimator is therefore more appropriate to estimate $\hat{\beta}_C$. In the just-identified case studied here, this estimator is the NLIV based on the residual function $r \equiv y\exp(-x'\beta - \beta_d d) - 1$ which, given an appropriate instrument z, implies the moment condition

$$\mathrm{E}(r \mid z) = \mathrm{E}[\exp(\varepsilon) - 1 \mid z] = 0.$$

Thus, writing the NLIV estimate of β_d as $\hat{\beta}_d^{\mathrm{NLIV}}$ and the corresponding Poisson PML estimate as $\hat{\beta}_d^{\mathrm{PPML}}$, a Hausman test for exogeneity can be based on the test statistic

$$h^1 = \frac{(\hat{\beta}_d^{\mathrm{PPML}} - \hat{\beta}_d^{\mathrm{NLIV}})^2}{\mathrm{Var}(\hat{\beta}_d^{\mathrm{NLIV}}) - \mathrm{Var}(\hat{\beta}_d^{\mathrm{PPML}})} \sim \chi^2_1. \tag{5}$$

Sometimes this Hausman test is implemented by additionally including all elements of β in the contrast, but both Creel (2004) and Chmelarova (2007) find that h^1 outperforms the full-β-version of the test in finite samples.

The denominator of h^1 results as a special case of the variance of a difference of estimates when the minuend is the efficient estimator, as then $\mathrm{Cov}(\beta_E, \beta_C) = \mathrm{Var}(\beta_E)$ (Hausman, 1978). There are two routes of potentially improving on h^1. The first would be to specify the distribution of ε and then calculating the corresponding ML estimator. For instance, if (4) holds, the model for y conditional on observables is a Poisson-log-normal (PLN) mixture. As the PLN estimator is efficient relative to the Poisson estimator in this model, a Hausman test statistic calculated by substituting the PPML estimates by PLN equivalents could perform better:

$$h^2 = \frac{(\hat{\beta}_d^{\mathrm{PLN}} - \hat{\beta}_d^{\mathrm{NLIV}})^2}{\mathrm{Var}(\hat{\beta}_d^{\mathrm{NLIV}}) - \mathrm{Var}(\hat{\beta}_d^{\mathrm{PLN}})} \sim \chi^2_1.$$

A second procedure in the vein of Weesie (1999) and Creel (2004) is to estimate $\mathrm{Cov}(\beta_E, \beta_C)$ directly instead of relying on the simplification under

asymptotic efficiency.[5] This implies rewriting the two optimization problems of the Poisson PML and the NLIV as a joint problem by stacking PPML's first-order conditions and the moment conditions of NLIV. The resulting test statistic is

$$h^3 = \frac{(\hat{\beta}_d^{\text{PPML}} - \hat{\beta}_d^{\text{NLIV}})^2}{\text{Var}(\hat{\beta}_d^{\text{PPML}}) + \text{Var}(\hat{\beta}_d^{\text{NLIV}}) - 2\text{Cov}(\hat{\beta}_d^{\text{PPML}}, \hat{\beta}_d^{\text{NLIV}})} \sim \chi_1^2.$$

If the errors follow a bivariate normal distribution, all three tests are asymptotically equivalent. If not, h^2 is inconsistent, but h^1 and h^3 retain their consistency. The performance of the two additional variants relative to h^1 is less clear in finite samples. For h^3 the potential gains depend crucially on the small sample properties of the covariance estimator. Likewise, for h^2 to outperform h^1 the higher precision of PLN relative to Poisson—which is an asymptotic result—needs to be visible enough in finite samples.

3.2. *Wald Tests*

There are alternatives to the Hausman contrast test for exogeneity. For instance, in the linear IV model, estimating a reduced form for the endogenous variable in order to obtain residuals which can be plugged into the structural equation leads to an asymptotically equivalent test for endogeneity (Hausman, 1978). Monte Carlo simulations in Chmelarova (2007) showed that Wald versions of the Hausman test often have better properties than the contrast version under a series of different conditions. However, the endogeneity in count data models in Chmelarova (2007) concerns continuous regressors, so that the residual inclusion technique is consistent. Residual inclusion in the framework discussed presently with an endogenous dummy, on the other hand, yields inconsistent estimates.[6] Nevertheless, a number of consistent Wald tests are available.

First, Wooldridge (1997) suggested that while the procedure yields inconsistent estimates, the test based on residual inclusion is consistent. Second, if one is willing to impose (4) and a distributional assumption for $y \mid x, d, \varepsilon$, one can recur to Terza's (1998) maximum likelihood estimator, which explicitly estimates the correlation coefficient of the bivariate normal distribution so that the hypothesis $\rho = 0$ can be tested directly. Relaxing the distributional assumption on the dependent variable still allows to estimate a scaled version of ρ based on (4), which can be used to test for endogeneity. Last, following the literature on inference using local approximations (cf. Chesher, 1991; Gourieroux and Visser, 1997), one can derive a test based on the inclusion of a generalized residual in the structural equation. While the second strategy yields consistent estimates for β_d under the alternative, the first and last do not. Their advantage, however, lies in their easy implementation, since only a standard Poisson regression is needed to carry out these tests.

[5]Creel's (2004) approach is optimal GMM, while Weesie (1999) does not use a second step weighting matrix. Clearly, in the just identified case under consideration both amount to the same as the choice of the weighting matrix does not affect the estimates.

[6]Terza et al. (2008) showed that residual inclusion in nonlinear models is inconsistent in general. Discussions of consistency of residual inclusion in Poisson PML models with continuous endogenous regressors and inconsistency with binary regressors can be found inter alia in Wooldridge (1997) and Winkelmann (2008).

3.2.1. Full Information Maximum Likelihood and Two-Stage Method of Moments Estimation. Assuming that (4) holds and that $y \mid x, d, \varepsilon$ follows a Poisson distribution with expectation (1), maximum likelihood estimation of the joint model proceeds by maximizing the sample log-likelihood function $\mathcal{L}(\beta_d, \beta, \gamma, \rho, \sigma) = \sum_{i=1}^{n} \log f(y_i, d_i \mid x_i, z_i)$, with $f(\cdot)$ denoting the probability density function, which given the assumptions is equal to (Terza, 1998)

$$\begin{aligned} f(y, d \mid x, z) &= \int_{-\infty}^{\infty} f(y \mid d, x, z, \varepsilon) \times f(d \mid x, z, \varepsilon) \times f(\varepsilon \mid x, z) d\varepsilon \\ &= \int_{-\infty}^{\infty} \exp(\lambda)\lambda^{y}(y!)^{-1} \times \Phi^{*}(\varepsilon)^{d}(1 - \Phi^{*}(\varepsilon))^{1-d} \times \sigma^{-1}\phi(\varepsilon/\sigma \mid x, z) d\varepsilon, \end{aligned}$$

where $\lambda \equiv \exp(x'\beta + \beta_d d + \varepsilon)$ and $\Phi^{*}(\varepsilon) \equiv \Phi\left(\frac{z'\gamma + \frac{\rho}{\sigma}\varepsilon}{\sqrt{1-\rho^2}}\right)$; $\Phi(\cdot)$ and $\phi(\cdot)$ denoting the cdf and pdf of the standard normal distribution, as usual. While the expression for $f(y, d \mid x, z)$ has no closed form solution, it is possible to approximate it through Gauss-Hermite quadrature. Given the ML estimate $\hat{\rho}$, the null hypothesis $H_0{:}\rho = 0$ is tested constructing the t-statistic

$$t^{1} = \frac{\hat{\rho} - 0}{s.e.(\hat{\rho})} \sim N(0, 1) \tag{6}$$

with $s.e.(\hat{\rho})$ indicating any usual asymptotically valid ML standard error of $\hat{\rho}$.

Terza (1998) also suggested a two stage estimation of this model which leaves $f(y \mid d, x, z, \varepsilon)$ unspecified. While the relaxation of assumptions is rather moderate as bivariate normality of the errors is maintained, the gains of such a procedure lie mostly in increased computational stability.[7] Consider (1) under assumption (4):

$$\begin{aligned} \mathrm{E}(y \mid x, d) &= \exp(x'\beta + \beta_d d)\mathrm{E}(\exp(\varepsilon) \mid x, d) \\ &= \exp(x'\beta + \beta_d d)\exp\left(\frac{\sigma_{\varepsilon}^2}{2}\right)\left[d\frac{\Phi(\theta + z'\gamma)}{\Phi(z'\gamma)} + (1 - d)\frac{1 - \Phi(\theta + z'\gamma)}{1 - \Phi(z'\gamma)}\right] \\ &\equiv \exp(x'\beta^{*} + \beta_d d)\psi(\theta, \gamma; z) \end{aligned}$$

with $\theta = \sigma\rho$. To estimate this model in stages, first a probit regression is performed to obtain estimates of γ, so that in a second stage estimation optimization proceeds with respect to (β, β_d, θ). Terza's (1998) suggestion is to implement the second stage as nonlinear least squares (NLS), or as nonlinear weighted least squares (NWLS) if the researcher wishes to incorporate a priory knowledge of the distribution of $y \mid d, x, z, \varepsilon$.

In the present work, however, the second stage estimation will also be implemented as a Poisson pseudo-ML regression, i.e., estimates of (β, β_d, θ) are obtained by maximizing a pseudo-log-likelihood function of the Poisson distribution with expectation $\tilde{\lambda} \equiv \exp(x'\beta^{*} + \beta_d d)\psi(\theta, \hat{\gamma}; z)$. This estimation strategy represents a compromise between NLS and NWLS, in the sense that it is bound to be

[7]An important aspect of leaving $f(y \mid d, x, z, \varepsilon)$ unspecified is that it broadens the class of models this estimator is applicable to other non count exponential CEF models. See, for instance, Egger et al. (2009) who applied such a model to bilateral trade.

more efficient for count data than NLS since it takes account of the inherent heteroskedasticity characteristic of count data,[8] while it avoids the computational difficulty of the more efficient NWLS procedure.

With an estimate of θ, the pertinent t-statistic of the test with null hypothesis $H_0{:}\theta = 0$ is

$$t^2 = \frac{\hat{\theta} - 0}{s.e.(\hat{\theta})} \sim N(0, 1). \tag{7}$$

3.2.2. Generalized Residual Inclusion. It is possible to approximate the estimation of the two-stage method described above without the need of estimating a Poisson regression with mean $\tilde{\lambda}$, which in general requires some extra programming as standard econometric software usually only allow to specify variables entering a linear index in the exponential function. This is related to Greene's (1995, 1998) work in the context of sample selection in count data models. The starting point of this approximation is again (1) under assumption (4), which written separately for the two possible outcomes of d is

$$\mathrm{E}(y \mid x, d = 1) = \exp(x\beta^* + \beta_d d)\frac{\Phi(\theta + z'\gamma)}{\Phi(z'\gamma)} = \exp(x\beta^* + \beta_d)Q_1 \quad \text{and}$$

$$\mathrm{E}(y \mid x, d = 0) = \exp(x\beta^*)\frac{1 - \Phi(\theta + z'\gamma)}{1 - \Phi(z'\gamma)} = \exp(x\beta^*)Q_0.$$

Taking logarithms of the endogeneity bias correction terms Q_0 and Q_1 allows to write them as part of the linear index in the exponential function. Furthermore, the first-order Taylor series expansion of $\log Q_0$ and $\log Q_1$ around $\theta = 0$ is

$$\log Q_1 \approx \theta\frac{\phi(z'\gamma)}{\Phi(z'\gamma)} \quad \text{and} \quad \log Q_0 \approx \theta\frac{-\phi(z'\gamma)}{1 - \Phi(z'\gamma)},$$

so that the second stage of the former estimator can be approximated by estimating a Poisson pseudo-ML regression with expectation

$$\mathrm{E}(y \mid x, d) \approx \exp(x'\beta^* + \beta_d d + \theta m), \quad \text{with} \quad m = d\frac{\phi(z'\gamma)}{\Phi(z'\gamma)} + (1 - d)\frac{-\phi(z'\gamma)}{1 - \Phi(z'\gamma)}$$

and replacing m with a consistent estimate $\hat{m}$ obtained with probit estimates $\hat{\gamma}$.[9]

Estimates of m represent generalized residuals in the sense that the first order conditions in the estimation of γ in the reduced form are a set of orthogonality conditions between m and z. Orme (2001), who introduced the same local approximation in the context of a dynamic probit model, proposed testing for the presence of an endogenous initial condition by using the estimated coefficient on the generalized residuals, $\hat{\theta}$. The same procedure can be applied here, suggesting a new test for exogeneity in the present count data context: if $\rho = 0$ the approximation is exact, so that the pseudo-ML estimates of θ will be consistent under the null hypothesis of exogeneity and the test statistic t^2 in (7) can be used.

[8]The argument for Poisson pseudo-MLE against NLS is presented extensively by Santos Silva and Tenreyro (2006) in the context of non count exponential CEF models.

[9]This technique has also been used by Angrist (2001) to approximate a Tobit MLE.

3.2.3. Residual Inclusion. While a glance at the pertinent literature shows that many resarchers are comfortable with assumption (4), the test proposed in Wooldridge (1997) is consistent under weaker distributional assumptions as it does not require bivariate normality. It does, however, in contrast to the Wald tests considered so far, require instruments.

The residual inclusion estimation procedure consists in including residuals from the reduced form equation for the endogenous variable in the linear index of the second stage exponential CEF. The two key assumptions for consistency of this technique are independence of the reduced form residuals from the instruments and linearity of the CEF of ε given v. The linear CEF condition holds if, as considered so far, the error terms are bivariate normally distributed. However, independence of the residuals from the instruments is unlikely to hold in the binary case. Nevertheless, as pointed out by Wooldridge (1997), the procedure is still valid to test for exogeneity, since under the null hypothesis of d being exogenous the two assumptions on the errors need not hold as then the CEF reduces to (2), i.e., while the procedure does not yield consistent estimates, it does provide a valid Hausman-type Wald test for endogeneity.

Starting with assumption (4), the CEF of ε given v is $\mathrm{E}(\varepsilon \mid v) = \theta v$, with $\theta = \sigma\rho$ as before. Therefore, it is always possible to write $\varepsilon = \theta v + error$, with this error being independent of v by construction. Thus, the suggested test would proceed by replacing ε in (1) with $\theta v + error$ and conditioning y on x, d, and v (instead of ε). That is, estimating

$$\mathrm{E}(y \mid x, d, v) = \exp(x'\beta + \beta_d d + \theta v)$$

by Poisson pseudo-ML, using $\hat{v} = d - \Phi(z'\hat{\gamma})$ for the unobserved v, where estimates for γ could be obtained from a probit regression or, alternatively, from other models for binary dependent variables such as the linear probability model, which would produce residuals $\hat{v} = d - z'\hat{\gamma}$. Again, the null hypothesis of exogeneity is expressed as $\theta = 0$ and the test statistic t^2 can be used.

4. A Monte Carlo Simulation Study

To assess finite sample properties of the tests discussed in the previous sections, a Monte Carlo simulation experiment is conducted. Bearing in mind the known limitations of such an approach, special care has been placed on addressing a variety of issues concerning the performance of the tests under different conditions, such as moderate misspecification and unavailability of instruments, as well as suitability of the tests for pretesting. All programming has been written in GAUSS, pseudo-random number generators, and other subroutines used were taken from GAUSS' libraries; code and a supplementary appendix containing more extensive results are available from the author on request.

4.1. *Experimental Design*

Every reported simulation proceeded by drawing a random sample of size n from two independent standard normally distributed variables, x and z. Next, the errors

ε and v were drawn from some joint distribution having 0 expectation and variance of v equal to 1. The endogenous binary variable, d was formed according to

$$d = \mathbb{1}(\gamma_z z + \gamma_x x + v \geq 0)$$

with $\mathbb{1}(\cdot)$ denoting the indicator function. Then, the conditional expectation of the count dependent variable y was constructed as

$$\lambda = \exp(-1 + 0.5x + d + \varepsilon)$$

so that, finally, y was obtained by random sampling from some count data distribution with expectation λ. Here, the effect of the dummy on the expectation of y is $\exp(1) - 1 \approx 1.71$ which might seem above what can be expected in some empirical applications, but adherence to the unit coefficient on d can be defended on the grounds of comparability to other studies.[10] Sample sizes (n) considered were 200, 500, and 1,000. Results for larger samples are not reported as then differences between tests even out quickly and they converge to their asymptotic limits. Smaller samples, on the other hand, were not investigated as microeconometric applications of this model with less observations are unlikely to be encountered in practice. Most Monte Carlo simulations were replicated 10,000 times, the significantly more computing-intensive routines for the tests based on full information maximum likelihood (FIML) estimates were performed with 2,000 and 1,000 replications. All tests were performed at a nominal significance level of 5%. Different data generating processes were obtained by varying the values of the vector γ, the joint distribution of the errors and the distribution of $y \mid x, d, \varepsilon$.

By assigning different values to γ, the strength of the instrument was manipulated. While in the linear IV model the concentration parameter provides an unequivocal summary measure of instrument strength (cf. Stock et al., 2002), there is no generic equivalent for nonlinear models. Trivially, the impact of the instrument is affected by the proportion of the variance of $(\gamma_z z + \gamma_x x + v)$ explained by $\gamma_z z$. Note that a given ratio can be obtained by either changing the variance of the error v with respect to the given variance of $(\gamma_z z + \gamma_x x)$, or by altering the relation $\mathrm{Var}(\gamma_z z)/\mathrm{Var}(\gamma_x x)$ with given relation of $\mathrm{Var}(\gamma_z z + \gamma_x x)$ to $\mathrm{Var}(v)$. While the two interventions amount to the same in the linear model, here results might differ.

The probability density function (pdf) $f(y \mid x, d, \varepsilon)$ was set to be either Poisson with mean λ or Negative Binomial I with mean λ and variance 2λ. With the exception of the test based on full information maximum likelihood, all tests should be invariant to the overdispersion introduced by the Negative Binomial I variant. The baseline specification for the error distribution was the bivariate normal distribution given in (4) with values of ρ ranging from 0–0.95 for most experiments. To assess sensitivity to misspecification of (4), (ε, v) were also generated from a bivariate Gaussian copula with an exponential Gamma marginal distribution for ε and a standard logistic marginal for v, inducing a Negative Binomial model for y conditional on observables and a logit model for d. Finally, the tests were conducted with the errors following the same exp-Gamma and logistic marginals but with joint distribution determined through the Frank copula.

[10]Monte Carlo studies of count data models with unit coefficient on endogenous variables include Creel (2004), Romeu and Vera-Hernandez (2005), and Chmelarova (2007).

Table 1
Details on the DGP of Monte Carlo simulations

Table	Columns	Distribution of $y \mid x, d, \varepsilon$	Distribution of (ε, v)	Reduced form parameters (γ_x, γ_z)
1	all	Poisson(λ)	BVN(0, 0, 1, 1, ρ)	$(\sqrt{0.50}, \sqrt{0.50})$
2	(1)	Poisson(λ)	BVN(0, 0, 1, 1, ρ)	$(\sqrt{0.75}, \sqrt{0.25})$
	(2)	Poisson(λ)	BVN(0, 0, 1, 1, ρ)	$(\sqrt{1.50}, \sqrt{0.50})$
	(3)	Poisson(λ)	BVN(0, 0, 1, 1, ρ)	$(\sqrt{0.25}, \sqrt{0.75})$
3	(1)	Poisson(λ)	BVN(0, 0, 1, 1, ρ)	$(\sqrt{0.50}, 0.00)$
	(2)	Poisson(λ)	BVN(0, 0, 1, 1, ρ)	$(\sqrt{2.00}, 0.00)$
4	(1), (2)	NegBin(λ, λ)	BVN(0, 0, 1, 1, ρ)	$(\sqrt{0.50}, \sqrt{0.50})$
	(3)	Poisson(λ)	$\varepsilon \sim \exp\text{Gamma}(1, 1)$, $v \sim \text{Logistic}(0, 3/\pi)$	$(\sqrt{0.50}, \sqrt{0.50})$
	(4), (5)	Poisson(λ)	Gaussian copula*	$(\sqrt{0.50}, \sqrt{0.50})$
	(6), (7)	Poisson(λ)	Frank copula*	$(\sqrt{0.50}, \sqrt{0.50})$
5	(1)	Poisson(λ)	BVN(0, 0, 1, 1, ρ)	$(\sqrt{0.50}, \sqrt{0.50})$
	(2)	Poisson(λ)	BVN(0, 0, 1, 1, ρ)	$(\sqrt{0.25}, \sqrt{0.75})$

*Marginal distributions of the copulae: $\varepsilon \sim \exp\text{Gamma}(1, 1)$, $v \sim \text{Logistic}(0, 3/\pi)$.

A table containing the descriptions of the precise data generating processes that were used in producing the results discussed below can be found in Table 1.

The next subsection discusses empirical size and power of the proposed tests under ideal assumptions on the data generating process, i.e., with assumption (4) holding. Next, the discussion centers on the tests that theoretically are able to identify exogeneity in the absence of instruments, assessing the goodness of their performance under this condition in the simulations. Results under misspecification of the data generating process are considered next, and the section closes considering the effect on the empirical size of tests on $\hat{\beta}_d$ after using endogeneity tests as pretests to choose between estimators for the model.

4.2. *Empirical Size and Power*

The first three columns of Table 2 contain simulation results for the empirical size of different tests for exogeneity with nominal size 5%. The table shows results for three different sample sizes of 200, 500, and 1,000 observations. The coefficients of the reduced form equation, γ_x and γ_z, were set to $\sqrt{0.5}$ each, so that the ratio $\text{Var}((\gamma_z z + \gamma_x x)/\text{Var}(v)$ equaled 1. With 10,000 replications, a 95% confidence interval for the estimated size of tests is $[0.05 \pm 1.96\sqrt{0.05 \times 0.95/10'000}] \approx [0.046, 0.054]$.[11]

The first three rows contain the rejection frequencies of the exogeneity hypothesis for the Hausman tests with test statistics h^1, h^2, and h^3 discussed previously. The test that contrasts PPML estimates with the NLIV estimates (H1)

[11]The corresponding confidence interval for 2,000 replications is approximately [0.405, 0.595].

Table 2
Rejection frequencies of tests for exogeneity—the effect of sample size

	$\rho = 0$			$\rho = 0.20$			$\rho = 0.50$		
Sample size:	200	500	1000	200	500	1000	200	500	1000
Hausman contrast tests									
H1	0.0365	0.0459	0.0517	0.0672	0.1168	0.2019	0.1796	0.4423	0.7451
H2	0.0287	0.0371	0.0432	0.0583	0.1050	0.1798	0.1708	0.4223	0.7239
H3	0.0038	0.0060	0.0084	0.0097	0.0265	0.0534	0.0363	0.1902	0.4788
Wald tests									
FIML	0.0540	0.0635	0.0640	0.0670	0.1600	0.2750	0.2070	0.6605	0.9160
TSM NLS	0.0893	0.0728	0.0627	0.0668	0.0638	0.0799	0.0790	0.1997	0.4376
TSM PPML	0.0739	0.0616	0.0561	0.0766	0.1079	0.1806	0.2046	0.4616	0.7620
GRI	0.0750	0.0603	0.0573	0.1047	0.1309	0.1958	0.2798	0.4971	0.7570
RI	0.0814	0.0605	0.0554	0.1060	0.1240	0.1706	0.2445	0.3964	0.5963
GRI-TSA	0.0509	0.0441	0.0420	0.0748	0.0999	0.1578	0.2188	0.4287	0.7043
RI-TSA	0.0711	0.0566	0.0535	0.0945	0.1192	0.1667	0.2272	0.3863	0.5931

Notes: Number of replications = 10,000 (FIML: 2,000 replications). Nominal test size = 0.05.

performs better than the two other Hausman tests. While underrejecting the true null hypothesis with 200 observations, H1 displays correct size for larger samples, while H2, which uses PLN estimates instead of PPML, underrejects slightly even for the largest sample. The test H3, which attempts to improve on H1 by estimating the covariance from the data instead of relying on the asymptotic simplification, has a serious underrejection problem for all sample sizes considered. Since estimated coefficients and their standard errors are the same as in H1, it follows that underrejection must be due to upward bias in the estimation of $\mathrm{Cov}(\beta_d^{PPML}, \beta_d^{NLIV})$. These results on the Hausman tests are opposite in sign to previous findings concerning continuous endogenous regressors (Creel, 2004; Chmelarova, 2007), were Hausman contrast tests tend to overreject H_0. As for results on power, Table 2 displays rejection frequencies of the false null hypothesis under $\rho = 0.2$ (columns 4–6) and $\rho = 0.5$ (columns 7–9). The performance of H1 and H2 are practically indistinguishable. This implies that there might be very small or even no gains at all from implementing H2 instead of the more robust H1, even under an ideal DGP for H2.

Turning to the Wald tests, results are presented for tests based on the FIML estimates (FIML), two-stage method of moments estimates implemented via NLS (TSM NLS) and PPML (TSM PPML), as well as for the new test derived from the generalized residual inclusion (GRI) and the test based on the residual inclusion procedure (RI). The TSM tests are based on two-stage adjusted standard errors. For GRI and RI, results are presented separately for tests using regular standard errors and two-stage adjusted standard errors (GRI-TSA and RI-TSA). Thus, GRI and RI are tests which virtually can be implemented by the practitioner in a matter of seconds, while the two-stage adjustment might take more time as it generally requires a minimum of custom programming.

Considering the empirical size of the Wald tests with samples of 200 observations, most of them overreject the null hypothesis by 2–4 percentage points,

with the exception of FIML and GRI-TSA, whose rejection frequencies are not significantly different from 5%. With increasing sample size, the other tests also gradually tend to the nominal size. As the results make evident, using two-stage adjusted standard errors improves noticeably the empirical size of the GRI and RI tests in small to moderate samples, although the GRI-TSA standard errors seem to be a little bit too large leading to slight underrejection in some cases. The TSM NLS test is the only one to overreject clearly even with sample size 1,000. It also performs comparatively poorly with respect to power. As expected, FIML has the largest power in this setting where it is the efficient estimator, followed by the TSM PPML and GRI(-TSA) tests. The RI(-TSA) tests are comparable in power to the H1 Hausman test.[12]

The DGP in Table 2 implied that $\mathrm{Var}(\gamma_z z)/\mathrm{Var}(\gamma_z z + \gamma_x x + v) = 0.25$, i.e., that the variance of the instrument determines one quarter of the total variance of the linear combination that determines d. Now, consider a change in instrument strength. By specifying a DGP which leaves $\gamma_z = \sqrt{0.5}$ as before, but with $\gamma_x = \sqrt{1.5}$, the fraction of the variance explained by the impact of the instrument, $\gamma_z z$, with respect to the whole systematic variance, $\mathrm{Var}(\gamma_z z + \gamma_x x)$, falls from 0.5 to 0.25, while the systematic variance relative to the error variance, $\mathrm{Var}(v)$, doubles. Taken together, the new instrument is weaker since $\mathrm{Var}(\gamma_z z)/\mathrm{Var}(\gamma_z z + \gamma_x x + v) \approx 0.167$. How does this change affect power and size of the tests? Comparing the columns with sample size 500 in Table 2 with columns labeled (2) in Table 3 gives an idea. While the Hausman and residual inclusion tests suffer severe power loss, TSM PPML and the generalized residual inclusion tests are barely affected. Figure 1

Table 3
Rejection frequencies of tests for exogeneity—the effect of instrument strength

	$\rho = 0$			$\rho = 0.20$			$\rho = 0.50$		
IV strength:	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
				Hausman contrast tests					
H1	0.0283	0.0449	0.0584	0.0639	0.0936	0.1541	0.2206	0.3042	0.5954
H2	0.0248	0.0375	0.0413	0.0602	0.0826	0.1286	0.2194	0.2995	0.5560
H3	0.0072	0.0086	0.0040	0.0222	0.0277	0.0217	0.1091	0.1325	0.2223
				Wald tests					
FIML	0.0820	0.0620	0.0555	0.1305	0.1525	0.2015	0.4935	0.6095	0.7710
TSM NLS	0.0819	0.0862	0.0695	0.0622	0.0744	0.0799	0.1100	0.2080	0.2905
TSM PPML	0.0666	0.0683	0.0566	0.0960	0.1071	0.1256	0.3111	0.4382	0.5841
GRI	0.0629	0.0640	0.0586	0.1009	0.1206	0.1528	0.3408	0.4543	0.6055
RI	0.0617	0.0633	0.0594	0.0951	0.0980	0.1494	0.2496	0.2665	0.5174
GRI-TSA	0.0451	0.0484	0.0419	0.0779	0.0964	0.1168	0.2835	0.3984	0.5460
RI-TSA	0.0533	0.0581	0.0577	0.0848	0.0908	0.1468	0.2315	0.2544	0.5130

Notes: Number of replications = 10,000 (FIML: 2,000 replications). Nominal test size = 0.05. IV-strength as detailed in text or Table 1.

[12]Some authors prefer to use what is called size-corrected power to make comparisons across tests. Here, no size-corrected power is presented, since the question addressed is how these tests work in practice and which are useful under given characteristics of the data generating process.

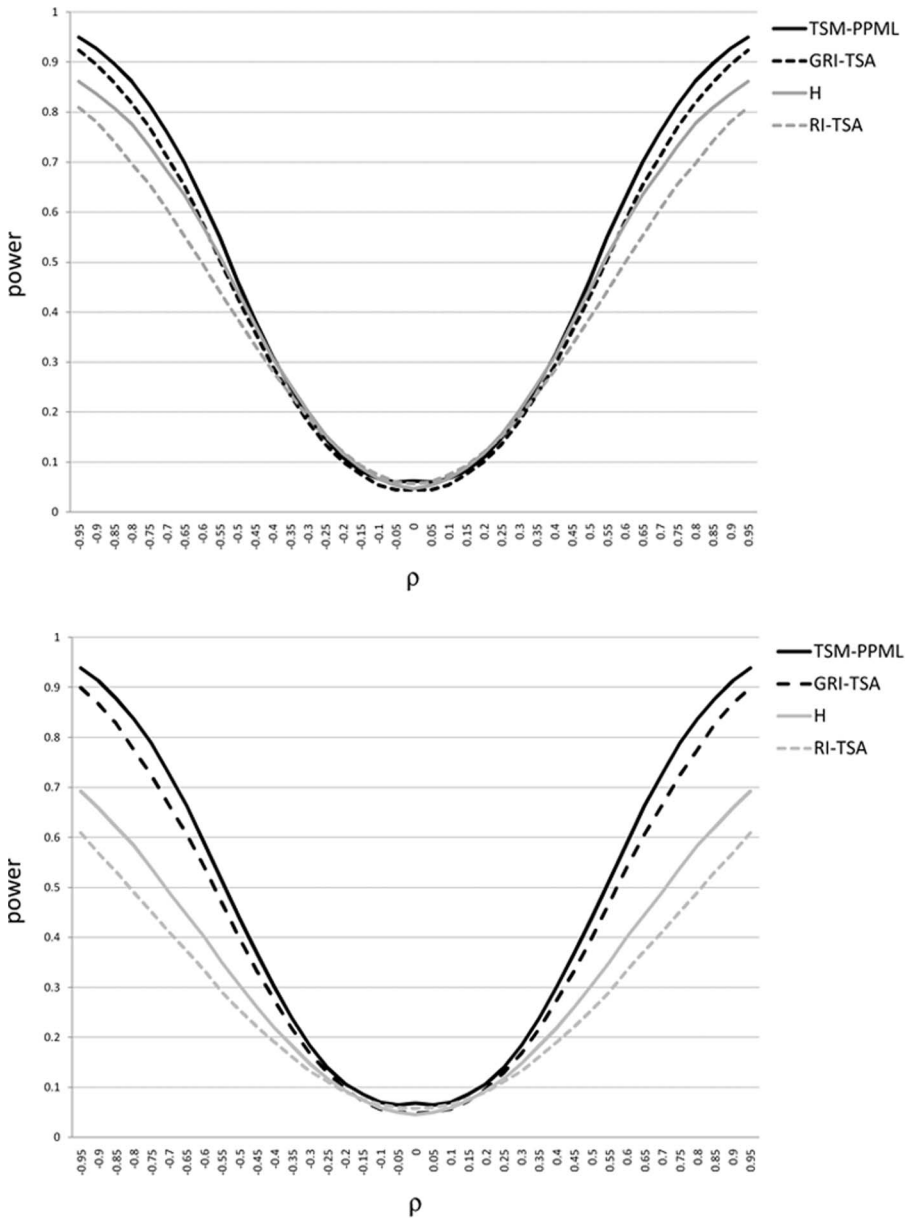

Figure 1. Empirical power of tests for exogeneity. *Notes*: Sample size = 500. Nominal test size = 0.05. Reduced form parameters: upper panel $(\gamma_x, \gamma_z) = (\sqrt{0.5}, \sqrt{0.5})$; lower panel $(\gamma_x, \gamma_z) = (\sqrt{1.5}, \sqrt{0.5})$. Graphs based on 20 points $\rho = 0, 0.05, 0.10, \ldots, 0.95$. Values for negative ρ mirrored symmetrically from corresponding positive points. Each point obtained from 10,000 replications.

details the circumstance graphically by plotting the power functions of H1, TSM PPML, GRI-TSA, and RI-TSA over the support of ρ for both DGPs. The difference in power grows with increasing absolute value of ρ and is over 20 percentage points at the extremes. The reason for this difference is that Hausman and residual inclusion tests rely only on the dependence between the instrument and the endogenous variable, which in this experiment was significantly weakened. Meanwhile, tests as TSM PPML and GRI seem to be able to compensate this loss with the increased variance of the systematic part which allows them to exploit more fully their functional form assumption.

The remaining columns in Table 3, labeled (1) and (3), show rejection frequencies of the null hypothesis for further instrument strength scenarios. Here, $\mathrm{Var}(\gamma_z z + \gamma_x x)$ is reset to unity as in Table 2, and only the fraction of it due to $\mathrm{Var}(\gamma_z z)$ is modified to 0.25 (1) and 0.75 (3), inducing a weaker and stronger instrument, respectively. The results show that only GRI-TSA and RI-TSA reach appropriate size in the weak instrument case. In the scenario with the strong instrument, results are very similar to Table 2, with FIML capitalizing on its efficiency, followed by a more equalized middle field including H1, TSM PPML, and their approximations GRI and RI. TSM NLS and H2 display markedly lower power, and H3 again falls prey to its strong underrejection problem.

Monte Carlo simulation studies always raise questions concerning the specificity of their results. To check that the presented results are not due to the particular choice of DGP, some sensitivity analysis has been conducted. First, orthogonal regressors are far from realistic in the social sciences. A further worry is the marginal distribution of the endogenous dummy, as in practice outcomes with 1 and 0 are often not balanced. Also, one may wonder if the tests are sensitive to a reduction of the effect of the dummy on the count variable. Finally, TSM and GRI are based on the null hypothesis $\theta = 0$, with $\theta = \sigma\rho$. Their positive performance could partly be due to the fact that in the shown DGP $\sigma = 1$ and so $\theta = \rho$. To address these concerns, separate simulations were carried out with $\mathrm{Corr}(x, z) = 0.5$, $\mathrm{E}(d \mid x, z) = 0.2$, $\beta_d = 0.1$ and $\sigma = \sqrt{2}$ (not reported). As it turns out, most results are by and large invariant to these alternatives. The exceptions are H1 and RI's reduced power when the regressors are correlated, as well as H1's when β_d is small. This latter finding is not surprising given that H1 is based on the contrast of estimates of β_d.

4.3. *Identification by Functional Form*

Having observed the performance of FIML, TSM PPML, and GRI-TSA under reduced impact of the instrument (cf. Fig. 1), a natural question is whether identification can be achieved by functional form alone, prescind from any instrument z. To this end, the DGP is specified as before, but setting $\gamma_z = 0$ and maintaining $\gamma_x = \sqrt{0.5}$. Results are shown in Table 4 in columns labeled (1) for sample sizes of 500 and 2,000 observations. The results prove to be rather discouraging, as both FIML and TSM PPML display empirical sizes that render the tests useless.[13] GRI-TSA's overrejection is not as pronounced, but the test

[13]Monfardini and Radice (2008) investigated exogeneity testing with no instruments in the bivariate probit model, which is related to the model under consideration through the bivariate normality assumption. The present results are in line with theirs, as they report high overrejection rates for Wald tests. They find likelihood ratio tests to have appropriate empirical size.

Table 4
Rejection frequencies of tests for exogeneity—identification by functional form

	(1)			(2)		
	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$
			$N = 500$			
FIML	0.1565	0.1750	0.2640	0.1375	0.1775	0.4155
TSM PPML	0.1783	0.1960	0.2473	0.1179	0.1181	0.2585
GRI-TSA	0.0729	0.0677	0.0860	0.0812	0.0838	0.2001
			$N = 2000$			
FIML	0.2340	0.2950	0.5700	0.1630	0.3490	0.8640
TSM PPML	0.1643	0.1700	0.2990	0.0780	0.1438	0.6538
GRI-TSA	0.0772	0.0714	0.1327	0.0610	0.1123	0.5546

Notes: Number of replications = 10,000 (FIML: 2,000 replications for $N = 500$, 1,000 replications for N = 2,000). Nominal test size = 0.05. IV-strength of columns (1) and (2) as detailed in text or Table 1.

lacks power in this setup. The exercise is repeated in columns (2) by strongly increasing the variance explained by the systematic part. To this end, γ_x is set to $\sqrt{2}$. However, little change is evident in the results for sample size 500. In the entries corresponding to the larger sample, on the other hand, some improvement is noticeable for TSM PPML and GRI-TSA, the latter's overrejection being only mild and showing increased power. Having empirical applications in mind, nevertheless, it seems that results from columns (1) represent a more realistic setting regarding instrument strength, so that the presence of an instrument in the DGP seems to be necessary for testing in finite samples.

4.4. *Results under Distributional Misspecification*

When specifying a parametric model, a natural concern relates to the robustness to distributional misspecification. In the context of count data, for instance, the overdispersion precluded from a Poisson distribution has been a major preoccupation which has led a portion of the empirical work to opt for the negative binomial regression model. Although under exogeneity the pseudo maximum likelihood properties of the Poisson model warrant consistency of the estimator, in the model with endogenous binary variable presented here, FIML, TSM and GRI are inconsistent if ε and v are not normally distributed. Moreover, in general, Terza's (1998) FIML estimator yields inconsistent estimates whenever $f(y \mid x, d, \varepsilon)$ does not follow a Poisson distribution. However, Romeu and Vera-Hernandez (2005) showed that in the case of the conditional distribution being Negative Binomial Type I (NegBinI), the FIML estimator remains consistent, suggesting that so does the FIML test.[14] The first two columns in Table 5 illustrates the performance of selected tests under the baseline DGP from Table 2 but with the modification

[14]Corollary 1 in Romeu and Vera-Hernandez (2005) established consistency of $(\hat{\beta}, \hat{\beta}_d)$ excluding the constant element, which is shifted. The estimate $\hat{\rho}$ is inconsistent for ρ but equals 0 whenever ρ does, securing consistency of the exogeneity test.

Table 5
Rejection frequencies of tests for exogeneity—sensitivity to distributional assumptions

	NegBinI			Gaussian copula		Frank copula	
	$\rho = 0$	$\rho = 0.5$	$\theta = 0$	$\theta^{GC} = 0.2$	$\theta^{GC} = 0.5$	$\theta^{FC} = 1$	$\theta^{FC} = 10$
H1	0.0382	0.3321	0.0415	0.0647	0.2930	0.0856	0.5833
FIML	0.1035	0.5970	0.0470	0.1445	0.5870	0.0820	0.7245
TSM PPML	0.0608	0.3747	0.0596	0.1546	0.5859	0.1008	0.9197
GRI-TSA	0.0451	0.6243	0.0400	0.1186	0.5359	0.0817	0.8317
RI-TSA	0.0582	0.5248	0.0582	0.1139	0.4404	0.0917	0.7064

Notes: Number of replications = 10,000 (FIML: 2,000 replications). Nominal test size = 0.05. Sample size = 500. IV-strength as detailed in text or Table 1.

$y \mid x, d, \varepsilon \sim$ NegBinI with expectation λ as before, and variance 2λ. Only GRI-TSA displays correct size. FIML overrejects quite severely, while TSM PPML does less so, but has noticeably less power than in the baseline case. H1 underrejects and ranks as the least powerfull among the compared tests.

To assess sensitivity of test size to the crucial assumption of bivariate normality, a DGP is implemented where the errors (ε, v) are independent and follow marginal distributions different from the normal. The chosen distributions are the exp-Gamma(1,1) for ε, which combined with a Poisson distribution for y conditional on observables and ε, yields a NegBinI distribution for y conditional on observables only; and a logistic distribution for v, scaled as to have unit variance, which gives a logit model for d. It might be argued that these modifications represent rather moderate departures from the distributional assumptions. However, there are at least two reasons for considering such a scenario. First, as mentioned before, there is a large body of empirical literature that uses NegBin and logit models, which consequently must imply either that there exists a large number of real-world problems where assuming negative binomial and logit processes is sensible, or that said literature's distributional assumptions are wrong. The former reason might find wider approval. Second, if the researcher has a strong belief in some form of significant departure from normality of the errors which goes beyond exp-Gamma or logit, she might as well opt to model this explicitly. Further, one might be interested in the performance of the tests under mild misspecification, since tests that do not conform to one's expectations even under these circumstances might as well be regarded as useless in view of the inherent uncertainty faced with respect to the 'true' data generating process. In other words, rather than her assumptions coinciding exactly with reality, all the applied econometrician might hope is that her assumptions approximate the underlying data generating process reasonably well.

Setting these concerns apart and considering the results of this analysis as shown in the third column in Table 5, the tests do present some minor size distortions, with H1 and GRI-TSA underrejecting, and TSM PPML and RI-TSA overrejecting H_0. FIML's overrejection is more substantial. In order to analyze empirical power of the tests under non normal marginals, dependence between the errors is induced by random sampling from copula functions. Columns 4 and 5 in Table 5 show rejection frequencies of the null hypothesis of exogeneity when

the errors' joint distribution is generated from a bivariate Gaussian copula with exp-Gamma and logistic marginals, with dependence parameter θ^{GC} equal to 0.2 and 0.5, respectively. Note that θ^{GC}, although having the same domain, is not a correlation coefficient as in the bivariate normal distribution, and thus comparisons to other tables are not valid. However, both columns reproduce the familiar pattern of the more parametric tests outperforming the supposedly more robust ones. Also, RI-TSA, which displayed power comparable to H1, clearly surpasses H1 in this setting. The last two columns in Table 5 contain results obtained by letting the joint distribution of the errors be determined by a Frank copula with the same non normal marginals as before. The Frank copula induces positive dependence between the variables through the parameter $\theta^{FC} \in (0, \infty)$, with independence resulting as a special case when $\theta^{FC} = 0$. The parameter is set to 1 in the sixth column and to 10 in the seventh column in Table 5. While for the weaker dependence power between the tests is rather similar, differences are considerably more pronounced for the case of stronger dependence. The ranking of the tests is almost the same as with the Gaussian copula, except for FIML falling back to third place. On the whole, these results seem to indicate that the tests relying on the bivariate normality assumption might perform equally well in non normal settings as the other tests. Furthermore, GRI-TSA's actual Type I error seems never to be larger than the level determined by the nominal size.

4.5. *Exogeneity Tests as Pretests: A Cautionary Note*

By far the most common use of tests for exogeneity is probably as pretests in order to choose between estimates. If a test rejects exogeneity, then estimates are obtained from an estimator that is consistent under endogeneity; while if the tests fails to reject the exogeneity hypothesis, estimates can be calculated from an estimator that is efficient under exogeneity, although inconsistent if the true DGP entails endogeneity. Thus, inference about a parameter of interest is conditional on the outcome of the exogeneity pretest.

The pretests or first stage tests to be considered are the exogeneity tests discussed so far, H1, FIML, TSM PPML, GRI-TSA, and RI-TSA. If the pretest fails to reject the null hypothesis, the model is estimated by Poisson MLE and a (second-stage) two-tailed t-test with null hypothesis H_0:$\beta_d = 1$ is conducted. Given rejection of exogeneity in the first stage test, the second-stage test of H_0:$\beta_d = 1$ is performed with NLIV estimates if the pretest was either H1 or RI-TSA. For TSM PPML and GRI-TSA pretests, second-stage tests are calculated with TSM PPML estimates, while FIML pretests use FIML estimates in the second stage.[15] In the DGP, the true β_d is left at 1 throughout all simulations, so that empirical rejection frequencies measure the finite sample size of the second-stage test.

Inspection of the results displayed in Table 6 suggests that the use of pretests for exogeneity leads to severe size distortions unless $\rho = 0$. Moreover, the overrejection is increasing over the range of ρ shown in the table, except for FIML. The reason for this is that for weaker levels of correlation, the weak power of the pretests leads to second-stage tests being performed with Poisson ML estimates whose bias for low ρ is sufficiently small as to not always reject H_0. Loosely speaking, as ρ increases, the

[15]Second-stage tests do not use RI-TSA and GRI-TSA estimates as these are inconsistent unless $\rho = 0$.

Table 6
Empirical size of second-stage tests of $\beta_d = 1$ using pretests for exogeneity

	(1)			(2)		
	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$
H1	0.0383	0.3596	0.5507	0.0409	0.3148	0.3960
FIML	0.0540	0.3565	0.3365	0.0520	0.3055	0.2270
TSM PPML	0.0516	0.3681	0.5327	0.0495	0.3326	0.4082
GRI-TSA	0.0493	0.3584	0.4981	0.0471	0.3219	0.3877
RI-TSA	0.0366	0.3578	0.6069	0.0382	0.3187	0.4767

Notes: Number of replications = 10,000 (FIML: 2,000 replications). Nominal test size = 0.05. Sample size = 500. IV-strength of columns (1) and (2) as detailed in text or Table 1.

bias in β_d increases faster than the power of the pretests, leading to higher rejection frequencies for all tests. Eventually, all second-stage tests' overrejection lowers, but except for FIML the turning point is after $\rho = 0.5$.

It is clear from the estimated rejection frequencies which are nowhere near the nominal size, that inference on structural parameters after pretesting in this model is likely to lead to false results and should thus be avoided. It should be stressed, however, that the pernicious effect of pretesting is due to interpreting the failure to reject exogeneity as that the variable in question is exogenous (*absence* of endogeneity). Obviously, exogeneity tests can be used to provide empirical evidence of the *presence* of endogeneity. This can be important in its own right, as for putting theories to test, and it can also provide ex-post empirical confirmation for a-priori concerns about potential endogeneity.

5. Conclusions

In this article, some tests for exogeneity of a binary variable in count data regression models, including the new GRI test, were examined for their finite sample properties through Monte Carlo simulations. The behavior of the tests under correct distributional specification was analyzed subjecting them to different sample sizes and levels of instrument strength. Test performances under data generating processes with no instrumental variables were reported, as well as under distributional misspecification. Finally, the use of these tests as pretests was assessed. Based on the results of the Monte Carlo experiments, a number of conclusions can be drawn which might provide some guidance for empirical practice.

The Hausman test which contrasts Poisson ML and NLIV estimates (H1) performs better than the other more refined versions based on Poisson-log-normal estimates (H2) or on estimation of the covariance between estimates (H3). Tests based on residual inclusion (RI) represent a very easy to implement alternative to H1, which in most scenarios display power comparable to H1, while outperforming Hausman contrast tests with respect to empirical size.

The other more parametric Wald tests which are based on the bivariate normality assumption generally present higher power than the Hausman tests, even in settings where they misspecify the DGP. The FIML test generally achieves the highest power of the tests. The more robust approximation to FIML, TSM, works

well when it is implemented through PPML instead of NLS, achieving power almost as high as FIML. The first order approximation to FIML, generalized residual inclusion (GRI), exhibits slightly lower power than TSM PPML, but still performs favorably compared to H1.

On the whole, therefore, these results suggest that using the simpler RI and GRI tests comes at virtually no cost in terms of test performance. Using two-stage adjusted standard errors noticeably improves the empirical size of the tests in smaller samples. Moreover, these tests show the best performances of all tests in the smallest samples and under the weakest instrument strength levels that were used in the simulations.

Two caveats have to be considered when testing for exogeneity. The first relates to the absence of exclusion restrictions in the DGP. Only with large samples and a very strong instrument does GRI-TSA come close to the nominal test size, the other tests perform worse. This suggests that there is little hope to test for endogeneity in practice if the structural model does not include any instruments.

The second issue concerns the use of these tests as pretests. In line with Guggenberger's (2008) finding of severe size distortions conditional on Hausman pretests in the classical linear model, large overrection rates render pretesting futile in the present count data model. The higher power of the Wald pretests clearly is not enough to result in acceptable second stage sizes. Therefore, practitioners are well advised to avoid using these tests as pretests. However, given that theoretical concerns about endogeneity have led a researcher to implement an estimation procedure that accounts for this, endogeneity tests can be used to obtain ex-post empirical evidence of these concerns having been justified.

Acknowledgments

The author wishes to thank João M.C. Santos Silva and an anonymous referee for helpful comments on this article. Special thanks to Rainer Winkelmann for extensive discussion and advise which significantly improved this article. Any remaining errors are the author's sole responsability.

References

Angrist, J. (2001). Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice. *Journal of Business and Economic Statistics* 19(1):2–16.

Chesher, A. (1991). The effect of measurement error. *Biometrika* 78(3):451–462.

Chmelarova, V. (2007). The Hausman Test, and Some Alternatives, with Heteroscedastic Data. Dissertation, Department of Economics, Louisiana State University.

Creel, M. (2004). Modified Hausman tests for inefficient estimators. *Applied Economics* 36:2373–2376.

Dagenais, M. G. (1999). Inconsistency of a proposed nonlinear instrumental variables estimator for probit and logit models with endogenous regressors. *Economics Letters* 63(1):19–21.

Degeorge, F., Derrien, F., Womack, K. L. (2007). Analyst hype in IPOs: explaining the popularity of bookbuilding. *Review of Financial Studies* 20(4):1021–1058.

Egger, P., Larch, M., Staub, K. E., Winkelmann, R. (2009). The trade effects of endogenous preferential trade agreements. Manuscript.

Gourieroux, C. S., Visser, M. (1997). A count data model with unobserved heterogeneity. *Journal of Econometrics* 79(2):247–268.

Greene, W. H. (1995), *Sample Selection in the Poisson Regression Model*. Working Paper, Department of Economics, Stern School of Business, New York University.

Greene, W. H. (1998). Sample selection in credit-scoring models. *Japan and the World Economy* 10(3):299–316.

Grogger, J. (1990). A simple test for exogeneity in probit, logit and Poisson regression models. *Economics Letters* 33(4):329–332.

Guggenberger, P. (2008). *The Impact of a Hausman Pretest on the Size of Hypothesis Tests*. Working Paper (first version: 2006), Department of Economics, University of California, Los Angeles.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica* 46:1251–1271.

Kenkel, D. S., Terza, J. V. (2001). The effect of physician advice on alcohol consumption: count regression with an endogenous treatment effect. *Journal of Applied Econometrics* 16:165–184.

Kozumi, H. (2002). A bayesian analysis of endogenous switching models for count data. *Journal of the Japanese Statistical Society* 32(3):141–154.

Masuhara, H. (2008). Semi-nonparametric count data estimation with an endogenous binary variable. *Economics Bulletin* 42(3):1–13.

Miranda, A. (2004). FIML estimation of an endogenous switching model for count data. *Stata Journal* 4(1):40–49.

Monfardini, C., Radice, R. (2008). Testing exogeneity in the bivariate Probit model: a Monte Carlo study. *Oxford Bulletin of Economics and Statistics* 70(2):271–282.

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

Nichols, A. (2007). *IVPOIS: Stata Module to Estimate an Instrumental Variables Poisson Regression via GMM*. Available online at http://ideas.repec.org/c/boc/bocode/s456890.html

Orme, C. D. (2001). Two-step inference in dynamic non linear panel data models. School of Economic Studies, University of Manchester, Manuscript.

Parrado, E. A., Flippen, C. A., McQuiston, C. (2005). Migration and relationship power among mexican women. *Demography* 42:347–372.

Quintana-Garcia, C., Benavides-Velasco, C. A. (2008). Innovative competence, exploration and exploitation: the influence of technological diversification. *Research Policy* 37(3):492–507.

Romeu, A., Vera-Hernandez, M. (2005). Counts with an endogenous binary regressor: a series expansion approach. *Econometrics Journal* 8:1–22.

Santos Silva, J. M. C., Tenreyro, S. (2006). The log of gravity. *Review of Economics and Statistics* 88(4):641–658.

Stock, J. H., Wright, J. H., Yogo, M. (2002). A survey of weak instruments and weak identification in generalized methods of moments. *Journal of Business and Economics Statistics* 20(4):518–529.

Terza, J. V. (1998). Estimating count data models with endogenous switching: sample selection and endogenous treatment effects. *Journal of Econometrics* 84(1):129–154.

Terza, J. V. (2006). Estimation of policy effects using parametric nonlinear models: a contextual critique of the generalized method of moments. *Health Services and Outcomes Research Methodology* 6(3–4):177–198.

Terza, J. V., Basu, A., Rathouz, P. J. (2008). Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics* 27:531–543.

Weesie, J. (1999). Seemingly unrelated estimation and the cluster-adjusted sandwich estimator. *Stata Technical Bulletin* 52:34–47.

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

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

Wooldridge, J. M. (1997). Quasi-likelihood methods for count data. In: Pesaran, M. H., Schmidt, P., eds. *Handbook of Applied Econometrics, Volume II: Microeconomics*. Massachusets, USA/Oxford, UK: Blackwell Publishers, pp. 352–406.