de Luna, Xavier; Johansson, Per

Working Paper

Testing exogeneity in cross-section regression by sorting data

Working Paper, No. 2000:2

Provided in Cooperation with:
IFAU - Institute for Evaluation of Labour Market and Education Policy, Uppsala

Suggested Citation: de Luna, Xavier; Johansson, Per (2000) : Testing exogeneity in cross-section regression by sorting data, Working Paper, No. 2000:2, Institute for Labour Market Policy Evaluation (IFAU), Uppsala

This Version is available at:
https://hdl.handle.net/10419/82223

Testing Exogeneity in Cross-section Regression by Sorting Data*

Xavier de Luna† and Per Johansson‡

18th April 2000

Abstract

We introduce a framework to test for the exogeneity of a variable in a regression based on cross-sectional data. By sorting data with respect to a function (sorting score) of known exogenous variables it is possible to utilize a battery of tools originally developed to detecting model misspecification in a time series context. Thus, we are able to propose graphical tools for the identification of endogeneity, as well as formal tests, including a simple-to-use Chow test, needing a minimum of assumptions on the alternative endogeneity hypothesis. Models of endogenous treatment and selectivity are utilized to illustrate the methods. With Monte Carlo experiments, including continuous and discrete response cases, we compare small sample performances with existing tests for exogeneity.

Keywords: Chow test; Endogenous treatment; Propensity score; Recursive residuals; Sample selection; Sorting score.

*Kurt Brännäs is gratefully acknowledged for his careful reading and comments on a previous version of this paper. We also thank Joseph Terza for kindly sharing with us his Gauss program to fit count data models. The first author has been financially supported by the Wikström Foundation.

†Department of Economics, Ume University, 901 87 Umeå, Sweden. E-mail: xavier@econ.umu.se.

‡IFAU - Office of Labour Market Policy Evaluation, 751 20 Uppsala, Sweden. Email: Per.Johansson@ifau.uu.se.

1 Introduction

In this paper we introduce a general framework to test the exogeneity of a variable in a cross-section regression context. The exogeneity of variables is essential for the interpretation of the parameters of interest, which can then be called structural (or causal). Exogeneity also reduces the parameter space and, as a consequence, the computational burden. The statistical definition of the exogeneity of a variable basically ensures that efficient inference can be performed on the parameters of interest while ignoring the marginal distribution of the exogenous variable (see Ericsson and Irons, 1994, for a survey). Exogeneity is essentially a model assumption and as such needs to be assessed. Tests for exogeneity are often easily constructed (e.g., Lagrange Multiplier tests, cf. Engle, 1984), under parametric models of endogeneity or sample selection.[1] An alternative test for exogeneity that does not, generally, need the distribution under the alternative hypothesis (endogeneity) is the Durbin, Wu and Hausman (DWH) test (cf. Hausman, 1978). This test needs, however, a consistent estimator under the alternative and another estimator that is efficient under the null hypothesis of exogeneity and inconsistent under the alternative. In small samples there is evidence of lack of power for this test (cf. Brännäs and Eriksson, 1997). Furthermore, in some situations it can be difficult to obtain the efficient estimator −consider, e.g., the case of heteroskedasticity under the null hypothesis,− as well as a consistent estimator under the alternative.

In order to avoid, a) a full parametric specification of the alternative endogeneity hypothesis, and b) the need to have two estimators, we propose a new approach based on the sorting score −most often a function of instruments,− whose function is to provide an ordering of the data which highlights the miss-specification arisen from the lack of exogeneity of a variable. The sorting score is related to the propensity score used to correct for selectivity bias (Rosenbaum and Rubin, 1983).[2] Moreover, it is common to correct for endogeneity and selectivity in econometric models by introducing a control function in the outcome equation (e.g. Heckman, 1978). When this is justified, the same control function constitutes a good sorting score. Once the observations are sorted, time series tools

[1] For a survey of sample selection models, see Vella (1998) and Winkelman (1998).

[2] The idea of constructing tests by sorting cross-sectional data is not new. For instance the Goldfeld and Quandt (1965) test of homoskedasticity is one example.

for miss-specification identification can be utilized. Thus, we propose, for instance, a graphical analysis of recursive residuals (standardized one-step ahead prediction errors) and a very simple-to-use Chow test. The advantage over the time series setting is that the observations are independent, thereby enhancing the constructions of tests. A drawback is that the ordering do not need to be a good one, in which case the corresponding tests may have low power.

The structure of the paper is the following. In Section 2 we define exogeneity using the conditional independence notation, and then continue by defining the sorting score. Section 3 presents different tools which, when associated to an ordering of the data, are helpful for diagnosing endogeneity, either by graphical analysis or with formal tests. In Section 4 specific models relating to sample selection and endogenous treatments are analyzed; sorting scores are displayed and examples of analysis are given with simulated data. Section 5 extends discusses non-parametric sorting scores. A Monte Carlo experiment, relating to the models in Section 4, is conducted in Section 6, and Section 7 concludes.

2 Exogeneity and the sorting score

2.1 General setting

We consider an observational study where independent observations are available for a response y together with a set of exogenous variables $\mathbf{x}$ and a possibly endogenous variable z (the effect on y of z is the issue of interest, hence in the sequel z is termed the treatment). A statistical model for these random variables $(y, \mathbf{x}, z) \in \Omega_y \times \Omega_x \times \Omega_z = \Omega$ is described by the conditional density $p(y, z|\mathbf{x}; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^t$. In the sequel, densities are denoted with p and distribution functions with P. The marginal density of $\mathbf{x}$, $p(\mathbf{x}; \boldsymbol{\delta})$, $\boldsymbol{\delta} \in D \subseteq \mathbb{R}^d$, is not needed for inference, because $\mathbf{x}$ is assumed exogenous, i.e. $p(y, z|\mathbf{x}; \boldsymbol{\theta})$ is not a function of $\boldsymbol{\delta}$ and $p(\mathbf{x}; \boldsymbol{\delta})$ is not a function of $\boldsymbol{\theta}$, or in other words $(y, z) \perp\!\!\!\perp \boldsymbol{\delta}|\mathbf{x}$ and $\mathbf{x} \perp\!\!\!\perp \boldsymbol{\theta}$, using the Dawid (1979) notation.[3] Similarly if the treatment z is exogenous, the study of its effect may be based on the reduced form $p(y|z, \mathbf{x}; \boldsymbol{\beta})$, $\boldsymbol{\beta} \in B \subseteq \mathbb{R}^p$, where $p(y, z|\mathbf{x}; \boldsymbol{\theta}) = p(y|z, \mathbf{x}; \boldsymbol{\beta})p(z|\mathbf{x}; \boldsymbol{\alpha})$, $\boldsymbol{\alpha} \in A \subseteq \mathbb{R}^q$ and $\Theta = B \times A$,

[3] The sign $\perp$ reads *independent of*. Here (y, z) is independent of δ given $\mathbf{x}$, and means that $P(y, z|\mathbf{x})$ is not a function of δ. Hence this notation is valid even when δ is not a random variable.

because $y \perp\!\!\!\perp \boldsymbol{\alpha}|\mathbf{x}, z$ and $z \perp\!\!\!\perp \boldsymbol{\beta}|\mathbf{x}$. However, in a typical observational study the exogeneity of the treatment z need to be assessed. Most of the existing tests for exogeneity in cross-sectional situations specify $p(z|\mathbf{x};\boldsymbol{\alpha})$ and how the conditional distribution of y given $(\mathbf{x}, z)$ is a function of $\boldsymbol{\alpha}$. The methodology proposed herein avoids such a strong specification of the possible endogeneity of z.

2.2 Sorting score

The methodology developed in this paper are based on an ordering of the n independent observations, $(y_i, \mathbf{x}_i, z_i), i = 1, \ldots, n$. The ordering is defined by sorting the observations with respect to the values taken by a function $s(\mathbf{x}, z) \in \Omega_s \subseteq \mathbb{R}$, which is such that $y \perp\!\!\!\perp s|\mathbf{x}, z$ under H_0 :"*z is exogenous*". We call $s(\mathbf{x}, z)$ a *sorting score*.[4]

Under H_0, $P(y|z, \mathbf{x};\boldsymbol{\beta})$, is well specified with $\boldsymbol{\beta}$ constant, and in particular $P(y|z, \mathbf{x}, s(\mathbf{x}, z) < c;\boldsymbol{\beta}) = P(y|z, \mathbf{x};\boldsymbol{\beta})$, for all $c \in \Omega_s$, because $y \perp\!\!\!\perp s|\mathbf{x}, z$ by the definition of $s(\mathbf{x}, z)$. On the other hand, when the treatment z is not exogenous, then two types of consequence may be distinguished. Most commonly, the non-acknowledgment of the endogeneity of z is accompanied by a structural miss-specification of $P(y|z, \mathbf{x};\boldsymbol{\beta})$. This situation is illustrated with a Garen model in the next section. The second possible consequence, occurring, for instance, when all the variables involved are jointly normal, is that $P(y|z, \mathbf{x};\boldsymbol{\beta})$ under H_0 or its alternative differ only in the value taken by $\boldsymbol{\beta}$. An illustration of this case is given in the next section by considering the classic demand equation from a model of market equilibrium.

2.3 Examples

In order to illustrate the above discussion we present two widely used models.

Demand equation: Consider the classic demand equation from a model of market equilibrium:

$$q_i = \beta x_{1i} + \gamma p_i + \varepsilon_i,$$

[4] Note that any function $s(\mathbf{x}, z)$ of $\mathbf{x}$ and z *only* is a potential sorting score under this definition, since then y is trivially independent of s given $\mathbf{x}$ and z. The condition $y \perp s|\mathbf{x}, z$, under H_0, is indeed not sufficient for the sorting score to be a useful one. We will characterize later on what a good sorting score is.

where the price p is endogenous, such that

$$p_i = \alpha x_{2i} + \nu_i,$$

with ε_i and ν_i correlated. In our previous notation the response y_i is here q_i, the treatment z_i is here p_i and $\mathbf{x}_i = (x_{1i}, x_{2i})$. Assuming $E(\varepsilon_i|\nu_i)$ linear in ν_i (e.g., bivariate normality), we have

$$E(q_i|\mathbf{x}_i, p_i) = \beta x_{1i} + \gamma p_i + \lambda(p_i - \alpha x_{2i}), \tag{1}$$

where $\lambda = 0$ if and only if ε_i and ν_i are uncorrelated. Let us set $s(x_{2i}) = (p_i - \alpha x_{2i})$ (linear in x_{2i}), as sorting score. This choice will be justified in Section 4.2. We have $E(q_i|x_{1i}, p_i, s(x_{2i}) < c) = \beta x_{1i} + \widetilde{\gamma} p_i + \alpha_c$, where $\widetilde{\gamma} = \gamma + \lambda$ and $\alpha_c = E(-\lambda\alpha x_{2i}|x_{1i}, p_i, s(x_{2i}) < c)$. Therefore, sorting the data with respect to $s(x_{2i})$ will lead to a varying coefficient α_c. The consequence is that $E(q_i|x_{1i}, p_i, s(x_{2i}) < c)$ is a biased predictor of q_i for $s(x_{2i}) \geq c$. Finally, note that $x_{2i} = x_{1i}$ would lead to the non-identifiability of γ.

Garen model: We consider now the selectivity model given by Garen (1984, 1988):

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\beta} + z_i\delta + z_i u_i + \varepsilon_i, \\ z_i &= f(\mathbf{x}_i^*) + \nu_i, \end{aligned}$$

where $E(\varepsilon_i|\mathbf{x}_i^*, z_i, u_i) = 0$ and $\mathbf{x}_i^*$ contains all the variables in $\mathbf{x}_i$ and possibly others. We have for this model

$$E(y_i|\mathbf{x}_i^*, z_i) = \mathbf{x}_i'\boldsymbol{\beta} + z_i\delta + z_i E(u_i|\nu_i). \tag{2}$$

Assuming $E(u_i|\nu_i)$ linear in ν_i (e.g., bivariate normality), we have $E(u_i|\nu_i) = \lambda(z_i - f(\mathbf{x}_i^*))$. Exogeneity of z_i corresponds to $\lambda = 0$, i.e. uncorrelated u_i and ν_i variables. Here, heteroskedasticity is present even if z_i is exogenous:

$$V(y_i|\mathbf{x}_i^*, z_i) = z_i^2 V(u_i|\nu_i) + \sigma_1^2, \tag{3}$$

where σ_1^2 is the variance of ε_i. In this case, neglecting the endogeneity of z_i leads to a miss-specification of the conditional expectation, by assuming it linear while $z_i E(u_i|\nu_i)$ is non-linear in $\mathbf{x}_i$ and z_i. We will see later (Section 4.2) that $z_i(z_i - f(\mathbf{x}_i))$ is the natural sorting score for this model.

3 Methods

In this section we present methods often used in a time series setting, where a natural ordering is available. These can be adapted to the cross-section setting by using the sorting score to order the data. We start by defining recursive residuals and then present graphical diagnostics and formal tests for exogeneity.

3.1 Recursive residuals

In order to introduce recursive residuals formally we assume first that the response y is continuous. For a set of independent observations $(y_i, \mathbf{x}_i, z_i)$, $i = 1, \ldots, n$, generated by a model with corresponding density $p(y|z, \mathbf{x}; \boldsymbol{\beta})$, it is assumed that, for each $k = q, \ldots, n-1$, a consistent estimate $\widehat{\boldsymbol{\beta}}_k$ of $\boldsymbol{\beta}$, based on $(y_i, \mathbf{x}_i, z_i)$, $i = 1, \ldots, k$, is available. Recursive residuals are then obtained by predicting y_j with $E(y_j|z_j, \mathbf{x}_j; \widehat{\boldsymbol{\beta}}_{j-1})$, $j = q+1, \ldots n$. This prediction is an estimate, based on observations $(y_i, \mathbf{x}_i, z_i)$, $i = 1, \ldots, j-1$, of the optimal (mean squared error sense) predictor $E(y_j|z_j, \mathbf{x}_j; \boldsymbol{\beta})$. The recursive residuals are then standardized prediction errors:

$$w_j = \frac{y_j - E(y_j|z_j, \mathbf{x}_j; \widehat{\boldsymbol{\beta}}_{j-1})}{Var(y_j - E(y_j|z_j, \mathbf{x}_j; \widehat{\boldsymbol{\beta}}_{j-1})|z_j, \mathbf{x}_j)}, \qquad j = q+1, \ldots n.$$

Under the model and assuming the involved moments exist, these recursive residuals are, at least asymptotically, independent and identically distributed with mean zero and variance one. These properties would hold exactly were $\boldsymbol{\beta}$ to be known.

Example 3.1 *The linear Gaussian model, $y_i = \mathbf{x}_i'\boldsymbol{\gamma} + \varepsilon_i$ with ε_i independently and normally distributed with mean zero and variance σ^2, is an important particular case for which recursive residuals were originally studied, e.g., by Brown et al. (1975). For this model we have, for $j = q+1, \ldots n$,*

$$w_j = \frac{y_j - \mathbf{x}_j'\widehat{\boldsymbol{\gamma}}_{j-1}}{\sigma(1 + \mathbf{x}_j'(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{x}_j)^{1/2}},$$

where $\mathbf{X}_{j-1} = (\mathbf{x}_1', \ldots, \mathbf{x}_{j-1}')'$. Assuming that $\mathbf{X}_{j-1}'\mathbf{X}_{j-1}$ are invertible, w_j are homoskedastic, independent, and with standard normal distribution (Brown et al., 1975). No asymptotic argument is needed here.

Recursive residuals can be generalized by considering predictive distributions instead of pointwise predictions: define $r_j = \Phi^{-1}(u_j)$, with $u_j = P(y_j|z_j, \mathbf{x}_j; \widehat{\boldsymbol{\beta}}_{j-1})$, $j = q+1, \ldots n$, where $\Phi(\cdot)$ and $P(y|z, \mathbf{x}; \boldsymbol{\beta})$ are the distribution functions corresponding to the standard normal density, $\phi(\cdot)$, and $p(y|z, \mathbf{x}; \boldsymbol{\beta})$, respectively. Such generalized recursive residuals are prediction errors in the sense that $P(y_j|z_j, \mathbf{x}_j; \widehat{\boldsymbol{\beta}}_{j-1})$ is the estimated predictive distribution of y evaluated at y_j when only $(y_i, \mathbf{x}_i, z_i)$, $i = 1, \ldots, j-1$, have been observed. When $p(y|z, \mathbf{x}; \boldsymbol{\beta})$ is well specified, u_j's are asymptotically independent and identically distributed uniform on $(0, 1)$ (Dawid, 1984). Thus, r_j's are independent and with standard normal distribution. These properties hold exactly when $\boldsymbol{\beta}$ is known.

When the response is continuous, it is advisable to use w_j's instead of r_j's because the former are only sensitive to the modelling of the conditional mean and variance. In the sequel we have therefore used w_j's when dealing with continuous responses.

For a discrete valued response y, recursive residuals w_j's are recursive counterparts of ordinary Pearson residuals. Recursive Pearson residuals are highly non-normal if the response y takes only a very limited number of values, e.g., binary variable, with as consequence that much larger samples are needed to obtain reliable diagnostics. Generalized recursive residuals, r_j's, are still asymptotically independent although neither identically nor uniformly distributed. These properties can, however, be recovered by randomization as was proposed in Smith (1985): Generate ξ_j independently uniform on $(0, 1)$ and compute $\widetilde{u}_j = \xi_j r_j + (1-\xi_j) P(y_j - 1|z_j, \mathbf{x}_j; \widehat{\boldsymbol{\beta}}_{j-1})$, for $j = q+1, \ldots n$. The resulting randomized generalized recursive residuals $\widetilde{r}_j = \Phi^{-1}(\widetilde{u}_j)$ are asymptotically independent and with standard normal distribution.

Although recursive residuals are often associated to time series applications and more particularly to the detection of structural changes (e.g., Brown et al., 1975, Tsay, 1998), they have also been found useful in model validation of cross-section regression models (see Kianifard and Swallow, 1996, for a review concerning the linear Gaussian model). Their use for non-linear and/or non-Gaussian models has been discussed by Dawid (1984), Smith (1985), and Harvey (1989). Their advantage over ordinary least squares residuals is that they are homoskedastic and independent (at least asymptotically) under the model assumptions, facilitating the developments of test statistics. On the other hand, they are not uniquely defined for cross-section data, depending as they are on the

ordering of the data set.

3.2 Graphical diagnostics and formal tests

Two forms of diagnostic tools associated to the recursive residuals are available. The first is obtained by graphically displaying recursive residuals. Their cumulative sum (CUSUM) is most useful. At this stage a useful family of sorting score may be characterized. We have seen that recursive residuals have mean zero for a well-specified model. When miss-specification arises, in our case when exogeneity of the treatment does not hold, the recursive residuals will typically have non-zero mean. In this situation we say that $s(\mathbf{x}, z)$ is a *monotone sorting score* if throughout the recursion the residuals have all positive (respectively negative) mean. A monotone sorting score is not always trivial to obtain, since we would need knowledge on $p(y, z|\mathbf{x}, \boldsymbol{\theta})$. In the time series context, the aim when inspecting cumulative sums is to detect a change over time of the parameter values. Most often this change is believed to be an abrupt structural change at a given time point. The endogeneity miss-specification is translated instead by small but systematic biases in predictions. Thus, a monotone sorting score is used in order for these biases to have the same sign. This guarantee the best visual effect when plotting the cumulative sum of the recursive residuals. These issues are clarified in the examples given in Section 4. The constancy of the bias sign is also important for the formal tests presented below to have power.

Three types of tests are proposed all using the sorted data set. Harvey and Collier (1979) proposed a simple test based on the sum of the recursive residuals to identify structural time changes. In our context, write

$$\overline{w} = \frac{1}{n-q} \sum_{i=q+1}^{n} w_i$$

the average of the recursive residuals. Then, under H_0, asymptotically (exactly under the normal model), $\overline{w}$ is normally distributed with mean zero and variance $1/(n-q)$. Thus, H_0 may be tested by testing the population mean of w_i to be zero. The result remains valid when w_i's are replaced by r_i's or $\widetilde{r}_i$'s. Recursive residuals may be computationally costly to obtain and/or very sensitive to modelling assumptions, mainly if the generalized versions are used. Both problems are avoided by using Chow's (1960) test.

To introduce Chow-type tests let us start by considering the following linear regression model

$$y_i = \mathbf{x}_i\boldsymbol{\delta} + z_i\gamma + \varepsilon_i,\ i = 1, ..., n, \tag{4}$$

and split the sample into n_1 and n_2 observations based on the sorting score. The Chow (1960) test is then given as

$$C^1 = \frac{[e_0'e_0 - e'e]/p}{e'e/(n-p)},$$

where e_0 is the $n \times 1$ vector of ordinary least square (OLS) residuals, p is the number of parameters in the equation (4) and $e'e = e_1'e_1 + e_2'e_2$, where e_1 and e_2 are the $n_1 \times 1$ and $n_2 \times 1$ vectors of OLS residuals from the two OLS estimates. If ε_i is $N(0, \sigma^2)$ the C^1 test is distributed as $F(p, n-p)$. Assume the more general model

$$E(y_i|\mathbf{x}_i, z_i) = g(\mathbf{x}_i'\boldsymbol{\delta} + z_i\gamma) = g(\mathbf{w}_i'\boldsymbol{\alpha}),\ i = 1, ..., n. \tag{5}$$

The parameters $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ for the two samples are estimated using a pseudo maximum likelihood (PML) estimator (cf. Gouriéroux, Montfort and Trognon, 1984). The covariance matrices of the PML estimator is given by

$$\mathbf{V}_j = \mathbf{J}_j^{-1}\mathbf{I}_j\mathbf{J}_j^{-1}, \quad j = 1, 2, \tag{6}$$

where

$$\mathbf{J}_j = E\left(-\frac{\partial^2\ell_j}{\partial\boldsymbol{\alpha}_j\partial\boldsymbol{\alpha}_j'}\right),\ \mathbf{I}_j = E\left(\frac{\partial\ell_j}{\partial\boldsymbol{\alpha}_j}\left(\frac{\partial\ell_j}{\partial\boldsymbol{\alpha}_j}\right)'\right), \quad j = 1, 2,$$

ℓ_j is the log-likelihood function from a member of the linear exponential family and expectations are taken over the distribution of y and the explanatory variables.[5] The covariance matrices can be estimated by inserting the estimated parameters $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\alpha}}_2$ into the corresponding empirical expressions i.e.

$$\widehat{\mathbf{J}}_j = \left(\sum_{i=1}^{n_j}\frac{\partial^2\ell_j}{\partial\boldsymbol{\alpha}_j\partial\boldsymbol{\alpha}_j'}\right)_{\boldsymbol{\alpha}_j=\widehat{\boldsymbol{\alpha}}_j} \text{ and } \widehat{\mathbf{I}}_j = \left(\sum_{i=1}^{n_j}\left(\frac{\partial\ell_j}{\partial\boldsymbol{\alpha}_j}\right)\left(\frac{\partial\ell_j}{\partial\boldsymbol{\alpha}_j'}\right)\right)_{\boldsymbol{\alpha}_j=\widehat{\boldsymbol{\alpha}}_j}, j = 1, 2.$$

[5] Note that the covariance matrix in (6) is the Eicker-White heteroskedasticity consistent covariance matrix estimator (cf. White, 1980). The robust covariance matrix can also be used in a quasi-LM test (cf. Engle, 1984).

The following Wald test

$$C^2 = (\widehat{\boldsymbol{\alpha}}_1 - \widehat{\boldsymbol{\alpha}}_2)'(\widehat{\mathbf{V}}_1 + \widehat{\mathbf{V}}_2)^{-1}(\widehat{\boldsymbol{\alpha}}_1 - \widehat{\boldsymbol{\alpha}}_2)$$

is asymptotically $\chi^2(p)$ distributed.[6] The advantage with this test is that it asymptotically yields the correct size under a general miss-specification of the probability distribution. The condition for unbiased estimates of $\widehat{\boldsymbol{\alpha}}_j$ is a correct specification of the conditional mean (Gouriéroux et al., 1984).

Finally, we consider the test of Nyblom (1989) and Hansen (1994).[7] Define the first order conditions for the k:th parameter as $f_{ki} = \partial \ell_i / \partial \alpha_k$ and let $\mathbf{f}_i = (f_{1i}, ..., f_{Ki})'$ then $\sum_{i=1}^{n} \mathbf{f}_i = \mathbf{0}$. Let

$$S_{kg} = \sum_{i=1}^{g} f_{ki},$$

and further define $\mathbf{S}_g = (S_{1g}, .., S_{Kg})'$ and $\mathbf{f}_g = (f_{1g}, ..., f_{Kg})'$ then a joint statistic for the stability of the parameters are

$$L_c = \frac{1}{n} \sum_{g=1}^{n} \mathbf{S}_g' \mathbf{V}^{-1} \mathbf{S}_g, \tag{7}$$

where $\mathbf{V} = \sum_g \mathbf{f}_g \mathbf{f}_g'$. This statistic has a non-standard distribution and asymptotic critical values are taken from Hansen (1994). Since $\mathbf{V}^{-1}$ is the Eicker-White heteroskedastic consistent covariance matrix we should expect this test to perform well under heteroskedasticity..

4 The sorting score for common models of endogeneity

The endogeneity of the treatment z_i may often be understood as an omitted variable problem. In such cases, the sorting score should provide an

[6] In a time series regression model with normally distributed errors, see Ohtani and Toyoda (1985) for a study of small sample performances.

[7] This test is for parameter constancy with the alternative specifying the parameter process as a martingale.

approximate ordering of the omitted variable in order to diagnose the miss-specification. We start by discussing the omitted variable issue, because it helps us latter to clarify the discussion on particular models of endogeneity.

4.1 Omitted variable

Assume that $y_i = z_i\gamma + u_i\xi + \varepsilon_i$ where u_i is unobserved (for simplicity we do not include extra exogenous variables, $\mathbf{x}_i$, in the model), and that $E(u_i|z_i)$ is linear in z_i. Then $E(y_i|z_i)$ is also linear in z_i and least squares estimation is consistent for $E(y_i|z_i) = z_i\widetilde{\gamma}$, where in general $\widetilde{\gamma} \neq \gamma$. Ordinary residuals (associated to least squares) are not able to diagnose the missing variable because the conditional expectation is well specified (linear function of z_i), unless the data can be sorted with respect to the ordering of u, in which case an index plot of the residuals may reveal the dependence, because $E(y_i - z_i\widetilde{\gamma}|z_i, u_i) \neq 0$. Similarly, the recursive residuals corresponding to the ordering of the variable u will uncover the missing variable because $E(y_i|z_i, u_i < c) \neq E(y_i|z_i)$. However, if $u_i = z_i\delta$, δ a constant, then ordinary and recursive residuals do not reveal any miss-specification because γ and $\delta\xi$ are not discernible (non-identifiable parameters).

When $E(u_i|z_i)$ is not linear in z_i, $E(y_i|z_i)$ is not either, and a linear regression model will be miss-specified. This should be apparent in the ordinary residuals plotted against z. However, recursive residuals with respect to a monotone sorting score, when one is available, are more appropriate. Indeed, the accumulation of the bias in the CUSUM of the residuals is then systematic in one direction, making visual identification of the problem easy. Moreover, the formal tests described in Section 3.2 can be used.

Of course, to use recursive residuals diagnostics we need an ordering (at least approximate) of the unobserved variable u. This is typically what the sorting score provides, when the confounder u is not observed, but some information on its distribution given observed exogenous variables is available. The following subsections illustrate this essential point.

4.2 Continuous response and treatment

Following the above discussion, the choice of sorting scores in the two examples in Section 2.3 are discussed and the potentials of a graphical analysis is illustrated.

Garen model: In this example, not taking into account the endogeneity of z_i corresponds to omitting the variable $z_i(z_i - f(\mathbf{x}_i))$ in (2), i.e. setting $\lambda = 0$. The omitted variable is here non-linear in $\mathbf{x}_i$ and z_i. This non-linearity of $E(y_i|\mathbf{x}_i, z_i)$ may often be hidden by the heteroskedastic noise when examining conventional residuals. On the other hand, recursive residuals are able to identify the systematic bias in predictions obtained with a monotone sorting score as is illustrated with the example below. The natural sorting score is here the omitted variable $s(\mathbf{x}_i) = z_i(z_i - f(\mathbf{x}_i))$. It ensures indeed monotonicity of the ordering.[8] Because f is unknown, an approximate sorting score must be used by estimating this function, yielding $z_i(z_i - \widehat{f}(\mathbf{x}_i))$. Notice that this framework allows us to proceed without specifying f but using instead a non-parametric estimate (see Section 5).

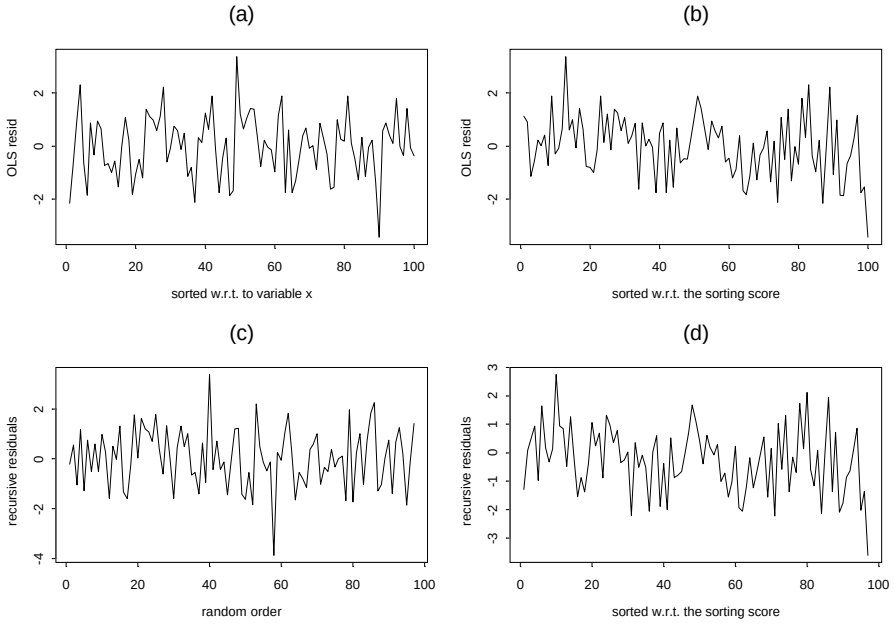
Example 4.1 *The Garen model is considered, and we simulate 100 observations with the following specifications: for* $i = 1, \ldots, 100$,

$$\begin{aligned} y_i &= 1 + 2x_{1i} + \gamma_i z_i + \varepsilon_i, \\ \gamma_i &= 1 + u_i, \\ z_i &= x_{1i} - x_{2i} + \nu_i, \end{aligned}$$

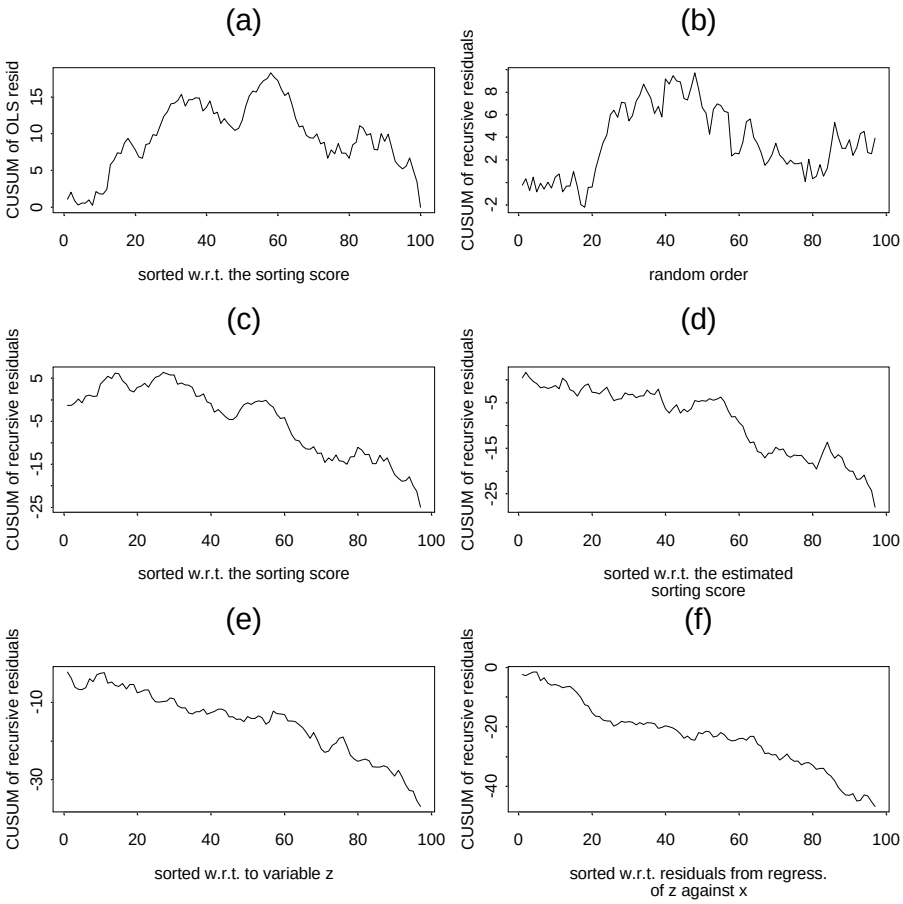
with $x_{1i} \sim U(0,1), x_{2i} \sim U(0,1), \varepsilon_i \sim N(0,1)$ *and* u_i *and* ν_i *bivariate normal with expectations zero, variances* 0.36 *and* 1 *respectively, and correlation* -0.5. *Assuming exogeneity* $E(y_i|x_{1i}, z_i) = x_{1i}\beta + z_i\gamma$ *is estimated with OLS. Several types of residual analyses are presented in Figures 1 and 2.*

From the residuals plots of Figure 1 there seems to be no severe heteroskedasticity. The miss-specification of the conditional mean is not straightforward to identify with these residual plots, although a trained eye may see some structure in the OLS residuals when sorted with respect to the omitted variable, graph (b), and in the recursive residuals obtained with this same sorting score, graph (d). The CUSUM plots in Figure 2 are more interesting. We note that the recursive residuals obtained with the monotone sorting score provides a clear sign of the miss-specification of the conditional mean of the model (endogenous treatment) by displaying

[8] Indeed, the non-linearity is either concave ($\lambda > 0$) or convex ($\lambda < 0$) in z_i (the omitted variable is independent of $\mathbf{x}_i$), thereby implying recursive residuals with, respectively, negative or positive mean.



Figure~1: Residuals from the Garen model of Example 2: (a) OLS residuals sorted w.r.t. the variable x; (b) OLS residuals obtained with the ordering of the omitted variable (optimal sorting score); (c) Recursive residuals obtained with a random ordering; (d) Recursive residuals obtained with the optimal sorting score.



Figure~2: CUSUM plots of various residuals from the Garen model of Example 2: (a) OLS residuals sorted w.r.t. the omitted variable (optimal sorting score) $-\mathrm{HC}(u_i \equiv 0) = 0.00$; (b) Recursive residuals obtained with a random ordering $-\mathrm{HC}(u_i \equiv 0) = 0.40$; (c) Recursive residuals obtained with the optimal sorting score $-\mathrm{HC}(u_i \equiv 0) = -2.53$; (d) ditto but with an estimate of the previous sorting score $-\mathrm{HC}(u_i \equiv 0) = -2.83$; (e) ditto but with the sorting score $s(x, z) = z$ $-\mathrm{HC}(u_i \equiv 0) = -3.75$; (f) ditto but with the OLS residuals from the regression of z on x as sorting score $-\mathrm{HC}(u_i \equiv 0) = -4.74$.

a systematic departure from zero of the CUSUM trajectory. This neat visual effect is due to the monotonicity of the sorting score. The values of the HC test (for $H_0 : u_i \equiv 0$) given in the caption of the figure confirm the visual impression. This sorting score cannot be used in practice because it is not observed, but graph (d) shows that the estimated sorting score gives a similar result. In this particular simulated example, simpler sorting scores also perform well, graph (e) and (f).

Demand equation: From (1) we note that $s(x_{2i}) = (p_i - \alpha x_{2i})$ plays the role of an omitted variable and is therefore the natural sorting score. It is, moreover, easily checked that this sorting score is monotone, because the predictor $E(q_i|x_{1i}, p_i, s(x_{2i}) < c)$ is always either positively or negatively biased when predicting q_i for $s(x_{2i}) \geq c$ (α_c is an increasing or decreasing value of c). The sorting score is unobserved, but a good approximation is $(p_i - \widehat{\alpha} x_{2i})$, where $\widehat{\alpha}$ is a consistent estimate.

Example 4.2 *We use data on U.S. consumption expenditures (c_t) disposable income (y_t) and government expenditure(g_t) for the years 1955-1986 in billions of 1982 dollars.*[9] *We assume:*

$$c_t = \gamma_0 + \gamma_1 y_t + \varepsilon_t,$$

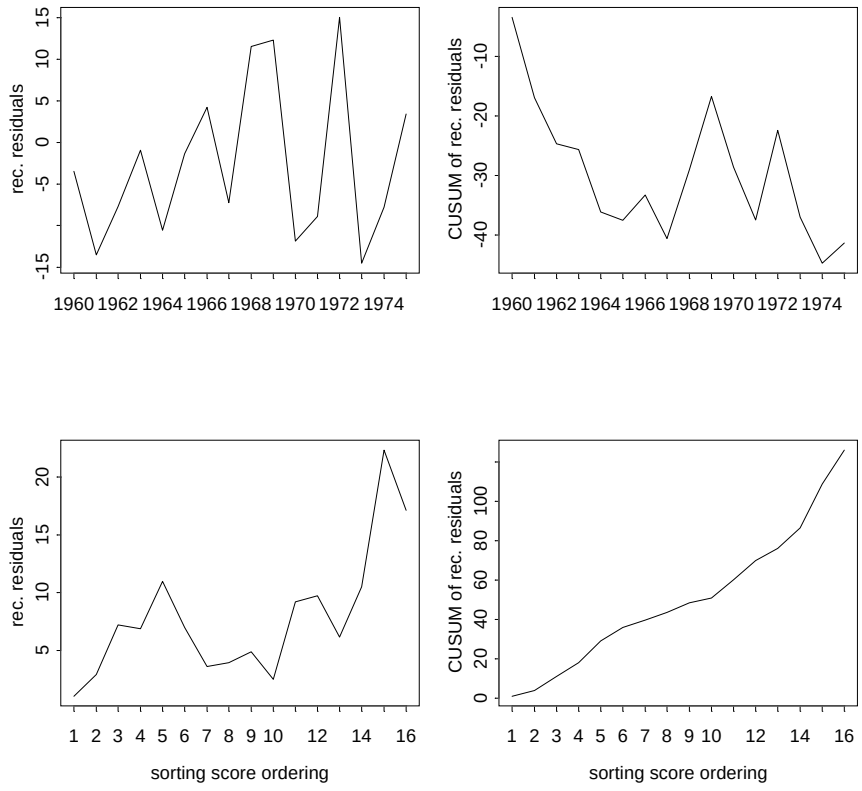
where y_t is endogenous and such that

$$y_t = \alpha_0 + \alpha_1 g_t + \nu_t. \tag{8}$$

Since there is a structural break at 1975 we use data for the period 1955 - 1975.[10] *Figure 3 shows how the endogeneity of y_t is revealed by using the residuals from 8 as sorting score.*

[9] This data is described in Hill, Griffiths and Judge (1997) and obtained from http:\\www.wiley.com.

[10] Using the data for the whole period the same pattern as in Figure 3 appears when the sorting score is used. Using the time ordering there is a visible structural break at 1975 since after that time the CUSUM of the recursive residuals starts to increase. However, the Harvey-Collier test statistics is significant ($HC = 6.57$) only when the sorting score is used, pointing out that the endogeneity problem is more serious than the structural break.



Figure˜3: Recursive residuals when regressing the consumption expenditures on the disposable income (y_t), using time ordering (above panels) $-$HC$= -1.09$; and the ordering using sorting score: residuals of regressing y_t on the government expenditure variable (below panels) $-$HC$= 5.67$.

4.3 Sample selection and binary treatment

The Heckman (1979) sample selection model is a classical specification of the selection problem in econometrics:

$$\begin{aligned} z_i^* &= \mathbf{x}_i^{*\prime}\boldsymbol{\alpha} + \varepsilon_{1i} \\ z_i &= I(z_i^* > 0) \end{aligned} \tag{9}$$

and

$$y_i = z_i(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_{2i}), \tag{10}$$

where z_i^* is an unobserved latent variable and sample selection arises when ε_{1i} and ε_{2i} are correlated. Assuming joint normality of ε_{1i} and ε_{2i}, and denoting $\sigma_1^2 = 1$, σ_2^2, ρ, their respective variances and correlation, we have

$$E(y_i|\mathbf{x}_i^*, z_i = 1) = \mathbf{x}_i'\boldsymbol{\beta} + \rho\sigma_2^{-1}\lambda_i \tag{11}$$

where $\lambda_i = \phi(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha})/(1 - \Phi(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}))$ can be thought of as a missing variable. Although λ_i is a non-linear function of $\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}$, it may be close enough to linearity to make the identification of the miss-specification difficult. Hence it is customary to assume that, at least, one variable in $\mathbf{x}_i^*$ is not included in $\mathbf{x}_i$ (see Section 4.1 and the discussion in Vella (1998, p. 135)).

Here the missing variable is not observed but can be estimated by evaluating λ_i at a consistent estimator $\widehat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$, yielding $\widehat{\lambda}_i$. The ordering of λ_i and of $\Phi(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}) = \Pr(z_i = 1|\mathbf{x}_i)$, the propensity score, are equivalent. Thus, sorting with respect to $\widehat{\lambda}_i$ is equivalent to sorting with respect to the estimated propensity score. Further, because λ_i enters linearly in (11), it is a monotone score.

The standard endogenous treatment model (cf. Heckman, 1978) is such that the choice is described by (9) and the outcome equation is:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + z_i\delta + \varepsilon_{2i}. \tag{12}$$

If ε_{1i} and ε_{2i} are bivariate normal and correlated we have that $E(\varepsilon_{2i}|\mathbf{x}_i, z_i)$ $= \rho\sigma_2^{-1}(\lambda_i z_i - \widetilde{\lambda}_i(1 - z_i))$ where $\widetilde{\lambda}_i = \phi(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha})/\Phi(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha})$. As above the missing variable can be estimated by replacing $\boldsymbol{\alpha}$ by a consistent estimator.

4.4 Discrete response and binary treatment

All the examples considered above have in common that the response is continuous. The framework proposed in this paper is, however, also applicable to discrete responses. As an example consider the case where the response variable y is a count, and $y_i|\mathbf{x}_i, z_i, u_i$ is Poisson distributed with expectation and variance

$$E(y_i|\mathbf{x}_i, z_i, u_i) = V(y_{it}|\mathbf{x}_i, z_i, u_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta} + z_i\delta + u_i),$$

and the treatment endogeneity is modelled by (9). If u_i and ε_i are independent z_i is exogenous, and we have a over-dispersed Poisson regression model. Estimation with an endogenous treatment has been discussed, e.g., by Windmeijer and Santos Silva (1997) and Terza (1998).

5 Non-parametric sorting score

All the examples presented so far have in common that they specify a parametric distribution for $z_i|\mathbf{x}_i$ as well as how the endogeneity arises. Such a full specification of the endogeneity alternative is often necessary to apply existing tests, the DWH test excepted. A non-parametric approach may be envisaged by fitting $E(z_i|\mathbf{x}_i) = f(\mathbf{x}_i)$ non-parametrically, and then making use of the corresponding residuals, or the propensity score if the treatment is binary valued, as a sorting score.

For the binary treatment case the propensity score provides the same ordering as the control function λ_i. It can also be noted (cf. Ameimiya, 1981) that the estimated parameters in a linear probability model (for z_i) is proportional to the parameters estimated with the probit or logit estimator. Hence, different distributional assumptions for the selection equations are likely to create approximately the same ordering and the sorting score based on estimates from a linear ("non-parametric") estimator will yield approximately the same ordering as when using probit or logit estimators. Consequently, the additive model for the propensity score, $E(z_i|\mathbf{x}_i) = \sum_j f_j(x_{ji})$, can be expected to perform well, as well as the generalization (Hastie and Tibshirani, 1990): $g(E(z_i|\mathbf{x}_i)) = \sum_j f_j(x_{ji})$.

6 Monte Carlo study

6.1 Experiments

In order to study the small sample performances of the exogeneity tests proposed in this paper, three types of experiments are conducted. In all three experiments we test for an exogenous treatment. In the first experiment we use the setup of Example 2 with a continuous response and continuous treatment. In the second experiment we have a continuous response and a discrete treatment and the third experiment is concerned with a discrete response (count) and a discrete treatment. In all tree experiments we study the power and size of the tests. We also give as reference alternative tests for exogeneity proposed earlier in the literature. In all three experiments we choose the sample size $n = 200, 400$ the number of replications N is 1000 and the exogenous variables are fixed within repeated samples. In all simulations we have two exogenous variables x_{1i} and x_{2i} which are uniform[0,1] variates. When calculating the recursive residuals q was set to 10. In the Chow tests C^1 and C^2 we split the sample in the to two equal parts, hence $n_1 = n_2 = 100$ and 200, respectively for $n = 200$ and 400.[11] The power of the test is calculated using the correlation $\rho = -0.75, -0.5, -0.25, 0.25, 0.5$ and 0.75. The sizes are given for $\rho = 0$. In both the Garen model (selection) and the count data model the model we estimate under H_0 is miss-specified and we therefore also calculate the size of the test under the assumption of correct marginal distribution under H_0.

The first experiment is the Example 2 in Section 4.2 hence

$$z_i = x_{1i} - x_{2i} + \nu_i, \tag{13}$$
$$y_i = 1 + 2x_{1i} + z_i + z_i u_i + \varepsilon_i. \tag{14}$$

Here, $\varepsilon_i \sim N(0,1)$ and u_i and v_i are bivariate normal with expectations zero and variances $\sigma_u^2 = 0.36$ and $\sigma_v^2 = 1$. We also perform the test where u_i, v_i and ε_i are all $\chi^2(1)$ (centered to have expectation zero). It is essential to note that heteroskedasticity is present even when $\rho = 0$. For that reason we also perform an experiment with $u_i \equiv 0$ for all i.

The second experiment is the conventional endogenous treatment model

[11] The small sample performance under normal errors and heteroskedasticity in a time series setting is studied in Ohtani and Toyoda (1985).

(See Section 4.3). Here we have the following setup.

$$\begin{aligned} z_i &= I((x_{1i} - x_{2i} + v_i) > 0), \\ y_i &= 1 + 2x_{1i} + z_i + u_i, \end{aligned} \tag{15}$$

where u_i and v_i are bivariate normal with expectations zero and variances $\sigma_u^2 = 0.36, \sigma_v^2 = 1$. We also perform the test where u_i and v_i are each $\chi^2(1)$ (centered to have expectation zero).

In the final experiment the response variable is a count and we generate the data such that $y_i|x_{1i}, z_i, u_i$ is Poisson distributed with expectation and variance

$$E(y_i|x_{1i}, z_i, u_i) = V(y_i|x_{1i}, z_i, u_i) = \exp(\ln(2)x_{1i} + z_i + u_i). \tag{16}$$

The same setup as in the endogenous treatment model above is used.[12] It should be observed that even when $\rho = 0$, the marginal distribution of the response is not Poisson. We thus perform an experiment under a Poisson marginal distribution, i.e. when $u_i \equiv 0$ for all i. When u_i is not constant the marginal distribution has no closed form (cf. Terza, 1998), although when $\rho = 0$ we know the first two unconditional expectations: $E(y_i|x_{1i}, z_i) = \mu_i = \exp(\beta_0 + \ln(2)x_i + z_i)$ and $V(y_i|x_{1i}, z_i) = \mu_i(1 + 0.36\mu_i)$, where $\beta_0 = -0.36/2$, such that $E(\exp(u_i)) = 1$.[13]

Garen suggests a two step method (TSM) to correct for selection: first estimate (13) with OLS and then estimate

$$y_i = \beta_0 + \beta_1 x_i + \gamma z_i + \tau(z_i \widehat{v}_i) + \varepsilon_i^*, \tag{17}$$

where $\widehat{v}_i$ is the residuals from the first step OLS estimator.[14] Thus, a $t-$test, using the heteroskedastic consistent covariance matrix (6), of $\tau = 0$ is a test for exogeneity of z. In our tests the sorting score $s(\mathbf{x}_i, z_i) = \widehat{v}_i$ and $s(\mathbf{x}_i, z_i) = \widehat{v}_i z_i$ were used.

In the two final experiments the sorting score is the predicted probabilities from a probit maximum likelihood estimator, i.e. $s(\mathbf{x}_i, z_i) =$

[12] The calculation of the recursive residuals are quite time consuming, that is why we only performed the Monte Carlo experiments under the (bivariate) normal error setup.

[13] If $u_i \sim N(\mu, \sigma^2)$ and $\mu = -\sigma^2/2$ then $E(\exp(u_i)) = \exp(\mu + \sigma^2/2) = 1$.

[14] Since Garen also assumes that the error terms in the two equations are correlated, he suggests the following model to be estimated $y_{it} = \alpha_0 + \alpha_1 x_i + \alpha_2 z_i + \alpha_3(z_i \widehat{v}_i) + \alpha_4 \widehat{v}_i + \varepsilon_i^*$.

$\Phi(\widehat{\alpha}_0 + \widehat{\alpha}_1 x_{1i} + \widehat{\alpha}_2 x_{2i})$.[15] In the second experiment with a dummy treatment model a standard test of selection is a t-test of $\tau = 0$ in

$$y_i = \beta_0 + \beta_1 x_i + \gamma z_i + \tau e_i + \varepsilon_i^*,$$

where $e_i = \widehat{\lambda}_i z_i + \widehat{\widehat{\lambda}}_i (1 - z_i)$ and $\widehat{\lambda}_i$ and $\widehat{\widehat{\lambda}}_i$ are obtained from an initial probit ML estimator. Since ε_i^* is heteroskedastic the covariance estimator (6) is used. For this model a DWH test is easy to perform since under H_0 the OLS estimator is efficient while under the TSM presented above is consistent under both hypothesis.

Terza (1998) suggests a TSM (similar to the Heckman correction in the Gaussian model) to correct for selection in the model (16): First estimate

$$y_i = h(\mathbf{x}_i, z_i, \theta, \widehat{\boldsymbol{\alpha}}) + \eta_i \tag{18}$$

with non-linear least squares (NLS). Here $h(\mathbf{x}_i, z_i, \theta, \boldsymbol{\alpha}) = \psi(\theta, \boldsymbol{\alpha}) \exp(\mathbf{x}_i \boldsymbol{\beta}^*)$, where $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1, \delta)$ and $\theta = \rho/\sigma_u$. The estimate $\widehat{\boldsymbol{\alpha}}$ is from an initial probit ML estimator, and

$$\psi(\theta, \widehat{\boldsymbol{\alpha}}) = z_i \left[\frac{\Phi(\theta + \mathbf{x}_i \widehat{\boldsymbol{\alpha}}_1)}{\Phi(\mathbf{x}_i \widehat{\boldsymbol{\alpha}}_1)} \right] + (1 - z_i) \left[\frac{1 - \Phi(\theta + \mathbf{x}_i \widehat{\boldsymbol{\alpha}}_1)}{1 - \Phi(\mathbf{x}_i \widehat{\boldsymbol{\alpha}}_1)} \right].$$

Let $\mathbf{b}_1 = (\boldsymbol{\beta}^{*\prime}, \theta)'$ then the NLS estimator $\widehat{\mathbf{b}}_1$ is asymptotically normal with expectation $\mathbf{b}_1$ and covariance matrix $\mathbf{D}$. A consistent estimator is

$$\widehat{\mathbf{D}} = \left(\mathbf{G}_1' \mathbf{G}_1\right)^{-1} \left(\mathbf{G}_1' \Psi \mathbf{G}_1 + \mathbf{G}_1' \mathbf{G}_2 \mathbf{V} \mathbf{G}_2' \mathbf{G}_1\right) \left(\mathbf{G}_1' \mathbf{G}_1\right)^{-1}, \tag{19}$$

where $\mathbf{G}_1$ and $\mathbf{G}_2$ are, respectively, the $n \times 4$ and $n \times 3$ matrices with typical row $g_{1i} = [\partial h_i / \partial \mathbf{b}_1] \,|\widehat{\mathbf{b}}_1$ and $g_{2i} = [\partial h_i / \partial \boldsymbol{\alpha}] \,|\widehat{\boldsymbol{\alpha}}$, $\Psi = \mathbf{diag}(\widehat{\eta}_i^2)$ and $\mathbf{V}$ is the covariance matrix of $\widehat{\boldsymbol{\alpha}}$. A more efficient procedure also suggested by Terza is to estimate

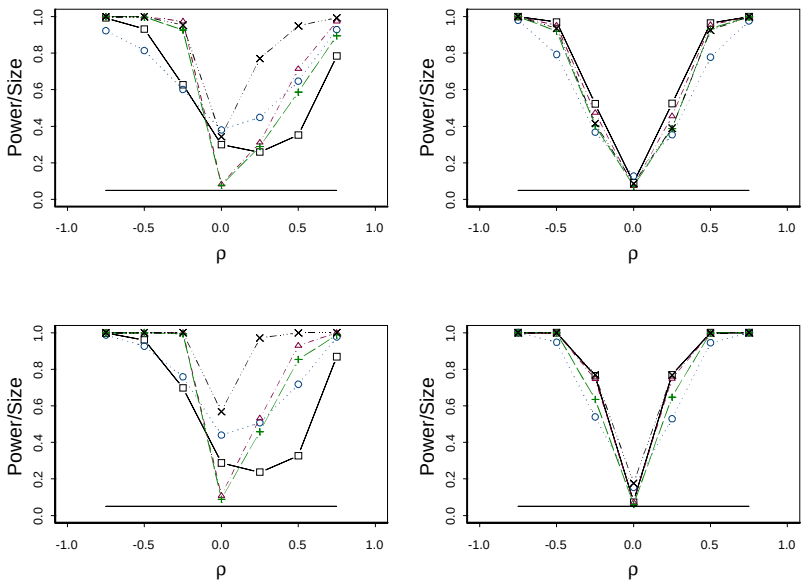
$$y_i = h(\mathbf{x}_i, z_i, \theta, \boldsymbol{\alpha}) + \eta_i,$$

with nonlinear weighted least square, using the notion that $V(\eta_i | \mathbf{x}_i, z_i) = h(\mathbf{x}_i, z_i, \theta, \boldsymbol{\alpha}) + \exp(2\mathbf{x}_i \boldsymbol{\beta}^*)(\exp\{\sigma_u^2 - 2\theta^2\}\psi_2 - \psi(\theta, \alpha)^2)$, where $\psi_2 = \exp(2\theta^2)$ $\psi(2\theta, \alpha)$. Test for endogeneity of z_i is then performed as a t-test of $\theta = 0$ denoted t (TSM) and t (NLWLS) in the sequel.

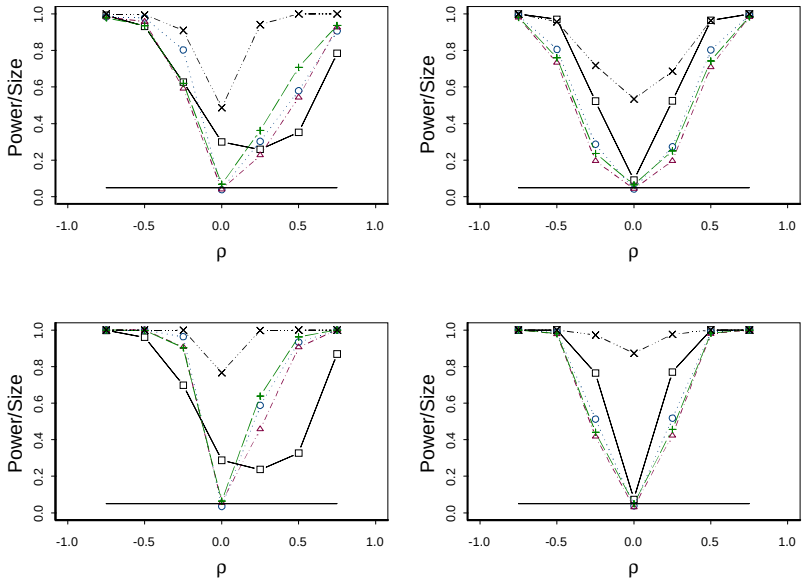
[15] Note that the predictions from a linear probability model or a logit (cf. Ameimiya, 1981) would give approximately the same ordering as using a probit estimator.

Table 1: Size of the test for exogeneity under the model by Garen (1984) and endogenous treatment model. (nominal level 5 percent).

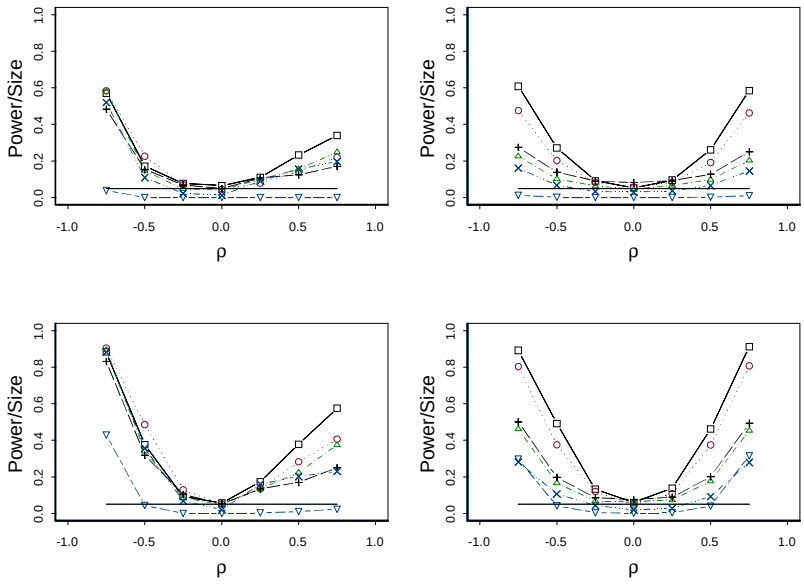
ε, u and v,	$\chi^2(1)$				$N(0,1)$			
n	200		400		200		400	
ρ	0	$u_i \equiv 0$	0	$u_i \equiv 0$	0	$u_i \equiv 0$	0	$u_i \equiv 0$
Model	$y_i = 1 + 2x_{1i} + z_i + z_i u_i + \varepsilon_i$ and $z_i = x_{1i} - x_{2i} + \nu_i$							
t	0.300	0.095	0.288	0.093	0.092	0.072	0.074	0.063
	$s(\mathbf{x}_i, z_i) = \widehat{v}_i$							
HC	0.380	0.069	0.439	0.063	0.128	0.043	0.154	0.045
C^1	0.082	0.053	0.108	0.061	0.074	0.044	0.072	0.045
C^2	0.078	0.068	0.089	0.065	0.073	0.063	0.061	0.061
L_c	0.344	0.098	0.569	0.085	0.086	0.023	0.176	0.022
	$s(\mathbf{x}_i, z_i) = z_i \widehat{v}_i$							
HC	0.039	0.050	0.035	0.049	0.042	0.048	0.037	0.036
C^1	0.040	0.049	0.056	0.046	0.043	0.051	0.030	0.037
C^2	0.068	0.052	0.064	0.057	0.066	0.063	0.050	0.048
L_c	0.487	0.049	0.767	0.046	0.535	0.022	0.873	0.023
Model	$y_i = 1 + 2x_{1i} + z_i + u_i$ and $z_i = I((x_{1i} - x_{2i} + v_i) > 0)$							
t	0.065		0.057		0.053		0.060	
DWH	0.000		0.000		0.000		0.001	
HC	0.045		0.051		0.056		0.047	
C^1	0.042		0.048		0.046		0.064	
C^2	0.049		0.053		0.081		0.076	
L_c	0.013		0.020		0.030		0.019	



Figure˜4: Power and size of the test for the Garen model and using sorting score $s(\mathbf{x}_i, z_i) = \widehat{v}_i$. Here $\square = t$, $\bigcirc = HC$, $\triangle = C_1, + = C_2$, $\times = L_c$ and nominal level (5 %) indicated by horizontal line. Left panels $\chi^2(1)$ distributions; right panels $N(0,1)$ distributions; top panels $n = 200$; and bottom panels $n = 400$.



Figure~5: Power and size ($\rho = 0$) of the test for the Garen model and using sorting score $s(\mathbf{x}_i, z_i) = z_i \widehat{v}_i$. Here $\square = t$, $\bigcirc = HC$, $\triangle = C_1$, $+ = C_2$, $\times = L_c$ and nominal level (5 %) indicated by horizontal line. Left panels $\chi^2(1)$ distributions; right panels $N(0,1)$ distributions; top panels $n = 200$; and bottom panels $n = 400$.



Figure˜6: Power and size of the test for the endogenous treatment model using sorting score $s(\mathbf{x}_i, z_i) = \Phi(\mathbf{x}_i^* \widehat{\boldsymbol{\alpha}})$. Here $\square = t$, $\bigcirc = HC$, $\triangle = C_1, + = C_2$, $\times = L_c$, $\triangledown = DWH$ and nominal level (5 %) indicated by horizontal line. Left panels $\chi^2(1)$ distributions; right panels $N(0,1)$ distributions; top panels $n = 200$; and bottom panels $n = 400$.

Table 2: Power and size of the test for exogeneity of a treatment variable in a count data regression model (nominal level 5 percent).

ρ	-0.75	-0.5	-0.25	0	$u_i \equiv 0$	0.25	0.50	0.75
n				200				
t (TSM)	0.135	0.072	0.034	0.014	0.015	0.009	0.010	0.015
t (NLWLS)	0.114	0.065	0.030	0.015	0.007	0.019	0.038	0.066
HC	0.264	0.166	0.217	0.272	0.072	0.377	0.468	0.577
C^2	0.142	0.098	0.094	0.076	0.071	0.087	0.087	0.130
L_c	0.028	0.012	0.011	0.006	0.007	0.009	0.012	0.021
n				400				
t (TSM)	0.406	0.169	0.075	0.021	0.023	0.016	0.032	0.095
t (NLWLS)	0.424	0.176	0.081	0.029	0.020	0.062	.118	.275
HC	0.323	0.182	0.210	0.368	0.044	0.557	0.745	0.819
C^2	0.235	0.147	0.078	0.087	0.039	0.056	0.120	0.185
L_c	0.054	0.016	0.007	0.005	0.002	0.009	0.016	0.044

6.2 Results

First we discuss the result from the Garen model, thereafter we discuss the result from the endogenous treatment model and lastly the results from the count data model.

- Figures 3 and 4 display the power and size for the model of Garen with the two different sorting scores $\widehat{v}_i$ and $z_i\widehat{v}_i$. The null hypothesis of exogeneity generally corresponds to $\rho = 0$. However, some of the tests implemented do not take into account the heteroskedasticity intrinsic to the Garen model and are therefore actually testing $u_i \equiv 0$: These are the Harvey-Collier[16] test and the C^1 test. For these the nominal size of 5 percent can only be expected under $u_i \equiv 0$ while for the other the nominal size is expected under both $\rho = 0$ and $u_i \equiv 0$. To support the evidence from the figures the size when $\rho = 0$ as well as when $u_i \equiv 0$ is presented in Table 1.

  - With Gaussian errors and $u_i \equiv 0$, the expected nominal size is obtained,[17] for all test statistics and both sorting scores except for the L_c test which yield too small a size.

[16] The HC test could have been adapted to take into account the heteroskedasticity by using (3) instead of assuming a constant variance ($u_i \equiv 0$).

[17] A 95% confidence interval for the empirical sizes is approximatively $\pm 1.4\%$.

- With centered $\chi^2(1)$ errors and $u_i \equiv 0$ most statistics give too large sizes. The exceptions are the HC, C^2 and C^1 tests under both sorting scores $z_i\widehat{v}_i$ and $\widehat{v}_i$.
- When $\rho = 0$ and Gaussian errors we have heteroskedasticity and most tests have too large sizes. The C^2-test has the correct size irrespective of the sorting score, while the HC and C^1 have correct size only when we have the sorting score $s(\mathbf{x}_i, z_i) = z_i\widehat{v}_i$.
- For $\rho = 0$ and centered $\chi^2(1)$ errors the heteroskedasticity is more severe than under the Gaussian errors and, as can be expected, the size of the tests are generally too large. The HC, C^1 and C^2 with $s(\mathbf{x}, z) = z_i\widehat{v}_i$ have the correct size and the C^2 using $s(\mathbf{x}, z) = \widehat{v}_i$ is not far from being of the correct size.
- Comparing empirical powers, we can see from Figure 3 and 4 that the C^2 test is not significantly worse than the Wald test (t) under normality while not being sensitive to the non-normal error terms considered.

With the estimated sorting score $z_i\widehat{v}_i$, the performances of all the tests improve with respect to size except for L_c that is still sensitive to heteroskedasticity. Most notable is the much better size of HC for $\rho = 0$. This is most surely due to a smoothly varying variance through the ordering obtained with this sorting score, rending the naive HC less sensitive to the heteroskedasticity.

- Figure 5 displays the result from the endogenous treatment model. The size of the test can also be studied in Table 1.
  - We see that the Wald test (t), the HC and C^1 tests always have the correct size. The L_c is conservative under the normal distribution but gives too large a size under the centered $\chi^2(1)$ error distributions. The opposite is observed for the C^2 test.
  - Looking at Figure 5 we can see that the DWH test has, as expected, very low power: no power at all when $n = 200$. When the error are normally distributed the Wald test is uniformly best, while when the errors follow the $\chi^2(1)$ distribution either the HC or the Wald test perform best.
- The results from the count data regression are displayed in Table 2.

- We can see that the Wald test (t) using the TSM and NLWLS both yields too small sizes, although this improves when the sample size increases. The HC and C^2 tests have correct size when there is no over-dispersion. When the marginal is miss-specified[18] the size is, as expected, much too large for HC. The C^2-test has a slightly too large size when over-dispersion is present, while when the marginal is Poisson distributed the size is correct for $n = 400$. The L_c test is always conservative and with almost no power at all.
- The power of the tests are generally low. The power is generally higher when $\rho < 0$ than when $\rho > 0$. The one exception is the HC test where the power is higher for the opposite situation, i.e. when $\rho > 0$.

The test with the best performance is, as expected, the Wald test using the NLWLS estimator. The C^2 test and Wald test using the TSM have comparable power.

7 Conclusion

We have introduced a new concept of testing an exogeneity assumption in cross-section regression by first sorting data and then using time series tests for miss-specification. This concept has allowed us to put forward two main diagnostic tools. First, what we believe is the first convincing graphical display allowing to detect endogeneity of a variable: the CUSUM of the recursive residuals obtained with a relevant sorting score. The second main contribution is a Chow test (C^2) which is simple to compute with ordinary econometrics packages and which has shown not to be sensitive to distributional assumptions in the experiments conducted.

By considering specific parametric models for endogeneity we have aimed at illustrating the usefulness of our framework. In particular, the conducted Monte Carlo experiments have allowed us to show that the proposed tests, including the Chow-type ones, have similar performances to tests making use of the parametric and distributional assumptions.

[18] Practically, a preliminary data analysis should be used to identify the over-dispersion issue. Recursive residuals calculated under e.g., a negative binomial regression model instead of Poisson are likely then to be less sensitive to the miss-specification of the marginal distribution.

Moreover, because we do not make strong assumptions on the endogeneity alternative hypothesis, the proposed tests, and most obviously the simple-to-use Chow test, are robust to the alternative distributional assumptions considered, both at the structural equation level and at the endogeneity modelling level.

We believe that the use of the sorting score is open to further potential applications than those highlighted in this article. For instance, the construction of non-parametric sorting scores as outlined in Section 5 but not further studied herein, seems to us to have promising prospects for successful applications.

References

Amemiya, T. (1981). "Qualitative Response Models: A Survey", *Journal of economic literature,* **XIX,** 1483–1536.

Brown, R.L., Durbin, J. and Evans, J. M. (1975). "Techniques for testing the constancy of regression relationships over time (with discussion)", *Journal of the Royal Statistical Society Series B*, **37**, 149-192.

Brännäs, K. and Eriksson, M. (1997). "Endogeneity in a Binomial Model", UES **440,** Department of economics, Umeå University.

Chow, G.C. (1960). "Tests of Equality Between Sets of Coefficients in two Linear Regressions", *Econometrica*, **28**, 591-605.

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

Dawid, A.P. (1979). "Conditional Independence in Statistical Theory", *Journal of the Royal Statistical Society Series B*, **41**, 1-31.

Dawid, A.P. (1984). "Statistical Theory: The Prequential Approach", *Journal of the Royal Statistical Society Series A*, **147**, 278-292.

Engle, R. F.(1984). "Wald, Likelihood Ratio and Lagrange Multiplier Tests in Econometrics", in Z. Griliches and M.D. Intriligator (eds.) *Handbook of Econometrics Volume II* Elsevier Publisher BV.

Engle, R. F., Hendry, D. F. and J-F. Richard (1994). "Exogeneity", in Neil R Ericsson and John S. Irons (eds.) *Testing Exogeneity* Oxford university Press, New York. Ericsson, N.R. (1994). "Testing Exogeneity: An introduction", in Neil R Ericsson and John S. Irons (eds.) *Testing Exogeneity* Oxford university Press, New York.

Ericsson, N.R. (1994). "Testing Exogeneity: An Introduction", in Neil R Ericsson and John S. Irons (eds.), *Testing Exogeneity,* Oxford university Press, New York.

Garen, J. (1984). "The returns to Schooling: A Selectivity Bias Approach with a Continuous Choice variable" *Econometrica*, **52**, 1199-1218.

Garen, J. (1988). "Compensating Wage Differentials and Endogeneity of Job Riskiness" *Review of Economics and Statistics*, **70**, 9-16.

Gourieroux, C., Montfort, A. and A, Trognon (1984). "Pseudo Maximum Likelihood Methods: Theory", *Econometrica*, **52**, 681-700.

Hansen, B. E. (1994). "Testing for parameter Instability in Linear Models" in Neil R Ericsson and John S. Irons (eds.), *Testing Exogeneity,* Oxford university Press, New York.

Goldfeld, S. and Quandt, R. (1965). "Some Test for Homoscedasticity" *Journal of the American Statistical Association*, **60**, 539-547.

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

Harvey, A. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

Harvey, A. and Collier G. (1977). "Testing for Functional Misspecification in Regression Analysis." *Journal of Econometrics*, **6**, 103-119.

Hastie, T.J. and Tibshirani R.J. (1990). *Generalized Additive Models*, Chapman & Hall, London.

Heckman, J.J. (1978). "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, **46**, 931-959.

Heckman, J.J. (1979). "Sample Selection Bias as a Specification Error," *Econometrica*, **47**, 153-161.

Hill, R.C., Griffiths, W.E. and G.G. Judge (1997). *Undergraduate Econometrics*, John Wiley & Sons, Inc. New York.

Kianifard, F. and Swallow W.H. (1996). "A Review of the Development and Application of Recursive Residuals in Linear Models," *Journal of the American Statistical Association*, **91**, 391-400.

Nyblom, J. (1989). "Testing for the Constancy of Parameters Over Time," *Journal of the American Statistical Association,* **84**, 223-230.

Ohtani, K. and T. Toyoda (1985). "Small Sample Properties of Test of Equality between Sets of Coefficients in Two Linear Regressions under Heteroscedasticity". *International Economic Review* **26**, 37-44.

Rosenbaum, P.R. and Rubin, D.B. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effect," *Biometrika*, **70**, 41-55.

Smith, J.Q. (1985). "Diagnostic Checks of Non-standard Time Series Models," *Journal of Forecasting*, 4, 283-291.

Terza, J.V. (1998). "Estimating Count Data Models with Endogenous Switching: Sample selection and Exogenous Treatments Effects," *Journal of Econometrics,* **84**, 129-154.

Tsay, R.S. (1998). "Testing and Modeling Multivariate Threshold Models," *Journal of the American Statistical Association*, **93**, 1188-1202.

Vella, F. (1998). "Estimating Models with Sample Selection Bias: A Survey", *Journal of Human Resources,* **38,** 127-169.

White, H. (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity", *Econometrica*, **48**, 817-838.

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

Windmeijer F. and Santos Silva, J. (1997). "Endogeneity in Count Data Models: An Application to demand for Health Care". *Journal of Applied Econometrics,* **12**, 281-294.