Australian School of Business Working Paper

Never Stand Still — Australian School of Business

Australian School of Business Research Paper No. 2014 ECON 34

Binary Choice Model with Endogeneity: Identification via Heteroskedasticity

Minxian Yang

Binary Choice Model with Endogeneity: Identification via Heteroskedasticity

Minxian Yang
School of Economics
UNSW Australia
(The University of New South Wales)
Sydney 2052
Australia
m.yang@unsw.edu.au

Abstract

The idea of identifying structural parameters via heteroskedasticity is explored in the context of binary choice models with an endogenous regressor. Sufficient conditions for parameter identification are derived for probit models without relying on instruments or additional restrictions. The results are extendable to other parametric binary choice models. The semi-parametric model of Manski (1975, 1985), with endogeneity, is also shown to be identifiable in the presence of heteroskedasticity. The role of heteroskedasticity in identifying and estimating structural parameters is demonstrated by Monte Carlo experiments.

Keywords: Qualitative response, Probit, Logit, Linear median regression, Endogeneity, Identification, Heteroskedasticity

JEL classification: C25, C35, C13

First Draft: November 2013

This Version: July 2014

1 Introduction

In the literature on endogeneity in binary choice models, with both parametric and semi-parametric approaches, attention has been focused on the cases where instruments, or exclusion restrictions, are available (see Heckman (1978), Rivers and Vuong (1988), Lewbel (2000), Blundell and Powell (2004) among others). The identification of structural parameters in binary choice models with endogeneity has not been available in the absence of instruments or exclusion restrictions.

Recently, a number of authors have exploited heteroskedasticity for the purpose of identifying structural parameters in linear simultaneous equation models where instruments or exclusion restrictions are not available (see Rigobon (2003), Klein and Vella (2010), Lewbel (2012) and Milunovich and Yang (2013) among others). Intuitively, variations in the conditional variances of regression errors bring out traces of the relationships amongst endogenous variables in the conditional mean, which in turn make it possible to identify structural parameters without relying on instruments or exclusion restrictions. Heteroskedasticity, which was traditionally regarded as a nuisance in regression analyses, turns out to be useful for dealing with endogeneity in linear models.

Extending the above literature, this article explores the idea of identifying structural parameters via heteroskedasticity for binary choice models that have an endogenous regressor. The models considered include popular probit and related parametric models, as well as the linear median model of Manski (1975, 1985) with endogeneity. In this article, the structural parameters in these models are shown to be locally identifiable in the presence of heteroskedasticity without relying on instruments or exclusion restrictions. Sufficient conditions for the identification of the structural parameters are formulated, in which a key requirement is described in terms of the conditional variances of error terms. Monte Carlo experiments are carried out to demonstrate the effect of heteroskedasticity on parameter identification and estimation. The framework and findings in this paper may be particularly useful for applications where instruments are unavailable and heteroskedasticity itself is of empirical interest. The results documented in this article are a new contribution to the existing literature.

The rest of the paper is organised as follows. Section 2 reports results about probit models, including some simulation evidence. Sections 3 and 4 discuss other parametric models and linear median models respectively. Section 5 is a conclusion. Proofs are collected in the appendix.

2 Probit Models

2.1 Probit model with endogeneity

Consider the following probit model

(1) $$y_{1i}^* = \beta_1' x_i + \gamma_1 y_{2i} + \varepsilon_{1i}\ , \qquad y_{1i} = \mathbf{1}(y_{1i}^* > 0)\ ,$$

$$y_{2i} = \beta_2' x_i + \sigma_{2i}\varepsilon_{2i}\ ,$$

$$\begin{bmatrix}\varepsilon_{1i}\\ \varepsilon_{2i}\end{bmatrix} \mid W_i \sim N(0,R)\ , \qquad R = \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix},$$

where i is the observation index, $\mathbf{1}(\cdot)$ is the indicator function, (x_i, y_{1i}, y_{2i}) are observable, W_i is the set of all observable exogenous or predetermined variables for index i, and $x_i \in W_i$ is a k-dementional vector of exogenous variables. The sense of endogeneity here is that the correlation ρ between ε_{1i} and y_{2i} is non-zero. The second equation in (1) is potentially heteroskedastic and the conditional variance σ_{2i}^2 may depend on W_i. As both equations involve the same exogenous regressors in x_i, the setup here covers the cases where neither instruments nor exclusion restrictions are available. In this setting, the instrument-variable related strategies (e.g., the control function method of Rivers and Vuong (1988)) are not applicable.

The joint normality of $(y_{1i}^*, y_{2i}|W_i)$ implies

(2) $$y_{1i}^*|W_i, y_{2i} \sim N(\mu_{1|2i},\ 1-\rho^2)\ , \qquad \mu_{1|2i} = (\beta_1 - \beta_2\rho/\sigma_{2i})' x_i + (\gamma_1 + \rho/\sigma_{2i}) y_{2i}\ .$$

Suppose for now that $\sigma_{2i} = \sigma_2$ is constant (y_{2i} is homoscedastic). In this case, when a single-equation probit model $y_{1i}^* = \beta_1' x_i + \gamma_1 y_{2i} + \varepsilon_{1i}$ is estimated with endogeneity being ignored, the estimators of β_1 and γ_1 will converge in probability to $(\beta_1 - \beta_2\rho/\sigma_2)(1-\rho^2)^{-1/2}$ and $(\rho/\sigma_2 + \gamma_1)(1-\rho^2)^{-1/2}$ respectively. Further, when (1) is treated as a simultaneous equation system, the parameters $(\beta_1, \gamma_1, \rho)$ are not identifiable without further restrictions, despite the fact that (β_2, σ_2^2) can be identified and consistently estimated from the second equation. In the following section, we analyse the identification of the structural parameters in the presence of heteroskedasticity.

2.2 Heteroskedasticity in Secondary Equation

Combining (1) and (2), we write the model as

(3) $$y_{1i}^*|W_i, y_{2i} \sim N(\mu_{1|2i},\ 1-\rho^2)\ , \qquad y_{1i} = \mathbf{1}(y_{1i}^* > 0)\ ,$$

$$y_{2i}|W_i \sim N(\beta_2' x_i,\ \sigma_{2i}^2)\ ,$$

$$\mu_{1|2i} = (\beta_1 - \beta_2\rho/\sigma_{2i})'x_i + (\gamma_1 + \rho/\sigma_{2i})y_{2i} ,$$

$$\sigma_{2i}^2 = \exp\big(F_2(z_i, \alpha_2)\big) , \qquad z_i \in W_i ,$$

where the function $F_2(z_i, \alpha_2)$ is continuously differentiable with respect to parameter vector α_2. Here z_i can be identical to x_i or a subset of x_i or a superset of x_i. It is now possible to identify and consistently estimate $(\beta_1, \gamma_1, \rho, \beta_2, \alpha)$ because x_i/σ_{2i} and y_{2i}/σ_{2i} (if not proportional to x_i and y_{2i}) serve as additional regressors in the conditional mean $\mu_{1|2i}$. In other words, the heteroskedasticity σ_{2i} plays the role of an “instrument”.

EXAMPLE 1. In the context of pooled cross-sections, suppose that data consist of random surveys from two different years (Year 0 and Year 1). The conditional variance is $\sigma_{2i}^2 = \exp(\alpha_{20} + \alpha_{21}z_{1i})$, where z_{1i} is a dummy variable that is one if observation i is from Year 1 and zero otherwise. Here z_{1i} may also be included in the conditional means of y_{1i}^* and y_{2i}. It can be verified that identification is achievable when $\alpha_{21} \neq 0$ (see Theorem 2 below).

We now formalise the conditions for parameter identification in the framework of maximum likelihood (ML). The joint conditional probability density (or mass) is given by

(4) $$P_i(y_{1i} = 1|W_i, y_{2i})\mathrm{pdf}(y_{2i}|W_i) = \Phi\big(\mu_{1|2i}(1-\rho^2)^{-1/2}\big)\phi\big((y_{2i} - \beta_2'x_i)/\sigma_{2i}\big)/\sigma_{2i} ,$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution and probability density functions of $N(0,1)$ respectively. The log likelihood function is

(5) $$L_n(\theta) = \sum_{i=1}^n \{y_{1i}\ln\Phi_i + (1-y_{1i})\ln(1-\Phi_i) - \tfrac{1}{2}\ln(2\pi\sigma_{2i}^2) - \tfrac{1}{2\sigma_{2i}^2}(y_{2i} - \beta_2'x_i)^2\} ,$$

where $\Phi_i = \Phi(\mu_{1|2i}/(1-\rho^2)^{1/2})$, n is the sample size, $\theta = [\beta_1', \gamma_1, \rho, \beta_2', \alpha']'$ is the vector of the structural parameters. The ML estimator of θ is the maximiser of (5). The log likelihood (5) can also be expressed in terms of the reduced-form parameter vector $h = [b_1', g_1, g_2, b_2', \beta_2', \alpha']'$ by writing

(6) $$\Phi_i = \Phi(m_{1|2i}) ,$$

$$m_{1|2i} = \mu_{1|2i}(1-\rho^2)^{-1/2} = b_1'x_i + g_1y_{2i} + g_2y_{2i}/\sigma_{2i} + b_2'x_i/\sigma_{2i} .$$

By the comparison of (3) and (6), the map from the structural parameter vector θ to the reduced-form parameter vector h is given by

(7) $\quad b_1 = \beta_1 q \ , \quad g_1 = \gamma_1 q \ , \quad g_2 = \rho q \ , \quad b_2 = -\beta_2 \rho q \ ,$

where $q = (1-\rho^2)^{-1/2}$, the reduced-form and structural (β_2, α_2) are identical.

We rely on Theorem 6 of Rothenberg (1971) to examine the local identification of the structural parameters. In essence, the structural parameters are locally identified if: (a) the Jacobian of the map in (7) is of full rank in an open neighbourhood of the true structural parameter point θ^*; (b) the reduced form parameters are locally identified at the true reduced-form parameter point h^* (corresponding to θ^*). Let Θ be the structural parameter space that excludes the points with $|\rho| = 1$. For a matrix $M(\theta)$ whose elements are continuous fuctions of θ, a parameter point θ is a regular point of the matrix if the matrix's rank does not change in an open neighbourhood of θ. The following theorem confirms that the Jacobian of (7) is of full rank.

THEOREM 1. *The Jacobian of the map in* (7) *is of full rank in* Θ. *If* θ^* *is an interior point, then it is a regular point of the Jacobian.*

Given Theorem 1, we look for conditions under which the reduced-form parameters are identified. Let $X_i' = [x_i', y_{2i}, y_{2i}/\sigma_{2i}, x_i'/\sigma_{2i}]$ and $h' = [B', \beta_2', \alpha_2']$, where $B' = [b_1', g_1, g_2, b_2']$. The score for the observation with index i is given by

(8) $$s_i(h) = \frac{\partial l_i(h)}{\partial h} = \lambda_i \frac{\partial m_{1|2i}}{\partial h} - \frac{1}{2\sigma_{2i}^2} \frac{\partial (y_{2i} - \beta_2' x_i)^2}{\partial h} + \frac{1}{2}\left[\frac{(y_{2i} - \beta_2' x_i)^2}{\sigma_{2i}^2} - 1\right] \frac{\partial F_2(z_i, \alpha_2)}{\partial h} \ ,$$

where $l_i(h)$ is the log likelihood of $(y_{1i}, y_{2i}|W_i)$, $\lambda_i = \frac{y_{1i}\phi_i}{\Phi_i} - \frac{(1-y_{1i})\phi_i}{1-\Phi_i}$ and $\phi_i = \phi(m_{1|2i})$. With the partition in $h' = [B', \beta_2', \alpha_2']$, the details of (8) are

$$\frac{\partial l_i(h)}{\partial B} = \lambda_i X_i \ ,$$

$$\frac{\partial l_i(h)}{\partial \beta_2} = \frac{1}{\sigma_{2i}^2} u_{2i}(\beta_2) x_i \ ,$$

$$\frac{\partial l_i(h)}{\partial \alpha_2} = \frac{1}{2}\left[\left(\frac{u_{2i}^2(\beta_2)}{\sigma_{2i}^2} - 1\right) - \frac{\lambda_i G_{2i}}{\sigma_{2i}}\right] f_{2i} \ ,$$

where $f_{2i} = \partial F_2(z_i, \alpha_2)/\partial \alpha_2$, $u_{2i}(\beta_2) = y_{2i} - \beta_2' x_i$ and $G_{2i} = g_2 y_{2i} + b_2' x_i$. Clearly, $E(\lambda_i | W_i, y_{2i}) = 0$ and $E(s_i(h)|W_i, y_{2i}) = 0$ at the true parameter point $h = h^*$. It follows that at $h = h^*$

(9) $\quad E\{E[s_i(h)s_i(h)'|W_i, y_{2i}]\} =$

$$E\begin{bmatrix} \Lambda_i X_i X_i' & 0 & -\frac{\Lambda_i G_{2i}}{2\sigma_{2i}} X_i f_{2i}' \\ 0 & \frac{1}{\sigma_{2i}^4} u_{2i}^2(\beta_2) x_i x_i' & 0 \\ -\frac{\Lambda_i G_{2i}}{2\sigma_{2i}} f_{2i} X_i' & 0 & \frac{1}{4}\left[\left(\frac{u_{2i}^2(\beta_2)}{\sigma_{2i}^2} - 1\right)^2 + \frac{\Lambda_i G_{2i}^2}{\sigma_{2i}^2}\right] f_{2i} f_{2i}' \end{bmatrix},$$

where $\Lambda_i = E(\lambda_i^2 | W_i, y_{2i}) = \frac{\phi_i^2}{\Phi_i} + \frac{\phi_i^2}{1-\Phi_i} > 0$ and the fact $E(\varepsilon_{2i}^3) = 0$ is used. Essentially, h^* is locally identifiable when the above matrix is invertible (Theorem 1 of Rothenberg (1971)). The following theorem provides a set of sufficient conditions for identification that are specific to our context.

THEOREM 2. *Assume that* (i) *the data* $(x_i, y_{1i}, y_{2i}, z_i)$ *are independent draws from* (3)*;* (ii) $E(x_i x_i'/\sigma_{2i}^2)$ *and* $E(f_{2i} f_{2i}')$ *are invertible at* h^**;* (iii) $P(\mathcal{A}) > 0$ *at* h^**, where the event* $\mathcal{A} = \{(d_1' + d_2'/\sigma_{2i})x_i \neq 0$ *for any non-zero constant vector* $[d_1', d_2']'\}$*;* (iv) θ^* *is an interior point of* Θ*. Then,* h^* *is locally identified. Further, together with Theorem 1,* θ^* *is also locally identified.*

Condition (ii) ensures the identification of (β_2, α_2) in the second equation of (3). Condition (iii) is the key requirement for identifying $(\beta_1, \gamma_1, \rho)$ in the first equation of (3). Obviously, if the model is homoscedastic, (iii) fails. Hence the presence of heteroskedasticity is necessary for identification in this context. Condition (iii) also place restrictions on the functional form of σ_{2i}. In particular, it fails when σ_{2i} is a linear function of x_i. The example below illustrates this possibility.

EXAMPLE 2. Let $x_i = [1, x_{1i}]'$ with $\mathrm{var}(x_{1i}) > 0$ and $z_i = x_i$. For the case $\sigma_{2i}^2 = \exp(\alpha_{20} + \alpha_{21} x_{1i})$ with $\alpha_{21} \neq 0$, conditions (ii) and (iii) can be verified. However, for the case $\sigma_{2i} = \alpha_{20} + \alpha_{21} x_{1i}$ with $\alpha_{21} \neq 0$ (positivity assumed for σ_{2i}), condition (iii) fails to hold because $P[(d_1' + d_2'/\sigma_{2i})x_i = 0] = 1$ for $d_1 = [-1, 0]'$ and $d_2 = [\alpha_{20}, \alpha_{21}]'$.

2.3 Heteroskedasticity in Both Equations

The model with heteroskedasticity in both equations can be written as

(10) $y_{1i}^* = \beta_1' x_i + \gamma_1 y_{2i} + \sigma_{1i}\varepsilon_{1i}$, $\quad y_{1i} = \mathbf{1}(y_{1i}^* > 0)$,

$y_{2i} = \beta_2' x_i + \sigma_{2i}\varepsilon_{2i}$,

$\begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{bmatrix} \mid W_i \sim N(0, R)$, $\quad R = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$,

$\sigma_{1i}^2 = \exp(F_1(z_i, \alpha_1))$, $\quad \sigma_{2i}^2 = \exp(F_2(z_i, \alpha_2))$, $\quad z_i \in W_i$,

where functions F_1 and F_2 are continuously differentiable with respect to α_1 and α_2 respectively. Define

$\bar{y}_{1i}^* = y_{1i}^*/\sigma_{1i} = \beta_1' x_i/\sigma_{1i} + \gamma_1 y_{2i}/\sigma_{1i} + \varepsilon_{1i}$.

The event $(y_{1i}^* > 0)$ is equivalent to $(\bar{y}_{1i}^* > 0)$. Clearly, if $F_1(z_i, \alpha_1)$ is unrestricted, $(\beta_1, \gamma_1, \sigma_{1i})$ will be observationally equivalent to $(c\beta_1, c\gamma_1, c\sigma_{1i})$ for any constant $c > 0$. A normalisation rule is needed to resolve this indeterminacy. Let A_{10} be the set of α_1 values for which $F_1(z_i, \alpha_1)$ does not depend on the non-constant elements of z_i. The normalisation we use is: $F_1(z_i, \alpha_1) = 0$ for all $\alpha_1 \in A_{10}$.

EXAMPLE 3. When $F_1(z_i, \alpha_1) = \alpha_{10} + \alpha_{11} z_{1i}$ with $\alpha_1 = [\alpha_{10}, \alpha_{11}]'$ and z_{1i} being non-constant, it is normalised as $F_1(z_i, \alpha_1) = \alpha_{11} z_{1i}$. When $F_1(z_i, \alpha_1) = \ln(\alpha_{10} + \alpha_{11} z_{1i})$ ($\alpha_{10} + \alpha_{11} z_{1i} > 0$ is assumed), the normalized version is $F_1(z_i, \alpha_1) = \ln(1 + \alpha_{11} z_{1i})$. In both cases, A_{10} is the set of α_1 values with $\alpha_{11} = 0$.

The model in (10) can be written as

(11) $\bar{y}_{1i}^* | W_i, y_{2i} \sim N(\mu_{1|2i},\ 1 - \rho^2)$, $\quad y_{2i} | W_i \sim N(\beta_2' x_i, \sigma_{2i}^2)$,

$\mu_{1|2i} = (\beta_1/\sigma_{1i} - \beta_2 \rho/\sigma_{2i})' x_i + (\gamma_1/\sigma_{1i} + \rho/\sigma_{2i}) y_{2i}$,

$\sigma_{1i}^2 = \exp(F_1(z_i, \alpha_1))$, $\quad \sigma_{2i}^2 = \exp(F_2(z_i, \alpha_2))$, $\quad z_i \in W_i$.

With the structural parameter vector $\theta = [\beta_1', \gamma_1, \rho, \alpha_1, \beta_2', \alpha_2']'$, the log likelihood expression in (5) is also valid for (11). The log likelihood can also be expressed in terms of the reduced-form parameter vector $h = [b_1', g_1, g_2, b_2', \alpha_1, \beta_2', \alpha_2']'$ by re-defining

(12) $\Phi_i = \Phi(m_{1|2i})$,

$$m_{1|2i} = \mu_{1|2i}(1-\rho^2)^{-1/2} = b_1'x_i/\sigma_{1i} + g_1 y_{2i}/\sigma_{1i} + g_2 y_{2i}/\sigma_{2i} + b_2'x_i/\sigma_{2i}\ .$$

A comparison of (11) and (12) shows that the mapping from θ to h is the same as (7), with the structural and reduced-form $(\alpha_1, \beta_2, \alpha_2)$ being identical. Again, the identification of θ^* in (11) is established if (a) the map from θ to h has a full-rank Jacobian in an open neighbourhood of θ^*; (b) h^* is locally identified. As (a) is established in Theorem 1, we need only consider (b) here. The identification conditions are summarised in the following theorem, where an additional symbol is $f_{1i} = \partial F_1(z_i, \alpha_1)/\partial \alpha_1$.

THEOREM 3. *Assume that* (i) *the data* $(x_i, y_{1i}, y_{2i}, z_i)$ *are independent draws from* (12)*;* (ii) $E(x_i x_i'/\sigma_{2i}^2)$ *and* $E(f_{2i} f_{2i}')$ *are invertible at* h^**;* (iii) $P(\mathcal{A}) > 0$ *at* h^*, *where the event* $\mathcal{A} = \{(d_1'/\sigma_{1i} + d_2'/\sigma_{2i} + d_3' f_{1i} b_1'/\sigma_{1i})x_i \neq 0$ *for any non-zero constant vector* $[d_1', d_2', d_3']'\}$*;* (iv) θ^* *is an interior point of* Θ. *Then,* h^* *is locally identified. Consequently, together with Theorem 1,* θ^* *is also locally identified.*

As in Theorem 2, condition (ii) ensures the identification of (β_2, α_2) in the second equation of (11). Condition (iii) is the key requirement for identifying $(\beta_1, \gamma_1, \rho, \alpha_1)$ in the first equation. It involves the conditional variances of both equations. Interestingly, condition (iii) may hold even when the second equation is homoscedastic (σ_{2i} is a constant). There are three scenarios where (iii) fails. First, it fails when σ_{1i}/σ_{2i} is a constant (i.e., $P(\mathcal{A}) = 0$ for $d_3 = 0$ and $d_1 = -d_2\sigma_{1i}/\sigma_{2i}$ for any $d_2 \neq 0$). Second, it fails when both $\sigma_{1i} = \alpha_1' x_i$ and $\sigma_{2i} = \alpha_2' x_i$ are linear functions of x_i (i.e., $P(\mathcal{A}) = 0$ for $d_3 = 0$ and $d_1 = -\alpha_1$ and $d_2 = \alpha_2$). Third, it also fails when σ_{1i} is not a constant and $f_{1i} b_1' x_i$ is linear in x_i or $f_{1i} b_1' x_i = M' x_i$ for a constant matrix M (i.e., $P(\mathcal{A}) = 0$ for $d_2 = 0$ and $d_1 = -M d_3$ for any $d_3 \neq 0$). The example below illustrates some of these scenarios.

EXAMPLE 4. Let $x_i = [1, x_{1i}]'$ and $z_i = x_i$, where x_{1i} is a continuous random variable with $\mathrm{var}(x_{1i}) > 0$. Suppose $F_2(z_i, \alpha_2) = \alpha_{20} + \alpha_{21} x_{1i}$ and $F_1(z_i, \alpha_1) = \alpha_{11} x_{1i}$ with $\alpha_{11} \neq 0$. As $f_{2i} = x_i$, condition (ii) is satisfied. Since $f_{1i} = x_{1i}$, $(d_1'/\sigma_{1i} + d_2'/\sigma_{2i} + d_3' f_{1i} b_1'/\sigma_{1i})x_i = 0$ with probability 1 for some $[d_1', d_2', d_3]' \neq 0$ if and only if: either $\alpha_{11} = \alpha_{21}$ (σ_{1i} becomes proportional to σ_{2i}); or $b_1' = [b_{10}, 0]$ (equivalently $\beta_1' = [\beta_{10}, 0]$). Hence, condition (iii) holds provided that σ_{1i} is not proportional to σ_{2i} and $\beta_1' = [\beta_{10}, \beta_{11}]$ with $\beta_{11} \neq 0$. The parameters are still locally identifiable when the second equation is homoscedastic with $\alpha_{21} = 0$.

2.4 Simulation Experiments

In this section, the role of heteroskedasticity in parameter identification and estimation is demonstrated in simulation experiments. The model in (3) is the data generating process (DGP), where $x_i = [1, x_{1i}]'$ and $z_i = x_i$ with x_{1i} being independent $N(0,1)$. The conditional variance of the second equation is specified as $\sigma_{2i}^2 = \exp(\alpha_2' x_i)$. The parameter values for the DGP are as follows

$$\alpha_2 = [\alpha_{20}, \alpha_{21}]' = [0,\ \alpha_{21}]', \quad \alpha_{21} = 0.1, 0.3, 0.5, 0.7, 0.9\ ,$$

$$\beta_2 = [\beta_{20}, \beta_{21}]' = [0.1, 0.2]', \quad \beta_1 = [\beta_{10}, \beta_{11}]' = [0.3, 0.4]',$$

$$\gamma_1 = 0.5, \quad \rho = -0.6.$$

Each sample consists of 500 independent observations on (x_i, y_{1i}, y_{2i}) drawn from the DGP (3). For each sample, the ML estimates are obtained by maximising the log likelihood (5). There are 500 independent samples. The biases and root mean squared errors (RMSE) are then computed from the 500 replications. The results are reported in Table 1 below.

Table 1. Biases and RMSEs of the ML Estimators

	γ_1	ρ	β_{10}	β_{11}	β_{20}	β_{21}	α_{20}	α_{21}
$\boldsymbol{\alpha_{21} = 0.1}$								
Bias	0.292	-0.299	-0.012	-0.037	-0.001	0.001	0.011	-0.006
RMSE	0.624	0.616	0.142	0.217	0.047	0.044	0.063	0.063
$\boldsymbol{\alpha_{21} = 0.3}$								
Bias	0.083	-0.082	-0.008	-0.018	-0.003	-0.003	0.006	-0.008
RMSE	0.325	0.302	0.112	0.171	0.046	0.044	0.064	0.063
$\boldsymbol{\alpha_{21} = 0.5}$								
Bias	0.025	-0.030	-0.002	-0.009	0.001	0.002	0.008	-0.003
RMSE	0.185	0.176	0.089	0.116	0.046	0.040	0.064	0.065
$\boldsymbol{\alpha_{21} = 0.7}$								
Bias	-0.001	-0.001	0.001	0.005	0.000	-0.001	0.008	-0.002
RMSE	0.128	0.118	0.071	0.093	0.051	0.036	0.066	0.066
$\boldsymbol{\alpha_{21} = 0.9}$								
Bias	-0.010	0.006	0.006	0.000	-0.003	0.000	0.007	0.003
RMSE	0.089	0.093	0.065	0.075	0.051	0.038	0.063	0.066

In Table 1, as expected, the estimates of (β_2, α_2) are accurate with small biases and RMSEs that vary little as α_{21} changes. On the other hand, the accuracy of the estimates of $(\beta_1, \gamma_1, \rho)$ crucially depends on the strength of heteroskedasticity measured by α_{21}. For weak heteroskedasticity ($\alpha_{21} = 0.1$) that is close to losing identification, the estimates of $(\beta_1, \gamma_1, \rho)$ exhibit large absolute biases and RMSEs. The estimation accuracy improves markedly as α_{21} increases and identification strengthens. When heteroskedasticity is strong ($\alpha_{21} = 0.9$), the estimates of $(\beta_1, \gamma_1, \rho)$ are almost as accurate as those of (β_2, α_2). The results in Table 1 indicate that heteroskedasticity, if present, may play an "instrumental" role in identifying and estimating the structural parameters in probit models when instruments and exclusion restrictions are not available.

The computation is carried out in R version 3.0.2 of R Core Team (2013). The function "optim" with the BFGS algorithm is used for maximising the log likelihood.

3 Logit and Other Parametric Models

The results in Section 2 depend on the joint normality of $[\varepsilon_{1i}, \varepsilon_{2i}]'$ to some extent. First, the reduced-form equation for y_{1i}^*, as in (3) and (11), is derived under the normality assumption. Second, the property $E(\varepsilon_{2i}^3) = 0$ is used in the proof of Theorems 2 and 3. Nonetheless, when ε_{1i} and ε_{2i} are linearly correlated, the approach used in Section 2 is applicable for cases with known distributions other than normal. For instance, consider a variant of (3)

$$(13) \qquad y_{1i}^* = \beta_1' x_i + \gamma_1 y_{2i} + \varepsilon_{1i} \,, \qquad y_{1i} = \mathbf{1}(y_{1i}^* > 0) \,, \qquad \mathrm{var}(\varepsilon_{2i}) = 1 \,,$$

$$y_{2i} = \beta_2' x_i + \sigma_{2i} \varepsilon_{2i} \,, \qquad \varepsilon_{2i} | W_i \sim N(0, 1) \,,$$

$$\varepsilon_{1i} = \rho \varepsilon_{2i} + e_i \,,$$

where e_i is independent of ε_{2i} and follows a symmetric distribution (not necessarily normal) with zero mean. Logit models fit the above description. The results in Section 2.2 are applicable to (13) because the log likelihood in (5) or (6) is valid as long as $\Phi(\cdot)$ is interpreted as the cumulative distribution function of $e_i/(1-\rho^2)^{1/2}$. Similarly, (11) can also be extended to accommodate alternative conditional distributions for $y_{2i}|W_i$.

4 Linear Median Binary Response Model

In the semi-parametric model of Manski (1975, 1985), the conditional median of the unobserved dependent variable y_{1i}^* is assumed to be a linear function of exogenous variables. We consider the case with an endogenous regressor

$$(14)\qquad y_{1i}^* = \beta_1' x_i + \gamma_1 y_{2i} + \varepsilon_{1i}\ , \qquad y_{1i} = \mathbf{1}(y_{1i}^* > 0)\ ,$$

$$y_{2i} = \beta_2' x_i + \sigma_{2i}\varepsilon_{2i}\ , \quad \varepsilon_{2i}|W_i \sim \mathcal{D}_2(0,1)\ , \quad \sigma_{2i}^2 = \exp\big(F_2(z_i,\alpha_2)\big)\ , \quad z_i \in W_i\ ,$$

$$\varepsilon_{1i} = \rho\varepsilon_{2i} + e_i\ , \qquad M(e_i|W_i,\varepsilon_{2i}) = 0\ ,$$

where $F_2(z_i,\alpha_2)$ is continuously differentiable with respect to α_2, $M(\cdot\mid\cdot)$ is the conditional median, and $\mathcal{D}_2(m,v)$ is an arbitrary distribution with mean m and variance v. Here, the conditional distribution of $y_{1i}^*|W_i,y_{2i}$ is specified only up to the conditional median. It is assumed that the support of $\mathcal{D}_2$ is the whole real line. The second equation is simply a regression model with heteroskedasticity and the parameters (β_2,α_2) can be consistently estimated by Gaussian quasi ML even when $\mathcal{D}_2$ is not normal (see Section 8.4.4 of Gourieroux and Monfor (1989) among others). Further, the semi-parametric approach of Klein and Vella (2010) may be adopted to describe σ_{2i}^2 such that the precise functional form $F_2(z_i,\alpha_2)$ need not be specified.

From (14), the conditional median of y_{1i}^* may be written as

$$(15)\qquad M(y_{1i}^*|W_i,y_{2i}) = (\beta_1 - \beta_2\rho/\sigma_{2i})' x_i + (\gamma_1 + \rho/\sigma_{2i})y_{2i}$$

$$= b_1 x_i + g_1 y_{2i} + g_2 y_{2i}/\sigma_{2i} + b_2 x_i/\sigma_{2i}$$

in terms of the structural and reduced-form parameters respectively. The sign of the conditional median (15) is intimately related to the conditional probability of observing the sign of y_{1i}^* (see Manski (1985)). In our context, a two-step approach may be used:

Step 1: estimate (β_2,α_2) from the second equation of (14) by the Gaussian quasi ML (with possibly semi-parametric σ_{2i}^2);

Step 2: estimate (β_1,γ_1,ρ) from (15) by the maximum score of Manski (1975, 1985) [1] , or the smoothed maximum score of Horowitz (1992), or the method of Lewbel (2000), using the estimated σ_{2i} and (β_2,α_2) from Step 1.

[1] Florios and Skouras (2008) suggest that the maximum score estimates be computed by mixed integer programming (MIP), which delivers the exact maximum point and computation efficiency.

The following theorem shows that the parameters in Step 2 are locally linear-median identified up to scale. Although the structure of (15) prevents a direct application of Lemma 2 of Manski (1985), his arguments there are closely followed in the proof of the following theorem. Notation-wise, define $\theta = [\beta_1', \gamma_1, \rho, \beta_2', \alpha_2']'$ and $h = [b_1', g_1, g_2, b_2', \beta_2', \alpha_2']'$ as the structural and reduced-form parameters respectively. True parameter values are stared.

THEOREM 4. *Assume that* (i) *the data* $(x_i, y_{1i}, y_{2i}, z_i)$ *are independent draws from* (14)*;* (ii) $E(x_i x_i'/\sigma_{2i}^2)$ *and* $E(f_{2i} f_{2i}')$ *are invertible at* h^**;* (iii) $P(\mathcal{A}) > 0$ at h^*, *where the event* $\mathcal{A} = \{(d_1 + d_2/\sigma_{2i} + d_3/\sigma_{2i}^2)' x_i \neq 0$ *for any non-zero constant vector* $[d_1', d_2', d_3']'\}$*;* (iv) θ^* *is an interior point with one of* (γ_1^*, ρ^*) *being non-zero. Then,* $[\beta_2^{*\prime}, \alpha_2^{*\prime}]'$ *is locally identified in Step 1 and* $[b_1^{*\prime}, g_1^*, g_2^*, b_2^{*\prime}]'$ *is locally identified up to scale in Step 2. Further,* $[\beta_1^{*\prime}, \gamma_1^*, \rho^*]'$ *is locally identified up to scale in Step 2.*

These sufficient conditions are derived by choosing y_{2i} as "the regressor with non-zero coefficient" (Manski's x_K). This choice appears natural for the cases with endogeneity (hence $\gamma_1^* \neq 0$). The same result can be obtained if an exogenous regressor in x_i (with certain additional assumptions) is chosen as Manski's "x_K".

Similar to the probit model in Section 2.2, condition (ii) provides the identification for the second equation in (14), whereas conditions (iii)-(iv) ensure the identification of (15). In addition to θ^* being an interior point, condition (iv) requires that one of (γ_1^*, ρ^*) be non-zero. Given that endogeneity is the focus, this requirement is not restrictive. Condition (iii) imposes certain restrictions (stronger than those in Theorem 2) on the functional form of σ_{2i}^2. It rules out the cases where σ_{2i} or σ_{2i}^2 is a linear function of x_i.

EXAMPLE 5. The statements of Example 2 are valid in the current context. Additionally, for the case $\sigma_{2i}^2 = \alpha_{20} + \alpha_{21} x_{1i}$ with $\alpha_{21} \neq 0$ (positivity assumed for σ_{2i}^2), condition (iii) in Theorem 4 fails to hold because $P[(d_1' + d_3'/\sigma_{2i}^2) x_i = 0] = 1$ for $d_1 = [-1, 0]'$ and $d_3 = [\alpha_{20}, \alpha_{21}]'$.

5 Conclusion

We show that heteroskedasticity may be exploited for identifying structural parameters in binary choice models with endogeneity, when instruments or exclusion restrictions are not available. Local identification is found to be generally achievable in the presence of heteroskedasticity. Formal conditions for local identification are provided for parametric models (such as probit and logit) and the semi-parametric model of Manski (1975, 1985). For these models, once parameter identification is established, estimation methods are available for inference purposes.

The results in this paper may be further extended to the cases where multiple endogenous regressors are included in the binary choice equation. One material issue in such cases is the handling of the correlations among these endogenous regressors. These correlations are expected to be relevant for the conditions for parameter identification.

6 Appendix

PROOF OF THEOREM 1. Notice the derivatives: $dq/d\rho = \rho q^3$ and $d(\rho q)/d\rho = q^3$. The Jacobian of the map (7) from the structural parameters $\theta = [\beta_1', \gamma_1, \rho, \beta_2', \alpha_2']'$ to the reduced-form parameters $h = [b_1', g_1, g_2, b_2', \beta_2', \alpha_2']'$ is

$$J = \frac{\partial h}{\partial \theta'} = \begin{vmatrix} I_k q & 0 & \beta_1 \rho q^3 & 0 & 0 \\ 0 & q & \gamma_1 \rho q^3 & 0 & 0 \\ 0 & 0 & q^3 & 0 & 0 \\ 0 & 0 & -\beta_2 q^3 & -I_k \rho q & 0 \\ 0 & 0 & 0 & I_k & 0 \\ 0 & 0 & 0 & 0 & I_{k_\alpha} \end{vmatrix}$$

where I_k is the k-dimensional identity matrix, k is the size of x_i and k_α the size of α. The conclusion holds because the size of J is $(3k + k_\alpha + 2) \times (2k + k_\alpha + 2)$ and there exists an invertible $(2k + k_\alpha + 2) \times (2k + k_\alpha + 2)$ sub-matrix (on the deletion of the fourth block row corresponding to b_2). The second claim is true because the Jacobian is continuous (matrix) function of θ. ■

PROOF OF THEOREM 2. The following statements are all made in reference to the true parameter point h^* (or θ^*). By assumption (i), the information matrix is

$$E\frac{\partial L_n}{\partial h}\frac{\partial L_n}{\partial h'} = E\sum_{i=1}^{n} s_i(h)s_i(h)' = nE\{E[s_i(h)s_i(h)'|W_i, y_{2i}]\},$$

which is a multiple of the matrix in (9). Given (ii), the matrix in (9) is invertible if and only if the sub-matrix on deleting the second block row and column is invertible, which can be written as (noting that $E(\varepsilon_{2i}^4) = 3$)

$$E\{\Lambda_i \begin{bmatrix} X_i \\ -\frac{G_{2i}}{2\sigma_{2i}} f_{2i} \end{bmatrix} \left[X_i', -\frac{G_{2i}}{2\sigma_{2i}} f_{2i}'\right]\} + \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{2}E(f_{2i}f_{2i}') \end{bmatrix}.$$

The above matrix is not invertible if and only if there is a non-zero constant vector $[C', D_2']'$ such that

$$E\{\Lambda_i \left(C'X_i - \frac{G_{2i}}{2\sigma_{2i}} D_2' f_{2i}\right)^2\} + \frac{1}{2}E\{(D_2' f_{2i})^2\} = 0,$$

which together with condition (ii) forces $D_2 = 0$. The information matrix is not invertible if and only if $C'X_i = 0$ with probability 1 for a non-zero constant vector $C = [c_1', c_2, c_3, c_4']'$, i.e.,

$$\left(c_1' + \frac{c_4'}{\sigma_{2i}}\right)x_i + \left(c_2 + \frac{c_3}{\sigma_{2i}}\right)y_{2i} = 0$$

with probability 1. The event $\mathcal{B} = \{c_2 + c_3/\sigma_{2i} \neq 0$ for any $[c_2, c_3] \neq 0\}$ is implied by the event $\mathcal{A} = \{(c_1' + c_4'/\sigma_{2i})x_i \neq 0$ for any $[c_1', c_4'] \neq 0\}$. Conditional on $\mathcal{A}$, the probability of $\mathcal{C} = \{C'X_i = 0$ for some $C \neq 0\}$ is zero because ε_{2i} is continuous and independent of W_i. Conditional on the complement $\bar{\mathcal{A}}$, the probability of $\mathcal{C}$ is one. It follows that $P(\mathcal{C}) = P(\mathcal{C}|\mathcal{A})P(\mathcal{A}) + P(\mathcal{C}|\bar{\mathcal{A}})P(\bar{\mathcal{A}}) = P(\bar{\mathcal{A}}) < 1$ by (iii). Therefore the information matrix is invertible at h^*. Because the determinant of a matrix is a continuous function of its elements and the elements in (9) are continuous functions of h, the true parameter point h^*, being an interior point of the parameter space by (iv), is a regular point of the matrix in (9). Therefore, by Theorem 1 of Rothenberg (1971), h^* is locally identified. Finally, by Theorem 1 and Rothenberg's Theorem 6, the structural parameter point θ^* is also locally identified. ■

PROOF OF THEOREM 3. The proof is similar to that of Theorem 2. Re-define $X_i' = [x_i'/\sigma_{1i}, y_{2i}/\sigma_{1i}, y_{2i}/\sigma_{2i}, x_i'/\sigma_{2i}]$ and $h' = [B', \alpha_1, \beta_2', \alpha_2']$, where $B' = [b_1', g_1, g_2, b_2']$. Expression (8) is applicable and the scores for the observation with index i are given by

$$\frac{\partial l_i(h)}{\partial B} = \lambda_i X_i \ , \qquad \frac{\partial l_i(h)}{\partial \alpha_1} = -\frac{\lambda_i G_{1i}}{2\sigma_{1i}} f_{1i} \ ,$$

$$\frac{\partial l_i(h)}{\partial \beta_2} = \frac{1}{\sigma_{2i}^2} u_{2i}(\beta_2) x_i \ , \qquad \frac{\partial l_i(h)}{\partial \alpha_2} = \frac{1}{2}\left[\left(\frac{u_{2i}^2(\beta_2)}{\sigma_{2i}^2} - 1\right) - \frac{\lambda_i G_{2i}}{\sigma_{2i}}\right] f_{2i} \ ,$$

where $\lambda_i = \frac{y_{1i}\phi_i}{\Phi_i} - \frac{(1-y_{1i})\phi_i}{1-\Phi_i}$, $G_{ji} = b_j' x_i + g_j y_{2i}$ and $f_{ji} = \partial F_j(z_i, \alpha_j)/\partial \alpha_j$ for $j = 1,2$. Also, $E(\lambda_i | W_i, y_{2i}) = 0$ and $E(s_i(h) | W_i, y_{2i}) = 0$ hold at θ^*. The information matrix at $h = h^*$ is

$$E\{E[s_i(h)s_i(h)' | W_i, y_{2i}]\} =$$

$$E\begin{bmatrix} \Lambda_i X_i X_i' & -\frac{\Lambda_i G_{1i}}{2\sigma_{1i}} X_i f_{1i}' & 0 & -\frac{\Lambda_i G_{2i}}{2\sigma_{2i}} X_i f_{2i}' \\ -\frac{\Lambda_i G_{1i}}{2\sigma_{1i}} f_{1i} X_i' & \frac{\Lambda_i G_{1i}^2}{4\sigma_{1i}^2} f_{1i} f_{1i}' & 0 & \frac{\Lambda_i G_{1i} G_{2i}}{4\sigma_{1i}\sigma_{2i}} f_{1i} f_{2i}' \\ 0 & 0 & \frac{1}{\sigma_{2i}^4} u_{2i}^2(\beta_2) x_i x_i' & 0 \\ -\frac{\Lambda_i G_{2i}}{2\sigma_{2i}} f_{2i} X_i' & \frac{\Lambda_i G_{1i} G_{2i}}{4\sigma_{1i}\sigma_{2i}} f_{2i} f_{1i}' & 0 & \frac{1}{4}[(\frac{u_{2i}^2(\beta_2)}{\sigma_{2i}^2} - 1)^2 + \frac{\Lambda_i G_{2i}^2}{\sigma_{2i}^2}] f_{2i} f_{2i}' \end{bmatrix},$$

where $\Lambda_i = E(\lambda_i^2 | W_i, y_{2i}) = \frac{\phi_i^2}{\Phi_i} + \frac{\phi_i^2}{1-\Phi_i} > 0$ and the fact $E(\varepsilon_{2i}^3) = 0$ is used. The following statements are all made in reference to the true parameter point $h = h^*$. The structure of the information matrix indicates that it is invertible if and only if the sub-matrix on deleting the third block row and column is invertible, given that $E(x_i x_i' / \sigma_{2i}^2)$ is of full rank by (ii). This sub-matrix can be expressed as (noting that $E(\varepsilon_{2i}^4) = 3$)

$$E\{\Lambda_i \begin{bmatrix} X_i \\ -\frac{G_{1i}}{2\sigma_{1i}} f_{1i} \\ -\frac{G_{2i}}{2\sigma_{2i}} f_{2i} \end{bmatrix} \left[X_i', -\frac{G_{1i}}{2\sigma_{1i}} f_{1i}', -\frac{G_{2i}}{2\sigma_{2i}} f_{2i}'\right]\} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} E(f_{2i} f_{2i}') \end{bmatrix}.$$

The information matrix is not invertible if and only there is a non-zero constant vector $[C', D_1', D_2']'$ such that

$$E\{\Lambda_i \left(C' X_i - \frac{G_{1i}}{2\sigma_{1i}} D_1' f_{1i} - \frac{G_{2i}}{2\sigma_{2i}} D_2' f_{2i}\right)^2\} + \frac{1}{2} E[(D_2' f_{2i})^2] = 0 \ ,$$

which together with condition (ii) forces $D_2 = 0$. Hence, it is not invertible if and only if

$$\left(\frac{c_1'}{\sigma_{1i}} + \frac{c_4'}{\sigma_{2i}} - \frac{D_1' f_{1i}}{2\sigma_{1i}} b_1'\right) x_i + \left(\frac{c_2}{\sigma_{1i}} + \frac{c_3}{\sigma_{2i}} - \frac{D_1' f_{1i}}{2\sigma_{1i}} g_1\right) y_{2i} = 0$$

with probability 1 for some constant vector $[C', D_1'] = [c_1', c_2, c_3, c_4', D_1'] \neq 0$. Note that the event $\mathcal{B} = \{c_2/\sigma_{1i} + c_3/\sigma_{2i} - \frac{1}{2} D_1' f_{1i} g_1/\sigma_{1i} \neq 0$ for any $[c_2, c_3, D_1'] \neq 0\}$ is implied by the event $\mathcal{A} = \{(c_1'/\sigma_{1i} + c_4'/\sigma_{2i} - \frac{1}{2} D_1' f_{1i} b_1'/\sigma_{1i}) x_i \neq 0$ for any $[c_1', c_4', D_1'] \neq 0\}$. Conditional on $\mathcal{A}$, the probability of the above display is zero because ε_{2i} is continuous and independent of W_i. Conditional on $\bar{\mathcal{A}}$, the probability of the above display is one. It follows that the unconditional probability of the above display is $P(\bar{\mathcal{A}}) < 1$ by (iii). Hence, the information matrix is invertible. Further, as in the proof of Theorem 2, h^* is a regular point of the information matrix by (iv). Hence, h^* is locally identified by Theorem 1 of Rothenberg (1971). Finally, θ^* is locally identified by Theorem 1 and Rothenberg's Theorem 6. ■

PROOF OF THEOREM 4. For Step 1, because $y_{2i} = \beta_2' x_i + \sigma_{2i}\varepsilon_{2i}$ is simply a linear regression with heteroskedasticity, $[\beta_2^{*\prime}, \alpha_2^{*\prime}]'$ is identified under (i) and (ii). For Step 2 identification, $[\beta_2^{*\prime}, \alpha_2^{*\prime}]'$ is fixed and the proof follows closely the proof of Lemma 2 of Manski (1985). Let $B^* = [b_1^{*\prime}, g_1^*, g_2^*, b_2^{*\prime}]'$ be the true parameter point and B be a local alternative point that is not a scalar multiple of B^*. The goal is to show that $P(\mathcal{B}_1) + P(\mathcal{B}_2) > 0$, where $\mathcal{B}_1 = \{B'X_i < 0 < B^{*\prime}X_i\}$ and $\mathcal{B}_2 = \{B^{*\prime}X_i < 0 < B'X_i\}$ are the events that B and B^* lead to different outcomes of y_{1i} conditional on $X_i = [x_i', y_{2i}, y_{2i}/\sigma_{2i}, x_i'/\sigma_{2i}]'$. Let $G_i = g_1 + g_2/\sigma_{2i}$ and $G_i^* = g_1^* + g_2^*/\sigma_{2i}$. All probability statements discussed below are conditional on the event $G_i^* > 0$ implicitly (the cases of $G_i^* < 0$ are similar). The conclusions hold unconditionally since $P(G_i^* \neq 0) > 0$ by (iii) and (iv). Let $\mathcal{A}^- = \{G_i < 0\}$, $\mathcal{A}^0 = \{G_i = 0\}$ and $\mathcal{A}^+ = \{G_i > 0\}$. We note

$$P(\mathcal{B}_1) = P(\mathcal{B}_1 \cap \mathcal{A}^-) + P(\mathcal{B}_1 \cap \mathcal{A}^0) + P(\mathcal{B}_1 \cap \mathcal{A}^+)$$

and similarly for $P(\mathcal{B}_2)$. Define $\tilde{X}_i = [x_i', x_i'/\sigma_{2i}]'$ and $\tilde{B} = [b_1', b_2']'$. First, when $P(\mathcal{A}^-) > 0$,

$$P(\mathcal{B}_1 \cap \mathcal{A}^-) = P\{-\tilde{B}'\tilde{X}_i/G_i < y_{2i},\ -\tilde{B}^{*\prime}\tilde{X}_i/G_i^* < y_{2i} | \mathcal{A}^-\}P(\mathcal{A}^-) > 0$$

and similarly $P(\mathcal{B}_2 \cap \mathcal{A}^-) > 0$ because $\varepsilon_{2i}|W_i$ has positive tail probabilities. Second, when $P(\mathcal{A}^+) > 0$, either

$$P(\mathcal{B}_1 \cap \mathcal{A}^+) = P\{-\tilde{B}^{*\prime}\tilde{X}_i/G_i^* < y_{2i} < -\tilde{B}'\tilde{X}_i/G_i | \mathcal{A}^+\}P(\mathcal{A}^+) > 0 \quad \text{or}$$

$$P(\mathcal{B}_2 \cap \mathcal{A}^+) = P\{-\tilde{B}'\tilde{X}_i/G_i < y_{2i} < -\tilde{B}^{*\prime}\tilde{X}_i/G_i^* | \mathcal{A}^+\}P(\mathcal{A}^+) > 0$$

holds because $P(\tilde{B}'\tilde{X}_i/G_i \neq \tilde{B}^{*\prime}\tilde{X}_i/G_i^*) > 0$ by the fact that B is not a scalar multiple of B^* and $P\{(b_1' + b_2'/\sigma_{2i})x_i/G_i \neq (b_1^{*\prime} + b_2^{*\prime}/\sigma_{2i})x_i/G_i^*\} > 0$. The latter probability is positive since the event in the brackets is equivalent to $Q(\sigma_{2i}^{-1})x_i \neq 0$, where $Q(\cdot)$ is a quadratic function with constant vector coefficients, and $P[Q(\sigma_{2i}^{-1})x_i \neq 0] > 0$ by (iii) and (iv). Then,

$$P(\mathcal{B}_1) + P(\mathcal{B}_2) \geq P(\mathcal{B}_1 \cap \mathcal{A}^-) + P(\mathcal{B}_1 \cap \mathcal{A}^+) + P(\mathcal{B}_2 \cap \mathcal{A}^-) + P(\mathcal{B}_2 \cap \mathcal{A}^+) > 0$$

because $P(\mathcal{A}^0) < 1$ (hence one of $P(\mathcal{A}^-)$ and $P(\mathcal{A}^+)$ must be positive) in a neighbourhood of θ^* by (iii) and (iv). Finally, as the map from $[\beta_1', \gamma_1, \rho]'$ to $[b_1', g_1, g_2, b_2']'$ is given by $b_1 = \beta_1$, $g_1 = \gamma_1$, $g_2 = \rho$ and $b_2 = -\rho\beta_2$ for given β_2, the Jacobian is clearly of full rank in an open neighbourhood of θ^* by (v). Hence, $[\beta_1^{*\prime}, \gamma_1^*, \rho^*]'$ is identified according to Rothenberg's Theorem 6. ■

7 Reference

Amemiya, T. (1985), Advanced Econometrics, Harvard University Press, Cambridge.

Florios, K. and S. Skouras (2008), Exact computation of max weighted score estimators, *Journal of Econometrics*, 146, 86-91.

Freedman, D.A. and J.S. Sekhon (2010), Endogineity in probit response models, *Political Science*, 18(2), 138-150.

Gourieroux, C., Monfort. A., (1989), Statistics and Econometric Models (v1), Cambridge University Press, Cambridge.

Heckman, J.J. (1978), Dummy endogenous variables in a simultaneous equation system, *Econometrica*, 46, 931-959.

Horowitz, J.L. (1992), A smoothed maximum score estimator for the binary response model, *Econometrica*, 60(3), 505-531

Klein, R. and F. Vella (2010), Estimating a class of triangular simultaneous equations models without exclusion restrictions. *Journal of Econometrics*, 154, 154-164.

Lewbel, A. (2000), Semiparametric quanlitative response model estimation with unknown heteroskedasticity or instrumental vairiables, *Journal of Econometrics*, 97, 145-177.

Lewbel, A. (2012), Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models, *Journal of Business and Economic Statistics*, 30, 67-80.

Manski, C.F. (1975), Maximum score estimation of the stochastic utility model of choice, *Journal of Econometrics*, 3, 205-228.

Manski, C.F. (1985), Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator, *Journal of Econometrics*, 27, 313-333.

Milunovich, G. and M. Yang (2013), Simultaneous equation systems with heteroskedasticity: identification, estimation, and stock price elasticities, Working paper

R Core Team (2013), R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/.

Rigobon, R. (2003), Identification through heteroskedasticity, *The Review of Economics and Statistics*, 85, 777-792.

Rivers, D. and Q. Vuong (1988), Limited information estimators and exogeneity tests for simultaneous probit models, *Journal of Econometrics*, 39, 347-366.

Rothenberg, T. J. (1971), Identification in Parametric Models, *Econometrica* 39, 577-591.