Simultaneous Sample Selection Models*

William Matcham†

June 27, 2022

Preliminary and Incomplete. Do not circulate.

Abstract

I extend sample selection models by allowing the outcome to affect selection directly. After a simple model microfoundation, I provide identification and estimation results for the case of jointly normal errors. The simultaneity between the outcome and selection generates additional endogeneity, and, unlike traditional sample selection models, my identification result requires an excluded regressor in the *outcome* equation. Simulations confirm the finite sample performance of the new estimator and show sizeable differences in parameters compared to models that do not account for the direct effect of the outcome on the selection decision. I finish with an application to the examination process for patents and their potential quality. I show that traditional sample selection methods understate the positive effect of the inventing firm's size on patent quality.

Keywords: Endogenous sample selection, Heckman Selection, Structural Parameters, Patent Application Process

JEL Classification: C01, C13, C21, C25, C31, C35, J22, J31, O31, O34

*I especially thank Mark Schankerman for many discussions relating to this paper. I also thank Tim Cejka, Roberto Maura-Rivero, Taisuke Otsu, and Marcia Schafgans for comments and suggestions.

†Department of Economics, London School of Economics. Email: w.o.matcham@lse.ac.uk

1 Introduction

Economic datasets are rarely a random sample from the population. Resultantly, a rich literature in economics and other social sciences explicitly models the sample selection process. The typical approach projects the outcome variable and sample selection utility onto a set of exogenous regressors and, in doing so, does not estimate a direct effect of the outcome on the selection decision. In this paper, I address this gap in the literature, analyzing sample selection models that include the dependent variable in the selection equation. My central and novel contribution is to provide sufficient conditions for identification of the complete set of model parameters alongside consistent and asymptotically normal estimation methods. Through an empirical example in the context of the patent application process, I illustrate the importance of my contribution.

As the natural starting point, I consider a model where the two structural equations are linear in parameters, the outcome variable is continuous, and the errors are jointly normal. The model contains the traditional sample selection issue: the researcher only observes the outcome for an endogenously selected subsample. The novelty is that the value of the outcome itself *directly* affects the selection decision. These two features combined create a simultaneity issue, rendering traditional identification results invalid. My identification result requires an excluded variable in the outcome equation, a condition stronger than the norm in sample selection models.

After identification, I explain how to estimate all parameters via likelihood methods. The likelihood is similar to the one from traditional sample selection estimation, except the contribution to the likelihood by those not selected consists entirely of reduced form parameters. Additionally, I show how to estimate the outcome equation parameters using a two-step regression method, analogous to Heckman (1976). I follow with simulations, which (1) confirm the finite-sample performance of the estimator and (2) illustrate substantial differences between the new estimator and existing methods when a direct effect exists.

I finish with an application relating to the patent system. In that context, researchers study patent quality (scope), *conditional* on obtaining a patent (Lanjouw and Schankerman, 2004; Kuhn and Thompson, 2019; Feng and Jaravel, 2020). But inventors only enjoy patent protection if the expected utility from a given level of scope exceeds the utility from alternative intellectual property protection such as trade secrecy. Allowing for a direct effect of potential patent quality on the selection decision, I find a more substantial positive impact of the inventing firm's size on patent quality. Existing methods, which correct only for *unobservables* driving both patent quality and selection, find a modest difference in patent quality between small and large firms because they miss the fact that, all else equal, small firms must have especially high-quality patents to be selected. My application emphasizes the importance of including a direct effect of the outcome on the selection equation where appropriate.

2 Related Literature

This paper relates to an influential literature on sample selection models, which started by analyzing women's labour force participation (Heckman, 1974; Gronau, 1974; Lewis, 1974).[1] Seminal work by Heckman (Heckman, 1976, 1979, 1990) formalized the idea in these empirical examples, motivating several extensions as surveyed in Vella (1998). My paper generalizes traditional selection models in a way made precise in Section 4.1.

In my setup, since the outcome directly affects the selection decision, and observing the outcome depends on the selection decision, the model relates to simultaneous equation models. Amemiya (1974) considers simultaneous equation models with truncated dependent variables rather than endogenously selected ones. Amemiya (1984) provides a general survey of Tobit models. The closest to my setup is Amemiya's "Type 2".[2] In another related paper, Heckman (1978) presents a broad range of simultaneous equation models, including one with endogenous dummy variables in a simultaneous equation system. My setup resembles Heckman's, except that the outcome is always observed in the latter.

Recent work on endogenous sample selection models splits into two broad categories. One part of the literature relaxes the linearity of the outcome and selection equations and the parametric distributional assumptions made on the error terms, studying semiparametric and nonparametric sample selection models (Ahn and Powell, 1993; Andrews and Schafgans, 1998; Das, Newey, and Vella, 2003). As a starting point, I focus on linear equations and parametric distributional assumptions for the error term, but nonlinear and nonparametric extensions are possible. I discuss extensions in the conclusion.

The second, and even more recent, strand of the literature on sample selection models focuses on robustness. For example, Bastos, Barreto-Souza, and Genton (2022) and Carlson (2022) account for heteroskedasticity, Marchenko and Genton (2012) addresses heavy tails through the use of student-t errors, Ogundimu and Hutton (2016) studies skewness of outcomes, and Bastos and Barreto-Souza (2021) uses the bivariate Birnbaum-Sanders distribution to address nonnegativity

[1]For a detailed textbook discussion of self-selectivity, see Maddala (1986) and Gourieroux (2000), or shorter summaries in Amemiya (1984, 1985).

[2]This is a model given by $y_{1i}^* = x_{1i}'\beta_1 + u_{1i}$ and $y_{2i}^* = x_{2i}'\beta_2 + u_{2i}$ with $y_{2i} = y_{2i}^* y_{1i}$ and $y_{1i} = 1(y_{1i}^* > 0)$.

of outcome variables. Roelsgaard and Taylor (2022) proposes a semiparametric machine learning estimator for sample selection models. All of these potential issues or novelties persist in my setup, and future work can check if the proposed solutions carry through.

There are numerous *applications* of traditional sample selection models.[3] In ongoing work, Matcham and Schankerman (2022) adds an equation for post-patenting outcomes to the two-equation model of patent scope and selection I estimate in my application.

3 Motivating Examples

Below, I describe four economic examples that motivate my model.

3.1 Wages and labor Force Participation

A common motivation for the sample selection model is that wages are only observed for those who participate in the labor force, and factors that determine labor force participation correlate with factors that determine wages. In particular, the original structural model motivating the sample selection literature *does* include the outcome directly in the selection equation. The three equations are

$$W_M = X'\beta_X + X_M'\beta_M + \varepsilon_M \tag{1}$$

$$W_R = X'\alpha_X + X_R'\alpha_R + \varepsilon_R \tag{2}$$

$$D = 1\left[W_M > W_R\right], \tag{3}$$

where W_M denotes market wage, W_R the reservation wage, D is an indicator of labor force participation, and $(\varepsilon_M, \varepsilon_R)$ are jointly normal. In particular, researchers only observe market wages when they exceed individuals' rarely-observed reservation wage. Substituting W_R from equation (2) into (3) yields

$$D = 1\Bigg[W_M\kappa_M + X'\kappa_X + X_R'\kappa_R - \varepsilon_R > 0\Bigg], \tag{4}$$

[3] Among others between 1970 and 1985, see: Heckman (1974, 1979), Nelson (1977), Cogan (1981), Hanoch (2014, 1976) Gordon and Blinder (1980), Griliches, Hall, and Hausman (1978), Kenny, Lee, Maddala, and Trost (1979), Willis and Rosen (1979), Lee (1978), Abowd and Farber (1982), Katz (1977), Nakosteen and Zimmer (1980), Poirier and Ruud (1981), Weisbrod (1980), Roberts, Maddala, and Enholm (1978), Trost (1977), Lee (1978), Rosen (1979), and King (1980).

with $(\kappa_M, \kappa_X, \kappa_R) = (1, -\alpha_X, -\alpha_R)$. This is similar to the econometric model I study. Further, substituting W_M from (1) into (4) gives

$$\begin{aligned} W_M &= X'\beta_X + X_M'\beta_M + \varepsilon_M \\ D &= 1\left[Z'\gamma + \varepsilon_D > 0\right], \end{aligned}$$

where $Z = (X, X_M', X_R')'$, $\gamma = (\beta_X' - \alpha_X', \beta_M', -\alpha_R')'$, and $\varepsilon_D = \varepsilon_M - \varepsilon_R$. This is the typical endogenous sample selection model, with X_R *and* all excluded regressors X_M added to the selection equation . The index $Z'\gamma + \varepsilon_D$ is a reduced form representation of the difference between market wage W_M and reservation wage W_R. On common regressors, the parameters of the selection equation are the difference of the structural parameters β_X and α_X. Further, because $\varepsilon_D = \varepsilon_M - \varepsilon_R$, the estimated covariance is not $\mathrm{cov}(\varepsilon_M, \varepsilon_R)$ but instead $\mathrm{cov}(\varepsilon_M, \varepsilon_M - \varepsilon_R)$. For the case of jointly normal errors, my model will estimate all parameters in the model given by (1) - (3) along with $\mathrm{cov}(\varepsilon_M, \varepsilon_R)$.

3.2 Entry Games

Consider an incumbent firm deciding whether to block ($D = 0$) or allow ($D = 1$) a potential firm's entry. If the new firm enters, they choose the optimal production quantity Y_N, and the incumbent chooses the quantity Y_I that maximizes their duopoly profit $\pi_d(Y_I, Y_N)$. If the potential entrant gets blocked, the incumbent obtains monopoly profit π_m, which includes a blocking cost. The incumbent blocks if $\pi_d(Y_I, Y_N) > \pi_m$. Hence the decision to block depends on Y_N. A researcher either analyzing multiple markets of this nature, or analyzing the same market over many periods will only observe the entrant's output Y_N if $D = 1$, and the value Y_N directly affects the incumbent's decision to block. Therefore, researchers can estimate the structural parameters of the incumbent's blocking choice and the potential entrant's output using the econometric model I analyze in this paper.

3.3 Patent Scope

To apply for a patent in the United States, inventors must submit an application to the United States Patent Trademark Office. First, the Office assigns an examiner to the application. Then, the examiner debates with the applicant (and their attorney) on the allowable level of patent protection (scope) to avoid infringing on existing patented inventions.[4] An often-missed detail about the patent application process is that the examiner cannot outright end the application process. Instead, the process can only end when the applicant meets the examiner's demands or the applicant abandons. Relating the process to sample selection models, we only observe patent

[4]See Graham, Marco, and Miller (2018) for more detail about the application process.

scope (Y) when the applicant does not abandon (D), and the applicant's decision to abandon is a function of the allowed patent quality, should it realize.

3.4 Survey Data Response

In settings where researchers collect sensitive survey data, the outcome variable may directly affect individuals' willingness and ability to be surveyed. Suppose a hospital is analyzing alcohol consumption in the community. The value of alcohol consumption directly affects willingness to participate, yet the hospital can only study the alcohol consumption of those who are willing and able to participate. As another example, research analyzing the correlates of domestic abuse know that domestic abuse itself has an effect on the willingness to report exposure.

4 Economic and Econometric Models

An individual decides whether to obtain an outcome variable, with their choice recorded in the dummy variable D. The potential amount of the outcome is $Y = g_Y(Z_Y, \varepsilon_Y)$, where Z_Y are exogenous features of the individual that affect the outcome quantity and ε_Y is a random shock affecting the outcome level.

The difference between the benefits and costs of realizing the outcome is $f(Y, Z_D, \varepsilon_B)$, where Z_D are exogenous features of the individual affecting benefits and costs, and ε_B is a random shock. The outside option delivers utility ε_D.[5] When the individual makes their decision, ε_D and ε_Y have realized but, since the individual is yet to know the full benefits, ε_B has not.[6] The individual chooses $D = 1$ if

$$\mathbb{E}_{\varepsilon_B|(Y,\varepsilon_D,Z_D,Z_Y)}\left[f(Y, Z_D, \varepsilon_B) - \varepsilon_D\right] > 0 \Longleftrightarrow g_D(Y, Z_D, \varepsilon_D) > 0.$$

The full model is thus

$$\begin{aligned} Y &= g_Y(X, X_Y, \varepsilon_Y) \\ D &= 1\left[g_D(Y, X, X_D, \varepsilon_D) > 0\right], \end{aligned}$$

[5]The outside option could be a function of exogenous variables as well as a random shock, but nothing is gained from this addition.

[6]For example, in the patent context, the inventor may know their level of allowable scope, but will not know the market returns that their invention will bring. Further, they won't know if their invention will become obsolete, so they won't know how long they will renew the patent rights for.

where I split Z_Y and Z_D into the common regressors X and excluded regressors X_Y and X_D.[7] The value of Y is only observed if $D = 1$, though I assume X_Y is always observed. This is the general model of the paper. Taking g_Y and g_D as linear (as I will in all that follows) the model is

$$\begin{aligned} Y &= \gamma_X' X + \gamma_Y' X_Y + \varepsilon_Y = \gamma' w_Y + \varepsilon_Y && (5) \\ D &= 1\left[\beta_X' X + \beta_D' X_D + \beta_Y Y + \varepsilon_D > 0\right] = 1\left[\beta_{-Y}' w_D + \beta_Y Y + \varepsilon_D > 0\right], && (6) \end{aligned}$$

where $\gamma = (\gamma_X', \gamma_Y')'$, $w_Y = (X', X_Y')'$, $\beta_{-Y} = (\beta_X', \beta_D')'$, and $w_D = (X', X_D')'$.

4.1 Relation to Standard Sample Selection Model

The remainder of the paper focuses on identification, estimation, and application of the model as given in (5) - (6). Before that it is befitting to compare this model to traditional endogenous sample selection models. To see this, substituting (5) into (6) yields

$$\begin{aligned} Y &= \gamma_X' X + \gamma_Y' X_Y + \varepsilon_Y \\ D &= 1\left[\kappa_X' X + \kappa_D' X_D + \kappa_Y' X_Y + \tilde{\varepsilon} > 0\right], \end{aligned} \tag{7}$$

where

$$\kappa_X = \beta_X + \beta_Y \gamma_X, \quad \kappa_D = \beta_D, \quad \kappa_Y = \beta_Y \gamma_Y \quad \tilde{\varepsilon} = \beta_Y \varepsilon_Y + \varepsilon_D.$$

This is a standard sample selection model, where the γ parameters are structural. But the parameters of the selection equation and the correlation between errors are reduced form are not structural. The traditional sample selection model will not estimate the β parameters. When $\beta_Y = 0$, the model collapses to the standard sample selection model, but the case of $\beta_Y \neq 0$ requires additional analysis. Further, if the researcher fails to include X_Y in the selection equation, traditional sample selection estimation methods will fail to estimate any of the structural parameters.

5 Identification

Now I turn to identification. Throughout, Y is continuous – I discuss identification with a discrete outcome briefly in Appendix B. I study identification of equations (5) and (6), where Y is observed if $D = 1$. The errors $(\varepsilon_Y, \varepsilon_D)$ are conditionally jointly normal distributed, and, letting $z = (X', X_D', X_Y')'$, I impose three normalizations: $\text{Var}(\varepsilon_D|z) = 1$ and $\mathbb{E}(\varepsilon_Y|z) = \mathbb{E}(\varepsilon_D|z) = 0$.

[7] I could combine exogenous regressors X and X_Y (respectively X and X_D) into a vector, say Z_Y (respectively Z_D), but for exposition, the identification result and proof that follows splits common and separate regressors.

Hence,

$$\begin{pmatrix} \varepsilon_Y \\ \varepsilon_D \end{pmatrix} \Bigg| z \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \\ \rho\sigma_Y & 1 \end{pmatrix} \right). \tag{8}$$

To understand the identification result, it helps to analyze the reduced form. Rearranging (7) for $\tilde{\varepsilon}$ and standardizing yields

$$D = 1\Bigg[\underbrace{\frac{\tilde{\varepsilon}}{\sqrt{v}}}_{\mathcal{N}(0,1)} > -\pi' z\Bigg] = 1\left[\frac{\beta_Y \varepsilon_Y + \varepsilon_D}{\sqrt{v}} > \tilde{z}\right], \tag{9}$$

where $v = \mathrm{Var}(\beta_Y \varepsilon_Y + \varepsilon_D) = \beta_Y^2 \sigma_Y^2 + 1 + 2\beta_Y \sigma_Y \rho$, $\pi = (\pi_X', \pi_D', \pi_Y')'$, $\tilde{z} = -\pi' z$, and

$$\pi_X = \frac{\beta_X + \beta_Y \gamma_X}{\sqrt{v}}, \quad \pi_D = \frac{\beta_D}{\sqrt{v}}, \quad \pi_Y = \frac{\beta_Y \gamma_Y}{\sqrt{v}}. \tag{10}$$

In what follows, I refer to π as the *standardized reduced form* parameters and z as the *exogenous* variables.

My identification result rests on two substantive conditions. First, I rule out perfect multicollinearity in exogenous regressors. Second, I require at least one regressor X_{Y_j} appearing in the equation for Y but not in the equation for D. This is different to the traditional sample selection model, which does not strictly require any excluded regressors. However, similar to traditional sample selection estimation methods, excluded regressors in the selection equation are also desirable, else identification only comes from the nonlinearity in the inverse Mills Ratio. The identification result finishes this section; I prove it in Appendix A.

Theorem 1 *Consider the model defined by (5), (6), and (8).*

Part 1 (γ): Assume that:

1. *(Moment Existence): Y is absolutely continuous with positive and finite variance*

2. *(No Perfect Multicollinearity): There are K_1 points denoted $\{z_{(k)}\}_{k=1}^{K_1}$ and $K_2 + 1$ points denoted $\{w_{Y(k)}, \lambda(z_{(k)})\}_{k=1}^{K_2+1}$ such that the matrices*

$$\mathcal{X}_1 = \begin{pmatrix} z_{(1)}' \\ z_{(2)}' \\ \vdots \\ z_{(K_1)}' \end{pmatrix}, \quad \mathcal{X}_2 = \begin{pmatrix} w_{Y(1)}' & \tilde{\lambda}(z_{(1)}) \\ w_{Y(2)}' & \tilde{\lambda}(z_{(2)}) \\ \vdots & \vdots \\ w_{Y(K_2+1)}' & \tilde{\lambda}(z_{(K_2+1)}) \end{pmatrix}$$

have full rank, where:

- $z_{(k)} = (X_k', X_{Dk}', X_{Yk}')' \in \mathbb{R}^{K_1}$
- $w_{Y(k)} = (X_k', X_{Yk}')' \in \mathbb{R}^{K_2}$
- $\tilde{\lambda}(z_{(k)}) = \dfrac{\phi(z_{(k)}'\pi)}{\Phi(z_{(k)}'\pi)} = \dfrac{\phi(\tilde{z}_{(k)})}{\Phi(-\tilde{z}_{(k)})}$ *and* π *are the standardized reduced form parameters as defined in (10)*

Then γ *is identified.*

Part 2 (F_ε *and* β*): (Excluded Regressor): Assume in addition that either* $\beta_Y = 0$*, or there exists* j *such that* $\gamma_{Yj} \neq 0$ *and* $X_{Yj} \neq 0$*. Then* β *and the parameters of the joint distribution of errors, denoted* F_ε*, (i.e.* σ_Y *and* ρ*) are identified.*

6 Estimation

In this section, I present a likelihood-based method to estimate the parameters of the econometric model as defined by (5) - (8), and I describe a regression-based approach to estimate the parameters of the outcome equation. I assume access to a random sample $\mathcal{W} = \{Y_i D_i, D_i, X_i, X_{Yi}, X_{Di}\}_{i=1}^n$ where observing $Y_i D_i$ is equivalent to observing Y_i only when $D_i = 1$.[8]

6.1 Likelihood Methods

Letting $\theta = (\gamma', \beta', \sigma_Y, \rho)'$, the conditional likelihood function for n independent observations is

$$\mathcal{L}(\theta; \mathcal{W}) = \prod_{i=1}^{n} \wp_{0i}^{1-D_i} \wp_{1i}^{D_i},$$

with the two $\wp$ terms defined below. The term $\wp_{0i}$ is the contribution to the likelihood when $D_i = 0$ and $\wp_{1i}$ represents the contribution when $D_i = 1$. Given the assumptions on $(\varepsilon_{Yi}, \varepsilon_{Di})$ and letting $z_i = (X_i', X_{Yi}', X_{Di}')'$, from (9)

$$\begin{aligned} \wp_{0i} &= \mathbb{P}\left(\frac{\beta_Y \varepsilon_{Yi} + \varepsilon_{Di}}{\sqrt{v}} \leq -\pi' z_i \middle| z_i \right) \\ &= \Phi\left(-\pi' z_i\right) \end{aligned} \tag{11}$$

[8] It is typical in sample selection models for the assigned value of Y when $D = 0$ to be irrelevant. That is the case here, though in a number of settings such as the patent scope example above, Y will take the value 0 when $D = 0$.

where the equality in (11) follows from

$$\beta_Y \varepsilon_{Yi} + \varepsilon_{Di} \sim \mathcal{N}_1\left(0, \underbrace{\beta_Y^2\sigma_Y^2 + 1 + 2\rho\beta_Y\sigma_Y}_{v}\right).$$

Equation (11) reveals that the contribution to the likelihood from individuals with $D_i = 0$ is a function of the standarized reduced form parameters π, so offers no assistance in separating structural parameters from reduced form. Next, as derived in section A.2,

$$\wp_{1i} = \frac{1}{\sigma_Y}\phi\left(\frac{Y_i - w'_{Yi}\gamma}{\sigma_Y}\right)\Phi\left(\frac{\beta'_{-Y}w_{Di} + \beta_Y Y_i + \frac{\rho}{\sigma_Y}(Y_i - w'_{Yi}\gamma)}{\sqrt{1-\rho^2}}\right).$$

The log-likelihood therefore has a closed form and yields, under standard MLE regularity conditions, consistent and asymptotically normal estimates of θ:

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{D} \mathcal{N}(0, V^{-1}),$$

with $V = -\mathbb{E}\left(\frac{\partial^2 \log(\mathcal{L})}{\partial\theta\partial\theta'}\right)$.

6.2 Regression Methods

Let $\lambda_i = \frac{\phi(\tilde{z}_i)}{\Phi(-\tilde{z}_i)}$, where $\tilde{z}_i = -\pi' z_i$. Following from the identification proof,

$$\mathbb{E}(Y_i | D_i = 1, z_i) = \gamma' w_{Yi} + \tau\lambda_i.$$

Resultantly, the outcome equation for the subset with $D_i = 1$ is

$$Y_i = \gamma' w_{Yi} + \tau\lambda_i + V_i, \tag{12}$$

where $\mathbb{E}(V_i | D_i = 1, z_i) = 0$. Hence, the following two-step procedure estimates γ and τ:

1. Run a probit regression of D on z. Let these reduced form estimates of π be $\hat{\pi}$.

2. Run a regression using data for $D = 1$ of Y on X, X_Y, and $\hat{\lambda}(z) = \frac{\phi(\hat{\pi}'z)}{\Phi(\hat{\pi}'z)} = \frac{\phi(\hat{\tilde{z}})}{\Phi(-\hat{\tilde{z}})}$.

In Appendix A, I provide further details on this procedure. In particular, under general conditions, discussed extensively in Amemiya (1973) and Jennrich (1969),

$$\sqrt{n_1}\begin{pmatrix} \hat{\gamma} - \gamma \\ \hat{\tau} - \tau \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, B\psi B'),$$

where asymptotics are $n, n_1 \to \infty$ with $\frac{n_1}{n} \to k \in (0,1)$. Here $n_1 = \sum_i D_i$ is the number of observations with $D_i = 1$ (the selected subsample),

$$B = \operatorname*{plim}_{n_1\to\infty}\left(\frac{W'_{Y+}W_{Y+}}{n_1}\right)^{-1},$$

where W_{Y+} is the matrix stacking X, X_Y and $\hat{\lambda}$, and

$$\psi = \operatorname*{plim}_{\substack{n\to\infty \\ n_1\to\infty}} \left[\frac{\sigma_Y^2}{n_1} \begin{pmatrix} \sum_i w_{Yi}w_{Yi}'\eta_i & \sum_i w_{Yi}\lambda_i\eta_i \\ \sum_i \lambda_i w_{Yi}'\eta_i & \sum_i \lambda_i^2\eta_i \end{pmatrix} + \right.$$
$$\left. \tau^2 \left(\frac{n_1}{n}\right) \begin{pmatrix} \frac{1}{n_1^2}\sum_i\sum_j w_{Yi}w_{Yj}'\vartheta_{ij} & \frac{1}{n_1^2}\sum_i\sum_j w_{Yi}\Upsilon_{ij} \\ \frac{1}{n_1^2}\sum_i\sum_j w_{Yi}'\Upsilon_{ij} & \frac{1}{n_1^2}\sum_i\sum_j \Omega_{ij} \end{pmatrix} \right],$$

where

$$\begin{aligned} \eta_i &= 1 + \frac{\tau^2(\tilde{z}_i\lambda_i - \lambda_i^2)}{\sigma_Y^2}, \\ \vartheta_{ij} &= \partial\lambda_i\,\partial\lambda_j\, z_i'V_{1st}z_j, \\ \Upsilon_{ij} &= \lambda_i\vartheta_{ij}, \\ \Omega_{ij} &= \lambda_i\lambda_j\vartheta_{ij}, \end{aligned}$$

V_{1st} is the asymptotic variance covariance matrix of $\hat{\pi}$ and $\partial\lambda_i = \partial\lambda_i/\partial\tilde{z}$ is the derivative of λ_i with respect to $\tilde{z}_i$. Conventional regression standard errors are invalid and must be adjusted to match $B\psi B'$ by replacing σ_Y, λ_i, and τ with appropriate estimators.

7 Simulation

Before an empirical application, I present Monte Carlo simulations. The simulations showcase the finite sample performance of the estimator compared to existing methods and confirm the intuitions I gave for the identification result. For three designs, I run 500 simulations of equations (5) - (8), each with a sample size of 2000.

Simulation 1: No Direct Effect

In all three simulations, variables are continuous and scalar-valued: X contains a constant and a regressor uniformly distributed on $[-5, 5]$, X_D is t-distributed with 10 degrees of freedom, and X_Y follows a logistic distribution with location and scale equal to two and one, respectively. I set $\beta_{-Y} = (\beta_0, \beta_X, \beta_D) = (1, 3, -8)'$, $\gamma = (\gamma_0, \gamma_X, \gamma_Y) = (3, -1, -2)$, $\rho = 0.5$, and $\sigma = 2$. To confirm that I nest traditional sample selection models, in Simulation 1 I set $\beta_Y = 0$ to rule out a direct effect of Y on the probability of treatment. Table 1 Panel A presents the simulation results and the selection probability. Row (1) provides means and standard deviation of the ML estimator as described in Section 6. Row (2) represents traditional sample selection methods, using MLE on a model imposing $\beta_Y = 0$.[9] Both methods estimate the parameters precisely, with small bias.

[9]Results from using the two-step Heckman estimator are almost identical to MLE results throughout.

Existing methods are marginally more precise: they benefit from not estimating β_Y, instead forcing it to equal 0.

Simulation 2: Direct Effect

The design of Simulation 2 mirrors Simulation 1, except β_Y is equal to two. Table 1 Panel B reports the results. Since the design meets the identification conditions, the new method estimates the parameters with minimal bias and standard deviation. Existing sample selection methods fail to estimate β. Appendix C Table 3 presents the results of a simulation similar to Simulation 2, except that X_Y is discrete. Since the identification result does not require continuous regressors, it is unsurprising that the new method still estimates parameters well, with slightly more noise owing to the reduced variation in the regressor.

Simulation 3: No Excluded Regressor

Finally, Simulation 3 mirrors Simulation 2 except with $\gamma_Y = 0$, so there is no excluded regressor. The results in Table 1 Panel C confirm the intuition of the identification result: both methods cannot estimate β_{-Y}, β_Y, or ρ. In particular, traditional sample selection methods estimate γ parameters correctly because $\gamma_Y = 0$ implies that the selection equation does not contain the omitted variable X_Y.

Table 1: Simulation Results

Panel A: No Direct Effect: $\bar{D} = 0.53$

Parameter:	σ	ρ	γ_0	γ_X	γ_Y	β_0	β_X	β_D	β_Y
True Value:	2	0.5	3	-1	-2	1	3	-8	0
(1) New:	2.06	0.49	3.00	-1.00	-2.00	1.04	3.08	-8.22	0.00
	(0.06)	(0.12)	(0.11)	(0.03)	(0.04)	(0.15)	(0.30)	(0.78)	(0.03)
(2) Existing:	2.00	0.50	3.00	-1.00	-2.00	1.03	3.07	-8.20	–
	(0.04)	(0.12)	(0.11)	(0.03)	(0.04)	(0.13)	(0.28)	(0.75)	

Panel B: Direct Effect: $\bar{D} = 0.47$

Parameter:	σ	ρ	γ_0	γ_X	γ_Y	β_0	β_X	β_D	β_Y
True Value:	2	0.5	3	-1	-2	1	3	-8	2
(1) New:	2.07	0.49	3.00	-1.00	-2.00	1.04	3.10	-8.27	2.07
	(0.08)	(0.13)	(0.09)	(0.02)	(0.04)	(0.26)	(0.51)	(1.31)	(0.35)
(2) Existing:	2.14	0.80	2.15	-1.00	-1.57	-0.12	0.12	-0.96	–
	(0.07)	(0.04)	(0.15)	(0.02)	(0.05)	(0.03)	(0.01)	(0.04)	

Panel C: No Excluded Regressor: $\bar{D} = 0.76$

Parameter:	σ	ρ	γ_0	γ_X	γ_Y	β_0	β_X	β_D	β_Y
True Value:	2	0.5	3	-1	0	1	3	-8	2
(1) New:	2.27	0.84	2.99	-1.01	-0.02	1.54	0.59	-2.62	0.31
	(0.39)	(0.18)	(0.23)	(0.25)	(0.24)	(0.64)	(0.88)	(1.64)	(0.74)
(2) Existing:	2.00	0.98	3.00	-1.00	0.00	1.53	0.22	-1.75	–
	(0.04)	(0.01)	(0.06)	(0.02)	(0.02)	(0.05)	(0.01)	(0.06)	

Notes: Table 1 reports the sample mean and sample standard deviations (in parentheses) of the estimates of the model parameters, over 500 repeated samples. The "New" row provide estimates from using the newly proposed model. The "Existing" row estimates a traditional sample selection model which forces $\beta_Y = 0$.

8 Application

Finally, I apply the methods described in this paper to estimate a model of patent quality (scope) and selection into patenting. Section 3.3 provides background on the patent application process, and explains why it fits into my model. I access data on approximately 300,000 patent applications filed to the United States Patent and Trademark Office between 2012-2014. For each application, I observe whether the applicant abandoned ($D = 0$) or obtained a patent ($D = 1$). As a measure of scope (Y), I use the number of independent claims on the granted patent.[10] Scope is not observed for patents that are not granted. I also observe a dummy for whether the firm applying has fewer than 500 employees (X), which is clearly determined before the realization of D. Firm size affects scope through many channels, for example, larger firms may have better quality inventions. Size also affects the decision to abandon because smaller firms may not be able to pay the legal fees associated with multiple rounds of negotiation with a patent examiner.

For excluded regressors I create two leniency measures for examiners assigned to applications. Leniency instruments are a common identification strategy in applied economic research.[11] In the case of patents, there is support in the literature that examiners are as good as randomly assigned to applications, that is, that applications of a certain type or quality are not assigned to specific examiners.[12] The first of the two leniency measures (X_D) is the average grant rate of the examiner assigned to the application. This is a common instrument used to explain the decision to patent (see e.g. Sampat and Williams (2019), Farre-Mensa, Hegde, and Ljungqvist (2020), Gaulé (2018), amongst others). The second measure of leniency is average number of independent claims on applications granted by the examiner assigned to the application. The "exclusion restriction" is that the leniency measure only affects the decision to patent through its effect on scope.

Table 2 provides the estimation results. A couple of findings warrant attention. First, the value of β_Y is statistically significant and, in being one fifth of the constant in the selection equation, implies that there is a direct effect of scope on the decision to abandon or obtain a patent. Second, the outcome equation coefficient on the small entity indicator in the general model is 9% larger than the respective coefficient in the traditional sample selection model. This implies that traditional sample selection models understate the negative effects of being a small firm on patent scope.

[10] This variable is one of the four indicators used in the index of patent quality in Lanjouw and Schankerman (2004).

[11] Kling (2006) is one of the original applications; see Frandsen, Lefgren, and Leslie (2019) for a non-exhaustive list of applications.

[12] I exclude examiners who have conducted fewer than 50 examinations, though results are robust to other threshold choices.

Traditional sample selection models ignore that the small firms who select into patenting must have especially high patent scope relative to small firms who abandon, because patent scope itself affects the patenting decision. Upon controlling for the direct effect, the positive effects of firm size are more apparent.

TABLE 2: Parameters: patent application process

Variable	**SS**	**New**
Patent issuance (D)		
CONSTANT (β_0)	-1.17 (0.01)	-1.63 (0.02)
SMALL ENTITY (β_X)	-0.38 (0.01)	-0.36 (0.01)
GRANT LENIENCY (β_{-D})	1.96 (0.01)	2.21 (0.03)
PATENT SCOPE (β_Y)	NA	0.30 (0.01)
Patent Scope (Y)		
CONSTANT (γ_0)	-1.21 (0.04)	-1.85 (0.02)
SMALL ENTITY (γ_X)	-0.67 (0.01)	-0.73 (0.01)
CLAIMS LENIENCY (γ_Y)	0.94 (0.01)	1.13 (0.01)
ρ	0.93 (0.00)	0.91 (0.00)
σ	1.74 (0.00)	1.82 (0.01)
n	332,199	332,199
$n_1 = \sum D_i$	149,523	149,523

Notes: Table 2 reports estimates from the patent application process model. The "SS" column provides estimates from using MLE on the standard sample selection model with no Y in the selection equation. The column labelled "New" presents estimates from the newly proposed estimation method as described in the paper. Standard errors are in parentheses next to coefficients.

9 Conclusion

In this paper, I study identification and estimation of the complete set of structural parameters in sample selection models where the outcome directly affects selection. I provide model identification conditions, along with maximum likelihood and two-step regression estimation methods. My simulations, and the empirical example relating to the patent application process, show the importance of including the outcome in the selection equation where appropriate. I recommend if practitioners believe that the outcome effects selection, they should use the methods I describe - even including all excluded regressors X_Y in the selection equation does not preserve the correlation or selection equation parameters.

This paper generates a couple of avenues for future research. First, as discussed in Appendix B, when the outcome variable Y is discrete, identification and estimation are more involved. Even with a simplification to the model so that Y^* enters the selection index D^* rather than Y, conditions for identification in this paper do not generalize. Also, though similar to derive, the likelihood requires simulation methods.

Second, throughout I considered *parametric* identification and estimation in a *linear* system. Extending the model to nonlinearity in Y and D, together with tackling semiparametric and nonparametric identification, is an avenue for further research.[13] I suspect that nonparametric identification requires more work: the first step towards identification in the parametric case is inverting the distribution of $\beta_Y \varepsilon_Y + \varepsilon_D$ to obtain the reduced form parameters. When $(\varepsilon_Y, \varepsilon_D)$ are jointly normal, the distribution of their linear form is normal. When the joint distribution of the errors is unknown (and even known but non-normal), this object is a convolution of unknown distributions and contains the unknown parameter β_Y. Making progress likely requires an entirely different approach.

[13] See Das, Newey, and Vella (2003) for nonparametric estimation of sample selection models.

References

The numbers at the end of every reference link to the pages citing the reference.

Abowd, J. M. and H. S. Farber (1982): "Job Queues and the Union Status of Workers," *ILR Review*, 35, 354–367. 4

Ahn, H. and J. L. Powell (1993): "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58, 3–29. 3

Amemiya, T. (1973): "Regression Analysis when the Dependent Variable is Truncated Normal," *Econometrica*, 41, 997–1016. 10, 25

——— (1974): "Multivariate Regression and Simultaneous Equation Models when the Dependent Variables Are Truncated Normal," *Econometrica*, 42, 999–1012. 3

——— (1984): "Tobit models: A survey," *Journal of Econometrics*, 24, 3–61. 3

——— (1985): *Advanced Econometrics*, Harvard University Press. 3

Andrews, D. W. K. and M. M. A. Schafgans (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," *The Review of Economic Studies*, 65, 497–517. 3

Bastos, F. and W. Barreto-Souza (2021): "Birnbaum–Saunders sample selection model," *Journal of Applied Statistics*, 48, 1896–1916. 3

Bastos, F. d. S., W. Barreto-Souza, and M. G. Genton (2022): "A Generalized Heckman Model With Varying Sample Selection Bias and Dispersion Parameters," *Statistica Sinica*. 3

Carlson, A. (2022): "gtsheckman: Generalized Two Step Heckman Estimator," *Unpublished Working Paper*. 3

Cogan, J. F. (1981): "Fixed Costs and Labor Supply," *Econometrica*, 49, 945–963. 4

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *The Review of Economic Studies*, 70, 33–58. 3, 16

Farre-Mensa, J., D. Hegde, and A. Ljungqvist (2020): "What Is a Patent Worth? Evidence from the U.S. Patent "Lottery"," *The Journal of Finance*, 75, 639–682. 14

Feng, J. and X. Jaravel (2020): "Crafting Intellectual Property Rights: Implications for Patent Assertion Entities, Litigation, and Innovation," *American Economic Journal: Applied Economics*, 12, 140–81. 2

Frandsen, B. R., L. J. Lefgren, and E. C. Leslie (2019): "Judging Judge Fixed Effects," *NBER Working Paper Series*. 14

Gaulé, P. (2018): "Patents and the Success of Venture-Capital Backed Startups: Using Examiner Assignment to Estimate Causal Effects," *The Journal of Industrial Economics*, 66, 350–376. 14

Gordon, R. H. and A. S. Blinder (1980): "Market wages, reservation wages, and retirement decisions," *Journal of Public Economics*, 14, 277–308. 4

Gourieroux, C. (2000): *Econometrics of Qualitative Dependent Variables*, Cambridge University Press. 3

Graham, S. J., A. C. Marco, and R. Miller (2018): "The USPTO Patent Examination Research Dataset: A window on patent processing," *Journal of Economics & Management Strategy*, 27, 554–578. 5

Greene, W. (2003): *Econometric Analysis*, Pearson Education. 22

Griliches, Z., B. H. Hall, and J. A. Hausman (1978): "Missing Data and Self-Selection in Large Panels," *Annales de l'inséé*, 137–176. 4

Gronau, R. (1974): "Wage Comparisons-A Selectivity Bias," *Journal of Political Economy*, 82, 1119–43. 3

Hanoch, G. (1976): *A Multivariate Model of Labor Supply: Methodology for Estimation*, Santa Monica, CA: RAND Corporation. 4

——— (2014): *Chapter 3. Hours And Weeks In The Theory Of Labor Supply:*, Princeton University Press, 119–165. 4

Heckman, J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679–94. 3, 4

——— (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement, Volume 5, number 4*, 475–492. 2, 3

——— (1978): "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 46, 931–59. 3

——— (1990): "Varieties of Selection Bias," *The American Economic Review: Papers and Proceedings*, 80, 313–318.

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161. 3, 4

Jennrich, R. I. (1969): "Asymptotic Properties of Non-Linear Least Squares Estimators," *The Annals of Mathematical Statistics*, 40, 633 – 643. 10, 25

Katz, A. (1977): "Evaluating Contributions of the Employment Service to Applicant Earnings," *Labor Law Journal*, 28, 472–8. 4

Kenny, L. W., L.-F. Lee, G. S. Maddala, and R. P. Trost (1979): "Returns to College Education: An Investigation of Self-Selection Bias Based on the Project Talent Data," *International Economic Review*, 20, 775–789. 4

King, M. A. (1980): "An Econometric Model of Tenure Choice and Demand for Housing as a Joint Decision," in *Econometric Studies in Public Finance*, National Bureau of Economic Research, Inc, NBER Chapters, 137–159. 4

Kling, J. R. (2006): "Incarceration Length, Employment, and Earnings," *American Economic Review*, 96, 863–876. 14

Kuhn, J. and N. Thompson (2019): "How to Measure and Draw Causal Inferences with Patent Scope," *International Journal of the Economics of Business*, 26, 5–38. 2

Lanjouw, J. O. and M. Schankerman (2004): "Patent Quality and Research Productivity: Measuring Innovation with Multiple Indicators," *The Economic Journal*, 114, 441–465. 2, 14

Lee, L.-F. (1978): "Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables," *International Economic Review*, 19, 415–33. 4

Lewis, H. G. (1974): "Comments on Selectivity Biases in Wage Comparisons," *Journal of Political Economy*, 82, 1145–55. 3

Maddala, G. S. (1986): *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press. 3

Marchenko, Y. V. and M. G. Genton (2012): "A Heckman Selection-t Model," *Journal of the American Statistical Association*, 107, 304–317. 3

Matcham, W. and M. Schankerman (2022): "On the Validity of the Leniency IV," *Working Paper*. 4

Nakosteen, R. A. and M. Zimmer (1980): "Migration and Income: The Question of Self-Selection," *Southern Economic Journal*, 46, 840–851. 4

Nelson, F. D. (1977): "Censored regression models with unobserved, stochastic censoring thresholds," *Journal of Econometrics*, 6, 309–327. 4

Ogundimu, E. O. and J. L. Hutton (2016): "A Sample Selection Model with Skew-normal Distribution," *Scandinavian Journal of Statistics*, 43, 172–190. 3

Poirier, D. J. and P. A. Ruud (1981): "On the appropriateness of endogenous switching," *Journal of Econometrics*, 16, 249–256. 4

Roberts, R. B., G. Maddala, and G. Enholm (1978): "Determinants of the Requested Rate of Return and the Rate of Return Granted in a Formal Regulatory Process," *Bell Journal of Economics*, 9, 611–621. 4

Roelsgaard, S. and L. Taylor (2022): "A Semiparametric Machine Learning Estimator for Sample Selection Models," *Unpublished Working Paper*. 4

Rosen, H. S. (1979): "Housing decisions and the U.S. income tax: An econometric analysis," *Journal of Public Economics*, 11, 1–23. 4

Sampat, B. and H. L. Williams (2019): "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome," *American Economic Review*, 109, 203–36. 14

Trost, R. (1977): "Demand for Housing: A Model Based on Inter-related Choices Between Owning and Renting," *Unpublished Ph.D dissertation, University of Florida*. 4

Vella, F. (1998): "Estimating Models with Sample Selection Bias: A Survey," *Journal of Human Resources*, 33, 127–169. 3

Weisbrod, B. (1980): *Wage Differentials Between the Private For-profit and Non-profit Sectors: The Case of Lawyers*, Discussion papers, University of Wisconsin-Madison, Institute for Research on Poverty. 4

Willis, R. J. and S. Rosen (1979): "Education and Self-Selection," *Journal of Political Economy*, 87, S7–S36. 4

A Proofs and Derivations

A.1 Proof of Theorem 1

Proof. The proof proceeds in four parts:

1. Identify the standardized reduced form parameters π as given in (10)

2. Identify γ

3. Identify σ_Y

4. Identify β_Y, ρ, and β_{-Y}

Each step works sequentially: I identify the standardized reduced form parameters first not because they are of interest *per se*, but because they aid in the identification of γ, σ_Y, ρ, and eventually β.

Part 1: Reduced Form Parameters

From (9),

$$\mathbb{P}(D=1|z) = \mathbb{P}\left(\frac{\beta_Y \varepsilon_Y + \varepsilon_D}{\sqrt{v}} > \tilde{z} \middle| z\right) = \Phi(-\tilde{z}) = \Phi(\pi' z),$$

so

$$\Phi^{-1}\left[\mathbb{P}(D=1|z)\right] = \pi' z.$$

Let z have dimension K_1. Then, using K_1 different vectors of z each denoted $z_{(k)}$ for $k = 1, \ldots, K_1$, there are K_1 equations

$$\underbrace{\Phi^{-1}\left[\mathbb{P}(D=1|z_{(k)})\right]}_{q_{(k)}} = \pi' z_{(k)}.$$

Let Q stack $q_{(k)}$ and χ_1 stack $z_{(k)}$. Then

$$Q = \mathcal{X}_1 \pi.$$

The matrices Q and $\mathcal{X}_1$ are known from the distribution of observables. By the assumptions of the theorem, $\mathcal{X}_1$ has full rank and thus π is identified. The standardized reduced form parameters π are the ones recovered from a probit regression of D on z.

Part 2: γ Parameters

Now I identify γ, given that I have identified π. Consider the conditional expectation of Y, conditional on $D = 1$ and z, i.e. $\mathbb{E}(Y|D = 1, z)$. Because I condition on $D = 1$, this object is known from the distribution of observables. Substituting Y from (5),

$$\mathbb{E}(Y|D=1,z) = \gamma' w_Y + \mathbb{E}(\varepsilon_Y|D=1,z), \tag{13}$$

where, for reference, $w_Y = (X', X_Y')'$. Recall that $D = 1$ if and only if $\dfrac{\beta_Y \varepsilon_Y + \varepsilon_D}{\sqrt{v}} > \tilde{z}$ where $\tilde{z}$ is identified because z is observed and π is identified. So, owing to the joint normality of ε_Y and $\dfrac{\beta_Y \varepsilon_Y + \varepsilon_D}{\sqrt{v}}$, there holds that (Greene, 2003)

$$\begin{aligned}
\mathbb{E}(\varepsilon_Y|D=1,z) &= \mathbb{E}\left(\varepsilon_Y \middle| \frac{\beta_Y \varepsilon_Y + \varepsilon_D}{\sqrt{v}} > \tilde{z}, z\right) \\
&= \operatorname{cov}\left(\varepsilon_Y, \frac{\beta_Y \varepsilon_Y + \varepsilon_D}{\sqrt{v}}\right) \frac{\phi(\tilde{z})}{1-\Phi(\tilde{z})} \\
&= \underbrace{\frac{\beta_Y \sigma_Y^2 + \rho \sigma_Y}{\sqrt{v}}}_{\tau} \underbrace{\frac{\phi(\tilde{z})}{\Phi(-\tilde{z})}}_{\lambda(\tilde{z})},
\end{aligned} \tag{14}$$

where $\lambda(\tilde{z})$ is the inverse Mills ratio. I also use the notation $\tilde{\lambda}(z)$ to denote the same quantity, i.e.

$$\tilde{\lambda}(z) = \lambda(\tilde{z}) = \frac{\phi(\tilde{z})}{\Phi(-\tilde{z})}$$

Putting (13) together with (14) yields

$$\mathbb{E}(Y|D=1,z) = \gamma' w_Y + \tau \tilde{\lambda}$$

This implies that a regression of Y on w_Y and $\tilde{\lambda}$ for observations with $D = 1$ delivers γ and τ. More formally, letting $(w_Y', \tilde{\lambda}(z))' \in \mathbb{R}^{K_2+1}$, then using $K_2 + 1$ different values of (w_Y', z), denoted $\chi_{2(k)}$ for $k = 1, \ldots, K_2 + 1$ we obtain $K_2 + 1$ equations

$$\underbrace{\mathbb{E}(Y|D=1,z_{(k)})}_{r(k)} = \underbrace{\begin{pmatrix} w_{Y(k)}' & \tilde{\lambda}(z_{(k)}) \end{pmatrix}}_{\chi_2(k)} \begin{pmatrix} \gamma \\ \tau \end{pmatrix}$$

Let R stack $r(k)$ and χ_2 stack $\chi_2(k)$. Then

$$R = \mathcal{X}_2 \begin{pmatrix} \gamma \\ \tau \end{pmatrix}.$$

By the assumptions of the theorem, $\mathcal{X}_2$ has full rank and thus γ and τ are identified.

Part 3: σ_Y parameter

It is natural next to look at $\operatorname{Var}(Y|D=1,z)$. Using properties of conditional multivariate normals (Greene, 2003),

$$\operatorname{Var}(Y|D=1,z) = \sigma_Y^2 + \tau^2 \tilde{z} \lambda(\tilde{z}) - \tau^2 \lambda(\tilde{z})^2.$$

From this σ_Y^2 is identified after rearrangement.

Part 4: β_Y, ρ and β_{-Y}

From (10) it follows that

$$\begin{aligned}\beta_X &= \pi_X\sqrt{v} - \beta_Y\gamma_X \\ \beta_D &= \pi_D\sqrt{v}.\end{aligned}$$

So, after showing that β_Y and ρ are identified (which implies that $v = \beta_Y^2\sigma_Y^2 + 1 + 2\beta_Y\sigma_Y\rho$ is identified), it is immediate that β_{-Y} is identified. Hence to finish the proof, I must show that β_Y and ρ are identified.

It follows from (10) and the assumptions of the theorem that for some j with $\gamma_{Yj} \neq 0$,

$$\frac{\beta_Y}{\sqrt{v}} = \frac{\pi_{Yj}}{\gamma_{Yj}} = \xi. \tag{15}$$

Additionally, recall that τ is identified, where

$$\tau = \frac{\beta_Y\sigma_Y^2 + \rho\sigma_Y}{\sqrt{v}}. \tag{16}$$

First, suppose $\xi = 0$ so that $\beta_Y = 0$. This implies that $v = 1$ and $\tau = \rho\sigma_Y$ or $\rho = \tau/\sigma_Y$. So when $\xi = 0$, β_Y and ρ are identified. From now on, suppose $\xi \neq 0$ so that $\beta_Y \neq 0$. Solving (15) and (16) simultaneously yields[14]

$$\beta_Y^2 = \frac{\xi^2}{1 + \sigma_Y^2\xi^2 - 2\tau\xi}$$

and

$$\rho = \beta_Y\left(\frac{\tau}{\sigma_Y\xi} - \sigma_Y\right).$$

From this, we know that

$$\beta_Y = \pm\sqrt{\frac{\xi^2}{1 + \sigma_Y^2\xi^2 - 2\tau\xi}}.$$

But, since $\beta_Y = \xi\sqrt{v}$, the sign of β_Y is the same as the sign of ξ. This means that β_Y is identified and therefore so is ρ.

$\square$

Back to Theorem 1

[14] Note that $1 + \sigma_Y^2\xi^2 - 2\tau\xi = v^{-1}$ so this term cannot be nonpositive.

A.2 Likelihood Term

For brevity I omit conditioning on z_i. Let $M_i = \beta_{-Y}' w_{Di} + \beta_Y \gamma_Y' w_{Yi}$. Then

$$
\begin{aligned}
\wp_{1i} &= \int_{-M_i-\beta_Y(Y_i - w_{Yi}'\gamma)}^{\infty} f_{\varepsilon_D,\varepsilon_Y}(\varepsilon_D, Y_i - w_{Yi}'\gamma)\, d\varepsilon_D \\
&= \int_{-M_i-\beta_Y(Y_i - w_{Yi}'\gamma)}^{\infty} f_{\varepsilon_D|\varepsilon_Y}(\varepsilon_D|Y_i - w_{Yi}'\gamma) f_{\varepsilon_Y}(Y_i - w_{Yi}'\gamma) d\varepsilon_D \\
&= \frac{1}{\sigma_Y}\phi\left(\frac{Y_i - w_{Yi}'\gamma}{\sigma_Y}\right) \int_{-M_i-\beta_Y(Y_i - w_{Yi}'\gamma)}^{\infty} \phi\left(\frac{\varepsilon_D - \frac{\rho}{\sigma_Y}(Y_i - w_{Yi}'\gamma)}{\sqrt{1-\rho^2}}\right) d\varepsilon_D \qquad (17) \\
&= \frac{1}{\sigma_Y}\phi\left(\frac{Y_i - w_{Yi}'\gamma}{\sigma_Y}\right)\left[1 - \Phi\left(\frac{-\beta_{-Y}' w_{Di} - (\beta_Y + \frac{\rho}{\sigma_Y})Y_i + \frac{\rho}{\sigma_Y} w_{Yi}'\gamma}{\sqrt{1-\rho^2}}\right)\right] \\
&= \frac{1}{\sigma_Y}\phi\left(\frac{Y_i - w_{Yi}'\gamma}{\sigma_Y}\right)\Phi\left(\frac{\beta_{-Y}' w_{Di} + (\beta_Y + \frac{\rho}{\sigma_Y})Y_i - \frac{\rho}{\sigma_Y} w_{Yi}'\gamma}{\sqrt{1-\rho^2}}\right).
\end{aligned}
$$

Equality (17) follows from

$$\varepsilon_D|\varepsilon_Y = s \sim \mathcal{N}\left(-\frac{\rho}{\sigma_Y}s, 1-\rho^2\right)$$

when $(\varepsilon_D, \varepsilon_Y)$ are joint normal as in (8).

Back to likelihood

A.3 Two-Step Estimation Details

Rewriting (12) with an estimated value of λ_i yields

$$Y_i = \gamma' w_{Yi} + \tau\hat{\lambda}_i + \underbrace{\tau(\lambda_i - \hat{\lambda}_i) + V_i}_{\check{V}_i},$$

so that the effective error term is $\check{V}_i = \tau(\lambda_i - \hat{\lambda}_i) + V_i$.

As mentioned, π is estimated using all n observations by a MLE probit, and so

$$\sqrt{n}(\hat{\pi} - \pi) \xrightarrow{d} \mathcal{N}(0, V_{1st}).$$

Since λ_i is a twice-continuously differentiable function of π, by the Delta-Method,

$$\sqrt{n}(\hat{\lambda}_i - \lambda_i) \xrightarrow{d} \mathcal{N}(0, \Sigma_i)$$

where $\Sigma_i = (\partial\lambda_i)^2 z_i' V_{1st} z_i$, for $\partial\lambda_i = \partial\lambda_i/\partial\tilde{z}_i$.

The task is to derive the asymptotic distribution of

$$\sqrt{n_1}\begin{pmatrix} \hat{\gamma} - \gamma \\ \hat{\tau} - \tau \end{pmatrix} = \begin{pmatrix} \frac{1}{n_1}\sum_i w_{Yi} w_{Yi}' & \frac{1}{n_1}\sum_i w_{Yi}\hat{\lambda}_i \\ \frac{1}{n_1}\sum_i w_{Yi}\hat{\lambda}_i & \frac{1}{n_1}\sum_i \hat{\lambda}_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{\sqrt{n_1}}\sum_i w_{Yi}\check{V}_i \\ \frac{1}{\sqrt{n_1}}\sum_i \hat{\lambda}_i\check{V}_i \end{pmatrix}.$$

All n observations are used to estimate the probit, but the regression in the second step only uses n_1 observations. As such, I take probability limits are taken with both $n, n_1 \to \infty$ and $n_1/n \to k \in (0,1)$. Under conditions on regressors discussed in Amemiya (1973) and Jennrich (1969),

$$\underset{n_1\to\infty}{\text{plim}} \begin{pmatrix} \frac{1}{n_1}\sum_i w_{Yi}w'_{Yi} & \frac{1}{n_1}\sum_i w_{Yi}\hat{\lambda}_i \\ \frac{1}{n_1}\sum_i w'_{Yi}\hat{\lambda}_i & \frac{1}{n_1}\sum_i \hat{\lambda}_i^2 \end{pmatrix}^{-1} = \underset{n_1\to\infty}{\text{plim}} \begin{pmatrix} \frac{1}{n_1}\sum_i w_{Yi}w'_{Yi} & \frac{1}{n_1}\sum_i w_{Yi}\lambda_i \\ \frac{1}{n_1}\sum_i w'_{Yi}\lambda_i & \frac{1}{n_1}\sum_i \lambda_i^2 \end{pmatrix}^{-1} = B$$

and

$$\begin{pmatrix} \frac{1}{\sqrt{n_1}}\sum_i w_{Yi}\check{V}_i \\ \frac{1}{\sqrt{n_1}}\sum_i \hat{\lambda}_i\check{V}_i \end{pmatrix} \xrightarrow{d} \mathcal{N}(0,\psi),$$

where

$$\psi = \underset{\substack{n\to\infty \\ n_1\to\infty}}{\text{plim}} \left[\frac{\sigma_Y^2}{n_1}\begin{pmatrix} \sum_i w_{Yi}w'_{Yi}\eta_i & \sum_i w_{Yi}\lambda_i\eta_i \\ \sum_i w'_{Yi}\lambda_i\eta_i & \sum_i \lambda_i^2\eta_i \end{pmatrix} + \right.$$
$$\left. \tau^2\left(\frac{n_1}{n}\right)\begin{pmatrix} \frac{1}{n_1^2}\sum_i\sum_j w_{Yi}w'_{Yj}\vartheta_{ij} & \frac{1}{n_1^2}\sum_i\sum_j w_{Yi}\Upsilon_{ij} \\ \frac{1}{n_1^2}\sum_i\sum_j w'_{Yi}\Upsilon_{ij} & \frac{1}{n_1^2}\sum_i\sum_j \Omega_{ij} \end{pmatrix}\right]$$

where

$$\begin{aligned} \eta_i &= 1 + \frac{\tau^2(\tilde{z}_i\lambda_i - \lambda_i^2)}{\sigma_Y^2} \\ \vartheta_{ij} &= \partial\lambda_i\,\partial\lambda_j\; z'_iV_{1st}z_j \\ \Upsilon_{ij} &= \lambda_i\vartheta_{ij} \\ \Omega_{ij} &= \lambda_i\lambda_j\vartheta_{ij}. \end{aligned}$$

Putting these two results together, and using the Cramer convergence theorem,

$$\sqrt{n_1}\begin{pmatrix} \hat{\gamma}-\gamma \\ \hat{\tau}-\tau \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, B\psi B'),$$

as required.

B Discrete Response

B.1 Model and Identification

When Y is discrete, say binary for simplicity, the model is

$$\begin{aligned} Y &= 1\left[\underbrace{\gamma'_X X + \gamma'_Y X_Y + \varepsilon_Y}_{Y^*} > 0\right] = 1\left[\gamma' w_Y + \varepsilon_Y > 0\right] \\ D &= 1\left[\underbrace{\beta'_X X + \beta'_D X_D + \beta_Y Y + \varepsilon_D > 0}_{D^*}\right] = 1\left[\beta_{-Y} w_D + \beta_Y Y + \varepsilon_D > 0\right] \end{aligned}$$

Where all terms are defined as in the main text. Now, when $D = 0$ we observe nothing on Y, and when $D = 1$ we observe $Y = 1$ or $Y = 0$.

Identification will revolve around two observable objects:

1. $\mathbb{P}(D = 0|z)$
2. $\mathbb{P}(D = 1 \cap Y = 1|z)$ (or alternatively $\mathbb{P}(Y = 1|D = 1, z)$)

Approaching identification as in Theorem 1 will not work because Y is not linear in parameters. Changing the model so that Y^* enters D^*, i.e.

$$
\begin{aligned}
Y &= 1\left[\underbrace{\gamma_X' X + \gamma_Y' X_Y + \varepsilon_Y}_{Y^*} > 0\right] = 1\left[\gamma' w_Y + \varepsilon_Y > 0\right] && (18)\\
D &= 1\left[\underbrace{\beta_X' X + \beta_D' X_D + \beta_Y Y^* + \varepsilon_D}_{D^*} > 0\right] = 1\left[\beta_{-Y} w_D + \beta_Y Y^* + \varepsilon_D > 0\right] && (19)
\end{aligned}
$$

makes progress possible, but identification is still not immediate. Though a probit regression for the D equation delivers standardized reduced form parameters, a probit regression for the Y equation for $D = 1$ does not deliver the γ parameters, and there are no higher moments of observables to identify the error correlation.

B.2 Estimation

The log likelihood for the model (18) - (19) with conditionally standard bivariate normal errors (with correlation ρ) is

$$
\log(\mathcal{L}_{\text{disc}}) = \sum_i (1 - D_i)\log(\wp_{0i}) + D_i Y_i \log(\wp_{11i}) + D_i(1 - Y_i)\log(\wp_{10i}),
$$

where $\wp_{0i} = 1 - \Phi(\pi' z_i)$ is the same as in the main text and

$$
\begin{aligned}
\wp_{11i} &= \mathbb{P}(D_i = 1 \cap Y_i = 1|z_i)\\
&= \underbrace{\mathbb{P}(Y_i = 1|D_i = 1, z_i)}_{\wp_{1|1i}}\underbrace{\mathbb{P}(D_i = 1|z_i)}_{\wp_{1i}}.
\end{aligned}
$$

Here

$$
\wp_{1i} = \Phi(\pi' z_i),
$$

Therefore, the log-likelihood is

$$\begin{aligned}\log(\mathcal{L}_{\text{disc}}) &= \sum_i (1-D_i)\log(\wp_{0i}) + D_i \log(\wp_{1i}) \\ &+ \sum_i D_i Y_i \log(\wp_{1|1i}) + D_i(1-Y_i)\log(\wp_{0|1i})\end{aligned}$$

The term $\wp_{0|1i}$ is given by

$$\begin{aligned}\wp_{0|1i} &= \mathbb{P}\left(\gamma' w_{Yi} + \varepsilon_{Yi} < 0 \,\middle|\, \frac{\beta_Y \varepsilon_{Yi} + \varepsilon_{Di}}{\sqrt{\upsilon}} > -\pi' z_i, z_i\right) \\ &= \int_{-\pi' z_i}^{\infty} \mathbb{P}\left(\varepsilon_{Yi} < -\gamma' w_{Yi} \,\middle|\, \frac{\beta_Y \varepsilon_{Yi} + \varepsilon_{Di}}{\sqrt{\upsilon}} = x, z_i\right)\phi(x)dx \\ &= \int_{-\pi' z_i}^{\infty} \Phi\left(\frac{-\gamma' w_{Yi} - \vartheta x}{\sqrt{1-\rho^2}}\right)\frac{\phi(x)}{1-\Phi(-\pi' z_i)}dx,\end{aligned}$$

where $\vartheta = \upsilon^{-1/2}(\beta_Y + \rho)$ (the same as τ previously except with $\sigma_Y = 1$) and similarly

$$\wp_{1|1i} = \int_{-\pi' z_i}^{\infty} \left[1 - \Phi\left(\frac{-\gamma' w_{Yi} - \vartheta x}{\sqrt{1-\rho^2}}\right)\right]\frac{\phi(x)}{1-\Phi(-\pi' z_i)}dx.$$

These integrals have no closed form but can be approximated with simulation methods. Results from simulated maximum likelihood estimation apply: the number of simulations must grow in proportion to the sample size to ensure that the asymptotic distribution of the estimator matches that of exact MLE.[15]

[15] Simulations of this model using Halton draws, which perform moderately well, are available upon request

C Extra Simulation Results

TABLE 3: Simulation Results Design 2 with Discrete Regressor

Design 2 with Discrete Regressor: $\bar{D} = 0.84$									
Parameter:	σ	ρ	γ_0	γ_X	γ_Y	β_0	β_X	β_D	β_Y
True Value:	2	0.5	3	-1	-2	1	3	-8	2
(1) New:	2.07	0.50	3.00	-1.00	-2.00	1.08	3.21	-8.58	2.14
	(0.06)	(0.16)	(0.07)	(0.02)	(0.04)	(0.36)	(0.68)	(1.74)	(0.47)
(2) Existing:	2.01	0.77	3.25	-1.00	-1.76	1.73	0.16	-1.26	–
	(0.04)	(0.04)	(0.06)	(0.02)	(0.05)	(0.06)	(0.01)	(0.06)	

Notes: Table 3 reports the sample mean and sample standard deviations (in parentheses) of the estimates of the model parameters, over 500 repeated samples. The "New" row provide estimates from using the newly proposed model. The "Existing" row estimates a model that assumes $\beta_Y = 0$.

Back to Simulation Section