Sample selection models with common endogeneity in the selection and outcome: Revisiting the family gap.

Grace Arnold [*] Riju Joshi[†]

March 8, 2022

Summary

In this paper, we develop a fully parametric estimation procedure for unbalanced panel data models with unobserved effects that allow for a common binary endogenous variable in both the selection equation and the outcome equation. We study the problem in the framework of a switching regression model and obtain correction terms that account for sample selection and the common endogenous dummy simultaneously and describe a two-step estimation procedure. We test the finite sample properties of the estimator using Monte Carlo simulations and find that our estimator performs better in the presence of high endogeneity and high sample selection compared to the estimators that ignore either or both of these issues. In addition, our estimator is also robust to distributional misspecification. We apply our econometric methodology to estimate the effect of fertility decisions on wages for white women using the National Longitudinal Survey of Youth 1979 (NLSY79) data from 1982-2006.

***JEL* classification:** C33, C34,J13, J22, J31
Keywords: endogeneity, sample selection, fertility, wage gap
Word count: 7325 (including figures)
The authors have no conflict of interest to declare.
The authors have no funding to disclose.

[*]Corresponding author. Email: garnold@pdx.edu, Portland State University, Department of Economics, Portland, OR.
[†]Email: riju@pdx.edu, Department of Economics, Portland State University, Portland, OR.

1 Introduction

Empirical researchers often face unbalanced panels when working with observational datasets. This situation is particularly prevalent when the cross-sectional unit is a firm, household, or individual. In some instances, as is the case with rotating panels, an unbalanced panel might result from the survey design where equally sized sets of sample units are brought in and out of the sample in some specified pattern. In other cases, incomplete panels might arise due to attrition.

In most circumstances, the panel is unbalanced due to nonrandom selection mechanisms. For example, empirical questions in labor economics often have unbalanced panels because the labor market outcome variable (i.e., wages) is observed only for individuals who choose to work. Thus we have nonrandom selection into the labor force, leading to sample selection bias when ignored.

The theoretical econometrics literature has developed numerous parametric and semi-parametric estimation procedures to correct for sample selection bias. Wooldridge (1995); Semykina and Wooldridge (2010); Semykina (2018); Kyriazidou (1997); Rochina-Barrachina (1999) are just a few notable examples that consider estimation in unbalanced panels for both linear and non-linear models with exogenous and endogenous covariates.

In this paper, we propose a fully parametric estimation procedure for panel data models with sample selection, with a common endogenous binary variable in both the outcome and the selection equations. Such models are especially useful in empirical research with unbalanced panels when we suspect a primary explanatory variable to be endogenous in both the outcome equation as well as the selection mechanism. One example of this situation arises in labor economics when estimating the effect of fertility (or marital) decisions on wages. When measuring the impact of fertility decisions on wages, for example, the decision to have kids is endogenous in both the wage (outcome) equation and the labor force participation (selection) equation. This scenario is not unique to labor

economics. For example, in health economics, we might suspect that an endogenous variable, such as health insurance choice, will affect hospital visits (the selection equation) and, subsequently, health care expenditure (the outcome variable).

il Kim (2006) considers similar models where the selection equation and the censored equation contain a common dummy variable. Specifically, il Kim (2006) describes a two-step estimation procedure that obtains correction terms to account for the endogenous binary variable. However, his paper only considers cross-sectional models. For many microeconometric applications, panel data is more desirable as it allows the researchers to explicitly incorporate unobserved time-constant heterogeneity, thus allowing for factors such as cognitive ability, motivation, etc. In addition, it is often desirable to allow this unobserved time-constant heterogeneity to be correlated with the other observed covariates arbitrarily. Given the increased availability and use of panel data models in empirical research, it is useful to obtain an estimation procedure specifically for panel data that allows researchers to simultaneously account for sample selection, common endogeneity, and unobserved heterogeneity.

Recently, Semykina (2018) considers the effect of young children on a binary outcome variable, women's self-employment, while addressing both endogeneity and sample selection in panel data models. She constructs and estimates a likelihood function that corrects for a common endogenous dummy variable and finds that this substantially affects the estimation results. Specifically, she finds that when corrected for both sample selection and endogeneity, the estimates roughly triples compared to the uncorrected estimators, emphasizing the importance of accounting and correcting for the common endogeneity of the variable of interest.

Our model and estimation procedure differs from Semykina (2018) in two key ways. First, we consider a continuous outcome variable. While this difference results in simpler model, in practice, many empirical researchers continue to need models for situations

where the outcome variables is roughly continuous, such as wages, hours worked, health care expenditure, etc. Additionally, researchers often use LPM when the outcome variable is binary, due to the ease of interpreting coefficients. In fact, we used the procedure proposed in this paper to replicate the empirical application in Semykina (2018). The findings from this exercise are quantitatively similar to those presented in Semykina (2018).

Secondly, we obtain a two-step estimation procedure for panel data models for continuous outcome variables with sample selection that allows for time-constant unobserved heterogeneity and contains a binary endogenous variable common to both the selection and outcome equation. Our two-step estimation procedure for panel data models relaxes the computational and non-convergence issues that frequently accompany maximizing a likelihood function. For example, the first step of our proposed procedure can be estimated using bivariate probit, and the second step is executed using pooled OLS. We obtain valid inference analytically by using Wooldridge (2010) or by formulating the estimation procedure as one big method of moment problem. This method accounts for the fact that pre-estimated parameters are used in the second step and correct the standard errors. Alternatively, a panel bootstrap routine is also straightforward and is how we calculate the standard errors in our empirical application presented in Section 4.

After developing our estimation procedure, we examine the finite sample properties of our proposed estimator through a series of Monte Carlo experiments and find that in cases where both common endogeneity and sample selection are severe, the estimator we recommend performs better than the estimators that ignore either endogeneity or sample selection, or both. In addition, our estimator is also robust to the distributional mispecifications in the model.

Finally, to illustrate how our procedure performs empirically, we revisit the effect of fertility decisions on the wages of white women using data for the years 1982-2006 from

the National Longitudinal Survey for Youth 1979 (NLSY1979). We present the findings from our estimation procedure alongside four frequently used estimation procedures: simple OLS, fixed effects (FE), fixed effects two-stage least squares (FE2SLS), and Heckman selection. We find that correcting for only one source of bias, either self-selection or endogeneity, may lead to either overestimating or underestimating of the negative effects of a third child on wages. When addressing both sources of biases, the impact of three children on the hourly wage rate of white women is estimated to be about -23.8%. This result is significantly larger than OLS, FE, and Heckman selection, but significantly smaller than FE2SLS.

The remainder of this paper is structured as follows. In Section 2, we describe the model and the estimation procedure. Section 3 reports the finite sample performance of the estimator. In Section 4, we consider the empirical application where we revisit the effect of fertility decisions on wages for women. Section 5 concludes.

2 Econometric Methodology

We begin with the model for the latent response variable y_{it}^* for an individual i at time t as:

$$y_{it}^* = x_{it1}\beta_1 + \delta d_{it} + c_{i1} + u_{it1} \tag{1}$$

where x_{it1} is a $1 \times K_1$ vector of exogenous variables, c_{i1} is the time constant unobserved heterogeneity in the outcome equation and u_{it1} is the idiosyncratic error. The covariates x_{it1} are assumed to be uncorrelated with the idiosyncratic errors. The key variable of interest is d_{it}, which is a scalar binary variable which will be allowed to be endogenous to both outcome equation and selection equation. As we will see later, d_{it} will indicate a third child in the household, which will be allowed to be endogenous in both the labor force participation and wage equation.

We specify a model for d_{it} as:

$$d_{it} = \mathbf{1}[d_{it}^* > 0] = \mathbf{1}[x_{it2}\beta_2 + c_{i2} + u_{it2} > 0] \tag{2}$$

where x_{it2} is a $1 \times K_2$ vector of observable characteristics assumed to be exogenous, c_{i2} contains time-constant unobserved heterogeneity and u_{it2} is idiosyncratic error. We can interpret equation (2) as the model for fertility (or fertility equation) hereafter.

If we interpret equation (1) as the wage equation, then the key issue is that although it is specified for the entire population, the outcome is observed only for the selected population of individuals who choose to work. Formally, we introduce this sample selection by defining a latent variable s_{it}^*:

$$s_{it} = \mathbf{1}[s_{it}^* > 0] = \mathbf{1}[x_{it3}\beta_3 + \gamma d_{it} + c_{i3} + u_{it3} > 0] \tag{3}$$

where x_{it3} is a $1 \times K_3$ vector of observable characteristics assumed to be exogenous, c_{i3} is time-constant unobserved effect and u_{it3} is the time-varying idiosyncratic error. Thus we can interpret s_{it} as the selection indicator that is equal to 1 if an individual i works in year t and is zero otherwise. Equation (3) could follow from a standard theoretical model of labor supply, where the labor force participation decision arises as a solution to a utility maximization problem subject to time and budget constraints. Note that we also have the binary indicator d_{it} in our selection model denoting that both constraints depend on the presence of this dummy variable (i.e having a third child.)

As a result of sample selection, we have the observed outcome (such as wage) as

$$y_{it} = s_{it} \times y_{it}^* \tag{4}$$

We define x_{it} as the union of mutually exclusive elements of $(x_{it1}, x_{it2}, x_{it3})$ that includes unity and assume that x_{it} and d_{it} are always observed. We also assume that conditional on

the unobserved effects, the vector of observable variables x_{it} is independent of the vector of idiosyncratic errors $\{u_{it2}, u_{it3}, u_{it1}\}$. However, we do not impose any restrictions on the serial dependence properties of the idiosyncratic errors. In addition, we allow for the time-constant unobserved effects in the three equations to be arbitrarily correlated with x_{it}.

We would like to allow for the unobserved time-constant heterogeneity to be correlated with the strictly exogenous variables. We use Mundlak (1978) approach to model the time-constant unobserved heterogeneity in terms of the time means of the exogenous variables. Mundlak device also known as correlated random effects is a widely used tool with linear as well as nonlinear panel data models. Recent example are: Semykina and Wooldridge (2018); Semykina (2018); Semykina and Wooldridge (2010); Wooldridge (2019); Jäckle and Himmler (2010), Maurer et al. (2011), Papke and Wooldridge (2008).

Thus we have:

$$c_{i1} = \overline{x}_i \xi_y + a_{i1} \tag{5}$$

$$c_{i2} = \overline{x}_i \xi_2 + a_{i2} \tag{6}$$

$$c_{i3} = \overline{x}_i \xi_3 + a_{i3} \tag{7}$$

with $\overline{x}_i \equiv T^{-1} \sum_{r=1}^{T} x_{ir}$; and a_{i1}, a_{i2}, a_{i3} are unobserved effects assumed to be independent of x_{it}. Defining composite errors as $v_{it1} \equiv a_{i1} + u_{it1}, v_{it2} \equiv a_{i2} + u_{it2}, v_{it3} \equiv a_{i3} + u_{it3}$ and upon substitution, we obtain

$$y_{it} = (x_{it1}\beta_1 + \delta d_{it} + \overline{x}_i \xi_1 + v_{it1}) \times s_{it} \tag{8}$$

$$d_{it} = \mathbf{1}[x_{it2}\beta_2 + \overline{x}_i \xi_2 + v_{it2} > 0] \tag{9}$$

$$s_{it} = \mathbf{1}[x_{it3}\beta_3 + \gamma d_{it} + \overline{x}_i \xi_3 + v_{it3} > 0] \tag{10}$$

2.1 A switching equation framework

To obtain the estimating equation that accounts for the endogeneity of the common dummy d_{it} in the selection and outcome model, we formulate our model in a switching regression framework. In particular, we posit the existence of two *regimes* defined as defined as $d_{it} = \{1, 0\}$. In studying the effect of fertility decisions on wages, for example, one could interpret the *regimes* as the one "*with-kids*" with $d_{it} = 1$ and one "*without-kids*" with $d_{it} = 0$. In this context, formulating our model in a switching regression framework allows us to differentiate the labor force participation decision as well as the outcome variable (such as wages) for those who have children and those who do not. Similarly, if one was studying the effect of the type of health insurance on health expenditure, one could interpret the *regimes* as one with an HMO plan (Health Maintenance Organization), for example, and the other with a PPO (Preferred Provider Organization) plan.

In this paper, we postulate the existence of selection and outcome models in two regimes. That is for $d_{it} = 0$, we have:

$$s^*_{it0} = x_{it3}\beta_3 + \overline{x}_i\xi_3 + v_{it3}; \;\; s_{it0} = \mathbf{1}[s^*_{it0} > 0] \tag{11}$$

$$y^*_{it0} = x_{it1}\beta_1 + \overline{x}_i\xi_1 + v_{it1} \tag{12}$$

and for $d_{it} = 1$, later interpreted as the regime "*with-kids*", the labor force participation and the wage equations are:

$$s^*_{it1} = x_{it3}\beta_3 + \gamma + \overline{x}_i\xi_3 + v_{it3}; \;\; s_{it1} = \mathbf{1}[s^*_{it1} > 0] \tag{13}$$

$$y^*_{it1} = x_{it1}\beta_1 + \delta + \overline{x}_i\xi_1 + v_{it1} \tag{14}$$

These equation are later interpreted as the labor force participation and wage equations in the regime "*without-kids*", and regime "*with-kids*" respectively. [1]

[1] One can allow separate errors for the selection and outcome equations in the two regimes. In other

Since we have $y_{it0} = s_{it0} \times y^*_{it0}$ and $y_{it1} = s_{it1} \times y^*_{it1}$ in regime 0 and regime 1 respectively, we can describe the observed outcome as

$$y_{it} = d_{it} y_{it1} + (1 - d_{it}) y_{it0}; \;\; t = 1, 2, ..., T \tag{15}$$

Next, we describe the correction terms for each regime, which are comprised of two parts: one corrects for the bias that arises from the *endogenous switching* and the other corrects for the sample selection bias for each *regime* characterized by the value of d_{it}. To do this, we first impose a distributional assumption on our error terms for $t = 1, ..., N$:

$$\begin{pmatrix} v_{it1} \\ v_{it2} \\ v_{it3} \end{pmatrix} \sim Normal\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1 & \rho_{13}\sigma_1 \\ & 1 & \rho_{23} \\ & & 1 \end{bmatrix} \right) \tag{16}$$

To define the correction terms, we consider the conditional mean equation:

$$\mathbb{E}[y_{it}|x_i, d_{it}, s_{it} = 1] = x_{ity}\beta_y + \delta_y d_{it} + \overline{x}_i \xi_y + d_{it}\mathbb{E}[v_{it1}|x_i, d_{it} = 1, s_{it} = 1] + (1 - d_{it})\mathbb{E}[v_{it1}|x_i, d_{it} = 0, s_{it} = 1] \tag{17}$$

Finally, using the normality assumption, we use Poirier (1980) to obtain the panel data

words, we can have $\{v_{it30}, v_{it31}\}; v_{it20} \neq v_{it21}$ in the selection equations and $\{v_{it10}, v_{it11}\}; v_{it10} \neq v_{it11}$ in the outcome equation. This gives us a generalized model where $(s_{it0}, s_{it1}) = (1, 0)$ even when $\gamma > 0$ and $y_{it0} > y_{it1}$ even when $\delta_y > 0$. Alternatively, $(s_{it0}, s_{it1}) = (0, 1)$ even when $\gamma < 0$ and $y_{it0} < y_{it1}$ even when $\delta_y < 0$. See il Kim (2006) for details.

versions of the correction terms in il Kim (2006). We have:

$$\mathbb{E}[v_{it1}|x_i, d_{it}=1, s_{it}=1] \equiv h_{it1} = \rho_{12}\sigma_1 \left[\frac{\phi(x_{it2}\beta_2 + \overline{x}_i\xi_2)\Phi\left(\frac{x_{it3}\beta_3+\gamma+\overline{x}_i\xi_3-\rho_{23}(x_{it2}\beta_2+\overline{x}_i\xi_2)}{\sqrt{1-\rho_{23}^2}}\right)}{\Phi(x_{it2}\beta_2 + \overline{x}_i\xi_2, x_{it3}\beta_3 + \gamma + \overline{x}_i\xi_3; \rho_{23})} \right] \tag{18}$$

$$+ \rho_{23}\sigma_1 \left[\frac{\phi(x_{it3}\beta_3 + \gamma + \overline{x}_i\xi_3)\Phi\left(\frac{x_{it2}\beta_2+\overline{x}_i\xi_2-\rho_{ds}(x_{it3}\beta_3+\gamma+\overline{x}_i\xi_3)}{\sqrt{1-\rho_{23}^2}}\right)}{\Phi(x_{it2}\beta_2 + \overline{x}_i\xi_2, x_{it3}\beta_3 + \gamma + \overline{x}_i\xi_3; \rho_{23})} \right]$$

$$\equiv \mu_{11}h_{it11} + \mu_{12}h_{it12} \tag{19}$$

and

$$\mathbb{E}[v_{it1}|x_i, d_{it}=0, s_{it}=1] \equiv h_{it0} = \rho_{12}\sigma_1 \left[-\frac{\phi(x_{it2}\beta_2 + \overline{x}_i\xi_2)\Phi\left(\frac{x_{it3}\beta_3+\overline{x}_i\xi_3-\rho_{23}(x_{it2}\beta_2+\overline{x}_i\xi_2)}{\sqrt{1-\rho_{23}^2}}\right)}{\Phi(-(x_{it2}\beta_2 + \overline{x}_i\xi_2), x_{it3}\beta_3 + \overline{x}_i\xi_3; -\rho_{23})} \right] \tag{20}$$

$$+ \rho_{13}\sigma_1 \left[\frac{\phi(x_{it3}\beta_3 + \overline{x}_i\xi_3)\Phi\left(\frac{-(x_{it2}\beta_2+\overline{x}_i\xi_2)+\rho_{23}(x_{it3}\beta_3+\overline{x}_i\xi_3)}{\sqrt{1-\rho_{23}^2}}\right)}{\Phi(-(x_{it2}\beta_2 + \overline{x}_i\xi_2), x_{it3}\beta_3 + \overline{x}_i\xi_3; -\rho_{23})} \right]$$

$$\equiv \mu_{01}h_{it01} + \mu_{02}h_{it02} \tag{21}$$

where $\mu_{11} \equiv \rho_{12}\sigma_1, \mu_{12} \equiv \rho_{12}\sigma_1, \mu_{01} \equiv \rho_{12}\sigma_1, \mu_{02} \equiv \rho_{13}\sigma_1$. $\phi(.)$ denotes univariate standard normal pdf, $\Phi(.)$ denotes univariate standard normal cdf and $\Phi(a, b; r)$ denotes bivariate standard normal cdf with correlation coefficient r evaluated for (a, b)

To obtain these correction terms, one can easily formulate a maximum likelihood for

bivariate probit model for panel data. In particular, we define

$$P_{00it} \equiv Prob(d_{it} = 0, s_{it} = 0,) = \Phi(-(x_{it2}\beta_2 + \overline{x}_i\xi_2), -(x_{it3}\beta_3 + \overline{x}_i\xi_3); \rho_{ds}) \tag{22}$$

$$P_{01it} \equiv Prob(d_{it} = 0, s_{it} = 1) = \Phi(-(x_{it2}\beta_2 + \overline{x}_i\xi_2), (x_{it3}\beta_3 + \overline{x}_i\xi_3); -\rho_{ds}) \tag{23}$$

$$P_{10it} \equiv Prob(d_{it} = 1, s_{it} = 0) = \Phi(x_{it2}\beta_2 + \overline{x}_i\xi_2, -(x_{it3}\beta_3 + \overline{x}_i\xi_3 + \gamma); -\rho_{ds}) \tag{24}$$

$$P_{11it} \equiv Prob(d_{it} = 1, s_{it} = 1) = \Phi(x_{it2}\beta_2 + \overline{x}_i\xi_2, x_{it3}\beta_3 + \overline{x}_i\xi_3 + \gamma; \rho_{ds}) \tag{25}$$

Based on these probabilities, the likelihood function is given as

$$\Omega_{it} = \left[P_{00it}^{(1-d_{it})(1-s_{it})} P_{00it}^{(1-d_{it})s_{it}} P_{10it}^{d_{it}(1-s_{it})} P_{11it}^{d_{it}s_{it}}\right] \tag{26}$$

Finally, to obtain our estimating equation, we simply plug-in the correction terms, as described in (20) and (18):

$$y_{it} = x_{it1}\beta_1 + \delta d_{it} + \overline{x}_i\xi_1 + d_{it}[\mu_{11}h_{it11} + \mu_{12}h_{it12}] + (1 - d_{it})[\mu_{01}h_{it01} + \mu_{02}h_{it02}] + \eta_{it} \tag{27}$$

where $\mathbb{E}[\eta_{it}|x_i, d_{it}, s_{it}] = 0$.

2.2 Two-step estimating procedure

We propose a two-step estimating procedure:

Procedure 2.1. 1. Using all the $N \times T$ observations and a full set of time dummies, estimate the selection equation and the endogenous dummy variable equation simultaneously by estimating the likelihood in equation (26)). Use the parameter estimates to obtain the estimates for the correction terms defined in (18) and (20). Denote the estimators by $\hat{h}_{it1}, \hat{h}_{it0}$.

2. Using the selected sample, that is for $s_{it} = 1$, run a Pooled OLS of y_{it} on $x_{it1}, d_{it}, \overline{x}_i, d_{it}\hat{h}_{it1}, (1-d_{it})\hat{h}_{it0}$.

This estimation procedure allows empirical researchers to effectively account for an endogenous dummy variable in both the selection equation as well as in the outcome equation and can be easily implemented in popular software packages. In Stata for example, the first step entails estimation of a bivariate probit and the second step entails OLS estimation on the selected sample with auxiliary correction terms. Furthermore, while excluded covariates for the selection equation and dummy variable equation would certainly increase the credibility of the estimates in the empirical application, they are not required for the estimation due to the highly nonlinear functional forms of the correction terms.

2.3 Asymptotics

Deriving the asymptotic properties for the estimator described in Procedure 2.1 is a straighforward extension of the consistency results and asymptotic variance formulas found in Newey and McFadden (1994) and therefore proofs are omitted. We begin by first formulating our two-step estimation procedure into a method of moment estimation procedure. In particular, let w_{it} and α denote the covariates and the parameters of the first step biprobit model respectively and define $g_{it} \equiv g_{it}(w_{it}, \alpha) \equiv \frac{\partial ln \wp_{it}}{\partial \alpha}$. Similarly, denote the covariates and the parameters in the second step by z_{it} and β respectively and define $m_{it} \equiv m_{it}(w_{it}, z_{it}, \alpha, \beta) \equiv s_{it} z_{it}'(y_{it} - z_{it}\beta)$. Thus for each i we have $g_i \equiv \begin{bmatrix} g_{i1}' & \dots & g_{iT}' \end{bmatrix}'$ and $m_i \equiv \begin{bmatrix} m_{i1}' & \dots & m_{iT}' \end{bmatrix}'$. The moment equations are given as

$$\tilde{g}_i(w_i, z_i, \theta) \equiv \begin{bmatrix} g_i(w_i, \alpha) \\ m_i(w_i, z_i, \alpha, \beta) \end{bmatrix} \tag{28}$$

Denote the true parameters as $\theta_* \equiv (\alpha_*', \beta_*')' \in \Theta$, where Θ is the parameter space and our estimators $\hat{\theta} \equiv (\hat{\alpha}, \hat{\beta})$ are defined as

$$\sum_{i=1}^{N} \tilde{g}_i(w_i, z_i, \hat{\theta}) \equiv \begin{bmatrix} \sum_{i=1}^{N} g_i(w_i, \hat{\alpha}) \\ \sum_{i=1}^{N} m_i(w_i, z_i, \hat{\alpha}, \hat{\beta}) \end{bmatrix} = \mathbf{0} \tag{29}$$

The following proposition is a direct application of Theorem 2.6 of Newey and McFadden (1994) and gives gives the consistency property of our estimator under standard identification, continuity and compactness assumptions

Proposition 2.1. (Consistency) For the outcome model described in equation (1), dummy variable model described in equation (2), sample selection model described in (3), under equations (5, 6, 7) and under the parametric assumption given in equation (16), suppose that (a) $\mathbb{E}[\tilde{g}_i(w_i, z_i, \theta_*) = 0$ and $\mathbb{E}[\tilde{g}_i(w_i, z_i, \theta) \neq 0$ for $\theta \neq \theta_*$; (b) Θ is compact; (c) $\tilde{g}_i(w_i, z_i, \theta)$ is continuous at each $\theta \in \Theta$ with probability one; (d) $\mathbb{E}[sup_{\theta \in \Theta} \| \tilde{g}_i(w_i, z_i, \theta) \| < \infty$ then our estimator obtained in Procedure 2.1 is consistent, that is $\hat{\theta} \rightarrow_p \theta_*$.

To obtain an expression for the asymptotic variance for β that corrects for the first step estimation, we apply Theorem 6.1 of Newey and McFadden (1994). In particular, we first define $G_{\alpha_*} \equiv E\left[\frac{\partial g_i}{\partial \alpha}\right]\Big|_{\alpha=\alpha*}; M_{\alpha_*} \equiv E\left[\frac{\partial m_i}{\partial \alpha}\right]\Big|_{\theta=\theta_*}; M_{\beta_*} \equiv E\left[\frac{\partial m_i}{\partial \beta}\right]\Big|_{\theta=\theta_*}$. Further letting $g*_i \equiv g_i(w_i, \alpha_*); m*_i \equiv m_i(w_i, z_i, \theta_*)$ and defining $V*_{gg} \equiv E[g_i g_i']; V*_{gm} \equiv E[g_i m_i']; V*_{mg} \equiv E[m_i g_i']; V*_{mm} \equiv E[m_i m_i']$ we obtain the expression for the asymptotic variance of our estimator that allow for heteroskedasticity and arbitrary serial autocorrelation over time:

Proposition 2.2. (Asymptotic Normality) For the outcome model described in equation (1), dummy variable model described in equation (2), sample selection model described in (3), under equations (5, 6, 7) and under the parametric assumption given in equation (16), suppose that conditions in Proposition 2.1 are true and suppose (a) $\theta_* \in$ *interior* Θ; (b) $\tilde{g}_i(w_i, z_i, \theta)$ is continuously differentiable in a neighborhood $\mathcal{N}$ of θ_* with probability approaching one; (c) $\mathbb{E}[\tilde{g}_i(w_i, z_i, \theta_*] = 0$ and $\mathbb{E}[\|\tilde{g}_i(w_i, z_i, \theta_*)\|^2$ is finite; (d)

$\mathbb{E}[sup_{\theta\in\mathcal{N}}||\tilde{g}_i(w_{it}, z_i, \theta)|| < \infty$; (e) $G_{\alpha_*}, M_{\alpha_*}, M_{\beta_*}$ are invertible; then $\hat{\alpha}$ and $\hat{\beta}$ are asymptotically normal and

$$\sqrt{N}(\hat{\beta} - \beta_*) \rightarrow_d Normal\,(0, V_\beta) \tag{30}$$

where $V_\beta \equiv M_{\beta*}^{-1} V *_{mm} M_{\beta*}^{-1'} + M_{\beta*}^{-1} M_{\alpha*} [G_{\alpha*}^{-1} V *_{gg} G_{\alpha*}^{-1'}] M'_{\alpha*} M_{\beta*}^{-1'}$

Condition (a) - (d) are standard assumptions that are needed for the moment conditions to be satisfied, condition (e) require the matrices to be invertible for the existence of the asymptotic variance. The asymptotic variance formula consists of two components. The first component is the asymptotic variance of $\hat{\beta}$ if α and thus h_{it1}, h_{it0} are assumed to be known. The second component is the additional variation due to the fact that α and thus h_{it1}, h_{it0} are estimated in the first step. Finally, we can obtain valid inference by estimating analytical standard errors that are robust to heteroskedasticity and arbitrary serial correlation as:

$$\hat{V}_\beta = \hat{M}_\beta^{-1} \hat{V}_{mm} \hat{M}_\beta^{-1'} + \hat{M}_\beta^{-1} \hat{M}_\alpha [\hat{G}_\alpha^{-1} \hat{V}_{gg} \hat{G}_\alpha^{-1'}] \hat{M}'_\alpha \hat{M}_\beta^{-1'} \tag{31}$$

where $\hat{G}_\alpha \equiv \sum_{i=1}^N \left[\frac{\partial g_i}{\partial \alpha}\right]\Big|_{\alpha=\hat{\alpha}}; \hat{M}_\alpha \equiv \sum_{i=1}^N \left[\frac{\partial m_i}{\partial \alpha}\right]\Big|_{\theta=\hat{\theta}}; \hat{M}_\beta \equiv \sum_{i=1}^N \left[\frac{\partial m_i}{\partial \beta}\right]\Big|_{\theta=\hat{\theta}}$. Further letting $\hat{g}_i \equiv g_i(w_i, \hat{\alpha}); \hat{m}_i \equiv m_i(w_i, z_i, \hat{\theta})$ and defining $\hat{V}_{gg} \equiv \sum_{i=1}^N [\hat{g}_i \hat{g}_i'] = \sum_{i=1}^N \sum_{t=1}^T [\hat{g}_{it} \hat{g}_{it}']; \hat{V}_{mm} \equiv E[\hat{m}_i \hat{m}_i'] = \sum_{i=1}^N \sum_{t=1}^T [\hat{m}_{it} \hat{m}_{it}']$

Alternatively, since biprobit and OLS are easily estimated, using a panel bootstrap routine is straightforward.

3 Monte Carlo Simulations

In this section, we study the finite sample properties of the proposed estimator using Monte Carlo experiments. y_{it}, d_{it}, s_{it} are generated as given in (8), (9), (10) with $\beta_1 = 1, \delta = 1, \beta_2 = 1, \beta_3 = 1, \gamma = 0.5, \xi_1 = (0, 0.1, 0.2, 0.1)', \xi_2 = (0, 0.2, 0.3, 0.4)', \xi_3 = (0, 0.2, 0.1, 0.1)'$.

The exogenous variables are generated as:

$$x_{it1} = b_{i1} + \epsilon_{it1}$$

$$x_{it2} = b_{i2} + \epsilon_{it2}$$

$$x_{it3} = b_{i3} + \epsilon_{it3}$$

where the unobserved effects b_{i1}, b_{i2}, b_{i3} are distributed as $Normal(0,1)$ and the correlation coefficient is 0.2 $\epsilon_{ti1}, \epsilon it2$ are idiosyncratic errors that are independent over i and are distributed as $Normal(0,1)$. The distribution of the errors $v_{it1}, v_{it2}, v_{it3}$ is given as: distribution of v_{it2} is $Normal(0,1)$, the distribution of v_{it3} is $\rho_{ds} * v_{it2} + (\sqrt{1-\rho_{ds}^2} * Norma(0,1)$ and v_{it1} is $Normal(0,1) + \rho_{sy} * v_{it3}$ where $\rho_{ds} = \rho_{sy}$ capture the severity of the endogeneity of dummy in the selection equation and severity of sample selection respectively and their values varies across designs.

1. DGP 1: In this design there is low sample selection and low endogeneity of the dummy variable, that is $\rho_{ds} = \rho_{sy} = 0.02$.

2. DGP 2: This design captures the case when both sample selection and common endogeneity of the dummy variable in both outcome equation and selection equation are sever. That is we have $\rho_{ds} = \rho_{sy} = 0.4$

3. DGP 3: Here we check for the robustness of our proposed estimator to the distributional assumptions. In particular, we consider Gamma distribution with Gaussian copula. Let $(e_{it1}, e_{it2}, e_{it3})$ be drawn from the following multivariate normal distribution:

$$\begin{pmatrix} e_{it1} \\ e_{it2} \\ e_{it3} \end{pmatrix} \sim Normal\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.25 & 0.3 \\ & 1 & 0.3 \\ & & 1 \end{bmatrix}\right) \tag{32}$$

Thus have $v_{itm} = (F^{-1}(\Phi(e_{itm}); 3.5, 1), m = 1, 2, 3$ where $F^{-1}(..; 3.5, 1)$ denotes the inverse

cumulative distribution function for the Gamma distribution with scale parameter 3.5 and shape parameter 1.

We consider the samples with $N = 1000$ and $T = 10$ and run 1000 replications. We compare the estimators obtained using Procedure 2.1 with several popular estimators. We estimate the parameters using Pooled OLS (POLS) that ignores the individual heterogeneity, sample selection and endogeneity, Fixed Effects (FE), that both ignores both sample selection and the endogenous dummy variable. We also estimate the parameters using Fixed Effects 2SLS (FE2SLS), which accounts for the endogenous dummy in the outcome equation but ignores endogenous sample selection. We also estimate the parameters using Heckman's procedure, which only corrects for sample selection and ignores the endogeneity of the dummy variable in both outcome equation and sample selection. The results from the simulation are reported in Table 1.

3.1 Results

In DGP 1, there is low sample selection and low endogeneity of the dummy variable in the selection model. Consequently, the Heckman estimator, in addition to the proposed, FE and FE2SLS are unbiased. Heckman estimator has the least RMSE followed closely by FE estimator. Our proposed estimator performs better than the FE2SLS estimator that ignores sample selection and endogeneity of the dummy variable in the selection model. The POLS estimator that ignores endogeneity and sample selection has serious bias, even in when both are mild.

In DGP 2 we have high sample selection and high endogeneity of the dummy variable in the selection model. In this case, only our proposed estimator is unbiased and has the smallest RMSE compared to all other considered estimators. In DGP 3 where we have non normal distribution, our proposed estimator again outperforms all the other estimators in terms of having lowest bias and lowest RMSE. This illustrates that our

Table 1: Simulation Results

	Bias	SD	RMSE
DGP 1: $\rho_{ds} = \rho_{sy} = 0.02$			
Procedure 2.1	-0.01087	0.06329	0.06419
POLS	0.69105	0.04464	0.69249
FE	-0.00285	0.03983	0.03991
FE2SLS	-0.00722	0.10097	0.10118
Heckman	0.00002	0.03627	0.03625
DGP 2: $\rho_{ds} = \rho_{sy} = 0.4$			
Procedure 2.1	-0.00377	0.07464	0.07470
POLS	0.69890	0.04605	0.70042
FE	0.10045	0.04471	0.10994
FE2SLS	-0.04347	0.10177	0.11062
Heckman	0.18193	0.04105	0.18649
DGP 3: NonNormal Distribution			
Procedure 2.1	-0.02034	0.14545	0.14679
POLS	1.11528	0.08423	1.11845
FE	0.53170	0.08509	0.53845
FE2SLS	-0.03591	0.38488	0.38636
Heckman	0.59290	0.08171	0.59850

proposed estimation procedure is robust to distributional misspecification.

4 Empirical Application

4.1 Revisiting the family gap

In Section 4.3, we will apply our estimation procedure (described in Section 2) to evaluate the effects of fertility decisions on the hourly earnings for white women. This empirical question has received substantial attention in labor economics. Even as the wage gap between men and women has been declining over time, many studies continue to find a persistent "family gap": the phenomenon that women with children, on average, earn lower wages than women without children. This wage gap persists even after one

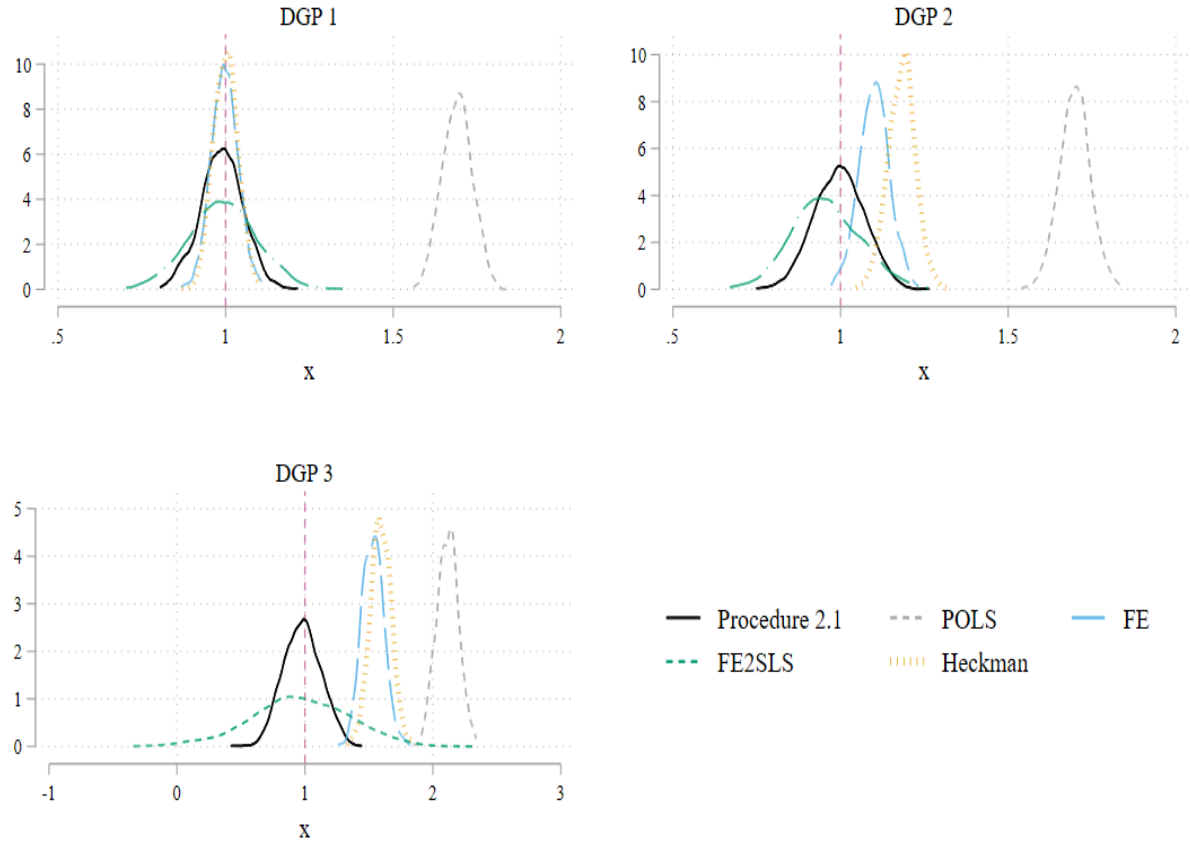

Figure 1: Kernel smoothed empirical distributions of estimators.

controls for differences in education, overall work experience, and full-time and part-time work experience, and some studies find that the gap is greater among highly-qualified women (Waldfogel, 1998, 1997; Bütikofer et al., 2018; Anderson et al., 2002).

The labor literature has long recognized the inherent endogeneity of childbearing decisions, and existing studies have used various creative instruments to account for the endogeneity of fertility in labor supply models. Commonly used instruments for childbearing include the occurrence of twin births, the sex composition of the first two children (Angrist and Evans, 1996), and infertility shocks (Aguero and Marks, 2008). In general, studies that have instrumented the childbearing variable of interest with twin births or sex composition of the first two children have found significant, negative effects of childbearing on wages, though the magnitude is generally smaller than OLS estimates.

There is a large body of research exploring the role of childbearing on parents' labor supply decisions, and whether these decisions can explain the observed wage penalty (Anderson et al., 2002; Lundberg and Rose, 2000; Angrist and Evans, 1996). Generally, studies have found that the presence of children decreases both labor market participation and labor market productivity (Lundberg and Rose, 2000; Anderson et al., 2002; Angrist and Evans, 1996). However, these decisions potentially explain only a portion of the observed wage gap. Additionally, studies have shown that women may choose jobs with lower wages in exchange for characteristics such as flexible schedules, unsupervised break time, and paid sick leave. This mechanism is reflected, for example, in the substantial increases in self-employment for women, especially among women with young children (Boden Jr, 1999; Lombard, 2001; Semykina, 2018).

To the best of our knowledge, the existing empirical investigations into the effects of fertility on labor market outcomes do not allow for the simultaneous endogeneity of childbearing decisions in both labor force participation and labor market outcomes while also accounting for sample selection. Failing to account for these concerns has the potential to lead to biased estimates due to omitted variables.

First, consider the effect of ignoring sample selectivity. Selection bias arises from the fact that only women who participate in the labor force have an observable wage. If the women for whom the negative impact of childbearing is most significant are the women most likely to remain out of the labor force, the analyses that ignores the labor force participation decision will potentially underestimate the family gap. Thus, correcting for sample selection could lead to more negative estimates. Alternatively, if lower-skilled women are more likely to leave the labor force in response to having a child, but higher-skilled women face a larger motherhood penalty (as suggested in Anderson et al. (2002); Bütikofer et al. (2018)), this could cause traditional estimates to overstate the family gap.[2]

[2]For example, there is evidence that highly educated women return to the workforce sooner than less educated women.

Overall, these two effects imply that ignoring sample selection could lead to biased results, but, a priori, the direction of this bias is ambiguous.

Secondly, let us consider the bias that might result from failing to account for endogenous childbearing in the wage equation. Specifically, consider an example where there is a negative economic shock that reduces labor market outcomes (such as wages) for individuals. This same shock could potentially motivate individuals to postpone childbearing until the economy improves. In this example, failing to account for endogenous childbearing in the wage equation is equivalent to omitting a variable that is negatively related to both childbearing and wages. Consequently, the bias would be positive, leading to an overestimation of the family gap. Alternatively, ignoring a shock that increases wages but postpones childbearing decisions (for example prospects of promotions etc) could underestimate family gap.

Finally, failing to account for endogenous fertility in the selection equation is similar to the bias resulting from failure to account for endogenous childbearing in the outcome equation as discussed above. Failure to allow for endogenous childbearing in the selection equation is comparable to omitting a variable such as a negative economic shock that reduces the likelihood of participating in the labor market, and the possibility of having kids would potentially overestimate the effect of childbearing on labor force participation. Alternatively, ignoring a shock that increases labor force participation but postpones childbearing (such as separation from the spouse) could underestimate the effect of childbearing on labor force participation.

Thus, it is reasonable to argue that the effects of ignoring the simultaneous endogeneity of fertility decisions (in both labor force participation and the wage equation) and the effect of ignoring sample selection on family gap remain theoretically ambiguous. This motivates us to estimate the family gap in a manner that helps us to account for sample selection as well as endogenous fertility decisions.

We will use the two-step procedure (Procedure 2.1) that accounts for sample selection into the labor force and allows for endogenous childbearing in both the outcome (hourly wages) and selection (labor force participation) equation. We compare our results to estimates obtained from estimation methods that fail to account for either sample selection or the endogeneity of fertility decisions (Pooled OLS and Fixed Effects) , only correct for endogenous fertility (FE2SLS), or only account for sample selection in the labor force (Heckman).

4.2 Data

We perform estimation of our model presented in Section 2 using data from the National Longitudinal Survey of Youth (NLSY79) data for 1982–2006. The NLSY79 is a nationally representative panel survey of men and women who were 14-22 years old in 1979. The survey was conducted annually from 1979-1994, then biannually thereafter. We restrict our analysis to non-Hispanic, white women who, at the time of the interview, were between 22-45 years of age, not in the military, and not attending school. We also exclude observations with missing values on any explanatory variables, such as region, education, and marital status. Because of how the NLSY calculates the hourly wage variable, some individuals receive extremely high or low wages. We follow the previous literature and exclude those observations where the reported hourly wage rate was less than $1 or more than $200[3]. Additionally, all analysis presented will show the results where the outcome variable is the log of the hourly wage. Finally, because we will use the sex composition of the first two children as an instrument, we exclude women who never have a child, who only have one child, or who have more than three children.[4] Our final sample includes 1,618 women, 20,837 person-year observations with 16,388 of those observations consisting of working women.

[3] All dollar values are expressed in real 1982-1984 dollars.

[4] In addition, we exclude women who suffered the loss of a child either before or during our survey period; this resulted in omitting approximately 852 observations (63 women) from the analysis

The NLSY79 contains cross-round variables for the birth date and the gender of each child. To capture fertility decisions, we construct a binary variable *Only one child present* that takes the value of one when the first child is born. The binary variable *Two children present* equals one once the second child is born, and the variable *Three children* equals one when the third child is born. We limit our sample to women who will have two or three children and thus treat the first and second child as an exogenous (predetermined) variable. Other explanatory variables include age, age squared, educational attainment indicators, an indicator for urban locales, marital status indicator, and region-specific indicators. We use the time means of all the exogenous variables to model the unobserved heterogeneity à la Mundlak (1978).

Table 2 presents the summary statistics for the sample, broken down by women who will and will not have three children. There are several differences to highlight between women who will have three children and those who will not. First, women who will have three children have a lower hourly wage rate (about 21% lower) and are less likely to participate in the labor force. Women with three children are also slightly less likely to live in an urban area or to have a college degree (or higher). Women with three kids also appear to disproportionately live in the north central region of the United States, and women with three children seem to have slightly lower standardized AFQT scores than those who will not have three kids. Finally, women with three children are more likely to have their first two kids of the same gender (roughly 17.2 percentage points). Other observable characteristics appear to be roughly balanced between the two groups.

Because we restrict our sample to women who will have two or three children, we treat "Three children present" as the endogenous variable. To instrument for our endogenous variable, we use the instrument proposed by Angrist and Evans (1996): sex composition of the first two children. This instrument has been widely used to study the impact of childbearing, specifically the decision to have a third child, on labor supply decisions. We

Table 2: Summary Statistics: White Women

	Never have 3 children	Will have 3 Children	Total
Hourly wage rate (in 1982-1984 dollars)	5.228	4.129	4.857
Labor Force Participation	0.814	0.733	0.786
North central	0.306	0.335	0.316
South	0.330	0.321	0.327
West	0.184	0.155	0.174
High school degree	0.456	0.481	0.464
Some college	0.198	0.197	0.198
College degree or greater	0.246	0.175	0.222
Age	30.960	30.878	30.932
Urban	0.703	0.658	0.688
Married	0.732	0.731	0.732
Samesex	0.259	0.431	0.317
Only one child present	0.233	0.155	0.206
Two children	0.528	0.267	0.440
Three children	0.000	0.418	0.141
AFQT	0.534	0.402	0.489
Observations	13803	7034	20837

Data is from NLSY79 1982-2006 for non-Hispanic, white women age 22-45 not currently enrolled in school. "Never have 3 children" is women who will only have two children throughout the survey period. "Will have 3 children" refers to women who will have three children during the survey period.

construct an indicator *samesex* that is equal to one if the first two children have the same gender. The panel nature of the data allows the instrument to vary over time. That is, *samesex* will turn to one after the woman has a second child of the same sex as her first. The underlying assumption is that the gender of the first two children should not impact the wages through any mechanism other than increasing the probability of having a third child.

As an exclusion restriction in the labor force participation equation, we use another commonly used instrument: an individual's Armed Forces Qualification Test (AFQT) score from 1979. (See Semykina (2018), Semykina and Wooldridge (2018)). Here, the underlying assumption is that a woman's ability, as captured by the AFQT score, should affect her decision whether or not to participate in the labor force; however, it should not affect her wages or her childbearing decisions after we account for other observable

variables such as educational attainment variables, age, and others. The AFQT scores are standardized to have mean zero and unit variance.

4.3 Empirical Results

The key variable of interest is the endogenous variable, "Three children." The coefficient on this variable captures the estimated effect of having a third child on a woman's log hourly wage. In Table 3, The estimate from Column (1) is -0.238, suggesting that the third child reduces women's hourly wage rate by approximately 23.8%. This result is approximately 43% larger than the estimate presented in Column (2). Column (2), which reports the pooled OLS results and fails to account for either endogeneity or sample selection, implies that the third child reduces women's hourly wage rate by approximately 16.7%. The results from FE estimation in Column (3) are much smaller than the estimates from Procedure 2.1, estimating the family gap of the third child at 7% percent. The result from the Fixed Effects 2SLS estimation, which fails to account for sample selection, in Column (4) is larger in magnitude than the OLS estimates. Column (4) also presents the F statistic for the instrument *samesex*. The F statistic is well above the "10" threshold generally recommended, and thus would not be considered a weak instrument.

Column (5) and (6) present the results from using Heckman's two-step procedure to correct for sample selection (but ignores the endogenous decision to have a third child). Column (5) presents the results from the outcome (wage) equation, and Column (6) shows the marginal effects of the selection equation. The coefficient on "Three children" in Column (5) is very similar to the result identified from using fixed effects. The coefficient on "Three children" implies that the third child is negatively related to a women's labor force participation. We also find that the AFQT score increases the likelihood of labor force participation and is statistically significant, suggesting that it is a useful instrument for selection.

Table 3: Effect of Children on Women's Log Wages: White Women

Estimation Strategy:	(1) Procedure 2.1	(2) OLS	(3) Fixed Effects	(4) Fixed Effects 2SLS	(5) Heckman Wage Eq.	(6) Marginal Effects Selection Eq.
High school degree	-0.059	0.189***	-0.021	-0.078**	-0.067*	0.033
	(0.040)	(0.015)	(0.030)	(0.032)	(0.039)	(0.024)
Some college	0.026	0.365***	0.079**	-0.024	0.023	0.053*
	(0.053)	(0.016)	(0.037)	(0.043)	(0.048)	(0.032)
College degree or greater	0.197***	0.622***	0.254***	0.149***	0.192***	0.087*
	(0.076)	(0.016)	(0.051)	(0.055)	(0.066)	(0.049)
Age	0.065***	0.060***	0.065***	0.102***	0.058***	-0.005
	(0.015)	(0.011)	(0.014)	(0.016)	(0.012)	(0.008)
Age squared	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Married	0.025*	0.045***	-0.001	0.031***	0.047***	-0.073***
	(0.013)	(0.009)	(0.009)	(0.011)	(0.012)	(0.009)
Only one child present	-0.007	-0.065***	0.005	-0.108***	0.037**	-0.174***
	(0.015)	(0.012)	(0.011)	(0.025)	(0.015)	(0.011)
Two children	-0.051***	-0.127***	-0.040***	-0.179***	0.020	-0.246***
	(0.019)	(0.012)	(0.012)	(0.030)	(0.016)	(0.010)
Three children	-0.238**	-0.167***	-0.071***	-0.406***	-0.073***	-0.249***
	(0.094)	(0.016)	(0.016)	(0.068)	(0.018)	(0.011)
Observations	20,837	16,388	16,388	16,347	20,837	20,837
F stat on *samesex*				934.49		

Data is from NLSY79 1982-2006 for age 22-45 not currently enrolled in school. All models include year indicators as well as indicators for the census region. The outcome variable is the log of women's hourly wage. Results from column (1), "Procedure 2.1", are from the econometric procedure described in 2. Standard errors are reported in parentheses. Standard Errors for the estimates from Procedure 2.1 are obtained using panel bootstrap with 1000 replications. * 0.10, ** 0.05 and *** 0.01 denote significance levels.

We find that the FE2SLS (which ignores the sample selection) tends to overstate the family gap and Heckman (which ignores the common endogeneity) estimates tend to understate the family gap compared to Procedure 2.1. This result highlights that one should exercise caution in applying standard instrumental variable estimation and/or applying standard Heckman correction methods when both nonrandom selection and a common endogenous dummy is present. If we ignore the estimates that address both the issues and simply compare the FE and FE2SLS estimates, one might mistakenly conclude that ignoring the endogenous fertility decisions severely underestimates the family gap. Similarly, if we only compare the FE and Heckman estimates, one might mistakenly conclude that ignoring sample selection does not dramatically impact estimates of the family gap. The estimates from Procedure 2.1 fall between the FE2SLS and the Heckman estimates.

When we control for the endogeneity of fertility decisions (both in wage equation and labor force participation equation) and sample selection, we find that the negative effect of having one child present to be about 0.7 percent, the negative effect of having two children present to be about 5 percent. The effects of having one child and two children as estimated from Procedure 2.1 are most similar to the fixed effects estimates.

To put our results in context, it may be helpful to compare our findings to those observed previously in the literature. A natural point of comparison, though not perfectly comparable, given the different datasets, is Angrist and Evans (1996). In Angrist and Evans (1996), the authors use the 1980 census, restricting their sample to women with more than two children present, and instrument the presence of a third child with the gender composition of the first two children. Angrist and Evans (1996) find that having more than two children is associated with a decrease in labor income for their sample of "all women" of approximately \$1,960[5], or a 27% reduction.[6] The estimates reported

[5]This estimate is reported in Table 7

[6]This estimate was calculated using the average labor income for "all women" (\$7,160), which is given in Table 2 of Angrist and Evans (1996)

from the FE2SLS estimates in Table 3 imply a larger reduction in the hourly wage rate of approximately 40.6%.[7] The estimated effect of three children using Procedure 2.1 suggests that the FE2SLS estimate overestimates the effect due to the failure to account for the sample selection.

Waldfogel (1997) uses the NLSY-Young Women cohort using data from 1968 through 1988 and finds that for all women, using pooled OLS, the effect of having one child is a roughly 5 percent reduction in wages and the effect of having more than one kids is near 14 percent reduction. She finds slightly larger effects when she restricts her sample to white women, 8 percent and 18 percent (comparable to our estimates), respectively. It is important to note that there are a few crucial differences between our data samples and estimation strategies that make a one-to-one comparison of point estimates difficult. For one, our sample is limited to women who eventually have two or three kids, thereby identifying the effect of the marginal, third child. Another important difference is that Waldfogel (1997) includes a number of mediator variables in her analysis in an attempt to identify potential mechanisms through which the family gap may operate. The goal of this paper is not to identify potential channels through which the family gap may operate, but instead to highlight the importance of accounting for endogeneity and sample selection using a new, computationally simple econometric methodology. Her finding that women with two or three children suffer a decrease in their wages by 18 percent (using the pooled OLS estimation method) is comparable with our pooled OLS estimates on two or three children thus suggesting that ignoring the endogeneity and selection underestimates the family gap.

[7] When we use the sample of women of all races (as is done in (Angrist and Evans, 1996), instead of the sub-sample of non-Hispanic, white women), the coefficient on "Three children" is -.289. This estimate is very close to findings in Angrist and Evans (1996). The coefficient on "Three children" using Procedure 2.1 is still smaller than the FE2SLS estimate at -0.176.

5 Concluding remarks

Frequently, empirical researchers find themselves working with panel data that suffers from nonrandom sample selection. Further complicating matters, researchers frequently need to incorporate an endogenous variable that simultaneously impacts the outcome and selection equation. Within labor economics alone, common examples of this phenomenon include studying the impact of fertility, marital, or educational decisions on wages. All of these decisions not only impact the wages observed by the researcher but may also impact the individual's decision to participate in the labor force. In this paper, we propose a computationally simple solution to this problem.

We develop a two-step estimation procedure for linear panel data models with unobserved effects that suffer from sample selection and have an endogenous dummy variable common to both the outcome and selection equations. We use a Mundlak (1978) device to model the unobserved effects in terms of the time means of the exogenous variables. The two-step procedure comprises estimating a simple biprobit model in the first step and calculating the correction terms as described in Section 2, then performing OLS estimation augmented with the correction terms in the second step. A bootstrap procedure easily obtains the standard errors.

Using Monte Carlo simulations, we show that in the presence of both nonrandom selection and common endogeneity, the proposed estimator performs better than the traditional estimates obtain through FE, FE2SLS, or Heckman (estimation methods that either ignore both endogeneity and selection or one of two the issues). Next, we illustrate our estimation procedure with an empirical application; specifically, we revisit the family wage gap- the phenomenon that women with children on average earn less than women without children. Specifically, we use NLSY data for white women from 1982-2006 and find that the proposed estimator yields results that are significantly smaller in magnitude than estimates obtained from FE2SLS or Heckman selection. The results imply ignoring

one source of bias, such as sample selection and/or common endogeneity issues, may lead to misleading conclusions.

One important limitation of the estimation method proposed in our paper is that we impose distributional assumptions on the idiosyncratic errors in the sense that the errors are assumed to have a trivariate homoskedastic normal distribution. This assumption is crucial as it not only helps us to obtain the correction terms, making our estimators are strictly parametric, but also makes the estimation procedure computationally simple. Further research in this area could involve extending the methodology is to relax this assumption and explore the nonparametric (or semiparametric) estimation procedures.

References

Aguero, J. M. and Marks, M. S. (2008). Motherhood and female labor force participation: evidence from infertility shocks. *American Economic Review*, 98(2):500–504.

Anderson, D. J., Binder, M., and Krause, K. (2002). The motherhood wage penalty: Which mothers pay it and why? *American economic review*, 92(2):354–358.

Anderson, D. J., Binder, M., and Krause, K. (2003). The motherhood wage penalty revisited: Experience, heterogeneity, work effort, and work-schedule flexibility. *ILR Review*, 56(2):273–294.

Andresen, M. and Nix, E. (2021). What causes the child penalty? evidence from adopting and same-sex couples.

Angrist, J. D. and Evans, W. N. (1996). Children and their parents' labor supply: Evidence from exogenous variation in family size. Technical report, National Bureau of Economic Research.

Bazen, S., Joutard, X., Périvier, H., et al. (2021). Measuring the child penalty early in a career: The case of young adults in france. Technical report, Institute of Labor Economics (IZA).

Boden Jr, R. J. (1999). Flexible working hours, family responsibilities, and female self-employment: Gender differences in self-employment selection. *American Journal of Economics and Sociology*, 58(1):71–83.

Budig, M. J., England, P., et al. (2001). The wage penalty for motherhood. *American sociological review*, 66(2):204–225.

Bütikofer, A., Jensen, S., and Salvanes, K. G. (2018). The role of parenthood on the gender gap among top earners. *European Economic Review*, 109:103–123.

Glauber, R. (2007). Marriage and the motherhood wage penalty among african americans, hispanics, and whites. *Journal of Marriage and Family*, 69(4):951–961.

Heckman, J. J. (1977). Dummy endogenous variables in a simultaneous equation system.

Hundley, G. (2000). Male/female earnings differences in self-employment: The effects of marriage, children, and the household division of labor. *ILR Review*, 54(1):95–114.

il Kim, K. (2006). Sample selection models with a common dummy endogenous regressor in simultaneous equations: a simple two-step estimation. *Economics Letters*, 91(2):280–286.

Jäckle, R. and Himmler, O. (2010). Health and wages panel data estimates considering selection and endogeneity. *Journal of Human Resources*, 45(2):364–406.

Kleven, H., Landais, C., Posch, J., Steinhauer, A., and Zweimuller, J. (2019). Child penalties across countries: Evidence and explanations. In *AEA Papers and Proceedings*, volume 109, pages 122–26.

Kunze, A. (2019). The effect of children on male earnings and inequality. *Review of Economics of the Household*, pages 1–28.

Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica: Journal of the Econometric Society*, pages 1335–1364.

Lombard, K. V. (2001). Female self-employment and demand for flexible, nonstandard work schedules. *Economic Inquiry*, 39(2):214–237.

Loughran, D. S. and Zissimopoulos, J. M. (2009). Why wait? the effect of marriage and childbearing on the wages of men and women. *Journal of Human resources*, 44(2):326–349.

Lundberg, S. and Rose, E. (2000). Parenthood and the earnings of married men and women. *Labour Economics*, 7(6):689–710.

Maurer, J., Klein, R., and Vella, F. (2011). Subjective health assessments and active labor

market participation of older men: evidence from a semiparametric binary choice model with nonadditive correlated individual-specific effects. *Review of Economics and Statistics*, 93(3):764–774.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society*, pages 69–85.

Newey, K. and McFadden, D. (1994). Large sample estimation and hypothesis. *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, pages 2112–2245.

Papke, L. E. and Wooldridge, J. M. (2008). Panel data methods for fractional response variables with an application to test pass rates. *Journal of econometrics*, 145(1-2):121–133.

Poirier, D. J. (1980). Partial observability in bivariate probit models. *Journal of econometrics*, 12(2):209–217.

Rochina-Barrachina, M. E. (1999). A new estimator for panel data sample selection models. *Annales d'Economie et de Statistique*, pages 153–181.

Semykina, A. (2018). Self-employment among women: Do children matter more than we previously thought? *Journal of Applied Econometrics*, 33(3):416–434.

Semykina, A. and Wooldridge, J. M. (2010). Estimating panel data models in the presence of endogeneity and selection. *Journal of Econometrics*, 157(2):375–380.

Semykina, A. and Wooldridge, J. M. (2018). Binary response panel data models with sample selection and self-selection. *Journal of Applied Econometrics*, 33(2):179–197.

Verbeek, M. and Nijman, T. (1992). Testing for selectivity bias in panel data models. *International Economic Review*, pages 681–703.

Waldfogel, J. (1997). The effect of children on women's wages. *American sociological review*, pages 209–217.

Waldfogel, J. (1998). Understanding the" family gap" in pay for women with children. *Journal of economic Perspectives*, 12(1):137–156.

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of econometrics*, 68(1):115–132.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wooldridge, J. M. (2019). Correlated random effects models with unbalanced panels. *Journal of Econometrics*, 211(1):137–150.