Double-Lasso Estimation of Heckman Sample Selection Model

Masayuki Hirukawa[*] **Di Liu**[†] **Irina Murtazashvili**[‡] **Artem Prokhorov**[§]

June 2022

Abstract

We extend the Heckman (1979) sample selection model by allowing for a large number of controls that are selected using lasso under a sparsity scenario. The standard lasso estimation is known to under-select causing an omitted variable bias in addition to the sample selection bias. We outline the required adjustments needed to restore consistency of lasso-based estimation and inference in such models. The adjustments include double lasso for both the selection equation and main equation and a correction of the variance matrix. We demonstrate the effect of the adjustments using simulations and we investigate the determinants of female labor market participation and earnings in the US using the new approach.

JEL Codes: C13
Key Words: Heckman, probit, double lasso, post selection inference

[*] Faculty of Economics, Ryukoku University, Kyoto, Japan; email: hirukawa@econ.ryukoku.ac.jp

[†] Stata Corp, College Station, Texas, USA; email: flyingliudi@gmail.com

[‡] Drexel University, LeBow College of business, Philadelphia, USA; email: im99@drexel.edu

[§] University of Sydney, Business School, NSW, Australia; email: artem.prokhorov@sydney.edu.au

1 Introduction

In this paper we consider an extension of the familiar Heckman (1979) sample selection model. In its traditional forms, it can be written as follows

$$
\begin{aligned}
y_{1i} &= \mathbf{x}_{1i}'\alpha + u_{1i}, &(1)\\
y_{2i} &= \mathbb{I}(\mathbf{x}_i'\beta + \mathbf{z}_i'\eta + u_{2i} \geq 0), &(2)
\end{aligned}
$$

where the outcome variable y_{1i} is observed only if the selection variable $y_{2i} = 1$. The main equation (1) contains a $k_1 \times 1$ vector of explanatory variables $\mathbf{x}_{1i}$ and we are interested in estimating and testing the coefficient vector α. In the selection equation (2), $\mathbb{I}(\cdot)$ denotes the indicator function, which takes the value one if its argument is true and zero otherwise. The explanatory variables of the selection equation are separated into two parts, $\mathbf{x}_i = (\mathbf{x}_{1i}, \mathbf{x}_{2i})$ and $\mathbf{z}_i$. In the traditional version of the model there is no distinction between $\mathbf{x}_{2i}$ and $\mathbf{z}_i$ – both represent the explanatory variables that are present in the selectivity model but not in the main equation of interest. These are the well known *exclusion restrictions* that facilitate identification of α. In our setting, for reasons that will become clear shortly we wish to differentiate between $\mathbf{x}_{2i}$ and $\mathbf{z}_i$.

Our extension is to introduce a high-dimensional vector of the explanatory variables in the selectivity model (2) which may or may not belong to the model. The vector $\mathbf{x}$ is a low-dimensional $k \times 1$ vector of selection determinants that we wish to keep in the model no matter what. The vector $\mathbf{z}$ is a high-dimensional $p \times 1$ vector of potential controls, where p can be as large as the (pre-selection) sample size N or larger and where we do not know which of these controls are important, if any. The vector β is a $k \times 1$ vector of coefficients on $\mathbf{x}$, which can be a target of inference too. The vector η, on the contrary, is just a $p \times 1$ nuisance parameter vector.

This extension has many empirical applications in economics where we have a well defined list of regressors for the main equation which has roots in economic theory (e.g., consumer and labor theory) while what determines selection into the sample is less certain (see, e.g., Roy, 1951; Heckman and Honore, 1990). The classic examples are the estimation of the female labor supply function and wage functions (see, e.g., Heckman, 1979; Arellano and Bonhomme, 2017), which may be subject to selection bias as determinants of the sample selection are confounded with the behavioral functions of interest. We return to women's labor force participation and labor supply decisions in our empirical application section.

Our objective is to consistently estimate α in the outcome equation (1) under a potential sample selection bias arising from the fact that in the observed sample

$$E(y_{1i}|\mathbf{x}_i, \mathbf{z}_i, y_{2i} = 1) = \mathbf{x}_{1i}'\alpha + E(u_{1i}|\mathbf{x}_i, \mathbf{z}_i, y_{2i} = 1) \neq \mathbf{x}_{1i}'\alpha,$$

unless $E(u_{1i}|\mathbf{x}_i, \mathbf{z}_i, y_{2i} = 1) = 0$, which is a questionable assumption in practice. Heckman (1979) assumed joint normality of (u_{1i}, u_{2i}) and showed that $E(u_{1i}|\mathbf{x}_i, \mathbf{z}_i, y_{2i} = 1) = \gamma\lambda(\mathbf{x}_i'\beta + \mathbf{z}_i'\eta)$, where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is known as the inverse Mills ratio. The two-step heckit procedure is (a) to run the MLE for the probit of y_{2i} on $(\mathbf{x}_i, \mathbf{z}_i)$ and use the estimates $(\widehat{\beta}, \widehat{\eta})$ to obtain $\widehat{\lambda}_i \equiv \lambda(\mathbf{x}_i'\widehat{\beta} + \mathbf{z}_i'\widehat{\eta})$ and then (b) to regress y_{1i} on x_{1i} and $\widehat{\lambda}_i$. Under correct specification, the resulting estimators $\widehat{\alpha}$ and $\widehat{\gamma}$ are consistent and the usual t-test on $\widehat{\gamma}$ can be used to test for selection bias. If the null of no bias is rejected, the standard errors of the second step have to be corrected for the first step estimation error which is done via a full MLE using normality of the errors or via an analytic correction to the variance in the second step.

The high-dimensionality of $\mathbf{z}_i$ poses a challenge in applying the traditional two-step procedure. First, we cannot include all the variables in $\mathbf{z}_i$ in the first step because there are too many variables. If p is larger than N, the probit with all $\mathbf{x}_i$ and $\mathbf{z}_i$ is infeasible, and even if p is substantially smaller than N but is large then including all these variables can cause difficulties in MLE convergence.

In order to make estimation feasible, it is common to impose a certain structure on η, known in the literature on regularized estimation as a sparsity scenario. In particular, we assume that only a few elements in the coefficient vector η are substantially different from zero. Although we assume η is sparse, we do not know which elements are non-zero and a consistent model selection technique is required. A popular approach to regularizing linear models is the least absolute shrinkage and selection operator (lasso) developed by Tibshirani (1996). The method penalizes the objective function with an l_l-norm of the coefficients. This shrinks the irrelevant coefficients to zero and thus serves as a model selection tool. However, even for purely linear models, this approach has well known challenges.

First, lasso makes mistakes. Failure to account for the fact that the covariates have been selected by lasso results in invalid inference. The reason is that lasso, like many other model selection techniques, does not always find all the relevant covariates especially when some coefficients are small. Model selection mistakes made by lasso cause the distribution of this naive estimator to be biased and nonnormal. For example, Leeb and Pötscher (2008a), Leeb and Pötscher (2008b), and Pötscher and Leeb (2009) showed that the normal approximation for the naive lasso estimator will produce misleading inference. Belloni, Chernozhukov, and Hansen (2014b), Belloni, Chernozhukov, and Wei (2016), and Chernozhukov et al. (2018) derive estimators that are robust to the mistakes made by lasso. Such robust estimators are often referred to as Neyman orthogonal (NO) estimators because they can be viewed as extensions of an approach proposed by Neyman (1959).

The second challenge is choosing the lasso tuning parameter. Lasso's ability to select relevant covariates depends on the method used to choose the tuning parameters. Belloni, Chernozhukov, and Hansen (2014b) propose a plug-in method and show that NO estimators perform well on linear models under that method. Belloni, Chernozhukov, and Wei (2016) extend the linear lasso to logit models and show good performance using a simplified version of the plug-in method. Drukker and Liu (2022) extend the plug-in method to cross-sectional generalized linear models and provide Monte Carlo evidence that their extension works well in finite samples.

In this paper, we develop NO estimation for the model in (1)-(2) which we call double-selection Heckman (DS-HECK) procedure. The DS-HECK estimator draws upon the classical two-step heckit estimator and the double-selection lasso for the high-dimensional generalized linear models proposed by Belloni, Chernozhukov, and Wei (2016). We detail the steps involved in the estimation, work out the estimator properties and derive the variance corrections.

The rest of the paper is organized as follows. Section 2 describes and studies the DS-HECK estimator. In Section 3, we present simulation results that demonstrate an excellent performance of DS-HECK in finite samples. In Section 4, we apply DS-HECK to estimate married women's wage using the 2013 PSID wave, in the presence of high-dimensional controls and potential sample selection bias. Finally, Section 5 concludes.

2 The DS-HECK estimator

2.1 Settings

We maintain the standard assumption of the Heckman sample selection model.

Assumption 1. *(a)* $(\mathbf{x}, \mathbf{z}, \mathbf{y}_2)$ *are always observed,* $\mathbf{y}_1$ *is observed only when* $\mathbf{y}_2 = 1$*; (b)* (u_1, u_2) *is independent of* $\mathbf{x}$ *and* $\mathbf{z}$ *with zero mean; (c)* $u_2 \sim N(0, 1)$*; (d)* $\mathbb{E}(u_1|u_2) = \gamma_1 u_2$*.*

Assumption 1 is in essence the same as in Wooldridge (2010, p. 803). Part (a) describes the nature of sample selection. Part (b) assumes that $\mathbf{x}$ and $\mathbf{z}$ are exogenous. Part (c) is restrictive but needed to derive the conditional expectation of $\mathbf{y}_1$ given that it is observed. Part (d) requires linearity in the conditional expectation of u_1 given u_2, and it holds when (u_1, u_2) is bivariate normal. However, it also holds under weaker assumptions when u_1 is not normally distributed.

Additionally, we impose a sparsity scenario on η.

Assumption 2. *η is sparse; that is, most of the elements of η are zeros. Namely, $||\eta||_0 \leq s$. We require s to be small relative to the sample size N. In particular, $\frac{s^2 \log^2(max(p,N))}{N} \longrightarrow 0$.*

This assumption follows Belloni et al. (2016). In the settings of generalized linear models, it allows for the estimation of the nuisance parameter in the selection equation at the rate $o(N^{-1/4})$ (see their Condition IR). In our settings, this rate is needed to guarantee the consistent estimation of β.

Under Assumption 1, it is easy to show that

$$\mathbb{E}(y_1|\mathbf{x}, \mathbf{z}, y_2 = 1) = \mathbf{x}_1'\alpha + \gamma\lambda(\mathbf{x}'\beta + \mathbf{z}'\eta), \tag{3}$$

where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the inverse Mills ratio. In essence, this is the classic formulation of Heckman (1979) where the presence of $\mathbf{x}_2$ and $\mathbf{z}$ in the selection equation (but not in the main equation of interest) permits estimation of the model even when the inverse Mills ratio is close to being linear in its argument .

We wish to explore the behavior of this conditional expectation with respect to potential errors in the choice of $\mathbf{z}$. It is easy to see from applying the mean-value theorem to the inverse Mills ratio evaluated at $\mathbf{x}'\beta$, that we can rewrite (3) as follows

$$\begin{aligned}\mathbb{E}(y_1|\mathbf{x}, \mathbf{z}, \mathbf{y}_2 = 1) &= \mathbf{x}_1'\alpha + \gamma\lambda(\mathbf{x}'\beta) + \gamma\lambda^{(1)}(q)\mathbf{z}'\eta \\ &= \mathbf{x}_1'\alpha + \gamma\lambda(\mathbf{x}'\beta) + \mathbf{z}'\omega \end{aligned} \tag{4}$$

where q is a point between $\mathbf{x}'\beta + \mathbf{z}'\eta$ and $\mathbf{x}'\beta$, $\lambda^{(1)}(\cdot)$ is the first-order derivative of $\lambda(\cdot)$, and $\omega = \gamma\lambda^{(1)}(q)\eta$. It is well known that $\lambda^{(1)}(\cdot)$ is monotone and bounded between -1 and 0 (see, e.g., Sampford, 1953). We note that ω depends on q, γ and η, and that if η sparse than ω is sparse too.

Proposition 1. *The vector ω in Eq. (4) inherits the same sparsity properties as the vector η, i.e. $||\omega||_0 = ||\eta||_0$.*

Proof. Sketches of the proofs of all less obvious propositions are given in the Appendix.

This Proposition makes it clear that a Heckman model with sparsity in the selection equation can be written as a heckit model with the same sparsity scenario in the main equation of interest.

Next we derive some conditions on the linear approximation of the inverse Mills ratio using $\mathbf{z}$ in the selected sample which is common for lasso-based model selection but new in the context of the Heckman model.

Let n denote the size of the selected sample, defined as follows,

$$n = \sum_{i=1}^{N} \mathbb{I}\{y_{2i} = 1\}.$$

Then, for the selected observations, we can write

$$y_{1i} = \mathbf{x}_{1i}'\alpha + g(\mathbf{x}_i, \mathbf{z}_i) + \epsilon_i,$$

where $g(\mathbf{x}_i, \mathbf{z}_i) = \gamma\lambda(\mathbf{x}_i'\beta + \mathbf{z}_i'\eta)$, $\mathbb{E}\epsilon_i|\mathbf{x}_i, \mathbf{z}_i = 0$ and $\mathbb{V}\epsilon_i|\mathbf{x}_i, \mathbf{z}_i = \sigma^2$. We follow Belloni et al. (2014b) and write $g(\mathbf{x}_i, \mathbf{z}_i)$ in a linear form subject to a bound on the approximation error:

$$y_{1i} = \mathbf{x}_{1i}'\alpha + \gamma\lambda(\mathbf{x}_i'\beta) + \mathbf{z}_i'\theta + r_i + \epsilon_i, \tag{5}$$

where $r_i, i = 1, \ldots, n$ is the approximation error such that $\sqrt{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}r_i^2} = O\left(\sqrt{\frac{s}{n}}\right)$. Additionally, we assume that the selected and pre-selection sample sizes are of the same order.

Assumption 3. *Equation 5 holds in the selected sample with $n = O(N)$ and*

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}r_i^2} = O\left(\sqrt{\frac{s}{n}}\right).$$

The assumption on the approximation error follows Belloni et al. (2014b). Similar to Assumption 2, Assumption 3 insures that we can estimate the nuisance parameter in the selected sample at the rate $o(n^{-1/4})$ rate. In the context of the Heckman model, it implies that $||\theta||_0 = ||\omega||_0 = ||\eta||_0 = s$ and that θ is also estimated at $o(N^{-1/4})$ since $n = O(N)$.

Next we investigate how to consistently estimate this model accounting for the high-dimensional nuisance parameter in both equations.

2.2 Estimation of the selection equation

Clearly if we knew the true value of β, we could treat $\lambda(\mathbf{x}'\beta)$ as a known variable and we could estimate α and γ, treating ω as nuisance parameters. So we start with a consistent estimation of β

in Eq. (2) using the approach of Belloni et al. (2016) combined with parameter tuning of Drukker and Liu (2022).

The estimation involves three steps:

Step 1 (post-lasso probit) We start by estimating a penalized probit of y_2 on $\mathbf{x}$ and $\mathbf{z}$ using the lasso penalty:

$$(\widehat{\beta}, \widehat{\eta}) = \arg\min_{\beta,\eta} \mathbb{E}_N(\Lambda_i(\beta, \eta)) + \lambda_1 ||(\beta, \eta)||_1,$$

where $\Lambda_i(\cdot)$ is the negative log-likelihood for the probit model, $||\cdot||_1$ is the lasso (l_1) norm of the parameters and λ_1 is a tuning parameter chosen using the plug-in method of Drukker and Liu (2022). This produces a subset of the variables in $\mathbf{z}$ indexed by $\mathit{support}(\widehat{\eta})$, where for a p-vector v, $\mathit{support}(v) := \{j \in \{1, ..., p\} : v_j \neq 0\}$. These variables are used in the post-lasso probit:

$$(\tilde{\beta}, \tilde{\eta}) = \arg\min_{\beta,\eta} \mathbb{E}_N(\Lambda_i(\beta, \eta)) : \mathit{support}(\eta) \subseteq \mathit{support}(\widehat{\eta})$$

As a result, we obtain the sparse probit estimates $(\tilde{\beta}, \tilde{\eta})$ where $\tilde{\eta}$ contains only a few non-zero elements. Belloni et al. (2016) propose using these estimates to construct weights $\widehat{f}_i = \widehat{w}_i / \widehat{\sigma}_i$, where $\widehat{w}_i = \phi(\mathbf{x}_i'\tilde{\beta} + \mathbf{z}_i'\tilde{\eta})$, and $\widehat{\sigma}_i^2 = \Phi(\mathbf{x}_i'\tilde{\beta} + \mathbf{z}_i'\tilde{\eta})(1 - \Phi(\mathbf{x}_i'\tilde{\beta} + \mathbf{z}_i'\tilde{\eta}))$, for $i = 1, \ldots, N$.

Step 2. We use the weights from Step 1, to run a weighted lasso regression in which for each variable x_j in $\mathbf{x}$, $j = 1, \ldots, k$, we run the penalized regression of $\widehat{f}_i x_{ij}$ on $\widehat{f}_i \mathbf{z}_i$,

$$\widehat{\theta}_j = \arg\min_{\theta_j} \mathbb{E}_N(\widehat{f}_i^2 (x_{ij} - \mathbf{z}_i'\theta_j)^2) + \lambda_2 ||\theta_j||_1,$$

where λ_2 is chosen by the plugin method of Drukker and Liu (2022). For each element of $\mathbf{x}$, this produces a selection from the variables in $\mathbf{z}$ indexed by $\mathit{support}(\widehat{\theta}_j), j = 1, \ldots, k$.

Step 3 (double selection probit). We use the variables selected from $\mathbf{z}$ in Steps 1 and 2 to run the probit of y_2 on $\mathbf{x}$ and the union of the sets of variables selected in Steps 1 and 2:

$$(\check{\beta}, \check{\eta}) = \arg\min_{\beta,\eta} \mathbb{E}_N(\Lambda_i(\beta, \eta) \widehat{f}_i / \widehat{\sigma}_i),$$

where $\mathit{support}(\eta) \subseteq \Big(\mathit{support}(\widehat{\eta}) \cup \mathit{support}(\widehat{\theta}_1) \cup \ldots \cup \mathit{support}(\widehat{\theta}_J)\Big)$.

Belloni et al. (2016) show that the double selection probit corrects for the omitted variable bias introduced by a naive application of lasso to Eq. (2). Under Assumption 2, $\check{\beta}$ is a consistent estimator of β and its variance can be obtained from Step 3 using the well known "sandwich" formula for probit. For example, in Stata it can be obtained using the `vce(robust)` syntax.

2.3 Estimation of the main equation

We can now return to the estimation of α and γ. Similar to Belloni et al. (2016), Belloni et al. (2014b) observed that the direct application of lasso to linear models with a large-dimensional nuisance parameter results in a biased estimation of the parameter of interest, which in their case is a scalar treatment effect. They propose a double selection procedure. We follow their approach subject to a few modifications that reflect the specifics of our main equation.

First, with a consistent estimator of β, a natural estimator of the inverse Mills ratio in Eq. (4) is as follows:

$$\widehat{\lambda(\mathbf{x}_i'\beta)} = \phi(\mathbf{x}_i'\check{\beta})/\Phi(\mathbf{x}_i'\check{\beta}).$$

It is also natural to account for the fact that this is a generated regressor when constructing the variance matrix, something we consider later.

Second, because the variables of interest $\mathbf{x}_1$ and $\lambda(\mathbf{x}'\beta)$ form a vector, we need to extend the original double selection lasso estimation to vectors. We provide the details of this extension using the NO arguments in Appendix A.

We can now discuss the estimation of the main equation which combines the double selection lasso of Belloni et al. (2014b) and parameter tuning by Drukker and Liu (2022). It proceeds in three steps:

Step 1. We run the lasso regression of $\mathbf{y}_1$ on $\mathbf{z}$

$$\check{\theta}_y = \arg\min_{\theta_y} \mathbb{E}_N \left[(y_{1i} - \mathbf{z}_i'\theta_y)^2\right] + \lambda_1||\theta_y||_1.$$

This produces a subset of $\mathbf{z}$ indexed by $support(\check{\theta}_y)$.

Step 2. For each variable x_{1j} in $\mathbf{x}_1$, $j = 1, \ldots, k_1$, we run the lasso regression of x_{1j} on $\mathbf{z}$:

$$\check{\theta}_j = \arg\min_{\theta_j} \mathbb{E}_N \left[(x_{1ij} - \mathbf{z}_i'\theta_j)^2\right] + \lambda_2||\theta_j||_1.$$

Additionally, we run the lasso regression of $\widehat{\lambda(\mathbf{x}_i'\beta)}$ on $\mathbf{z}$:

$$\check{\theta}_j = \arg\min_{\theta_\lambda} \mathbb{E}_N \left[(\widehat{\lambda(\mathbf{x}_i'\beta)} - \mathbf{z}_i'\theta_j)^2 \right] + \lambda_2 ||\theta_\lambda||_1.$$

This step produces subsets of $\mathbf{z}$ indexed by $support(\check{\theta}_j), j = 1, \ldots, k_1$, and $support(\check{\theta}_\lambda)$.

Step 3. We run the regression of y_{1i} on $\mathbf{x}_{1i}$, $\widehat{\lambda(\mathbf{x}_i'\beta)}$, and the union of the sets selected in Steps 1 and 2:

$$(\widehat{\alpha}, \widehat{\gamma}, \widehat{\theta}) = \arg\min_{\alpha,\gamma,\theta} \mathbb{E}_N \left[(y_{1i} - \mathbf{x}_{1i}'\alpha - \widehat{\lambda(\mathbf{x}_i'\beta)}\gamma - \mathbf{z}_i'\theta)^2 \right],$$

where $support(\theta) \subseteq support(\check{\theta}_y) \cup support(\check{\theta}_1) \cup \ldots \cup support(\check{\theta}_J)$

Proposition 2. *Under Assumptions 1-2, the double selection heck-it estimation in Steps 1-3 above is consistent for α and γ.*

The DS-Heckit estimator corrects the bias generated by applying the lasso directly to Eq. (4). The simulation experiments we report in Section 3 illustrate the size of the biases.

Following Belloni et al. (2014b), we can claim that inference about the vector (α', γ) is valid but, unlike Belloni et al. (2014b), it is valid up to the variance matrix correction reflecting the post-lasso probit estimation of β.

2.4 Variance matrix estimation

We start with some new notation. Let $\widehat{\lambda}_i = \widehat{\lambda(\mathbf{x}_i'\beta)} = \lambda(\mathbf{x}_i'\check{\beta})$ and define

$$\xi_i = \widehat{\lambda}_i(\widehat{\lambda}_i + \mathbf{x}_i'\check{\beta}),$$

where $\check{\beta}$ is obtained by the double selection probit. Let e denote the vector of residuals from the last step of the double-selection lasso estimation, with typical element $e_i, i = 1, \ldots, n$. That is,

$$e_i = y_{1i} - \mathbf{x}_{1i}'\widehat{\alpha} - \widehat{\lambda}_i\widehat{\gamma} - \mathbf{z}_i'\widehat{\theta},$$

where $support(\theta) \subseteq support(\check{\theta}_y) \cup support(\check{\theta}_1) \cup \ldots \cup support(\check{\theta}_J)$. Let W denote the matrix containing $\mathbf{x}_1$, the $n \times 1$ vector of $\widehat{\lambda}_i$'s, and the variables in $\mathbf{z}$ that survived the double selection.

Let R be a $n \times n$ diagonal matrix, with diagonal elements $(1 - \widehat{\rho}^2 \xi_i)$, where $\widehat{\rho} = \widehat{\gamma}/\widehat{\sigma}$ and $\widehat{\sigma}^2 = (e'e + \widehat{\gamma}^2 \sum_i \xi_i)/n$.

Proposition 3. *A consistent estimator of the variance matrix of the double selection heckit estimator* $(\widehat{\alpha}', \widehat{\gamma}, \widehat{\theta}')$ *is*

$$V = \widehat{\sigma}^2 (W'W)^{-1}(W'RW + Q)(W'W)^{-1},$$

where

$$Q = \widehat{\rho}^2 (W'D\mathbf{x}) V_b (\mathbf{x}'DW)$$

where V_b *is the "sandwich" variance matrix for the double selection probit estimator* $\check{\beta}$ *and* D *is the diagonal matrix with diagonal elements* ξ_i*.*

The variance for $\widehat{\alpha}$ and $\widehat{\gamma}$ is the upper $(k_1 + 1) \times (k_1 + 1)$ submatrix of V.

3 Monte Carlo Simulations

To evaluate the finite-sample performances of the double selection heckit estimator, we conduct a simulation study using four estimators: (i) ordinary least squares on the selected sample (*OLS*), (ii) Heckman two-step estimator based on the true model (*Oracle*), (iii) Heckman two-step estimator using lasso to select variables in Eq. (2) (*Naive*), and the proposed double selection Heckman estimator (*DS*) [1]

OLS is inconsistent unless there is no sample selection bias, i.e. $\gamma = 0$. *Naive* is inconsistent due to error made by lasso. Moreover, *Naive* does not provide valid inference as it is not robust to the model selection bias. In contrast, *DS* is expected to retain consistency in the presence of sample selection biases and show robustness against the model selection bias. *Oracle* is expected to behave like the standard Heckman estimator under the true model but, in practice, *Oracle* is infeasible since we do not know the true model.

[1] The double selection Heckman estimator is implemented as a Stata command `dsheckman`. The syntax is described in C.

3.1 Setup

Our data generating process is as follows

$$\begin{aligned} y_1 &= 1 + x_1 + x_2 + u \\ y_2 &= \mathbf{1}(w_0\eta_0 + v > 0), \end{aligned}$$

where $v \sim N(0,1)$ and where $u = \gamma v + \epsilon$, with $\epsilon \sim N(0,1)$. We vary the strength of the selection bias by setting γ to be $0.4, 0.5, 0.6, 0.7$, and 0.8 and we observe y_1 only when $y_2 = 1$.

The selection equation is generated using nine non-zero variables in $\mathbf{z}$ of which four have a relative large effect and five relatively small:

$$w_0\eta_0 = -1.5 + x_1 - x_2 + z_1 - z_2 + 0.046z_3 + z_5 - 0.046z_{10} - 0.046z_{11} + 0.046z_{12} - 0.046z_{15} + z_{20}.$$

The value 0.046 is chosen so that it violates the so called "beta-min" condition and causes lasso to make model selection mistakes (see, e.g., Liu et al., 2020; Drukker and Liu, 2022). The sample size is 2000. The number of replications is 1000.

We consider two scenarios for p, the dimension of $\mathbf{z}$: (i) $p = 1000$, fewer variables than observations; (ii) $p = 2100$, more variables than observations. The variables are generated using a Toeplitz correlation structure with decreasing dependence. In particular, let Z be the matrix of dimension $N \times p$ containing $\mathbf{z}$, then

$$Z = ML'$$

where M is $N \times p$ and has the typical element $(\zeta_{ij} - 15)/\sqrt{30}$, where $\zeta_{ij} \sim \chi^2(15)$ and where L is the Cholesky decomposition of a symmetric Toeplitz matrix V of dimension $p \times p$ such that its elements obey the following laws: $V_{i,j} = V_{i-1,j-1}$ and $V_{1,j} = j^{-1.3}$.

The variables $\mathbf{x}$ are also correlated and they are generated as functions of $\mathbf{z}$. In particular,

$$\begin{aligned} x_1 &= z_3 + z_{10} + z_{11} + z_{12} + z_{15} + \epsilon_{x_1} \\ x_2 &= 0.5(z_3 + z_{10} + z_{11} + z_{12} + 2z_{15}) + \epsilon_{x_2} \end{aligned}$$

where ϵ_{x_1} and ϵ_{x_2} follow a Toeplitz structure similar to Z.

As a result, for the selected sample, the true model is

$$y_1 = 1 + x_1 + x_2 + \gamma\lambda(w_0\eta_0) + u$$

where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ and the true parameter values are $\beta_1 = \beta_2 = 1$ and $\gamma = 0.4, 0.5, 0.6, 0.7$, or 0.8.

3.2 Results

For each estimator, we report the following measures: (i) true value of parameter (*True*), (ii) mean squared error (*MSE*), (iii) average of estimates across simulations (*Mean*) (iv) standard deviation of estimates across simulations (*SD*), (v) average standard error across simulations ($\overline{SE}$) and (vi) rejection rate for the H_0 that the parameter equals its true value against the nominal 5% level of significance (*Rej. rate*).

We report the simluation results for β_1, β_2, and γ in Tables (1), (2), and (3), respectively. Several observations are clear from the tables. First, *OLS* is badly biased and the bias is greater when selection is stronger. Second, *Naive* is also biased and it fails to provide valid inference at any value of γ and p. This demonstrates that *Naive* is not robust to the model selection errors. The rejection rate and MSE increase as γ increases, which is expected because greater γ value indicate a greater sample selection bias. Third, *Oracle* shows consistency and rejection rates close to the nominal 5% significance level. Fourth, *DS* performs similarly to *Oracle* for all values of γ and p. In particular, its MSE is consistently smaller than for *Naive* and *OLS*, $\overline{SE}$ is close to *SD*, which shows that the proposed variance adjustment works well. Finally, *Rej. rate* for *DS* is near the 5% significance level, which supports that our estimator offers valid inference.

TABLE 1 ABOUT HERE

TABLE 2 ABOUT HERE

TABLE 3 ABOUT HERE

4 Application to Female Earnings Estimation

4.1 Labor force participation and earnings

Estimation of the female earnings equation is a topic of long standing interest among economists. Early investigations of the labor supply decisions including both participation and hours by married women date back to Gronau (1974) and Heckman (1974), who were among the first to highlight sample selection bias stemming from the labor supply decision. Labor market decisions by women over the later decades have been studied and documented by Mroz (1987); Ahn and Powell (1993); Neumark and Korenman (1994); Vella (1998); Devereux (2004); Blau and Kahn (2007); Mulligan and Rubinstein (2008); Cavalcanti and Tavares (2008); Bar et al. (2015), among others.

Numerous applied works have extensively scrutinized the empirical aspects of the sample selection problem when estimating female labor market behavior. The determinants of the earnings equations for married women are similar to those of men and have been mostly agreed upon. These determinants traditionally include women's education, experience, age, tenure and location. The hallmark of correcting for sample selection bias is finding some appropriate exclusion restriction(s) (i.e., variable(s) affecting selection but not the earnings) in order to ensure proper identification of the model parameters. Two main competing choices of such exclusion restrictions have been exploited for estimation of labor market earnings for married women: non-wife/husband's income and the existence or number of (young) children. The underlying argument is that these two sets of variables affect the labor supply decision of married women but not their earnings. Huber and Mellace (2014) provide an overview of the related literature on sample selection bias in the female earnings equations.

Cavalcanti and Tavares (2008) provide an alternative view and argue that the declining price and wider availability of home appliances play a crucial role in explaining the rise in female labor force participation. This suggests a long list of potential exclusion restrictions. Furthermore, the exact nature and functional form of the chosen exclusion restriction(s) in the selection equation is uncertain. For example, should labor work experience include years of part-time employment? Should educational attainment be measured in full years of completed education or in millstones such as high school or college, as a replacement or complement to years of education? Similarly, should age enter the model in a linear or quadratic form?

Our goal in this section is to illustrate the performance of the double-selection Heckman procedure on the following earnings equation :

$$\log(earnings) = \alpha_0 + \alpha_1 education + \alpha_2 \mathbf{x}_1 + \alpha_3 state\ dummies + u_1, \tag{6}$$

where $\log(earnings)$ is the natural logarithm of the individual's total annual labor income, *education* is the person's completed years of education, *state dummies* is a vector containing a full set of state dummies, and u_1 is an idiosyncratic error. The vector $\mathbf{x_1}$ varies across the exact specification we consider and can contain age and/or work experience.

To address the potential self-selection bias we employ DS-HECK as a data-driven procedure for choosing the explanatory variables and functional form (among high-dimensional options provided) in the following labor force participation equation:

$$inlf = \mathbf{x}\beta + \mathbf{z}\eta + u_2, \tag{7}$$

where *inlf* is the dummy variable that is equal to one for those women who are in the labor force at the time of the interview and zero otherwise, and u_2 is an idiosyncratic error. The vector $\mathbf{x}$ includes all the explanatory variables from Eq. (6) (both in a liner and quadratic functional form) as well as exclusion restrictions.

In practice, $\mathbf{x}$ is constructed as follows. To simplify notation, denote all the explanatory variables from Eq. (6) as $\mathbf{x_1}$. First, running lasso probit of *inlf* on high-dimensional controls $\mathbf{w}$, where $\mathbf{w}$ includes $\mathbf{x_1}$ and some other high-dimensional controls. Denote the selected variables as $\mathbf{x_2}$. Second, $\mathbf{x}$ is union between all $\mathbf{x_1}$ and $\mathbf{x_2}$. All the non-selected controls in $\mathbf{w}$ are used as $\mathbf{z}$.

4.2 Sample construction

We obtain our sample from the 2013 wave of the Panel Study of Income Dynamics (PSID) where we focus on the sub-population of white married women. The choice of explanatory variables reflects their availability in the PSID and the traditional set of regressors used in the existing literature on female labor force participation and earnings. Specifically, the explanatory variables we collect from the PSID include information on the educational attainment of the individual (both as the number of years completed and as a set of indicators for milestone achievements in education), a set of indicators for whether the individual obtained her education in the USA, outside the USA, or both, as well as a set of indicators for the educational levels of the individual's parents, work experience of the individual, age and geographical location of the individual (captured by a set of dummy variables for the current state where the individual is located), a set of indicators reflecting the Beale-Ross rural-urban continuum code for the individual's current residence, and an indicator for whether the individual owns a vehicle. Table 4 contains a description of the key explanatory variables.

TABLE 4 ABOUT HERE

The set of (potential) exclusion restrictions includes the number of children in the household under 18 years of age, an indicator for whether there are any children age 15 years old or younger in the individual's household, annual labor income of the husband, child care expenses, and household major expenditure (i.e. expenditure on household furnishings and equipment, including household textiles, furniture, floor coverings, major appliances, small appliances and miscellaneous housewares). While admittedly far from ideal, this last variable is the closest information we find in the PSID to capture household expenditure on major household appliances, which allows us to test the argument of Cavalcanti and Tavares (2008). Finally, the dependent variables for the earnings and selection equations are (the natural logarithm of) the individual's total annual labor income and the indicator for whether the individual is in the labor force, respectively.

Our sample contains 1,989 white married women, of whom 1,294 are in the labor force and 695 are not. Table 5 reports summary statistics for key variables in the dataset. A set of dummy variables for the current state as well as a set of indicators reflecting the Beale-Ross rural-urban continuum code are omitted to save space. A total of 46 states are present in our sample, with Delaware, District of Columbia, Hawaii, New Mexico, and Rhode Island omitted from our sample during data cleaning. We note that some women report being neither employed nor (temporarily) unemployed while also reporting non-zero labor income during that time. There are 161 such women in the sample. We treat these individuals as not being in the labor force.

TABLE 5 ABOUT HERE

4.3 Empirical findings

We consider several specifications when estimating the earnings equation subject to sample selection bias. Table 6 reports the key results for both equations obtained using DS-HECK. The top panel provides coefficient estimates (as well as their standard errors) for α_1, α_2 and the coefficient on the inverse Mills ratio while the bottom panel provides estimates (and standard errors) for β. In addition to the reported estimates, each specification contains two more sets of of estimates which we do not report in the table to save space. First, we do not report the coefficients on the full set of state and urban-rural dummies included in both equations. Second, for each specification there are the selected controls in both equations; the number of such controls is reported at the bottom of the respective panels but the coefficients themselves are not reported.

We note that following the original Heckman specification, the explanatory variables present in the earnings equation are always kept in the labor force participation equation. The lasso selection is not applied to them, only to the additional controls. Columns (1) and (2) report the estimates when work experience enters the two equations, with and without the quadratic form, while age is not included. Columns (3) and (4) report the estimate when age is included but experience is not. Columns (5) and (6) contain both but differ in whether experience squared is included. Column (7) reports the estimates of the traditional Heckman specification where no selection is done over the additional controls.

TABLE 6 ABOUT HERE

As Table 6 suggests, the signs of all the reported coefficient estimates are as expected, and they are highly significant for the most part. We note that from the five potential exclusion restrictions, the lasso selects child care expenditure as the relevant covariate for the labor force participation equation - this variable is not present in the earnings equation. Finally, we note that the results for the labor force participation equation reported in column (7) are similar to those reported in columns (5) and (6) except that the three additional exclusion restrictions turn out to be statistically significant in the traditional setting.

Next we focus on the estimates of the labor income equation.According to the results reported in Table 6, we conclude that the educational level of the individual plays a crucial role in explaining labor income for white married women in 2012. When statistically significant, the estimated rate of return to education ranges from 5.6% to almost 9% depending on the specification. Furthermore, there is evidence that the individual's age is more important, both statistically and practically, than work experience for explaining the individual's labor income in our sample. We note that when the individual's age (in any functional form) is used in the labor income equation, the rate of return to education is statistically significant.

Most importantly, Table 6 suggests that the inverse Mills ratio is highly statistically significant in all specifications implying that the correction for the sample selection was needed. Given the economic interpretation of the estimated coefficients, their signs and economic as well as statistical significance, specification (5) seems most attractive in light of the existing studies on the topic. Interestingly, the traditional Heckman specification produces results that are close to those reported in column (5).

5 Conclusion

We have proposed an extension to the traditional Heckman sample selection model by incorporating it into a model selection framework with many controls. A double application of the lasso to each part of the model permits valid inference on the parts of the model that is of interest to the empirical economist. We detail the steps involved in a consistent estimation with valid standard errors and we provide a new Stata command to implement it.

Lasso and double selection in linear models have been recently subject of scrutiny in cases when lasso under-selects controls in finite samples even under a sparsity scenario and the double selection estimators have severe omitted variable biases (see, e.g., Wuthrich and Zhu, 2021; Lahiri, 2021). This happens when the signal from the variables lasso works on is weak and they do not get selected by either of the two selection procedures. The solution proposed by Wuthrich and Zhu (2021) is to resort in such cases to the regular OLS estimation using a high-dimensional variance matrix computations which is computationally difficult and works only when $p < n$.

We showed how substantial the errors committed by lasso can be and we provide an application to a class problem in labor economics where using our method leads to a few new insights. We provide a user-friendly and versatile Stata command, which can help empirical economists use the proposed methodology. The command as well as the simulation and application data are made available on the authors web pages.

Finally we note that the results of this paper can be extended to other consistent methods of model selection beyond lasso, such as the Dantzig Selection Candes and Tao (2007).

Table 1: **Simulation results for** β_1

	$p = 1000$						$p = 2100$				
	True	MSE	Mean	SD	$\overline{SE}$	Rej. Rate	MSE	Mean	SD	$\overline{SE}$	Rej. Rate
$\gamma = 0.4$											
Oracle	1	0.0014	0.9984	0.0369	0.0374	0.0450	0.0013	0.9985	0.0362	0.0374	0.0510
DS	1	0.0025	0.9977	0.0501	0.0484	0.0540	0.0022	0.9991	0.0464	0.0486	0.0380
Naive	1	0.0049	0.9424	0.0395	0.0350	0.4200	0.0052	0.9379	0.0360	0.0349	0.4310
OLS	1	0.0046	0.9422	0.0350	0.0350	0.3890	0.0045	0.9421	0.0339	0.0350	0.3610
$\gamma = 0.5$											
Oracle	1	0.0014	0.9993	0.0380	0.0387	0.0430	0.0015	0.9980	0.0389	0.0387	0.0480
DS	1	0.0025	1.0000	0.0504	0.0502	0.0570	0.0027	0.9979	0.0518	0.0502	0.0600
Naive	1	0.0067	0.9289	0.0409	0.0362	0.5240	0.0075	0.9229	0.0389	0.0361	0.5840
OLS	1	0.0063	0.9291	0.0353	0.0362	0.4930	0.0066	0.9277	0.0373	0.0363	0.5190
$\gamma = 0.6$											
Oracle	1	0.0016	0.9995	0.0395	0.0401	0.0440	0.0015	0.9984	0.0393	0.0400	0.0520
DS	1	0.0025	1.0001	0.0504	0.0519	0.0460	0.0027	0.9981	0.0522	0.0521	0.0500
Naive	1	0.0089	0.9183	0.0475	0.0376	0.6110	0.0101	0.9081	0.0413	0.0373	0.6880
OLS	1	0.0087	0.9148	0.0374	0.0376	0.6190	0.0087	0.9144	0.0370	0.0376	0.6150
$\gamma = 0.7$											
Oracle	1	0.0019	0.9969	0.0435	0.0416	0.0690	0.0018	1.0001	0.0419	0.0419	0.0510
DS	1	0.0032	0.9956	0.0564	0.0539	0.0570	0.0028	1.0008	0.0525	0.0543	0.0410
Naive	1	0.0126	0.8999	0.0508	0.0389	0.7400	0.0135	0.8930	0.0455	0.0390	0.7640
OLS	1	0.0118	0.8993	0.0401	0.0392	0.7170	0.0115	0.9005	0.0397	0.0393	0.7110
$\gamma = 0.8$											
Oracle	1	0.0019	0.9962	0.0437	0.0434	0.0470	0.0019	0.9968	0.0438	0.0434	0.0510
DS	1	0.0031	0.9965	0.0560	0.0558	0.0470	0.0034	0.9966	0.0582	0.0562	0.0530
Naive	1	0.0165	0.8835	0.0544	0.0406	0.8020	0.0173	0.8777	0.0484	0.0404	0.8360
OLS	1	0.0152	0.8838	0.0408	0.0409	0.8150	0.0148	0.8858	0.0414	0.0408	0.7870

Table 2: **Simulation results for** β_2

		$p = 1000$					$p = 2100$				
	True	MSE	Mean	SD	$\overline{SE}$	Rej. Rate	MSE	Mean	SD	$\overline{SE}$	Rej. Rate
$\gamma = 0.4$											
Oracle	1	0.0028	1.0018	0.0527	0.0524	0.0530	0.0027	1.0014	0.0518	0.0524	0.0500
DS	1	0.0033	1.0015	0.0576	0.0584	0.0390	0.0034	1.0025	0.0580	0.0587	0.0530
Naive	1	0.0145	1.1018	0.0641	0.0530	0.5210	0.0151	1.1093	0.0564	0.0532	0.5340
OLS	1	0.0057	1.0541	0.0523	0.0512	0.1820	0.0054	1.0538	0.0506	0.0512	0.1760
$\gamma = 0.5$											
Oracle	1	0.0029	0.9998	0.0536	0.0542	0.0470	0.0031	1.0029	0.0558	0.0542	0.0620
DS	1	0.0036	1.0005	0.0600	0.0603	0.0380	0.0038	1.0030	0.0613	0.0606	0.0600
Naive	1	0.0200	1.1249	0.0659	0.0547	0.6420	0.0224	1.1365	0.0614	0.0548	0.7020
OLS	1	0.0070	1.0651	0.0523	0.0530	0.2460	0.0077	1.0683	0.0546	0.0530	0.2540
$\gamma = 0.6$											
Oracle	1	0.0031	1.0014	0.0555	0.0562	0.0510	0.0030	1.0020	0.0548	0.0560	0.0450
DS	1	0.0038	1.0034	0.0620	0.0623	0.0450	0.0038	1.0038	0.0615	0.0628	0.0410
Naive	1	0.0271	1.1453	0.0775	0.0566	0.7150	0.0307	1.1626	0.0653	0.0567	0.8000
OLS	1	0.0095	1.0804	0.0550	0.0550	0.3100	0.0092	1.0795	0.0536	0.0549	0.2950
$\gamma = 0.7$											
Oracle	1	0.0036	1.0031	0.0596	0.0583	0.0620	0.0034	0.9997	0.0582	0.0585	0.0430
DS	1	0.0042	1.0034	0.0649	0.0648	0.0590	0.0043	0.9994	0.0653	0.0657	0.0430
Naive	1	0.0377	1.1755	0.0833	0.0587	0.8170	0.0414	1.1902	0.0725	0.0591	0.8680
OLS	1	0.0121	1.0934	0.0583	0.0573	0.3740	0.0118	1.0925	0.0572	0.0575	0.3640
$\gamma = 0.8$											
Oracle	1	0.0037	1.0060	0.0606	0.0607	0.0440	0.0037	1.0016	0.0611	0.0608	0.0520
DS	1	0.0049	1.0064	0.0698	0.0676	0.0630	0.0049	1.0010	0.0703	0.0683	0.0530
Naive	1	0.0505	1.2059	0.0898	0.0611	0.8690	0.0519	1.2143	0.0777	0.0613	0.8990
OLS	1	0.0158	1.1106	0.0598	0.0598	0.4510	0.0147	1.1052	0.0600	0.0598	0.4250

Table 3: **Simulation results for γ**

	$p = 1000$						$p = 2100$				
	True	MSE	Mean	SD	$\overline{SE}$	Rej. Rate	MSE	Mean	SD	$\overline{SE}$	Rej. Rate
Oracle	0.4	0.0092	0.3948	0.0959	0.0948	0.0490	0.0087	0.3969	0.0933	0.0952	0.0430
DS	0.4	0.0096	0.3873	0.0970	0.0943	0.0580	0.0091	0.3877	0.0946	0.0942	0.0510
Naive	0.4	0.0115	0.3658	0.1017	0.0982	0.0710	0.0108	0.3646	0.0979	0.0988	0.0590
Oracle	0.5	0.0102	0.4940	0.1011	0.0973	0.0540	0.0093	0.4948	0.0961	0.0972	0.0460
DS	.05	0.0107	0.4843	0.1022	0.0969	0.0650	0.0095	0.4837	0.0963	0.0964	0.0550
Naive	0.5	0.0135	0.4558	0.1077	0.1010	0.0820	0.0125	0.4537	0.1017	0.1013	0.0760
Oracle	0.6	0.0092	0.5952	0.0958	0.1000	0.0390	0.0098	0.5945	0.0988	0.1000	0.0460
DS	0.6	0.0098	0.5835	0.0977	0.0998	0.0420	0.0102	0.5810	0.0995	0.0993	0.0500
Naive	0.6	0.0135	0.5454	0.1024	0.1041	0.0780	0.0142	0.5450	0.1058	0.1050	0.0800
Oracle	0.7	0.0116	0.6903	0.1072	0.1030	0.0600	0.0104	0.6999	0.1020	0.1031	0.0410
DS	0.7	0.0124	0.6765	0.1091	0.1029	0.0770	0.0111	0.6830	0.1040	0.1025	0.0510
Naive	0.7	0.0172	0.6359	0.1145	0.1080	0.1170	0.0153	0.6453	0.1111	0.1091	0.0740
Oracle	0.8	0.0117	0.7912	0.1078	0.1064	0.0510	0.0117	0.7831	0.1071	0.1066	0.0560
DS	0.8	0.0131	0.7762	0.1119	0.1065	0.0720	0.0130	0.7657	0.1088	0.1062	0.0820
Naive	0.8	0.0200	0.7260	0.1204	0.1124	0.1380	0.0202	0.7213	0.1182	0.1132	0.1240

Table 4: **Description of key PSID variables**

Variable	Type	Definition
Controls		
Education	Continuous	Years of education
High school education	Categorical	Graduated from high school (3 categories)
Enrolled in school	Categorical	If currently enrolled in regular school (2 categories)
College attendance	Categorical	If attended college (2 categories)
Other degree or certificate	Categorical	If received other degree/certificate (2 categories)
US education	Categorical	If the individual obtained her education in the USA, outside the USA, or both (3 categories)
Father's education	Categorical	Educational level of the individual's father (8 categories)
Mother's education	Categorical	Educational level of the individual's mother (8 categories)
If owns a vehicle	Categorical	If the individual owns a vehicle (2 categories)
Current state	Categorical	Geographical location of the individual (46 states)
Rural-urban location	Categorical	Beale-Ross rural-urban continuum code for the individual's current residence (9 categories)
Potential exclusion restrictions		
Number of kids	Continuous	The number of children in the household under 18 years of age
If kids less than or equal to 15 years old	Categorical	If there are any children age 15 years old or younger in the individual's household (2 categories)
Child care expenditure	Continuous	Child care expenses (in thousand dollars)
Husband's labor income	Continuous	Annual labor income of the individual's husband (in thousand dollas)
Household major expenditure	Continuous	Expenses on household furnishings and equipment, including household textiles, furniture, floor coverings, major appliances, small appliances and miscellaneous housewares (in thousand dollars)

Table 5: **Key sample characteristics**

	In the labor force		Not in the labor force		All	
	Mean	SD	Mean	SD	Mean	SD
Controls						
Age	43.84	11.74	53.63	16.17	47.26	14.24
Experience	9.86	6.64	12.21	9.30	10.68	7.75
Education (in years)	14.78	1.98	13.96	2.16	14.49	2.08
Graduated from high school (1-3 range)	1.05	0.27	1.14	0.46	1.09	0.35
If currently enrolled in regular school	0.03	0.17	0.03	0.16	0.03	0.17
If attended college	0.82	0.39	0.68	0.47	0.77	0.42
If received other degree/certificate	0.21	0.41	0.20	0.40	0.20	0.40
Received education in the USA (1-3 range)	1.04	0.30	1.04	0.27	1.04	0.29
Father's educational level (1-8 range)	5.08	1.80	4.56	1.92	4.90	1.86
Mother's educational level (1-8 range)	4.96	1.61	4.54	1. 64	4.81	1.64
If owns a vehicle	0.99	0.09	0.98	0.15	0.99	0.12
Exclusion Restrictions						
Number of kids under 18 years old	1.00	1.17	0.82	1.27	0.93	1.21
If kids less than or equal to 15 years old	0.47	0.50	0.35	0.48	0.43	0.49
Child care expenditure	1.47	3.91	0.25	1.64	1.04	3.35
Husband's labor income	61.49	81.16	62.76	195.65	61.93	132.85
Household major expenditure	1.49	2.90	1.20	2.69	1.39	2.83
Outcomes						
Labor income	45,575.65	43,701.59	12,439.52	15,746.09	41,909.04	42,822.50
Log (Labor income)	10.37	0.98	8.52	1.58	10.17	1.21
If in the labor force	1	0	0	0	0.65	0.48
Number of Observations (N)	1,294		695		1,989	

Table 6: **Estimation results**

Earnings equation							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Education (years)	0.053	0.047	0.089***	0.056*	0.089***	0.060*	0.075***
	(0.034)	(0.040)	(0.023)	(0.034)	(0.023)	(0.033)	(0.016)
Experience	0.026***	0.006			-0.000	-0.018	0.001
	(0.007)	(0.018)			(0.015)	(0.021)	(0.014)
Experience^2		0.001				0.001	0.000
		(0.001)				(0.001)	(0.000)
Age			0.019***	-0.090	0.018***	-0.086	0.002
			(0.004)	(0.057)	(0.005)	(0.055)	(0.022)
Age^2				0.001*		0.001*	0.000
				(0.001)		(0.001)	(0.000)
λ	-1.336***	-1.463***	-0.838***	-1.584***	-0.924***	-1.581***	-0.942***
	(0.388)	(0.484)	(0.146)	(0.456)	(0.171)	(0.455)	(0.156)
Number of observations (N)	1294	1294	1294	1294	1294	1294	1294
Number of controls	35	34	34	34	35	34	
Number of selected controls	19	19	20	21	21	21	

Labor force participation equation							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Education (years)	0.066***	0.066***	0.074***	0.074***	0.070***	0.069***	0.067***
	(0.024)	(0.024)	(0.025)	(0.025)	(0.025)	(0.025)	(0.026)
Experience	0.012	0.012			0.042***	0.041***	0.042***
	(0.013)	(0.013)			(0.015)	(0.015)	(0.015)
Experience^2	-0.001***	-0.001***				-0.001**	-0.001**
	(0.000)	(0.000)				(0.000)	(0.000)
Age			0.168***	0.168***	0.160***	0.159***	0.164***
			(0.023)	(0.023)	(0.022)	(0.023)	(0.020)
Age^2			-0.002***	-0.002***	-0.002***	-0.002***	-0.002***
			(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Child care expenditure	0.073***	0.073***	0.075***	0.075***	0.072***	0.072***	0.085***
	(0.025)	(0.025)	(0.026)	(0.026)	(0.026)	(0.025)	(0.016)
If kids under 15							-0.288**
							(0.130)
Number of kids							-0.195***
							(0.047)
Husband's income							-0.001**
							(0.000)
Major expenses							0.014
							(0.012)
Number of observations (N)	1989	1989	1989	1989	1989	1989	1989
Number of controls	35	34	34	34	35	34	
Number of selected controls	21	21	22	22	23	22	

Appendix

A Neyman orthogonal estimation for vectors

Consider an extension of the partial linear framework of Belloni et al. (2014a):

$$\begin{aligned}
Y &= D'\theta_0 + g_0(X) + U & E(U|D, X) &= 0 \\
D_1 &= m_{0,1}(X) + V_1 & E(V_1|X) &= 0 \\
&\vdots & & \\
D_k &= m_{0,k}(X) + V_k & E(V_k|X) &= 0
\end{aligned}$$

where D and θ are $k \times 1$-vectors, rather then scalars. The functions $g_0(X)$ and $m_{0,j}(X)$ are unknown but we can use approximately sparse linear models to approximate these functions so that the approximation errors is small enough. We know that if θ is a scalar, we can apply the double selection lasso to achieve valid inference for θ. Here, we will show that the same arguments can be applied when when θ is a low-dimensional vector.

We follow Chernozhukov et al. (2018) and show that

1. the moment condition implied by the Robinson-style “partialling-out” is Neyman orthogonal.
2. the application of Theorem 3.1 of Chernozhukov et al. (2018) provides the asymptotic distribution (note that cross-fitting only permits to relax sparsity requirement, and the “partialling-out” approach still provides valid inference for θ even without cross-fitting).
3. the “partialling-out’ approach is equivalent to the double selection lasso.

We start with a general definition of Neyman orthogonal moments (see, e.g., Chernozhukov et al., 2018, Definition 2.1). Suppose we wish to estimate a low-dimensional parameter θ_0 using the moment conditions

$$\mathbb{E}\left[\psi(W; \theta_0, \eta_0)\right] = 0, \tag{8}$$

where W contains the random variables and η_0 is a nuisance parameter vector, which can be high-dimensional. Suppose we use machine learning techniques to estimate η_0. In essence, Neyman orthogonality means that small mistakes in the estimation of η will not disturb the consistent estimation of θ_0.

Formally, Neyman orthogonality depends on the concept of pathwise (Gateux) derivative. Let D_r denote the Gateux derivative of the moment condition ψ with respect to η in direction r. Then,

$$D_r[\eta - \eta_0] = \partial_r \left\{ \mathbb{E}\left[\psi(W; \theta_0, \eta_0 + (\eta - \eta_0)r)\right] \right\} \tag{9}$$

for all $r \in [0, 1)$. When D_r is evaluated at $r = 0$, we denote it as $D_0[\eta - \eta_0]$.

Definition 1. *The moment condition* $\mathbb{E}\left[\psi(W; \theta_0, \eta_0)\right] = 0$ *is Neyman orthogonal if*

$$D_0[\eta - \eta_0] = 0.$$

Now we show Neyman orthogonality of the "partialing-out" approach. For notational simplicity, we group the equations for $D_j \quad (j = 1, \ldots, k)$.

$$Y = D'\theta_0 + g_0(X) + U, \qquad E(U|D, X) = 0, \tag{10}$$

$$D = m_0(X) + V, \qquad E(V|X) = 0, \tag{11}$$

where $D = (D_1, D_2, \ldots, D_k)'$, $m_0(X) = (m_{0,1}(X), m_{0,2}(X), \ldots, m_{0,k}(X))'$, and $V = (V_1, V_2, \ldots, V_k)'$. Plugging Eq. 11 into 10, we obtain the reduced form for Y:

$$Y = l_0(X) + B, \qquad E(B|X) = 0, \tag{12}$$

where $l_0(X) = m_0(X)\theta_0 + g_0(X)$ and $B = U + V'\theta_0$. Now we can show that the following Robinson-style "partialling-out" moment condition is Neyman orthogonal.

Proposition 4. *The moment condition using the function*

$$\psi\left[W; \theta, \eta\right] = (Y - l(X) - (D - m(X))'\theta)(D - m(X)) \tag{13}$$

where $W = (Y, X, D)$ *and* $\eta = (l, m)$*, is Neyman orthogonal.*

Proof. First we show that $\mathbb{E}\left[\psi\left(W;\theta_0,\eta_0\right)\right]=0$:

$$\begin{aligned}\mathbb{E}\left[\psi\left(W;\theta_0,\eta_0\right)\right]&=\mathbb{E}\left[(Y-l_0(X)-(D-m_0(X))'\theta_0)(D-m_0(X))\right]\\&=\mathbb{E}\left[(B-V'\theta_0)V\right]\\&=\mathbb{E}\left[(U+V'\theta_0-V'\theta_0)V\right]\\&=\mathbb{E}\left[UV\right]\\&=0,\end{aligned}$$

where the last equality holds because $\mathbb{E}(U|D,X)=0$

Next, we prove that $D_0\left[\eta-\eta_0\right]=0$. First note that

$$\begin{aligned}&\mathbb{E}\left[\psi(W;\theta_0,\eta_0+(\eta-\eta_0)r)\right]\\&\quad=\mathbb{E}\left[(Y-l_0(X)-r(l(X)-l_0(X))-(D-m_0(X)-r(m(X)-m_0(X)))'\theta)(D-m_0(X)-r(m(X)-m_0(X))\right]\end{aligned}$$

Thus, we can compute $D_r[\eta-\eta_0]$ as follows:

$$\begin{aligned}D_r[\eta-\eta_0]=&-\mathbb{E}\left[(Y-l_0(X)-r(l(X)-l_0(X))-(D-m_0(X)-r(m(X)-m_0(X)))'\theta_0)(m(X)-m_0(X))\right]\\&-\mathbb{E}\left[(l(X)-l_0(X))-(m(X)-m_0(X))'\theta_0)(D-m_0(X)-r(m(X)-m_0(X))\right]\end{aligned}$$

Now, set $r=0$ and evaluate $D_0[\eta-\eta_0]$:

$$\begin{aligned}D_0[\eta-\eta_0]&=-\mathbb{E}\left[\left(Y-l_0(X)-(D-m_0(X))'\theta_0\right)\left(m(X)-m_0(X)\right)\right]\\&\quad-\mathbb{E}\left[\left(\left(l(X)-l_0(X)\right)-(m(X)-m_0(X))'\theta_0\right)\left(D-m_0(X)\right)\right]\\&=-\mathbb{E}\left[U(m(X)-m_0(X)\right]-\mathbb{E}\left[(l(X)-l_0(X))-(m(X)-m_0(X))'\theta_0)V\right]\\&=0,\end{aligned}$$

where the last equality holds because of the law of iterated expectation.

Therefore, the moment condition is Neyman orthogonal. □

Next, we provide the asymptotic distribution. Let $\widehat{\theta}$ be a solution to $\frac{1}{n}\sum\psi_i(W;\theta,\eta)=0$, given η. By Theorem 3.1 of Chernozhukov et al. (2018), $\widehat{\theta}$ is a consistent estimator of θ_0, and it is asymptotically normal with the rate of $\sqrt{N}$. Its variance is

$$\Sigma = J_0^{-1}\,\mathbb{E}(\psi(W;\theta_0,\eta_0)\psi(W;\theta_0,\eta_0)')(J_0^{-1})', \tag{14}$$

where $J_0 = \mathbb{E}[\partial\psi/\partial\theta|_{\theta=\theta_0}]$. We can simplify this expression to arrive at the following formula:

$$\Sigma = (\mathbb{E}(VV'))^{-1}\,\mathbb{E}(VV'U^2)(\mathbb{E}(VV'))^{-1}. \tag{15}$$

Eq. (15) leads to the following straightforward estimator of Σ:

$$\widehat{\Sigma} = \left(\frac{1}{n}\sum_i \widehat{V_i}\widehat{V_i}'\right)^{-1}\left(\frac{1}{n}\sum_i \widehat{V_i}\widehat{V_i}'\widehat{U_i}^2\right)\left(\frac{1}{n}\sum_i \widehat{V_i}\widehat{V_i}'\right)^{-1}, \tag{16}$$

where $\widehat{V_i}$ and $\widehat{U_i}$ are the residuals from Eqs. (10) and (11), respectively. Apparently, the variance estimator is the classic heteroskedasticity-consistent estimator.

The last step is to show that double-selection lasso is equivalent to the "partialling-out". We start with the equation

$$y = D'\theta_0 + g_0(X) + U.$$

If we partial out X from both Y and D, the equation becomes

$$y - l_0(X) = (D - m_0(X))'\theta_0 + U \tag{17}$$

It is easy to see that the solution for θ in Eq. 17 comes from the Neyman orthogonal moment condition using the moment function in Eq. 13.

Therefore, double-selection is equivalent to the Robinson-style "partialling-out" approach.

B Proofs of Propositions 2 and 3

Proof of Proposition 2.

Proof of Proposition 3.

C `dsheckman`: Stata command to estimate the double-selection Heckman model

The syntax of `dsheckman` is

`dsheckman` *depvar indepvars* [`if`] [`in`], `selection(`*depvar_s* `=` *indepvars_s*`)` [selvars(*varlist*)]

where

- *depvar* specifies the dependent variable in the main equation, which corresponds to y_1 in Eq. (1)

- *indepvars* specifies the independent variables in the main equation, which corresponds to $\mathbf{x_1}$ in Eq. (1).

- *depvar_s* specifies the dependent variable in the selection equation, which corresponds to y_2 in Eq. (2)

- *indepvars_s* specifies the independent variables in the selection equation, which corresponds to $\mathbf{x}$ and $\mathbf{z}$ in Eq. (2).

- `selvars()` specifies $\mathbf{x}$ in Eq. (2). If this option is not specified, $\mathbf{x}$ is constructed in two steps. First, run lasso probit of *devpar_s* on *indepvars_s* using the plugin penalty. Denote the selected variables as $\mathbf{x_2}$. Second, $\mathbf{x}$ is the union of $\mathbf{x_1}$ and $\mathbf{x_2}$.

References

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

Arellano, M. and S. Bonhomme (2017): "Quantile selection models with an application to understanding changes in wage inequality," *Econometrica*, 85, 1–28.

Bar, M., S. Kim, and O. Leukhina (2015): "Gender Wage Gap Accounting: The Role of Selection Bias," *Demography*, 52, 1729–1750.

Belloni, A., V. Chernozhukov, and C. Hansen (2014a): "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 28, 29–50.

——— (2014b): "Inference on treatment effects after selection among high-dimensional controls," *The Review of Economic Studies*, 81, 608–650.

Belloni, A., V. Chernozhukov, and Y. Wei (2016): "Post-Selection Inference for Generalized Linear Models With Many Controls," *Journal of Business & Economic Statistics*, 34, 606–619.

Blau, F. D. and L. M. Kahn (2007): "Changes in the Labor Supply Behavior of Married Women: 1980–2000," *Journal of Labor Economics*, 25, 393–438.

Candes, E. and T. Tao (2007): "The Dantzig Selector: Statistical Estimation When p Is Much Larger than n," *The Annals of Statistics*, 35, pp. 2313–2351.

Cavalcanti, T. V. d. V. and J. Tavares (2008): "Assessing the "Engines of Liberation": Home Appliances and Female Labor Force Participation," *The Review of Economics and Statistics*, 90, 81–88.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68.

Devereux, P. J. (2004): "Changes in Relative Wages and Family Labor Supply," *The Journal of Human Resources*, 39, 696–722.

Drukker, D. and D. Liu (2022): "Finite-sample results for lasso and stepwise Neyman-orthogonal Poisson estimators," *Econometric Reviews*, Forthcoming.

Gronau, R. (1974): "Wage Comparisons–A Selectivity Bias," *Journal of Political Economy*, 82, 1119–1143.

Heckman, J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679–694.

Heckman, J. J. (1979): "Sample selection bias as a specification error," *Econometrica*, 47, 153–161.

Heckman, J. J. and B. E. Honore (1990): "The Empirical Content of the Roy Model," *Econometrica*, 58, 1121–1149.

Huber, M. and G. Mellace (2014): "Testing exclusion restrictions and additive separability in sample selection models," *Empirical Economics*, 47, 75–92.

Lahiri, S. (2021): "Necessary and sufficient conditions for variable selection consistency of the LASSO in high dimensions," *Annals of Statistics*, 49(2), 820–844.

Leeb, H. and B. M. Pötscher (2008a): "Can one estimate the unconditional distribution of post-model-selection estimators?" *Econometric Theory*, 24, 338–376.

——— (2008b): "Sparse estimators and the oracle property, or the return of Hodges' estimator," *Journal of Econometrics*, 142, 201–211.

Liu, H., X. Xu, and J. J. Li (2020): "A Bootstrap lasso + partial ridge method to construct confidence intervals for parameters in high-dimensional sparse linear models," *Statistica Sinica*, 30, 1333–1355.

Mroz, T. A. (1987): "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55, 765–799.

Mulligan, C. B. and Y. Rubinstein (2008): "Selection, Investment, and Women's Relative Wages over Time," *The Quarterly Journal of Economics*, 123, 1061–1110.

Neumark, D. and S. Korenman (1994): "Sources of Bias in Women's Wage Equations: Results Using Sibling Data," *The Journal of Human Resources*, 29, 379–405.

Neyman, J. (1959): "Optimal asymptotic tests of composite statistical hypotheses," in *Probability and Statistics: The Harald Cramer Volume*, ed. by U. Grenander, 213–234.

Pötscher, B. M. and H. Leeb (2009): "On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding," *Journal of Multivariate Analysis*, 100, 2065–2082.

Roy, A. D. (1951): "Some thoughts on the distribution of earnings 1," *Oxford Economic Papers*, 3, 135–146.

Sampford, M. R. (1953): "Some Inequalities on Mill's Ratio and Related Functions," *The Annals of Mathematical Statistics*, 24, 130 – 132.

Tibshirani, R. (1996): "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

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

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

Wuthrich, K. and Y. Zhu (2021): "Omitted Variable Bias of Lasso-Based Inference Methods: A Finite Sample Analysis," *The Review of Economics and Statistics*, 1–47.