




# Regularization and variable selection in Heckman selection model

Emmanuel O. Ogundimu<sup>1</sup> 

Received: 4 September 2020 / Revised: 1 June 2021  
© The Author(s) 2021

## Abstract

Sample selection arises when the outcome of interest is partially observed in a study. A common challenge is the requirement for exclusion restrictions. That is, some of the covariates affecting missingness mechanism do not affect the outcome. The drive to establish this requirement often leads to the inclusion of irrelevant variables in the model. A suboptimal solution is the use of classical variable selection criteria such as AIC and BIC, and traditional variable selection procedures such as stepwise selection. These methods are unstable when there is limited expert knowledge about the variables to include in the model. To address this, we propose the use of adaptive Lasso for variable selection and parameter estimation in both the selection and outcome submodels simultaneously in the absence of exclusion restrictions. By using the maximum likelihood estimator of the sample selection model, we constructed a loss function similar to the least squares regression problem up to a constant, and minimized its penalized version using an efficient algorithm. We show that the estimator, with proper choice of regularization parameter, is consistent and possesses the oracle properties. The method is compared to Lasso and adaptively weighted  $L_1$  penalized Two-step method. We applied the methods to the well-known Ambulatory Expenditure Data.

**Keywords** Coordinate descent · Non-random selection · Penalized regression · Variable selection · Missing data

## 1 Introduction

Sample selection arises when the outcome of interest is non-randomly missing for a subset of the sample, resulting in a sample that is not representative of the population under study. This problem is ubiquitous in empirical economics, social sciences

---

✉ Emmanuel O. Ogundimu  
emmanuel.ogundimu@durham.ac.uk

<sup>1</sup> Department of Mathematical Sciences, Durham University, Durham, UK

and medical research. Consider, as an example, the ambulatory expenditures data (Cameron and Trivedi (2010)), where the amount of money spent on the medical services is expected to be linked with the decision to spend. The analysis of the data could proceed in two ways: analysis of only the positive expenditures without taking into account the zero expenditures or analysis of both the positive and zero expenditures. These methods will lead to biased estimates and efficiency loss. An optimal solution based on the use of sample selection model incorporates the decision to spend in order to model the amount of expenditures.

Heckman (1976) introduced a model for sample selection and several extensions in the parametric framework have been proposed (Marchenko and Genton (2012), Ogundimu and Hutton (2016), Lee (1983)). The estimation method is often based on the full information maximum likelihood (FIML) due to the maximization of the joint likelihood function of both the outcome and selection submodels simultaneously. An alternative specification of the model is the two-step procedure, where the problem is treated as a model misspecification problem due to omitted covariates (Heckman (1979)). A key drawback in the use of the Two-step estimator (and to a lesser extent FIML estimator) is its susceptibility to collinearity in the absence of an exclusion restriction (Leung and Yu (2000)). An exclusion restriction implies that there are variables in the selection submodel that are absent in the outcome equation. This is to avoid multicollinearity as a consequence of the linearity of the inverse Mills ratio over a wide range of its support. In the absence of an exclusion restriction, model identifiability relies on the non-linearity of the inverse Mills ratio. In general, both estimators have the same set of assumptions, and two-step estimator is indeed less sensitive to the assumptions in a very specific case of measurement error (see Stapleton and Young (1984) and Leung and Yu (2000)).

Sartori (2003) noted that when theory points to identical covariates in both components of the model, a common practice by applied researchers is a “mad” search for an exclusion restriction. This practice is dangerous because including extraneous variables may lead to specification error. Further, the impact of extraneous variables as exclusion restrictions was demonstrated in Ogundimu and Collins (2019). It was shown that the use of these variables in the selection submodel can constitute a nuisance in the model estimation instead of alleviating the problem of collinearity. In particular, if the extraneous variable(s) for exclusion restriction does not affect selection in the population but it is correlated with a true covariate and with selection in the sample, including it in the equation can bias the estimates of the effect of the covariates. Also, the use of an exclusion restriction that is not based on clear theoretical knowledge has been shown to produce results that vary both in effect size and precision (Genbäck et al. (2015)).

Genbäck et al. (2015) used set identification to compute bounds on regression parameters in sample selection model. Although the method does not impose any exclusion restriction and distributional assumption, correct coverage of the bounds relies on specifying an interval for the correlation parameter that contains the true value. We took a different approach from Genbäck et al. (2015) by keeping parametric assumption and develop penalized regression that can shrink variables that are not true exclusion restriction to zero while simultaneously selecting variables that are associated with the outcome and selection submodels.

Penalized regression with convex and non-convex penalties such as least absolute shrinkage and selection operator (Lasso - Tibshirani (1996)), elastic net (Zou and Hastie (2005)), adaptive Lasso (Zou (2006)), smoothly clipped absolute deviation (SCAD - Fan and Li (2001)) and minimax concave penalty (MCP - Zhang (2010)) have been widely applied for variable selection in generalized linear models and survival models. To the best of our knowledge, limited work has been done in penalized variable selection in sample selection settings. An example that is so close and yet so far to what we pursue here is, perhaps, the paper of Caner and Fan (2010), where adaptive Lasso was used in two-stage least squares regression for removing weak instrumental variables. Another important contribution is the use of the so-called outcome-adaptive Lasso for selecting covariates for inclusion in propensity score models to account for confounding bias (Shortreed and Ertefaie (2017)). While the propensity score models and the sample selection models are correction methods for selection bias, it has been opined that the latter should be preferred when the error terms in the substantive outcome submodel and the selection process are correlated (Antonakis et al. (2010)). This is the motivation for the new approach that we propose in the present article. Theoretically, the direct application of the penalty terms to the likelihood function of sample selection model is possible but it is computationally challenging due to the two components of the model and the possibility of different covariates in the components.

To circumvent this, we propose adaptive Lasso (ALasso) penalized sample selection model for the selection of variables and efficient estimation of parameters for both the outcome and selection submodels of the sample selection model. The proposed method is based on the use of second-order Taylor series expansion to approximate the likelihood function with respect to the maximum likelihood estimator (MLE). The resulting least squares approximation is then solved subject to adaptively weighted  $L_1$  penalty using the coordinate descent algorithm. The ultimate goal, from application perspective, is to identify and consistently estimate parameters in the components. We also extend the method to Lasso and adaptively weighted  $L_1$  penalized Two-step method. These methods are compared with the standard significance testing at  $\alpha = 0.05$  (*P-value*). That is, variables with a *P-value* higher than 5% is deemed non-significant and removed from the model.

The rest of the paper is organized as follows. Section 2 introduces the Heckman selection model and potential specification issues that are germane to the proposed approach. A penalized version of the model is presented in Sect. 3, while the computational routine and asymptotic properties of the estimator are evaluated in Sect. 4. In Sect. 5, the finite sample properties of the proposed model are studied and applied to the Ambulatory Expenditure data set. A discussion is provided in Sect. 6. Technical details are given in the “Appendix”.

## 2 Sample selection model

The model consists of two equations: an outcome equation and a selection equation

$$Y_i^* = \boldsymbol{\beta}^T \mathbf{x}_i + \sigma \varepsilon_{li}, \quad i = 1, \dots, n. \quad (1)$$

$$S_i^* = \boldsymbol{\gamma}^T \mathbf{w}_i + \varepsilon_{2i}, \quad i = 1, \dots, n, \quad (2)$$

respectively, where  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$  and  $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \dots, \gamma_q)$  are unknown parameters with corresponding covariates  $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})$  and  $\mathbf{w}_i = (1, w_{i1}, \dots, w_{iq})$ ,  $\sigma$  is the variance, and  $(\varepsilon_{1i}, \varepsilon_{2i})$  are random errors with means zero and correlation  $\rho$ . It is possible for  $\mathbf{x}_i$  and  $\mathbf{w}_i$  to overlap. We observe the indicator  $S_i = I(S_i^* > 0)$  such that the outcome  $Y_i = Y_i^* S_i$  is the observed part of the selected sample. Notice that the variance of  $S_i^*$  is set to 1 because we only observed its sign and it is not identifiable in (2). The conditional density of the observed data,

$$f(y|\mathbf{x}, \mathbf{w}, S^* > 0) = \frac{f(y, S^* > 0|\mathbf{x}, \mathbf{w})}{P(S^* > 0|\mathbf{w})} = \frac{f(y|\mathbf{x})P(S^* > 0|\mathbf{w})}{P(S^* > 0|\mathbf{w})}, \quad (3)$$

is the basis of the unification of sample selection problems as skew distributions given by Arellano-Valle et al. (2006).

Under the additional assumption of bivariate normal errors,

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim N_2 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\},$$

it is straightforward to show that the PDF in (3) is

$$f(y|\mathbf{x}, \mathbf{w}, S = 1; \boldsymbol{\theta}) = \frac{1}{\sigma} \phi\left(\frac{y - \boldsymbol{\beta}^T \mathbf{x}}{\sigma}\right) \Phi\left(\frac{\boldsymbol{\gamma}^T \mathbf{w} + \rho\left(\frac{y - \boldsymbol{\beta}^T \mathbf{x}}{\sigma}\right)}{\sqrt{1 - \rho^2}}\right) / \Phi(\boldsymbol{\gamma}^T \mathbf{w}), \quad (4)$$

where  $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma, \rho)$ . The parameter  $\rho \in (-1, 1)$  determines the correlation between  $Y_i^*$  and  $S_i^*$ , and hence the nature and severity of the selection process. The complete density of the sample selection model has a continuous component, with conditional density given by (4), and a discrete component. The discrete component is often modeled with probit regression as  $P(S = 1|\mathbf{w}) = \{\Phi(\boldsymbol{\gamma}^T \mathbf{w})\}^s \{1 - \Phi(\boldsymbol{\gamma}^T \mathbf{w})\}^{1-s}$ . The log-likelihood function for  $n$  observations is therefore

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^n S_i \left( \ln f(y_i|\mathbf{x}_i, \mathbf{w}_i, S_i = 1; \boldsymbol{\theta}) \right) + \sum_{i=1}^n S_i (\ln \Phi(\boldsymbol{\gamma}^T \mathbf{w}_i)) \\ &\quad + \sum_{i=1}^n (1 - S_i) \ln \Phi(-\boldsymbol{\gamma}^T \mathbf{w}_i). \end{aligned} \quad (5)$$

If the assumed model is correct and the Gaussian assumption holds, then the MLE based on equation (5) is  $\sqrt{n}$ -consistent for  $\boldsymbol{\theta}$  and asymptotically normal under general conditions ( Nicoletti and Peracchi (2001)). These properties are essential for the validity of the proposed method in Sect. 3.

The Two-step estimator is derived from the conditional expectation of the observed data and is given by

$$E(Y^*|\mathbf{x}, \mathbf{w}, S^* > 0) = \int_{-\infty}^{\infty} yf(y|\mathbf{x}, \mathbf{w}, S = 1; \boldsymbol{\theta}) dy = \boldsymbol{\beta}^T \mathbf{x} + \sigma\rho\Lambda(\boldsymbol{\gamma}^T \mathbf{w}), \tag{6}$$

where  $\Lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$  is the inverse Mills ratio,  $\phi(\cdot)$  and  $\Phi(\cdot)$  are standard normal density and cumulative distribution function respectively. To obtain the estimates of  $\rho$  and  $\sigma$  from (6), the average of the conditional variance,

$$\text{Var}(Y^*|\mathbf{x}, \mathbf{w}, S^* > 0) = \sigma^2[1 - \rho^2\Lambda(\boldsymbol{\gamma}^T \mathbf{w})\{\boldsymbol{\gamma}^T \mathbf{w} + \Lambda(\boldsymbol{\gamma}^T \mathbf{w})\}], \tag{7}$$

is equated to the observed residual variance in the second-stage least square regression.

The penalized regression in Sect. 3 also depends on the correct specification of the model in (5). A possible indication of model misspecification is the non-convergence of the model as a result of the violation of the requirements that  $|\rho\sigma| \leq \sigma$  and  $\sigma \geq 0$  (Copas and Li (1997)). That is, the estimates of  $\rho$  is close to  $\pm 1$ . Even if reparametrization of  $\rho$  allows for model convergence, equation (5) is not, in general, globally concave and the model can converge to a local maximum. Further, when the non-intercept variables in  $\mathbf{w}$  have coefficients equal to zero,  $\Lambda(\boldsymbol{\gamma}^T \mathbf{w})$  is a constant resulting in the failure of the regression model.

### 3 Penalized sample selection model

The general form of a penalized estimator for sample selection model is given by

$$\min_{\boldsymbol{\theta}} \left\{ -l(\boldsymbol{\theta}) + \sum_{j=1}^p p_{\lambda_1}(|\beta_j|) + \sum_{k=1}^q p_{\lambda_2}(|\gamma_k|) \right\}, \tag{8}$$

where  $p_{\lambda_1}(\cdot)$  and  $p_{\lambda_2}(\cdot)$  are penalty functions with tuning parameters  $\lambda_1$  and  $\lambda_2$ . In this case, the regression coefficients for the two equations have different penalties and  $p_{\lambda_1}(\beta_0) = p_{\lambda_2}(\gamma_0) = 0$  to ensure the intercepts of the equations are unpenalized. This article considers the case where  $p_{\lambda_1}(\cdot) = p_{\lambda_2}(\cdot) = p_{\lambda}(\cdot)$ . We can re-write (8) as

$$\min_{\boldsymbol{\theta}} \left\{ -l(\boldsymbol{\theta}) + \sum_{d=1}^s p_{\lambda}(|\theta_d|) \right\},$$

where  $(p+q) = s$  is the combined dimension of  $\boldsymbol{\beta}$  and  $\boldsymbol{\gamma}$ , and  $\sigma$  and  $\rho$  are unpenalized.

The Lasso penalized regression proposed by Tibshirani (1996) is based on

$$p_{\lambda}(|\theta_d|) = \lambda|\theta_d|, \quad d = 1, \dots, s. \tag{9}$$

The R.H.S. of (9) is the  $L_1$ -penalty, which shrinks small coefficients to zero to obtain sparse representation of the solution. Here,  $\lambda \geq 0$  is a tuning parameter controlling the

amount of shrinkage. Since the Lasso penalizes all the regression coefficients equally, it over-penalizes the important variables thereby resulting in biased estimators. The lack of the oracle property (Fan and Li (2001)) of Lasso prompted the development of the adaptive Lasso (Zou (2006)) with this property. The oracle property implies the method is consistent in variable selection, unbiased and asymptotically normal.

The adaptive Lasso estimator is a generalization of the Lasso penalty. Unlike Lasso, the adaptive Lasso penalty function penalizes the coefficients of different covariates at a different rate by using adaptive weights. This is accomplished by the introduction of individual weights,  $\tau_d$  in the penalty as

$$p_\lambda(|\theta_d|) = \lambda \tau_d |\theta_d|, \quad d = 1, \dots, s,$$

where  $\boldsymbol{\tau} = (\tau_1, \tau_2, \dots, \tau_s)$  are chosen in a data dependent way. The adaptive Lasso estimator for the sample selection model is the solution of

$$\min_{\boldsymbol{\theta}} \left\{ -l(\boldsymbol{\theta}) + \lambda \sum_{d=1}^s \tau_d |\theta_d| \right\}. \quad (10)$$

Often, the weights,  $\tau_d$ , is set to  $1/|\tilde{\theta}_d|^\delta$  for some appropriately chosen  $\delta > 0$ . For simplicity we set  $\delta = 1$  and  $\tilde{\boldsymbol{\theta}}$  as the MLE.

In addition to the proposed method, we adapted the adaptive Lasso approach to the Two-step estimator in (6). For this, a standard probit model is penalized as

$$\min_{\boldsymbol{y}} \left\{ -\sum_{i=1}^n S_i \ln \Phi(\boldsymbol{y}^T \mathbf{w}_i) - \sum_{i=1}^n (1 - S_i) \ln \Phi(-\boldsymbol{y}^T \mathbf{w}_i) + \sum_{k=1}^q p_{\lambda_2}(|\gamma_k|) \right\}.$$

Let  $\hat{\boldsymbol{y}} = (\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2)$ , where  $\hat{\boldsymbol{y}}_1$  corresponds to the  $h$  nonzero components ( $h \leq q$ ) of  $\hat{\boldsymbol{y}}$  and  $\hat{\boldsymbol{y}}_2$  are its zero elements. Next, the quantity,  $\Lambda(\hat{\boldsymbol{y}}_1^T \mathbf{w}^h)$  is formed and taken as an additional covariate in the second stage least squares regression with adaptive Lasso penalty. The coefficients of  $\Lambda(\hat{\boldsymbol{y}}_1^T \mathbf{w}^h)$  is left unpenalized in the second stage regression. The covariance matrix generated in the second stage regression is inconsistent, but will not be addressed in this paper.

An interesting question is whether the model in equation (10) (or its approximation that we proposed next) is without loss of generality. We justified the adequacy of the model by analysing directly equation (8) using a smooth approximation to the  $L_1$ -norm. Specifically, we used  $\lim_{\epsilon \rightarrow 0} \sqrt{\theta_d^T \theta + \epsilon} = \|\theta_d\|_1$ , where the LHS of the equation is the required approximation. We choose  $\epsilon = 10^{-6}$ . This allows for the derivation of the analytical score and Hessian. The optimization routine is based on the Newton-Raphson method with the tuning parameter selected using BIC. The minimum  $\lambda$  is taken over two-dimensional grid search. Initial empirical evaluation of the performance of this method with our proposal shows that the former does not offer significant improvement over the latter. In particular, the computational time grows as the number of variables increases. Consequently, we did not investigate this approach any further.

## 4 Computational algorithm and asymptotic results

### 4.1 The optimization routine

We approximate the penalized function in (10) by a second-order Taylor expansion with respect to  $\tilde{\theta}$ , the unpenalized MLE in (5). Thus,  $-l(\theta)$  can be approximated by the quadratic form  $2^{-1}(Y - \theta'X)^T(Y - \theta'X)$ , where  $Y = (X^T)^{-1}\{\nabla^2l(\theta)\theta - \nabla l(\theta)\}$  is the pseudo response vector,  $\nabla^2l(\theta) = X^T X$  such that  $X$  is derived as the Cholesky decomposition of  $\nabla^2l(\theta)$ . The notations  $\nabla l(\theta) = -\partial l(\theta)/\partial \theta$  is the gradient vector and  $\nabla^2l(\theta) = -\partial^2l(\theta)/\partial \theta \partial \theta^T$  is the Hessian matrix. Thus, the minimization problem in (10) becomes

$$\min_{\theta} \left\{ \frac{1}{2} (Y - \theta'X)^T (Y - \theta'X) + \lambda \sum_{d=1}^s \tau_d (|\theta_d|) \right\},$$

which is in the form of least squares problem subject to adaptive Lasso penalization. This can be solved using efficient minimization algorithms such as least angle regression (lars - Efron et al. (2004)) and the coordinate descent algorithm (Friedman et al. (2010)). We adopt the latter, and update the parameters in the model as

$$\hat{\theta}_d = \frac{S \left\{ \sum_{i=1}^s x_{id} \left[ y_i - \sum_{l \neq d}^s \hat{\theta}_l^T x_{il} \right], \lambda \tau_d \right\}}{\sum_{i=1}^s x_{id}^2}, \quad d = 1, \dots, s,$$

where  $S(z, \delta)$  is the soft threshold operator with

$$S(z, \delta) = \begin{cases} z - \delta & \text{if } z > 0 \text{ \& } \delta < |z| \\ z + \delta & \text{if } z < 0 \text{ \& } \delta < |z| \\ 0 & \text{if } \delta \geq |z| \end{cases}$$

Recall that four parameters in  $\theta$  ( $\sigma$ ,  $\rho$  and the intercepts of outcome and selection models) are unpenalized. Although we estimated all the parameters in the model, these four parameters are not set to zero in  $S(z, \delta)$ . The optimal tuning parameter,  $\lambda$  can be estimated by using AIC, BIC and GCV (generalized cross-validation) criteria. The BIC criterion is known to identify the true model with probability tending to one. We therefore use BIC to select optimal  $\lambda$ :

$$\text{BIC}(\lambda) = -2l(\hat{\theta}) + df_{\lambda} \log(n),$$

where  $0 \leq df_{\lambda} \leq s$  is the degree of freedom corresponding to the number of nonzero coefficients of  $\hat{\theta}$ . The optimal value of  $\lambda$  is computed over a grid of candidate values of  $\lambda$  between  $\lambda = 0$  and  $\lambda = \lambda_{\max}$ , with step size of 0.1, where  $\lambda_{\max}$  is the value of  $\lambda$  for which the entire vector of  $\hat{\theta}$  is zero. We allowed optimal  $\lambda = 0$  for the unregularized solution since degenerate cases are uncommon in FIML estimator.

## 4.2 Variance estimation

In low dimensional setting such as the case here, it is possible to first select covariates by using Lasso and adaptive Lasso, and thereafter obtain parameter estimates and their standard errors with maximum likelihood method. The disadvantage of this is the loss of the benefit of the estimation of parameters and selection of variables simultaneously, which is inherent in estimators with oracle properties.

Following from Fan and Li (2002) and Zhang and Lu (2007), let  $\hat{\theta}_1$  (with  $r$  elements,  $r \leq s$ ) be non-vanishing component of  $\hat{\theta}$  base on the optimal tuning parameter  $\lambda_1$ . Define  $A(\hat{\theta}) = \text{diag}\{1/\hat{\theta}_{11}, \dots, 1/\hat{\theta}_{1s}\}$  and  $B(\theta) = \text{diag}\{I(\hat{\theta}_{11} \neq 0)/\hat{\theta}_{11}^2, \dots, I(\hat{\theta}_{1s} \neq 0)/\hat{\theta}_{1s}^2\}$ . The estimator of the covariance matrix of the adaptive Lasso is given by the sandwich formula

$$\widehat{\text{cov}}(\hat{\theta}_1) = \left\{ \nabla^2 l(\hat{\theta}_1) + \lambda_1 A(\hat{\theta}_1) \right\}^{-1} \Sigma_{\lambda_1}(\hat{\theta}_1) \left\{ \nabla^2 l(\hat{\theta}_1) + \lambda_1 A(\hat{\theta}_1) \right\}^{-1}, \quad (11)$$

where  $\Sigma_{\lambda_1}(\hat{\theta}_1) = \{\nabla^2 l(\hat{\theta}_1) + \lambda_1 B(\hat{\theta}_1)\} \{\nabla^2 l(\hat{\theta}_1)\} \{\nabla^2 l(\hat{\theta}_1) + \lambda_1 B(\hat{\theta}_1)\}$ . The submatrix,  $\widehat{\text{cov}}(\hat{\theta}_1)$ , corresponding  $\hat{\theta}_1$  can then be obtained. An alternative formula based on the decomposition of the Hessian matrix,  $\nabla^2 l(\hat{\theta})$  can also be used. Let  $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ , where  $\hat{\theta}_1$  is as defined previously, and  $\hat{\theta}_2$  are the zero elements of  $\hat{\theta}$ . Then,

$$M = \nabla^2 l(\hat{\theta}) = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix},$$

where  $M_{11}$  corresponds to the first  $r \times r$  submatrix of  $M$ . Further, let  $A_{11}$  be the first  $r \times r$  submatrix of  $A(\hat{\theta})$ . Define  $E = M_{22} - M_{21} M_{11}^{-1} M_{12}$  and  $\tilde{M}_{11} = M_{11} + \lambda_1 A_{11}$ . Then,

$$\widehat{\text{cov}}(\hat{\theta}_1) = M_{11}^{-1} + \left( M_{11}^{-1} - \tilde{M}_{11}^{-1} \right) M_{12} E^{-1} M_{21} \left( M_{11}^{-1} - \tilde{M}_{11}^{-1} \right).$$

We estimated the variance for the penalized model using (11) and compared its performance with the sample standard deviation in the simulation study.

## 4.3 Asymptotic properties

We study the asymptotic properties of the adaptive Lasso estimator  $\hat{\theta}$  obtained by maximizing the penalized likelihood function based on  $n$  samples:

$$Q_n(\theta) = l(\theta) - n\lambda_n \sum_{d=1}^s \tau_d(|\theta_d|), \quad (12)$$

with respect to  $\theta$ . Denote the true value of  $\theta$  by  $\theta_0 = (\theta_{10}^T, \theta_{20}^T)^T$ , where  $\theta_{10} = ((\beta_{10}^T, \gamma_{10}^T)^T)$  is an  $r \times 1$  vector whose components are nonzero and  $\theta_{20}$  is the  $(s - r)$  remaining zero components. The maximizer of equation (12) can be written as  $\hat{\theta} =$



$(\hat{\theta}_1^T, \hat{\theta}_2^T)^T$ . Also, let  $I(\theta_0)$  be the information matrix based on the likelihood function in equation (5) and let  $I_1(\theta_{10}) = I_{11}(\theta_{10}, \mathbf{0})$ , where  $I_{11}(\theta_{10}, \mathbf{0})$  is the leading  $r \times r$  submatrix of  $I_1(\theta_0)$  with  $\theta_{20} = \mathbf{0}$ .

The following regularity conditions are necessary to establish the asymptotic properties of the adaptive Lasso estimator for the model.

- C1. Identifiability:** Suppose that  $y_i, i = 1, \dots, n$  are *i.i.d* with a mixed probability mass and density function  $f(y_i; \theta_0)$ . The parameters in the model are identifiable if  $\theta \neq \theta_0$  implies  $f(y_i; \theta) \neq f(y_i; \theta_0)$  with probability 1
- C2.** The parameter space  $\Theta$  is compact and  $\theta_0 \in \Theta$
- C3.**  $E[\sup_{\theta \in \Theta} |\ln f(y_i; \theta)|] < \infty$
- C4.**  $E[\nabla l(\theta_0)] = 0$  and  $E[\nabla^2 l(\theta_0)]$  is finite and positive definite.

Since the information matrix is nonsingular at the true parameter vector, the local identifiability of the model parameters can be inferred (see Appendix A.1 in the supplementary material for the derivation of the expected information matrix). Inherent in the identification condition is the assumption that there is no multicollinearity among the variables in  $\mathbf{x}_i$  and  $\mathbf{w}_i$  and that there is no perfect correlation between  $\mathbf{x}_i$  and  $\Lambda$ , the inverse Mills ratio. In particular, conditions 1 - 3 ensure that  $\hat{\theta}$  is consistent while conditions 1 - 4 is required for its asymptotic normality.

The main asymptotic results are obtained in a similar version to Fan and Li (2001) by using the regularity conditions:

**Theorem 1 (Consistency of  $\hat{\theta}$ ).** *If  $\sqrt{n}\lambda_n = O_p(1)$ , then there exists a local maximizer  $\hat{\theta}$  of  $Q_n(\theta)$  in equation (12) such that  $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2})$ , where  $\|\cdot\|$  denotes the Euclidean norm.*

**Theorem 2 (Oracle properties).** *If  $\sqrt{n}\lambda_n \rightarrow 0$  and  $n\lambda_n \rightarrow \infty$ , as  $n \rightarrow \infty$  then  $\hat{\theta}$  has the following properties:*

- (i) *Sparsity:  $\hat{\theta}_2 = \mathbf{0}$ ;*
- (ii) *Asymptotic normality:  $\sqrt{n}(\hat{\theta}_1 - \theta_{10}) \xrightarrow{d} N(\mathbf{0}, I_1^{-1}(\theta_{10}))$ .*

Note that the consistency and sparsity of  $\hat{\theta}$  in Theorems 1 and 2(i) imply that the adaptive lasso estimator is consistent in variable selection. Theorem 2(ii), on the other hand, implies that the adaptive lasso estimator for the nonzero coefficients is efficient as if the irrelevant covariates are known. The proof of both theorems are given in Appendix A.2.

## 5 Numerical studies

In this section we use simulation and a real data to compare the finite sample performance of the adaptive Lasso, Lasso and the Two-step estimators for sample selection models.

### 5.1 Simulation study

The data is generated under two scenarios - strong and weak signals.

### Scenario 1: Large effect

The data for the outcome submodel was generated from  $Y_i^* = \boldsymbol{\beta}^T \mathbf{x}_i + \sigma \varepsilon_{1i}$ , where  $\boldsymbol{\beta}^T = (0.5, 1, 1, 1.5, 0, 0, 0, 0)$ ,  $\beta_0 = 0.5$ . For the setting with exclusion restriction, we generated data for the selection process as  $S_i^* = \boldsymbol{\gamma}^T \mathbf{w}_i + \varepsilon_{2i}$ , where  $\boldsymbol{\gamma}^T = (1.5, 1, 1, 1, 0, 0, 0, 1)$ ,  $\gamma_0 = 1.5$ . We exclude the last element of  $\boldsymbol{\gamma}$  for the case with no exclusion restriction (that is,  $\mathbf{x} = \mathbf{w}$ ). The covariates  $\mathbf{x}_i$  and  $\mathbf{w}_i$  are independent of the error terms  $\varepsilon_i^T = (\varepsilon_{1i}, \varepsilon_{2i})$ . The error terms are generated from a bivariate normal distribution with  $\Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}$ , where  $\sigma = 1$  and the correlation  $\rho = \{0, 0.3, 0.5, 0.7\}$ . The covariates  $x_1, \dots, x_8$  are generated such that their distribution are marginally standard normal with pairwise correlations  $\text{corr}(x_j, x_k) = \rho^{|j-k|}$ . We take  $\rho = 0.5$  to allow for moderate correlation between the covariates. We only observe the values of  $Y_i^*$  when  $S_i^* > 0$ . This leads to approximately 30% unobserved cases in the realized sample. Two sample sizes,  $n = 500$  and 1000 are evaluated using 1000 replications. We also evaluated the impact of error distribution misspecification, where the error terms are generated from a bivariate Student's  $t$  distribution with degree of freedom  $\nu = 5$ . For this, we consider the setting under the absence of the exclusion restriction variable and sample size of 1000.

### Scenario 2: Small effect

For this scenario, we consider only the case with the absence of exclusion restriction and sample size of 1000. We generated the true parameters as follows:  $\boldsymbol{\beta}^T = (0.5, 0.2, 0.2, 0.2, 0, 0, 0, 0)$ ,  $\beta_0 = 0.5$  and  $\boldsymbol{\gamma}^T = (0.58, 0.2, 0.2, 0.2, 0, 0, 0, 0)$ ,  $\gamma_0 = 0.58$ . This ensures roughly 30% of the data is missing. The remaining parameters are simulated as in scenario 1.

### Scenario 3: Shrinkage of false exclusion restriction variable

We investigate the performance of adaptive lasso in detecting non valid exclusion restriction. The data was generated as in scenario 1 with a true exclusion restriction. We consider the case where two variables ( $X_9, X_{10}$ ) are used as exclusion restriction whereas these variables are not associated with the outcome model and not predictive of missingness (*no\_cor\_outcome*). The second data is generated such that the two variables are associated with the outcome but not predictive of missingness (*cor\_outcome*).

The accuracy of the estimators are evaluated using mean squared errors,  $(\hat{\boldsymbol{\theta}}_d - \boldsymbol{\theta}_d)^T V (\hat{\boldsymbol{\theta}}_d - \boldsymbol{\theta}_d)$ , where  $V$  is the population covariance matrix and the summary of the median over 1000 replications is obtained. We also assess the performance of the methods using sensitivity (mean of proportion of nonzero coefficients that were correctly identified) and specificity (mean of proportion of zero coefficients that were correctly identified) as well as the comparison of the model based and the empirical standard errors.

For computational convenience, all the model parameters are estimated with the support  $(-\infty, \infty)$ . We used estimation metric  $\text{atanh} \rho = \ln\{(1 + \rho)/(1 - \rho)\}/2$  for  $\rho$ , where  $\text{atanh} \rho$  is the inverse hyperbolic tangent of  $\rho$ , and logarithm of  $\sigma$ . *P-value* is based on significance testing at 5% level of significance. The "oracle" estimate is based on the model fitted using the variables that have non-zero coefficients.

Tables 1 and 2 summarize the sensitivity, specificity and the mean square error (MSE) for the models with sample sizes of 500 and 1000 in the presence of the

**Table 1** Simulation results for exclusion restriction- sample size =500

Method	Sensitivity Selection Equation	Specificity Selection Equation	Sensitivity Outcome Equation	Specificity Outcome Equation	Selection MSE	Outcome	Combined
$\rho = 0$							
Oracle	1	1	1	1	0.063	0.011	0.081
Lasso	1	0.673	1	0.372	0.195	0.024	0.222
ALasso	1	0.960	1	0.950	0.092	0.014	0.110
Two-step	1	0.951	1	0.985	0.084	0.013	0.101
P-value	1	0.948	1	0.948	0.145	0.031	0.184
$\rho = 0.3$							
Oracle	1	1	1	1	0.065	0.012	0.081
Lasso	1	0.666	1	0.368	0.188	0.024	0.219
ALasso	1	0.960	1	0.959	0.088	0.014	0.110
Two-step	1	0.951	1	0.986	0.084	0.013	0.103
P-value	1	0.945	1	0.944	0.145	0.031	0.183
$\rho = 0.5$							
Oracle	1	1	1	1	0.067	0.012	0.080
Lasso	1	0.661	1	0.373	0.183	0.022	0.214
ALasso	1	0.966	1	0.962	0.085	0.013	0.104
Two-step	1	0.954	1	0.986	0.083	0.013	0.102
P-value	1	0.946	1	0.948	0.138	0.030	0.176
$\rho = 0.7$							
Oracle	1	1	1	1	0.059	0.010	0.073
Lasso	1	0.652	1	0.385	0.179	0.021	0.205
ALasso	1	0.949	1	0.956	0.082	0.012	0.100
Two-step	1	0.954	1	0.989	0.083	0.013	0.103
P-value	1	0.945	1	0.943	0.133	0.030	0.168

**Table 2** Simulation results for exclusion restriction- sample size = 1000

Method	Sensitivity Selection Equation	Specificity	Sensitivity Outcome Equation	Specificity	Selection MSE	Outcome	Combined
$\rho = 0$							
Oracle	1	1	1	1	0.029	0.006	0.038
Lasso	1	0.670	1	0.405	0.116	0.012	0.129
ALasso	1	0.982	1	0.975	0.040	0.007	0.049
Two-step	1	0.977	1	0.991	0.037	0.006	0.046
P-value	1	0.952	1	0.946	0.064	0.016	0.083
$\rho = 0.3$							
Oracle	1	1	1	1	0.030	0.005	0.038
Lasso	1	0.708	1	0.421	0.116	0.011	0.128
ALasso	1	0.983	1	0.977	0.040	0.006	0.048
Two-step	1	0.977	1	0.993	0.037	0.006	0.045
P-value	1	0.952	1	0.947	0.065	0.015	0.082
$\rho = 0.5$							
Oracle	1	1	1	1	0.029	0.005	0.037
Lasso	1	0.700	1	0.424	0.113	0.010	0.123
ALasso	1	0.981	1	0.981	0.038	0.006	0.047
Two-step	1	0.976	1	0.994	0.039	0.006	0.047
P-value	1	0.949	1	0.951	0.064	0.015	0.081
$\rho = 0.7$							
Oracle	1	1	1	1	0.028	0.005	0.034
Lasso	1	0.686	1	0.425	0.103	0.009	0.115
ALasso	1	0.982	1	0.980	0.035	0.006	0.043
Two-step	1	0.981	1	0.994	0.039	0.006	0.047
P-value	1	0.952	1	0.949	0.060	0.014	0.076

**Table 3** Simulation results for absence of exclusion restriction- sample size =500

Method	Sensitivity Selection Equation	Specificity Selection Equation	Sensitivity Outcome Equation	Specificity Outcome Equation	Selection MSE	Outcome	Combined
$\rho = 0$							
Oracle	1	1	1	1	0.046	0.012	0.064
Lasso	1	0.759	1	0.464	0.144	0.026	0.175
ALasso	1	0.970	1	0.968	0.059	0.015	0.081
Two-step	1	0.968	1	0.985	0.058	0.014	0.076
P-value	1	0.947	1	0.942	0.113	0.031	0.152
$\rho = 0.3$							
Oracle	1	1	1	1	0.046	0.012	0.064
Lasso	1	0.782	1	0.485	0.149	0.024	0.176
ALasso	1	0.973	1	0.972	0.062	0.014	0.079
Two-step	1	0.967	1	0.989	0.061	0.013	0.080
P-value	1	0.946	1	0.949	0.114	0.030	0.151
$\rho = 0.5$							
Oracle	1	1	1	1	0.044	0.011	0.060
Lasso	1	0.775	1	0.490	0.140	0.021	0.166
ALasso	1	0.974	1	0.969	0.060	0.013	0.078
Two-step	1	0.968	1	0.987	0.060	0.014	0.080
P-value	1	0.951	1	0.947	0.111	0.029	0.147
$\rho = 0.7$							
Oracle	1	1	1	1	0.042	0.011	0.057
Lasso	1	0.758	1	0.492	0.131	0.020	0.156
ALasso	1	0.968	1	0.970	0.054	0.012	0.070
Two-step	1	0.970	1	0.989	0.058	0.013	0.075
P-value	1	0.948	1	0.947	0.103	0.028	0.136

Table 4 Simulation results for absence of exclusion restriction- sample size =1000

Method	Sensitivity Selection Equation	Specificity Selection Equation	Sensitivity Outcome Equation	Specificity Outcome Equation	Specificity	Selection MSE	Outcome	Combined
$\rho = 0$								
Oracle	1	1	1	1	1	0.022	0.006	0.031
Lasso	1	0.765	1	1	0.490	0.079	0.013	0.093
ALasso	1	0.985	1	1	0.988	0.027	0.007	0.036
Two-step	1	0.982	1	1	0.994	0.026	0.007	0.035
P-value	1	0.943	1	1	0.950	0.053	0.015	0.071
$\rho = 0.3$								
Oracle	1	1	1	1	1	0.024	0.006	0.031
Lasso	1	0.764	1	1	0.481	0.074	0.011	0.090
ALasso	1	0.985	1	1	0.986	0.027	0.006	0.035
Two-step	1	0.981	1	1	0.991	0.025	0.006	0.034
P-value	1	0.946	1	1	0.948	0.054	0.015	0.072
$\rho = 0.5$								
Oracle	1	1	1	1	1	0.021	0.006	0.029
Lasso	1	0.754	1	1	0.477	0.074	0.011	0.085
ALasso	1	0.986	1	1	0.985	0.026	0.006	0.033
Two-step	1	0.981	1	1	0.993	0.025	0.006	0.034
P-value	1	0.946	1	1	0.947	0.052	0.014	0.068
$\rho = 0.7$								
Oracle	1	1	1	1	1	0.021	0.005	0.028
Lasso	1	0.733	1	1	0.470	0.069	0.010	0.080
ALasso	1	0.986	1	1	0.987	0.024	0.006	0.031
Two-step	1	0.980	1	1	0.994	0.025	0.006	0.033
P-value	1	0.943	1	1	0.948	0.049	0.014	0.064

exclusion restriction variable. The three estimators correctly identified all the nonzero coefficients (sensitivity = 1) in both submodels. The advantage of ALasso and Two-step methods are more pronounced for specificity, in which case at least 95% of zero coefficients in both submodels are correctly identified (specificity ranges between 95% and 98%). The larger the sample size, the better the performance of the methods in the consistency of variable selection across the two submodels. Lasso has higher specificity under the selection submodel than the outcome submodel, and its performance is generally better with larger sample size. There is no clear distinction in the effect of correlation ( $\rho$ ) on sensitivity and specificity. Overall, ALasso and Two-step estimators perform better than the Lasso in terms of MSE. The impact of  $\rho$  slightly manifested here: the accuracy of the estimated nonzero coefficients increases as  $\rho$  increases for both Lasso and ALasso. In particular, MSE improves as sample size increases. The results also show that ALasso is consistently superior to significance testing at  $\alpha = 5\%$  level (*P-value*).

Tables 3 and 4 summarize the simulation results for the models with sample sizes of 500 and 1000 in the absence of the exclusion restriction variable. The patterns are similar to the corresponding results in Tables 1 and 2. Interestingly, the performance of the measures without the exclusion restriction variable is consistently better than the corresponding results under the exclusion restriction. This result has practical implications - if the model relies on identification through the inverse Mills ratio, then ALasso and Two-step estimators would identify the true variables in the model without exclusion restrictions with probability tending to 1.

The results of fitting the models to a data generated from a Student's  $t$  error distribution in the absence of exclusion restrictions can be found in Appendix A.3 of the supplementary material. The MSE of the parameter estimates in the selection submodel are lower than the corresponding results under the data generated from the normal error model (Table 4). This, however, does not affect the specificity of the estimators. In particular, the specificity of the ALasso and the Two-step methods are higher than the corresponding results in Table 4. Further, the specificity under the outcome model for ALasso is slightly lower than the corresponding results in Table 4. Overall, ALasso is more robust to misspecification of the error distribution than the Two-step estimator.

The impact of covariates with weak effect on the proposed methods can be found in Appendix A.3. As expected, Lasso performs better than the other methods in terms of sensitivity in both the selection and outcome equations. This result is not surprising as Lasso is generally known to include true covariates, but also some irrelevant covariates (Meinshausen and Bühlmann (2006)). A striking result is the poor performance of the two-step estimator in terms of sensitivity but superior performance with regards to specificity in the outcome equation. A possible reason for this is the collinearity between the inverse Mills ratio and covariates with weak effects in the second stage regression as these covariates can be easily pushed to zero.

Table 5 shows the results of the number of times the variables that are used for exclusion restrictions are selected. The performance of ALasso estimator is better when the false exclusion restriction variables ( $X_9$ ,  $X_{10}$ ) are not associated with the outcome than when the variables are associated with it. Overall, the true exclusion restriction variable ( $X_8$ ) is not shrunk to zero.

**Table 5** Simulation results for the number of times the true and false exclusion restriction variables are selected using adaptive lasso (out of 1000)

Correlation type	$X_8$	$X_9$	$X_{10}$
cor_outcome	1000	26	25
no_cor_outcome	1000	15	15

## 5.2 Ambulatory expenditure data

The data on ambulatory expenditure contains 3,328 observations of which 526 (15.8%) of the outcome of interest (expenditure) is missing. Apart from expenditure, which is highly skewed, other explanatory variables such as age, gender, education status (educ), ethnicity (blhisp), number of chronic conditions (totchr), insurance status (ins) and income are available in the data. We use log expenditure (lambexp) as the outcome variable due to skewness in line with previous applications of the data (Marchenko and Genton 2012; Ogundimu and Collins 2019). Since the decision to spend is likely to be related to the spending amount, the statistical analysis method that was used by previous authors is sample selection model. The outcome equation, which is often the model of interest, contains  $x = (1, \text{age}, \text{female}, \text{educ}, \text{blhisp}, \text{totchr}, \text{ins})$  while the selection equation,  $w = (x, \text{income})$ . Income is included for the exclusion restriction criteria although its use for this purpose is debatable (see Cameron and Trivedi 2010; Marchenko and Genton 2012).

Table 6 shows the results of the application of the proposed methods to the data. We obtain the same results for the selection normal model and the Lasso estimator. This is in consonance with the simulation result where Lasso tends to retain more irrelevant variables in the model than adaptive Lasso. Both adaptive Lasso and Two-step estimators set two variables from the outcome model (education and insurance status) to zero. The two variables are also reported as non-significant at 5% significance level under the classical selection normal model. As noted earlier, the regularization methods eliminate the need to retain or remove covariates based on arbitrary thresholds.

An important hypothesis of no sample selection is  $H_0 : \rho = 0$  (equivalently  $H_0 : \text{atanh}\rho = 0$  and  $\sigma\rho$  for the adaptive Lasso and Two-step methods respectively). The reported Wald test for the hypothesis under the classical selection model gave a *P-value* of 0.380. This hypothesis will not be rejected at 5% significance level. The implication of this is that the amount of money spent is unrelated with the decision to spend, and the outcome can be analyzed separately. As noted by Marchenko and Genton (2012), and previous authors who analyzed the data, this conclusion is not plausible.

The inference for sample selection bias under adaptive Lasso suggests the existence of selection bias at 5% significance level ( $p = 0.039$ ). Although the proposed adaptive Lasso estimator was developed under the selection normal model, the result is in agreement with the analysis of the same data by using the selection-t model proposed in Marchenko and Genton (2012). In particular, the coefficient estimate of  $\text{atanh}\rho$  under the selection normal model is -0.131 whereas adaptive Lasso estimate is -0.323. This estimate is close to the estimate of  $\text{atanh}\rho$  ( $\text{atanh}\rho = -0.322$ ) that was reported in Marchenko and Genton (2012). Similarly, the corresponding parameter estimate



**Table 6** Penalized likelihood for Ambulatory Expenditure Data

	Selection Normal		Lasso		Adaptive Lasso		Two-step	
	Estimate	S.E.	Estimate	S.E	Estimate	S.E	Estimate	S.E
Selection Equation								
(Intercept)	-0.671	0.194	-0.671	0.194	-0.564	0.191	-0.669	0.194
age	0.088	0.027	0.088	0.027	0.077	0.027	0.087	0.028
female	0.663	0.061	0.663	0.061	0.630	0.060	0.664	0.061
educ	0.062	0.012	0.062	0.012	0.062	0.012	0.062	0.012
blhisp	-0.364	0.062	-0.364	0.062	-0.350	0.061	-0.366	0.062
totchr	0.797	0.071	0.797	0.071	0.776	0.071	0.796	0.071
ins	0.170	0.063	0.170	0.063	0.114	0.062	0.169	0.063
income	0.003	0.001	0.003	0.001	0.001	0.001	0.003	0.001
Outcome Equation								
(Intercept)	5.044	0.228	5.044	0.228	5.438	0.135	5.561	0.121
age	0.212	0.023	0.212	0.023	0.199	0.023	0.192	0.019
female	0.348	0.060	0.348	0.060	0.284	0.060	0.248	0.051
educ	0.019	0.011	0.019	0.011	0	-	0	-
blhisp	-0.219	0.060	-0.219	0.060	-0.165	0.061	-0.148	0.050
totchr	0.540	0.039	0.540	0.039	0.507	0.039	0.475	0.033
ins	-0.030	0.051	-0.030	0.051	0	-	0	-
ln sigma	0.240	0.015	0.240	0.015	0.247	0.019		
atanhρ	-0.131	0.150	-0.131	0.150	-0.323*	0.156		
IMR							-0.665**	0.181

IMR -  $\sigma\rho$  parameter for inverse Mills ratio.

\* $P$ -value = 0.039; \*\* $P$ -value = 0.000

for the inverse Mills ratio ( $\sigma\rho$ ) under the Two-step method resulted in  $p = 0.000$ . In addition, a naive analysis based on missing not at random (MAR) assumption ( $\rho = 0$ ) for non-penalized and penalized models did not remove any variable in the outcome equation.

## 6 Concluding remarks

This paper proposed a method for variable selection and estimation of covariate effects in sample selection models. Data sets that provide information about sample selection are becoming increasingly common in many fields. If these data sets are analyzed and variable selection is carried out using conventional methods, then the conclusions can be misleading. Although many statistical procedures have been developed in the literature for the analysis of selected samples, there is no existing research on regularized variable selection methods for this model. This is, perhaps, due to the association between covariates and the outcome submodel as well as the association between covariates and the selection submodel.

Based on the results of our simulation and data analysis, we recommend the use of the ALasso estimator especially in the presence of weak covariate effects. Application of the ALasso estimator to the Ambulatory expenditure data further corroborate the usefulness of the method. A key advantage of the proposed estimator is that there is no need to specify arbitrary thresholds in order to determine the importance of covariates in the model, which is not the case with the use of *P-value*.

The applicability of our method depends on correct specification of the sample selection model. If the profile likelihood of  $\rho$  is very flat, then there is limited information about selectivity in the data. It is possible to extend our proposal to high dimensional data settings by avoiding the non-differentiability of Lasso and ALasso penalty functions using approximation of different norms. Also, it is possible to accomplish model and variable selection simultaneously by using various parametric extensions of sample selection models such as the selection-t model (Marchenko and Genton 2012) and the selection skew-normal model (Ogundimu and Hutton 2016), robust extensions (Zhelonkin et al. 2016) or by using alternative estimation techniques such as the EM-algorithm (Zhao et al. 2020). In particular, the oracle property of the penalized estimators is a pointwise asymptotic feature and does not necessarily hold for all the points in the parameter space (Leeb and Pötscher (2005)). Consequently, the problem of post selection inference for non-random sample deserves further investigation.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s00362-021-01246-z>.

**Acknowledgements** The author thanks the Associate Editor and the referees for their constructive comments, which led to improvements in the manuscript.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Antonakis J, Bendahan S, Jacquart P, Lalive R (2010) On making causal claims: a review and recommendations. *Leadersh Q* 21(6):1086–1120
- Arellano-Valle RB, Branco MD, Genton MG (2006) A unified view of skewed distributions arising from selections. *Can J Stat* 34:581–601
- Cameron AC, Trivedi PK (2010) *Microeconometrics using Stata*, revised edn. Stata Press, College Station, TX
- Caner M, Fan Q (2010) The adaptive lasso method for instrumental variable selection. North Carolina State University, Tech. rep
- Copas JB, Li H (1997) Inference for non-random samples. *J R Stat Soc B* 59:55–95
- Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. *Ann Stat* 32(2):407–499
- Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 96(456):1348–1360

- Fan J, Li R (2002) Variable selection for Cox's proportional hazards model and Frailty model. *Ann Stat* 30(1):74–99
- Friedman J, Hastie T, Tibshirani RJ (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 33(1):1–22
- Genbäck M, Stanghellini E, de Luna X (2015) Uncertainty intervals for regression parameters with non-ignorable missingness in the outcome. *Stat Pap* 56(3):829–847
- Heckman J (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Ann Econ Soc Meas* 5:475–492
- Heckman J (1979) Sample selection bias as a specification error. *Econometrica* 47:153–161
- Lee L (1983) Generalized econometric models with selectivity. *Econometrica* 51(2):507–512
- Leeb H, Pötscher BM (2005) Model selection and inference: facts and fiction. *Economet Theory* 21(1):21–59
- Leung SF, Yu S (2000) Collinearity and two-step estimation of sample selection models: problems, origins and remedies. *Comput Econ* 15:173–199
- Marchenko YV, Genton MG (2012) A Heckman Selection-t Model. *J Am Stat Assoc* 107:304–317
- Meinshausen N, Bühlmann P (2006) High dimensional graphs and variable selection with the lasso. *Ann Stat* 34(3):1436–1462
- Nicoletti C, Peracchi F (2001) Two-step estimation of binary response models with sample selection. Technical report, Faculty of Economics, Tor Vergata University
- Ogundimu EO, Collins GS (2019) A robust imputation method for missing responses and covariates in sample selection models. *Stat Methods Med Res* 28:102–116
- Ogundimu EO, Hutton JL (2016) A sample selection model with skew-normal distribution. *Scand J Stat* 43:172–190
- Sartori AE (2003) An estimator for some binary-outcome selection models without exclusion restrictions. *Polit Anal* 11(2):111–138
- Shortreed SM, Ertefaie A (2017) Outcome-adaptive lasso: variable selection for causal inference. *Biometrics* 73:1111–1122
- Stapleton DC, Young DJ (1984) Censored normal regression with measurement error on the dependent variable. *Econometrica* 52:737–760
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B* 58(1):267–288
- Zhang CH (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann Stat* 38(2):894–942
- Zhang HH, Lu W (2007) Adaptive Lasso for Cox's proportional hazards model. *Biometrika* 94(3):691–703
- Zhao J, Kim HJ, Kim HM (2020) New EM-type algorithms for the Heckman selection model. *Comput Stat Data Anal* 146:106930
- Zhelonkin M, Genton M, Ronchetti E (2016) Robust inference in sample selection models. *J R Stat Soc B* 78:805–827
- Zou H (2006) The adaptive lasso and its oracle properties. *J Am Stat Assoc* 101(476):1418–1429
- Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Stat Soc Ser B* 67(2):301–320