Addressing Sample Selection Bias for Machine Learning Methods: Supplemental Appendix

Dylan Brewer* and Alyssa Carlson†

June 29, 2023

Appendix A considers an alternative prediction goal than what was considered in "Addressing Sample Selection Bias for Machine Learning Methods." In the main paper, the goal is to predict for a sample unaffected by the sample selection mechanism. Here we consider the case in which predicting the unobserved sample ($S_i = 0$) is the goal.

Appendix B investigates the sensitivity of the CF approaches to their underlying selection assumption. We provide simulations for when the instrument is not excluded and the errors are not normally distributed.

Appendix C contains supplementary tables for the application.

A Predicting the unobserved sample

Beginning with the same sample-selection framework from the main paper:

$$Y_i = f(X_i) + U_i \tag{1}$$

$$S_i = 1\{g(X_i, Z_i, \delta) + V_i > 0\} \tag{2}$$

where Y_i is only observed in the training sample when $S_i = 1$. When considering the setting of a selected sample for training, there are three possible applications of the trained learners.

*School of Economics, Georgia Institute of Technology. Email: brewer@gatech.edu

†Department of Economics, University of Missouri. Email: carlsonah@missouri.edu

The first case is a prediction sample conditional on $S_i = 1$, in this cases, sample selection does not effect the quality of prediction as the training sample and prediction sample are selected the same way. The second case, considered in the main paper, is when the prediction sample is unbiased and unaffected by sample selection. The last case is when the prediction sample is the unobserved sample in which $S_i = 0$. In this case, the training sample is biased due to selection into the training sample, while the prediction sample is also biased due to selection out of the training sample.

So how does the analysis of the paper, which focuses on prediction for an unbiased sample, change when we are instead interested in predicting the unobserved sample? We find that there are some nuanced differences between predicting an unbiased sample and the unobserved sample, especially with respect to implementation, which requires the researcher to recognize predicting the unobserved sample should not be treated as the same predicting an unbiased sample. We also apply the methods in simulation and find very similar patterns and conclusions from the main paper result.

A.1 Strategies to address selection on unobservables

In this section, we outline how the importance weighting and control function procedures can be altered to predict the unobserved sample.

When considering predicting the unobserved, the loss function for the importance weighting approach must be altered slightly.

$$\arg\min_{\theta} \sum_{i=1}^{n} \frac{1 - P(S_i = 1|X_i)}{P(S_i = 1|X_i)} S_i L\left(\hat{f}(X_i, \theta), Y_i, \alpha\right) \tag{3}$$

where in place of $P(S_i = 1)$ in the numerator, we have the probability of being unobserved, $P(S_i = 0|X_i) = 1 - P(S_i = 1|X_i)$.

This means, that instead of re-weighting just by the inverse of the probability of being observed, we also weight by the probability of being part of the prediction sample (unobserved). Intuitively, this is utilizing the information that the prediction sample was unobserved, $S_i = 0$, which adds more information to the prediction process.

By weighting the sample by the inverse of the proportion of be observed, we are informing

the algorithm which observations are more important to use for training (because they were under-sampled) versus those we should not place a lot of importance to (because they are over-sampled). By also weighting by the proportion of being part of the prediction sample (unobserved), we are telling the algorithm how likely we will see a similar observation in the prediction sample and therefore we should place more importance on that observation.

When applying the two CF methods, CF-VA and CF-PO, changing from predicting an unbiased sample to predicting the unobserved, the training process does not change. This means the calculation of the control function (inverse Mills ratio) and the implementation of the CF methods (either variable addition or partialling out) stays the same. The notable difference in predicting the unobserved sample is how we apply the trained learner to the prediction set to obtain the best possible prediction outcome.

For example, if we are interested in predicting an unbiased sample, as done in the main paper, then the best predictor (in the mean-squared-error sense) is the conditional mean,

$$E(Y_i|X_i, Z_i) = f(X_i)$$

whereas, if we are interested in obtaining predictions of the unobserved, then the best predictor is still the conditional mean,

$$E(Y_i|X_i, Z_i, S_i = 0) = f(X_i) + E(U_i|X_i, Z_i, S_i = 0) \tag{4}$$

but we can condition on the knowledge that the outcome is unobserved in the training sample. This adds information and predictive power. The second term in the above conditional mean can be derived based on the model assumptions. For example, When assuming the errors are jointly normal and independent of the features, then

$$E(U_i|X_i, Z_i, S_i = 0) = -\rho\lambda(-g(X_i, Z_i, \delta)) \tag{5}$$

where $\lambda(\cdot)$ is the inverse Mills ratio. By incorporating the information from the control function when predicting the unobserved sample, we can achieve additional gains in predictive accuracy. How this is executed, depends on which CF approach is applied.

First let us understand how this effects CF-VA approach. Recall that these learners

estimate the following population object

$$E(Y_i|X_i, Z_i, S_i = 1) = f(X_i) + \rho\lambda(g(X_i, Z_i, \delta)) \tag{6}$$

where we use $\hat{h}([X_i, \hat{\lambda}_i], \gamma)$ to model $E(Y_i|X_i, Z_i, S_i = 1)$ in equation (6) as a function of both X_i and $\hat{\lambda}_i = \lambda(\hat{g}(X_i, Z_i, \hat{\delta})$ where $\hat{\delta}$ is estimated in a first stage. Then in this case, an estimate of $E(Y_i|X_i, Z_i, S_i = 0)$ is recovered from $\hat{h}(X_i, -\lambda(-\hat{g}(X_i, Z_i, \hat{\delta})), \hat{\gamma})$. We expect that the inclusion of the second term adds information to the learner and could possibly result in a more accurate predictor, even compared to an unbiased baseline.

Unlike the CF-VA approach, applying the CF-PO approach to predicting the unobserved sample requires several steps for implementation. Recall that $\tilde{X}_i$ and $\tilde{Y}_i$ denote the partialled out attributes and outcome respectively. Then the learners are estimating $E(\tilde{Y}_i|\tilde{X}_i, S_i = 1)$ with $\hat{f}(\tilde{X}_i, \theta)$ where $\hat{f}(\cdot, \theta)$ is an approximation to the conditional mean function $f(\cdot)$ following linear projection arguments.

To apply the learner to predict in an unbiased sample we use, $\hat{f}(X_i, \hat{\theta})$. Note that when using the learner for prediction we evaluate using the un-partialled out attributes. The partialled out random variables can be thought of as the sample selection bias adjusted random variables: adjusted as if there was no selection on unobservables.

To predict in the unobserved sample, we would like to incorporate the information captured by $E(U_i|X_i, Z_i, S_i = 0)$ in equation (4). Implementation of this takes several steps after training the learner and requires model assumptions to derive $E(U_i|X_i, Z_i, S_i = 0)$. For example, consider the jointly normal case where the conditional mean is defined in equation (5. To predict the unobserved sample, we would first, estimate ρ by regressing $Y_i - \hat{f}(X_i, \hat{\theta})$ on $\lambda(\hat{g}(X_i, Z_i, \hat{\delta}))$ in the observed training sample. Then in a second step, predict the unobserved conditional mean, $E(Y_i|X_i, Z_i, S_i = 0)$ with

$$\hat{f}(X_i, \hat{\theta}) - \hat{\rho}\lambda(-\hat{g}(X_i, Z_i, \hat{\delta}) \tag{7}$$

By incorporating the additional information for the unobserved sample, we also expect the CF-PO approach could result in a more accurate predictor than the unbiased baseline.

A.2 Simulation

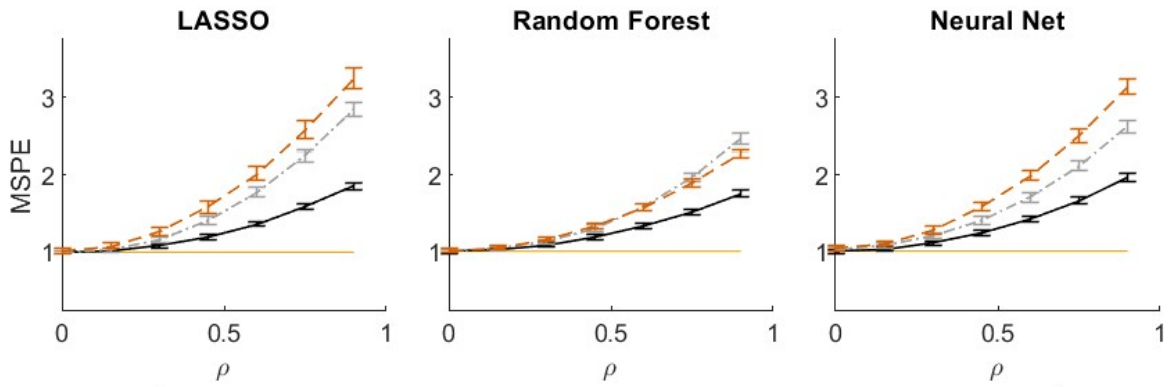
This section reports the simulations results for the data generating processes and estimators described in section 4 of the main paper, but applied to predicting the unobserved sample.

Figure 1 reports the simulation results for the three considered DGPs. As a reference, we include an unbiased baseline equal to the median MSPE of the Ignore Sample Selection (1) when $\rho = 0$. The patterns and conclusions from these figures mimic the patterns and conclusions in the main paper. First, selection on unobservables does matter with respect to prediction quality, in that both ignoring sample selection and using a weighting approach worsen as the correlation increases. Second, including the instrument as an attribute can result in worse prediction accuracy. Third, weighting will predict worse than ignoring sample selection at high levels of correlation, even if there are gains to be had when there is only selection on observables.

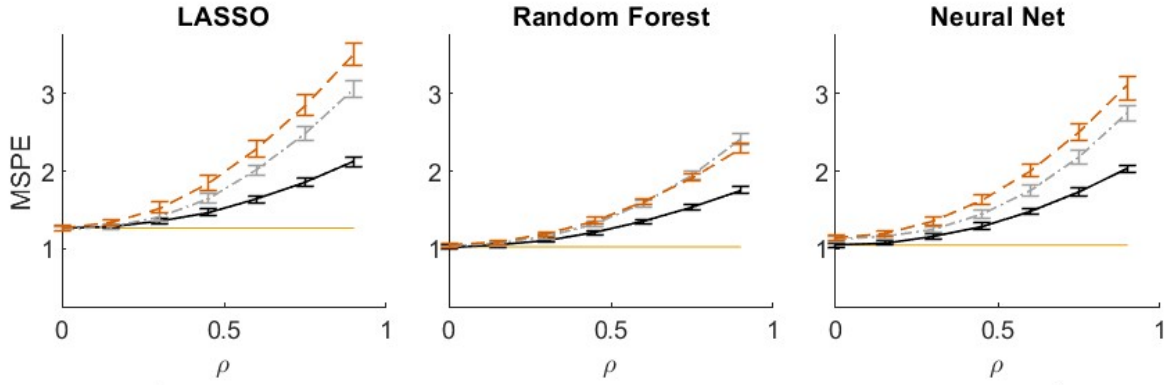
Figure 2 reports the simulation results for the two CF approaches over the three considered DGPs. There are several notable patterns that depart from the main paper. First CF-PO is fairly consistently downward sloping as correlation increases, and in some most cases, dips below the unbiased baseline. This make sense because in our prediction process we capitalize on the additional information predicting on the unobserved sample provides. In contrast, the quality of CF-VA is more variable across DGPs and learners. For instance, with LASSO, the CF-VA appears to perform similar to the CF-PO approach unless there is a weak instrument (DGP3). For random forest and neural nets, CF-VA does offer minor improvements over ignore SS (1), but does not always achieve the same downward slope that CF-PO does.

Consequently, the advice to the applied researcher stays very much the same whether you are predicting in an unbiased sample or the unobserved sample. In general, the CF-PO approach will provide better prediction quality. If the correlation is rather low, so selection does not depend much on the unobservable component, then the CF-PO does not provide much improvement.

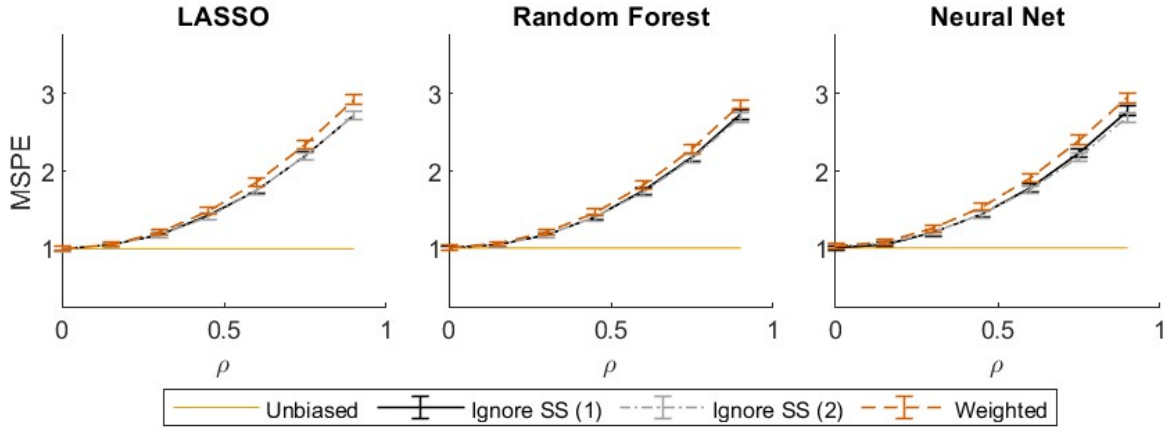
(a) DGP1



(b) DGP2



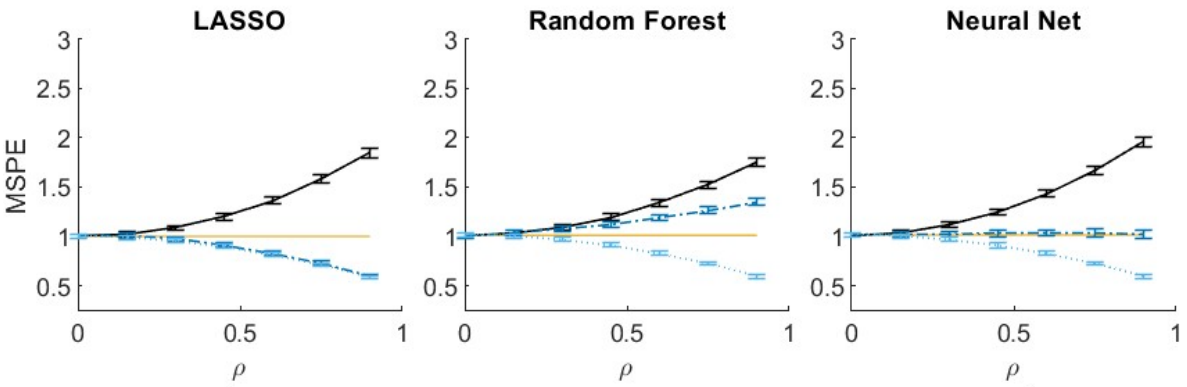
(c) DGP3



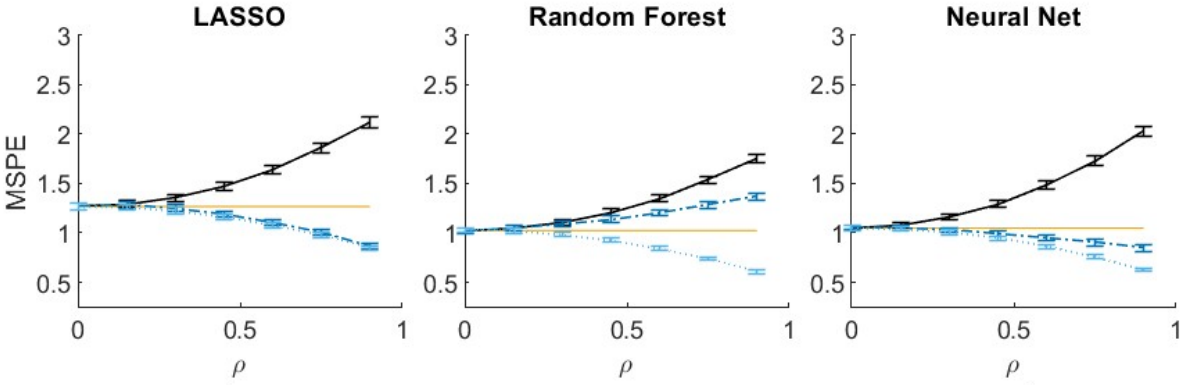
Compares unobserved sample MSPE for weighted and un-weighted approaches to sample selection with error bars corresponding to 10^{th} and 90^{th} percentiles across simulations for (a) DGP1 , (b) DGP2, and (c) DGP3.

Figure 1: Unobserved sample: Weighted and un-weighted approaches to sample selection.

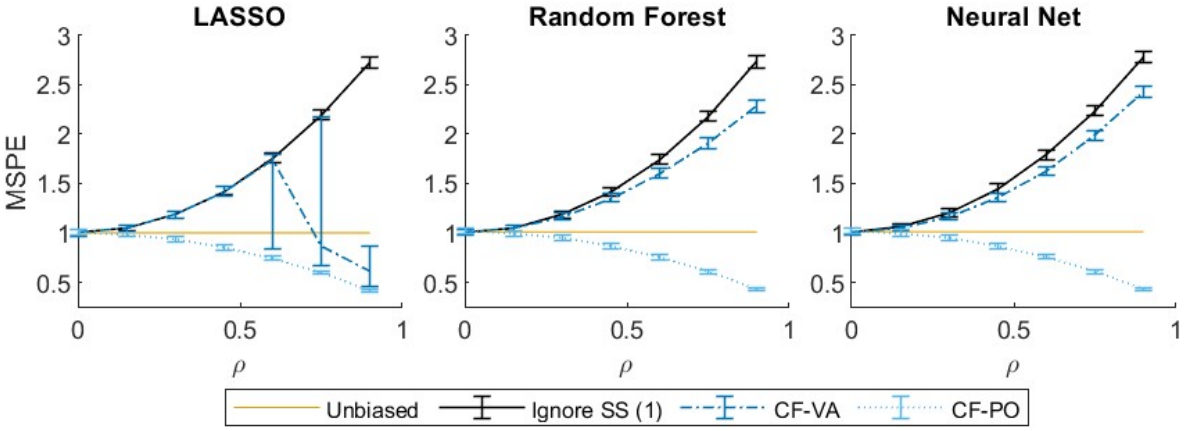
(a) DGP1



(b) DGP2



(c) DGP3



Compares unobserved sample MSPE for Heckman CF approaches to sample selection with error bars corresponding to 10^{th} and 90^{th} percentiles across simulations for (a) DGP1 , (b) DGP2, and (c) DGP3.

Figure 2: Unobserved sample: CF approaches to sample selection

B Simulations for control function approaches with misspecification

Fundamental to the accuracy and validity of the control function approaches are the following three assumptions:

1. Excluded Instrument: the instrument Z_i has no impact on the outcome directly, e.g., $E(Y_i|X_i, Z_i) = E(Y_i|X_i)$
2. Instrument Relevance: the instrument is informative for the sample selection process, e.g., $g(X_i, Z_i, \delta) \neq g(X_i, \delta)$.
3. Correct CF functional form: $E(U_i|X_i, Z_i, S_i = 1) = E(U_i|g(X_i, Z_i, \delta)) = m(g(X_i, Z_i, \delta))$ is well approximated by a linear combination of $m^k(\cdot) = (m_{1k}(\cdot), ..., m_{kk}(\cdot))$

The main paper considers (near) violations of the second assumption with weak instruments in the third DGP in the simulation. In this section, we consider violations of the other two assumptions and investigate the performance of the CF estimators.

B.1 Not excluded instrument

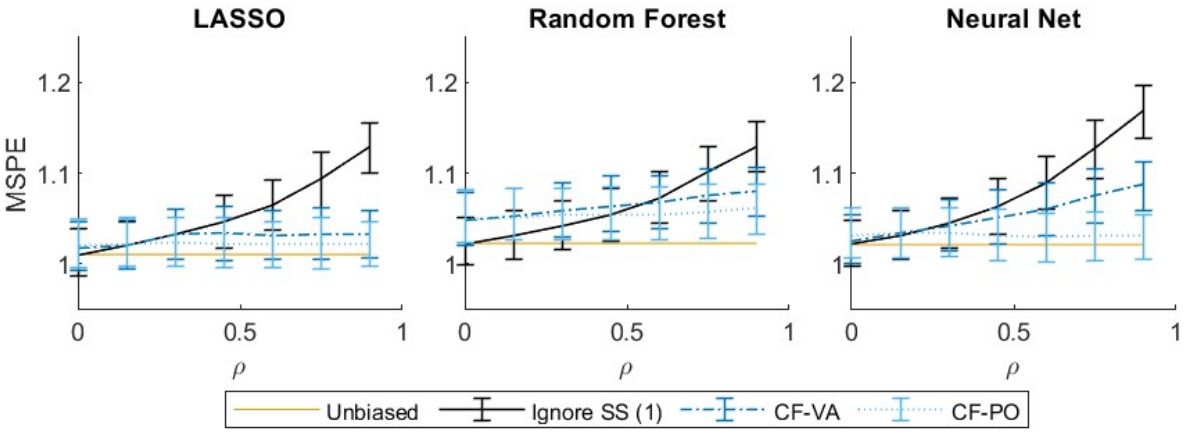
The instrument is not excluded if it has a direct impact on the outcome Y_i in the structural equation. This is problematic for the control function approaches since including Z_i in the conditional mean model for outcome would result in a loss of identification. Specifically, let the conditional mean also be a function of the instrument, $E(Y_i|X_i, Z_i) = f(X_i, Z_i)$, then there is no new identifying power in the control function $m(g(X_i, Z_i, \delta))$ resulting in a lack of identification. On the other hand, if a researcher excludes Z_i from the conditional mean model for the outcome when it truly has an effect, then there is omitted variable bias. We will investigate the consequence of presuming our instrument is excluded, while in reality it has a direct impact on the outcome.

The data generating process is the same as the first DGP in the main paper, but the instrument has a direct impact on the outcome,

$$f(X_i, Z_i) = \sum_{j=1}^{100} X_{ji}\theta_j + 0.1Z_i. \tag{8}$$

Note that with a coefficient of 0.1, Z_i has the second largest effect on Y_i among all attributes. The simulations consist of 500 iterations with training sample sizes of around 20,000 and prediction sample sizes of 5,000.

Figure 3 reports the MSPE results for out of sample prediction performance of the two CF approaches as well as the Ignore SS (1) as a comparison. Both the control function approaches and the Ignore SS (1) tend to do worse with an invalid instrument compared to the case in which the instrument is valid, DGP 1 in the main paper. This is because the exclusion of Z_i as a predictor for the outcome creates omitted variable bias when the instrument truly has a direct effect on the outcome. In addition, the CF approaches tend to perform worse relative to the Ignore SS (1), particularly at low levels of correlation. This is because the estimators are not only suffering from omitting the impact of Z_i but they are also misappropriating the impacts of the instrument Z_i through the control function. This means that both of the CF approaches tend to have higher MSPE relative to the unbiased baseline. Finally, we find that even with the poorer performance from the CF estimators, when the level of endogenous sample selection is very high, then the CF approaches can still offer some improvements for prediction.



Compares MSPE for Heckman CF approaches to sample selection with error bars corresponding to 10^{th} and 90^{th} percentiles when the instrument is not excluded.

Figure 3: CF approaches to selection with instrument not excluded.

B.2 Incorrect CF functional form

Specifying the incorrect CF functional form can occur for a variety of reasons. For example, when we use the inverse Mills ratio as the CF, then CF can be misspecified if the true underlying distribution is not jointly normal. Alternatively, if the errors are heteroskedastic (also a consequence of heterogeneous coefficients) then $E(U_i|X_i, Z_i, S_i = 1) = E(U_i|X_i, Z_i, g(X_i, Z_i, \delta)) \neq E(U_i|g(X_i, Z_i, \delta))$ so the conditional mean of the unobserved error conditional on selection cannot be controlled for only as a function of the model for the sample selection process. We investigate the consequences of misspecification along the lines of the first example. The second example of misspecification has been shown to be quite consequential to the validity of the CF approaches (Schaffner 2002, Carlson & Joshi 2022). Recently, a "generalized" CF approach proposed by Carlson & Joshi (2022) and Carlson (2022) allows for heteroskedasticity by including interactions between X_i and the CF in linear models and we leave it to future research to extend it to the ML setting.

The data generating process is the same as the first DGP in the main paper but instead of jointly normal errors, we generate errors as follows,

$$V_i \sim 0.8N(-0.5, 0.2) + 0.2N(2, 0.4) \tag{9}$$

$$U_i \sim \rho V_i + \sqrt{1-\rho^2}\left[0.8N(-0.5, 0.2) + 0.2N(2, 0.4)\right]. \tag{10}$$

where both V_i and U_i are drawn from a mixture of normal distributions (with 80% probability, a $N(-0.5, 0.2)$, and with 20% probability, $N(2, 0.4)$), which produces bimodal and skewed distributions with mean 0 and variance approximately 1. Note that the correlation between U_i and V_i is created through ρV_i when generating U_i.

We present two variations of the CF estimators. The first variation incorrectly assumes normality and therefore uses the inverse Mills ratio as the control function

$$\hat{m}_i = \lambda(\hat{g}(X_i, Z_i, \hat{\delta})). \tag{11}$$

The second variation attempts to allow for departures from normality by using a series

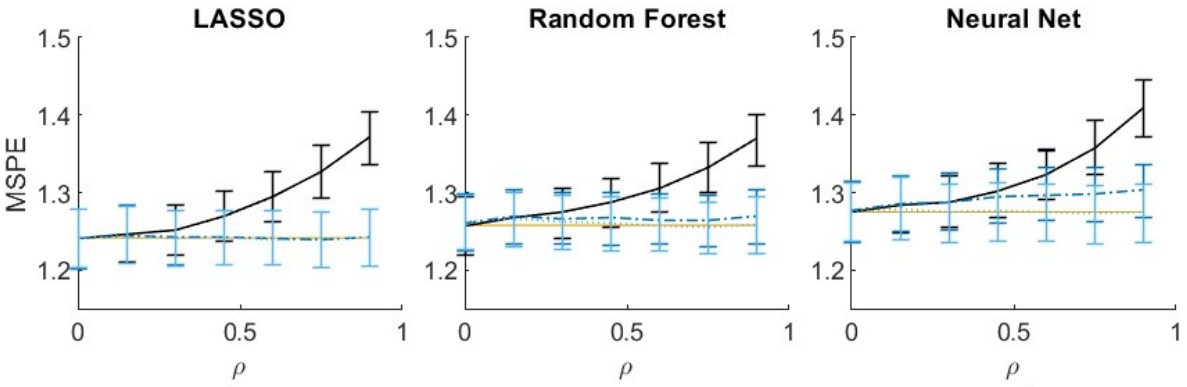
approximation. Specifically, we use polynomials up to order 4 of the inverse Mills ratio

$$\hat{m}_i = \big(\lambda(\hat{g}(X_i, Z_i, \hat{\delta})), [\lambda(\hat{g}(X_i, Z_i, \hat{\delta}))]^2, [\lambda(\hat{g}(X_i, Z_i, \hat{\delta}))]^3, [\lambda(\hat{g}(X_i, Z_i, \hat{\delta}))]^4\big). \tag{12}$$

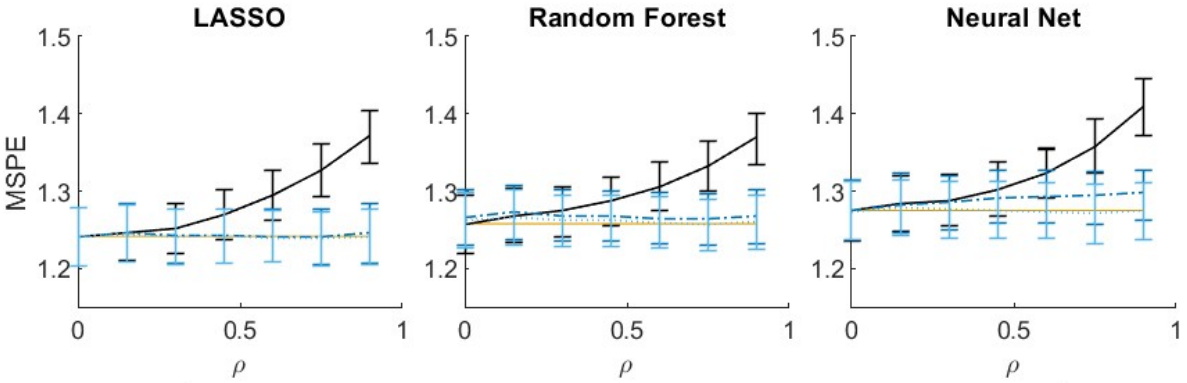
Alternatively one could consider using splines instead of polynomials, or take series of $\hat{g}(X_i, Z_i, \hat{\delta}))$ or $\Phi(\hat{g}(X_i, Z_i, \hat{\delta}))$ instead (see Newey (2007) for descriptions of other possible variations).

Figure (4) reports the results. We find that even when the distribution of the errors deviate far from normality, using the inverse Mills ratio for the control function, displayed in panel (a) still works quite well. This has been noted in the literature before, Van der Klaauw & Koning (2003) observed in their simulations that "departures from normality do not cause serious bias." Utilizing the non-parametric CF, displayed in panel (b), also works well, but given the already strong performance of the inverse Mills ratio, it does not offer much, if any, improvement.

(a) Inverse Mills ratio control function



(b) Non-parametric control function



Compares MSPE for Heckman CF approaches to sample selection with error bars corresponding to 10^{th} and 90^{th} percentiles for non-normal DGP using (a) inverse Mills ratio or (b) non-parametric control functions.

Figure 4: CF approaches to selection with parametric vs non-parametric CF.

C Supplemental tables for application

Table 1: Application summary statistics

	(1) Incumbent Mean/SD	(2) Open Mean/SD	(3) Difference Diff./t-stat
Per-capita dividends	4,889.93	4,843.95	45.98
	(1,279.95)	(1,399.04)	(0.38)
Per-capita earnings	30,802.06	30,722.95	79.11
	(5,342.46)	(5,182.02)	(0.17)
Per-capita household earnings	16,980.64	16,778.80	201.85
	(3,850.09)	(3,900.43)	(0.58)
Per-capita transfer benefits	332.86	355.75	-22.89
	(131.80)	(148.69)	(-1.79)
Per-capita non-farm proprieters income	23,050.77	22,931.51	119.26
	(6,628.85)	(6,325.43)	(0.20)
Per-capita retirement income	2,918.58	3,013.68	-95.10
	(947.42)	(1,094.85)	(-1.02)
Per-capita unemployment insurance	121.90	144.92	-23.02**
	(81.06)	(108.62)	(-2.62)
Per-capita farm proprieters income	19,611.15	20,614.29	-1,003.14
	(18,838.67)	(21,728.91)	(-0.54)
Presidential election year	0.23	0.23	0.00
	(0.42)	(0.42)	(0.03)
Farm employment	45,086.60	42,905.31	2,181.29
	(44,598.66)	(37,613.53)	(0.59)
Non-farm employment	415109.35	425139.25	-10029.90
	(554009.76)	(555372.92)	(-0.20)
Wage employment	2.26e+06	2.23e+06	37,498.32
	(2.59e+06)	(2.44e+06)	(0.17)

Republican incumbent	0.46	0.41	0.04
	(0.50)	(0.49)	(1.00)
Democrat incumbent	0.53	0.57	-0.04
	(0.50)	(0.50)	(-0.85)
Third-party challenger	0.11	0.18	-0.08*
	(0.31)	(0.39)	(-2.35)
Incumbent's previous vote share	0.56	0.58	-0.02***
	(0.07)	(0.09)	(-3.35)
Observations	285	217	502
Counties	17095	12895	29990

t-statistics assume unequal variances, ** p<0.01, * p<0.05

Table 2: Application first stage Probit coefficient estimates.

VARIABLES	(1)	(2)	(3)
Term limit	-1.125***	-1.135***	-1.144***
	(0.122)	(0.123)	(0.124)
Third party present		-0.162**	-0.115*
		(0.0650)	(0.0690)
Unemployment insurance per capita			-0.152**
			(0.0707)
Constant	-0.00267	-0.00407	-0.00594
	(0.0825)	(0.0832)	(0.0835)
Observations	502	502	502

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1
Column 3 contains all covariates selected in the L1-penalized Probit and columns 1-2 show how the coefficient on the instrument changes when removing the other two selected covariates.

Table 3: Application mean-squared prediction errors

	(1) Ignore SS	(2) Weighted	(3) CF-PO
LASSO	0.0085	0.0090	0.0081
Random forest	0.0046	0.0041	0.0164
Neural net	0.0028	0.0026	0.0031

Ten-fold mean-squared prediction error of the learners for approaches ignoring selection, weighting to address selection on observables, and using a control function approach (CFPO).

References

Carlson, A. (2022), 'gtsheckman: Generalized two-step heckman estimator', *2022 Stata Conference, Stata Users Group* . https://www.stata.com/meeting/us22/slides/US22_Carlson.pdf.

Carlson, A. & Joshi, R. (2022), 'Sample selection in linear panel data models with heterogeneous coefficients'. https://ideas.repec.org/p/umc/wpaper/2203.html.

Newey, W. K. (2007), 'Nonparametric continuous/discrete choice models', *International Economic Review* **48**(4), 1429–1439.

Schaffner, J. A. (2002), 'Heteroskedastic sample selection and developing-country wage equations', *Review of Economics and Statistics* **84**(2), 269–280.

Van der Klaauw, B. & Koning, R. H. (2003), 'Testing the normality assumption in the sample selection model with an application to travel demand', *Journal of Business & Economic Statistics* **21**(1), 31–42.