Project of CIML

The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions

Sergio Blanco Piñeiro

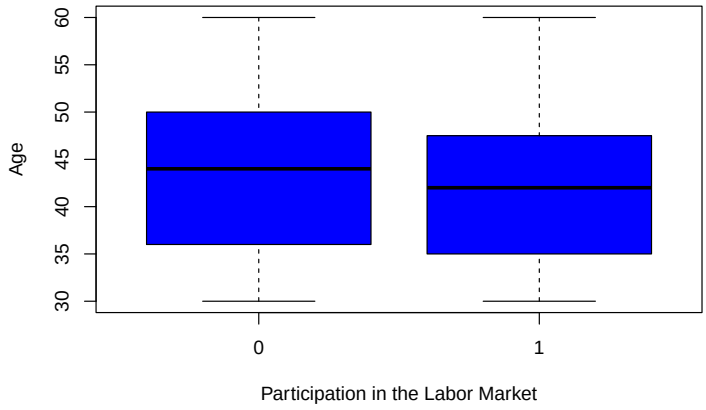
2023-03-31

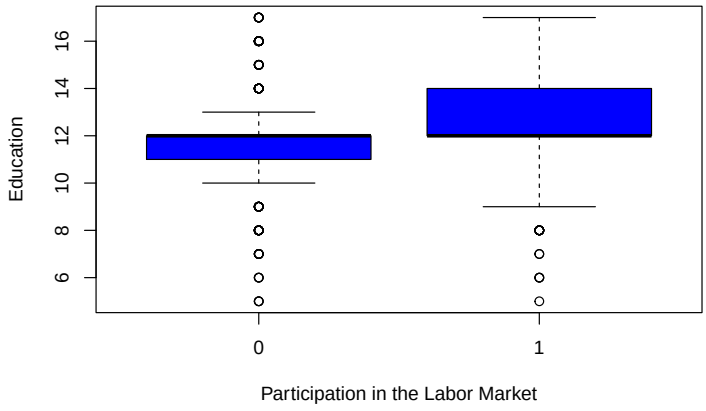
Abstract

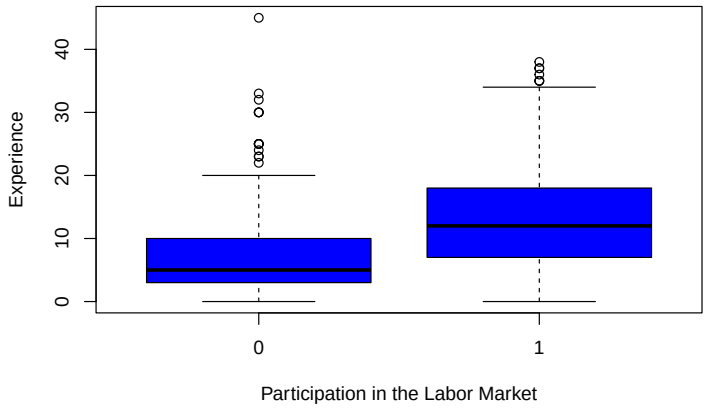
Our goal will be to follow Mroz's multistage procedure to estimate a structural labor supply equation. We will employ the data set used in Mroz (1987), which contains data for 753 working-age, married, woman, in 1975. The data set is a Cross-section data originating from the 1976 Panel Study of Income Dynamics (PSID), based on data for the previous year, 1975.

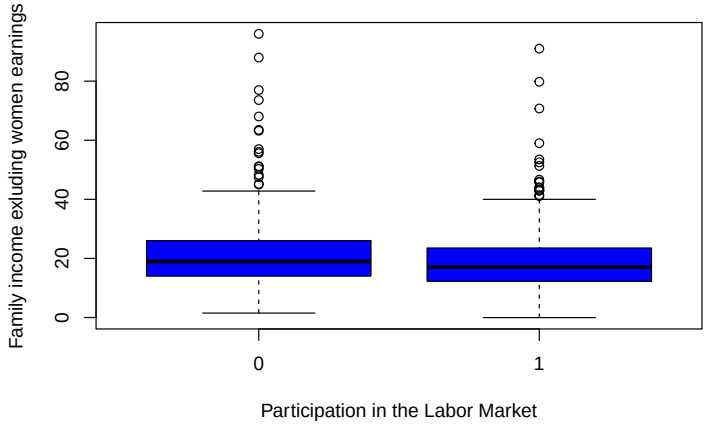
Data loading and visualization

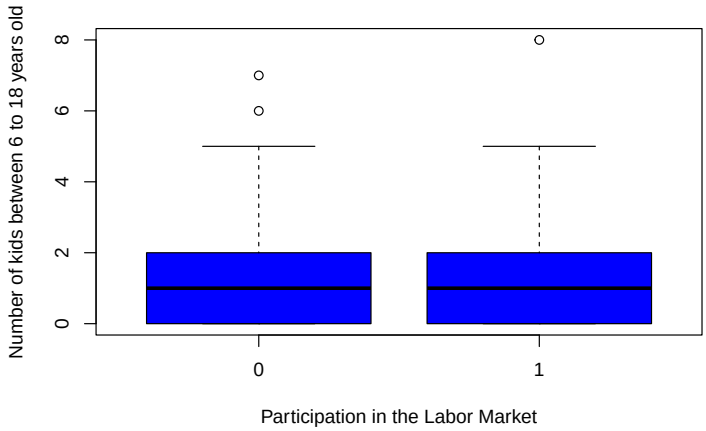
We will compute some basic descriptive statistics for the variables of this data set. In the following boxplots we can see that we have a Sample Selection issue since the Education, Age, Experience, Family income excluding women earnings and number of kids differ across women that participate and not on the Labour Market.

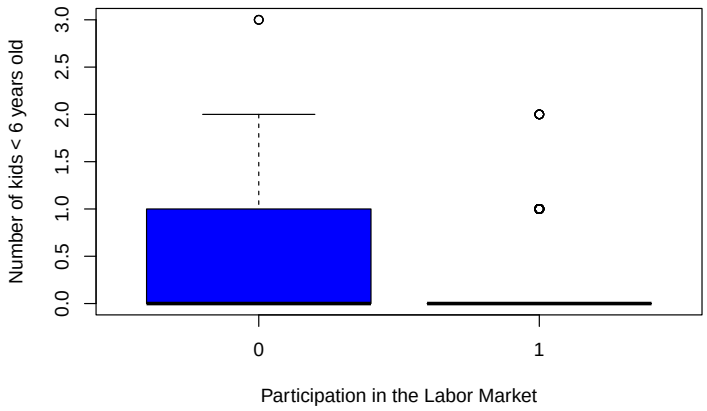





Experience
0
10
20
30
40
0
1
Participation in the Labor Market


Family income exluding women earnings
0
20
40
60
80
0
1
Participation in the Labor Market


Number of kids between 6 to 18 years old
0
2
4
6
8
0
1
Participation in the Labor Market



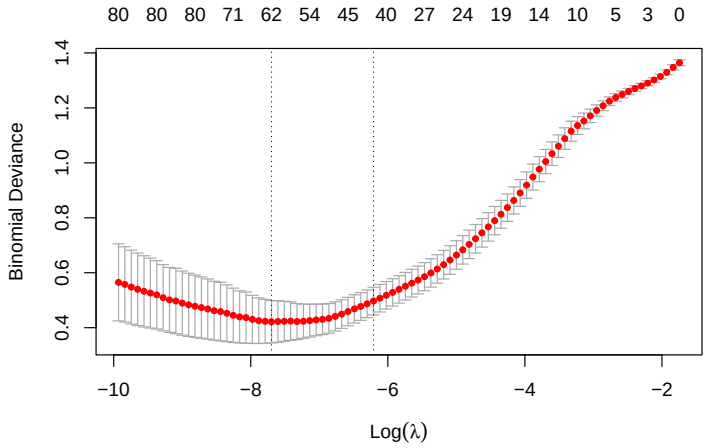
Imputation of log(Wage)

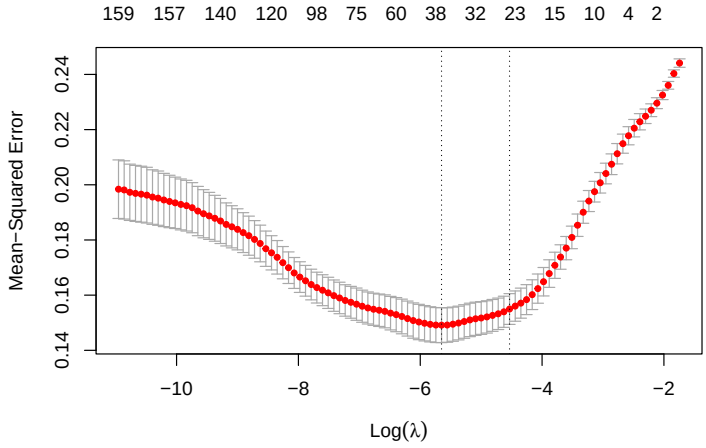
Lets compute a naive regression for the subpopulation of women that participate in the labor market. We will estimate the following equation: $log(Wage)_i = \beta_0 + \beta_1 * Age_i + \beta_2 * Education_i + \beta_3 * Experience_i + \beta_4 * Experience_i{}^2 + \beta_5 * City_i + \epsilon_i$ where we have as explanatory variables woman's age, education, labor market experience and its square, and the binary variable on living in a large city and the log of wages as a dependent variable. After we estimate this naive regression as our objective in this part is to generate an imputed log(Wages) (called, i_log(Wages)) Once we achieve that we will use that prediction in order to impute the i_log(Wage) for the non-participant women and us the prediction as the i_log(Wage) for the participant women. Even though we imputed this log(Wage) by applying the Heckman two-step-procedure we will not use this i_log(Wages) as this method accounts for Sample Selection.

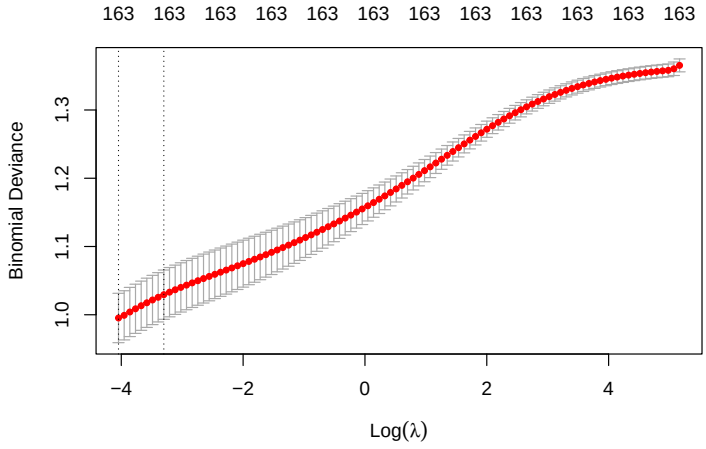
First Step estimation

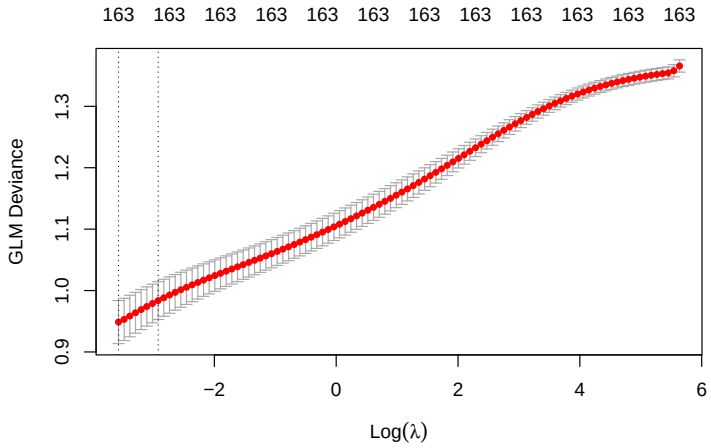
Once, we have created our variable i_log(Wage). Lets estimate the following equation $D_i := 1\{inlf_i = 1\} = F(\mathbf{Z}_i^T\gamma)$ using a set of covariates $\mathbf{Z}_i$. In order to estimate this model we will use firstly a Linear Probability model, Logit model and a Probit model using a naive regression with a small set of covariates. But before we estimate the previous model, we will create a HD setting.

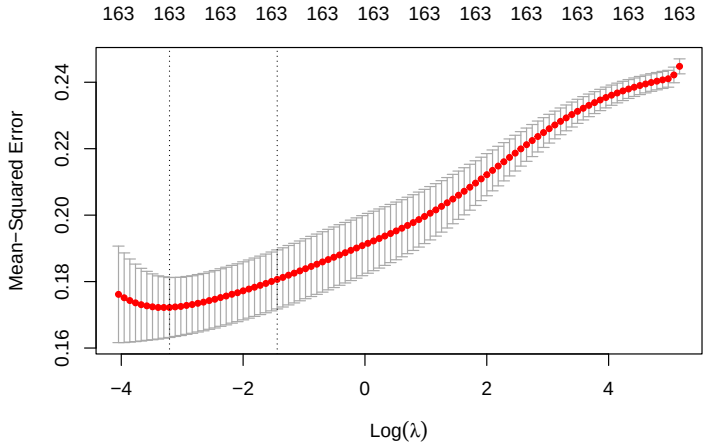
Once we have computed the naive First Step regression (notice that we will only care about prediction) we will make use of the machine learners in order to predict the participation of women in our sample. We will also exploit the HD created by interactions and powers of the covariates in our sample. For that, we will use a set of learners that will be trained in the i-th training set and tested in the i-th test set, where $i \in \{1, .., 5\}$:

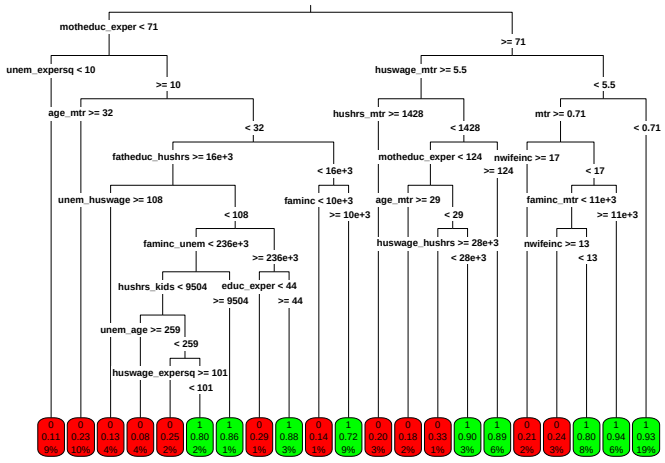

80 80 80 71 62 54 45 40 27 24 19 14 10 5 3 0
Binomial Deviance
0.4 0.6 0.8 1.0 1.2 1.4
−10 −8 −6 −4 −2
Log(λ)

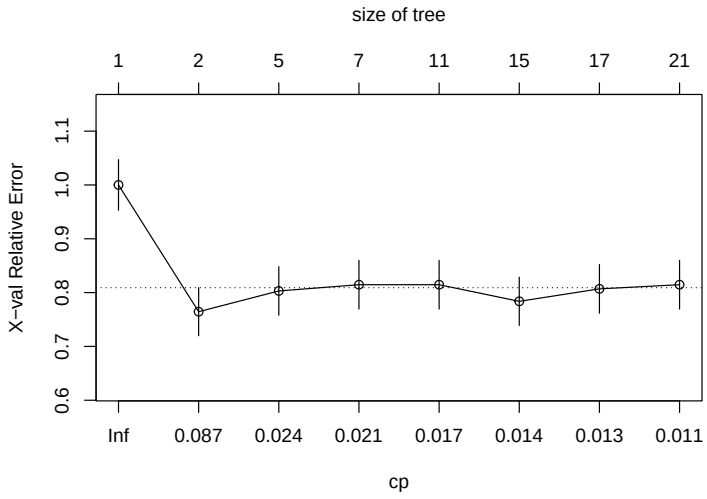

159 157 140 120 98 75 60 38 32 23 15 10 4 2
Mean−Squared Error
0.14 0.16 0.18 0.20 0.22 0.24
−10 −8 −6 −4 −2
Log(λ)


163 163 163 163 163 163 163 163 163 163 163 163
Binomial Deviance
1.0 1.1 1.2 1.3
−4 −2 0 2 4
Log(λ)


163 163 163 163 163 163 163 163 163 163 163 163
GLM Deviance
0.9 1.0 1.1 1.2 1.3
−2 0 2 4 6
Log(λ)


163 163 163 163 163 163 163 163 163 163 163 163
Mean−Squared Error
0.16 0.18 0.20 0.22 0.24
−4 −2 0 2 4
Log(λ)


motheduc_exper < 71
>= 71
unem_expersq < 10
>= 10
huswage_mtr >= 5.5
< 5.5
age_mtr >= 32
< 32
hushrs_mtr >= 1428
< 1428
mtr >= 0.71
< 0.71
fatheduc_hushrs >= 16e+3
< 16e+3
motheduc_exper < 124
>= 124
nwifeinc >= 17
< 17
unem_huswage >= 108
< 108
faminc < 10e+3
>= 10e+3
age_mtr >= 29
< 29
faminc_mtr < 11e+3
>= 11e+3
faminc_unem < 236e+3
>= 236e+3
huswage_hushrs >= 28e+3
< 28e+3
nwifeinc >= 13
< 13
hushrs_kids < 9504
>= 9504
educ_exper < 44
>= 44
unem_age >= 259
< 259
huswage_expersq >= 101
< 101



In this plots (notice that we only plot the first splits) we can see that taking into account $\lambda = 0 \Rightarrow log(\lambda) = \infty$ that even though we do not have $p > n$ given our number of covariates using Lasso or Ridge may have some justification as we do not archive the minimum for the measure of performing for lambdas equal to 0. Take into account that this will vary across our 5 splits of the training sample. We will display also some measures of performance of the Decision Tree learner and the a graph of this Decision Tree for the 5 splits.

Table 1: RCC

	Split 1	Split 2	Split 3	Split 4	Split 5	Average
Naive Logit	0.6206897	0.6078431	0.6184211	0.6602564	0.5730337	0.6160488
Naive Probit	0.6068966	0.6078431	0.6250000	0.6602564	0.5730337	0.6146060
Naive LPM	0.6137931	0.5882353	0.6315789	0.6602564	0.5617978	0.6111323
Logit Lasso	0.9517241	0.9346405	0.9210526	0.9423077	0.9382022	0.9375854
Probit Lasso	0.9379310	0.9281046	0.9144737	0.9102564	0.9269663	0.9235464
LPM Lasso	0.8275862	0.8562092	0.8157895	0.8333333	0.8089888	0.8283814
Logit Ridge	0.7517241	0.7647059	0.7302632	0.7564103	0.7808989	0.7568005
Probit Ridge	0.7517241	0.7450980	0.7236842	0.7307692	0.7640449	0.7430641
LPM Ridge	0.7517241	0.7450980	0.7236842	0.7884615	0.8202247	0.7815215
LDA	0.7448276	0.7777778	0.7763158	0.7820513	0.7640449	0.7651626
Decision Tree	0.7517241	0.7450980	0.7828947	0.6987179	0.6853933	0.6791865
Random Forest	0.6413793	0.6928105	0.6776316	0.7884615	0.7528090	0.7375350
K-Nearest Neighbors (K=24)	0.7103448	0.7189542	0.7171053	0.5705128	0.5786517	0.5765515
K-Nearest Neighbors (K=25)	0.5793103	0.5490196	0.6052632	0.6089744	0.5674157	0.5887661

In the previous matrix we can see the Rate of Correct Classification (RCC) for the five splits of the test set and for the average value, that will be the measure of out-of-sample performing.

```
##
##       0  1
##   0 66  7
##   1  0 72
```

```
##
##       0  1
##   0 66  9
##   1  0 70
```

As we can see here the confusion matrix (notice that we only plot the first splits) for the Logit Lasso and the Probit Lasso tell us something similar to what the RCC is telling us. Hence, we will use the measure of RCC as a measure of performing out-of-sample since this measure is consistent with the confusion matrix.

Once we have tested all our learners, by computing the Rate of Correct Classification (RCC) and also the confusion matrices of each split, we will proceed by selecting two different models for our estimation of this First Stage. These model are the Logit Lasso and the Probit Lasso, as they report a RCC of 0.95 and 0.93, respectively. So, once we have selected the models we will use these predictions to compute the Second Stage. In order to do show, we will compute a similar procedure to the one following by Mroz (1987) in the known Mroz's multistage procedure to estimate a structural labor supply equation. This procedure is as follows, we will have the following equations, where the setting is as stated in Weerasooriya (2018), that in our setting is the following one:

$D_i := 1\{inlf_i = 1\} = F(\mathbf{Z}_i^T\gamma) + e_i$ (1)

$log(Wage_i) = \mathbf{X}_{1i}^T\beta + v_{1i}$ (2)

$h_i = \mathbf{X}_{2i}^T\beta + v_{2i}$ (3)

where equation (1) is the Selection equation in the model and equation (2) is the Second Stage. Hence, both equations (1) and (2) accounts for the Heckman two-step procedure. We will have that $\mathbf{Z}_i$ in Mroz (1987) are the following covariates, for two different specifications:

(i) : family composition (children), powers of woman's age and education and their interactions between them to approximate a third-order polynomial in these two variables,education of the woman's father and mother, local unemployment rate, binary variable on living in a large city, and family income excl. woman's earnings.

(ii) : the variables in (i) plus labor market experience and its square.

Whereas in Mroz (1987) this is the set of covariates used in equation (1), we will follow a more data driven way and thus, make use of the HD setting to predict by using machine learner. The two best learners are Logit Lasso and the Probit Lasso. We will compute the Inverse Mill's ratio to account for the sample Selection in our Second Stage. Once we have computed an estimate of $\lambda(\mathbf{Z}_i^T\gamma)$, say $\hat{\lambda}(\mathbf{Z}_i^T\hat{\gamma})$, where we can apply the Heckman two-step approach assuming joint normality of $(e_i, v_{1i}) \sim \mathcal{N}_{\in}(0, , \Sigma)$ or a non-parametric version where we compute $\hat{\lambda}(\mathbf{Z}_i^T\hat{\gamma}) = \dfrac{f_n(\mathbf{Z}_i^T\hat{\gamma})}{F_n(\mathbf{Z}_i^T\hat{\gamma})}$ where $f_n(\mathbf{Z}_i^T\hat{\gamma})$ is the empirical pdf from the predicted values

of the machine learner estimate, while $F_n(\mathbf{Z}_i^T\hat{\gamma})$ is the empirical cdf from the predicted values of the machine learner estimate. Once we have compute our estimate of the Inverse Mill's ratio, we will use it in equation (2) where $\mathbf{X}_{1i} = (\mathbf{Z}_i, \hat{\lambda}(\mathbf{Z}_i^T\hat{\gamma}))$ and by that we will account for the selection bias.

Once, we have solve this selection issue in equation (2) we will proceed by predicting the log(Wage) using in this step also some learners.

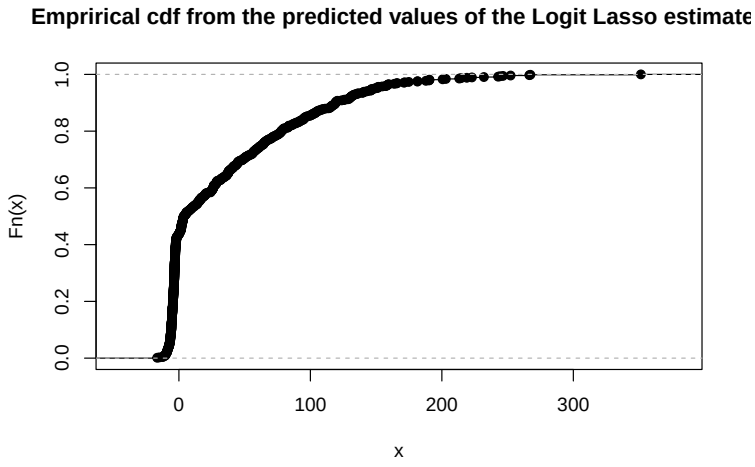
Finally, the contents of $\mathbf{X}_{2i}$ will be $\hat{log}(Wage)_i$ from equation (2), the non-wife income, number of older and younger children, the age and education of the wife and $\hat{\lambda}(\mathbf{Z}_i^T\hat{\gamma})$ from equation (1).

Second Step estimation

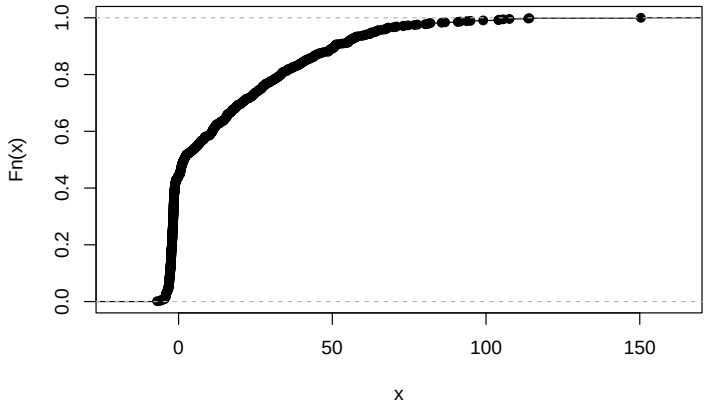
In this step we will proceed by using the estimates of the probit lasso model and the logit lasso model, as these two learners are the one that predicts the better out-of-sample. Given that, we will use both learners to compute the empirical lambda of Heckman we will use for the following formula for computing the empirical cdf:

$$F_{\mathbf{Z}i^T\hat{\gamma}}(b) = \frac{1}{n}\sum_{i=1}^{n} 1\{\mathbf{Z}_i^T\hat{\gamma} \leqslant b\}$$

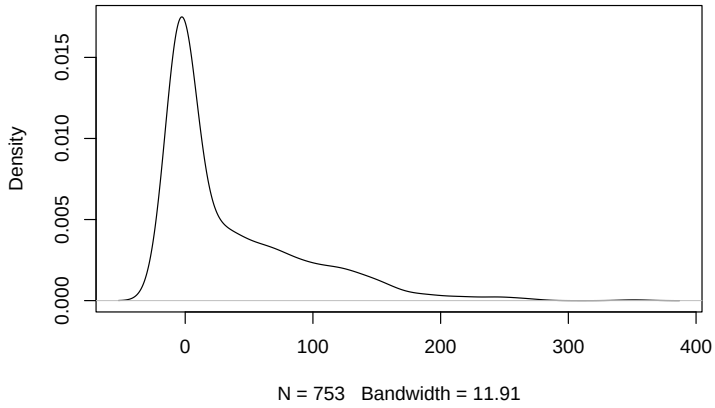
and using the Kernel Density Estimation in order to estimate the empirical pdf. Once, we have compute both empirical functions the empirical lambda de Heckman is just the ratio between both functions.



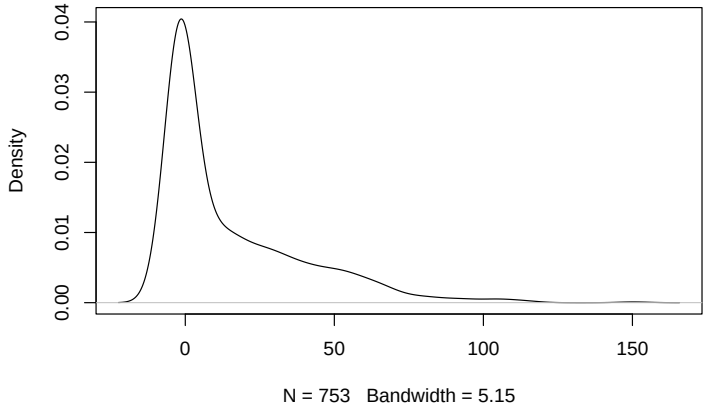
Emprirical cdf from the predicted values of the Probit Lasso estimat

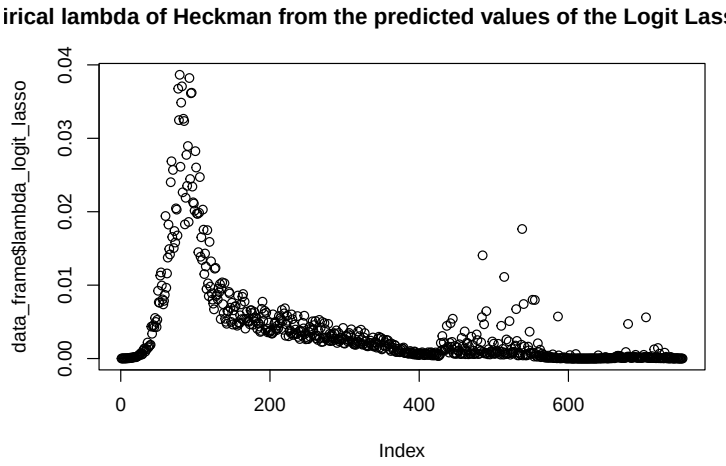


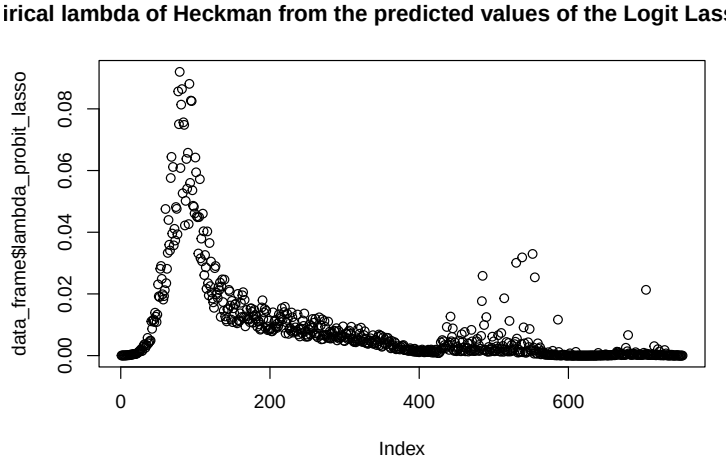
Emprirical pdf from the predicted values of the Logit Lasso estimate



Emprirical pdf from the predicted values of the Probit Lasso estimat







Now, we will proceed by training some learners in order to predict log(Wage) in equation (2) and in order to do so accounting for sample selection we will use the empirical lambda of Heckman. Hence, as we proceed by computing two versions of the empirical lambda of Heckman, from this point until the estimation of the structural labour supply equation for women we will have two different versions of the predicted log(Wage) and two versions of the the structural labour supply equation for women, say two versions of equation (3).

In the following step we will use the five divisions that we already compute in previous steps and use the two train our learners and in the following step test them. Take into account that previously we have equation (1) where the outcome is a binary variable, which implies that we use as a measure to test the learners the RCC, as stated in James, Witten, Hastie, & Tibshirani (2013). But in equation (2) we would not have classification, as our outcome is a continuous variable, then our measure in order to test the learners is going to be the MSE.

Once we have trained in five splits the following models:

(i) OLS (ii) OLS Lasso (iii) OLS Ridge (iv) OLS with $\alpha = 0.5$, which corresponds to the Elastic Net (v) Decision Tree (vi) Random Forest (vii) Neural Networks

We will use Neural Networks to test whether in this second equation performs as in Paliouras & Jessen (1999), even though these authors compute NN for equation (1). The following step is to test the the learners using the MSE as a measure of performing out-of-sample.

Table 2: Mean Squared Error (MSE) for empirical lambda of Heckman for Logit Lasso

	Split 1	Split 2	Split 3	Split 4	Split 5	Average
OLS	0.5374274	0.3962991	2.9495570	1.6460311	12.8785866	3.6815802
OLS Lasso	0.2951904	0.2125511	0.3537131	0.2611471	0.3358735	0.2916950
OLS Ridge	0.3448815	0.2485945	0.4033669	0.2866134	0.3989835	0.3364880
OLS Elastic Net	0.3078681	0.2106040	0.3555900	0.2676939	0.3525626	0.2988637
Decision Tree	0.4954084	0.3453911	0.5388070	0.3810834	0.4918682	0.4505116
Random Forest	0.3797217	0.2858148	0.4220522	0.3609050	0.4079058	0.3712799
Neural Networks	0.4404758	0.4065796	0.5693793	0.4557084	0.4804247	0.4705136

Here, for the case in which we use the empirical lambda of Heckman for the predicted values of the Logit Lasso model, we have 3 learners that predict the best out-of-sample. These learners are the OLS Lasso (Average MSE=0.2916950), the OLS EN $\lambda = 0.5$ (Average MSE=0.2988637) and the OLS Ridge (Average MSE=0.3364880). In this case, we will use the OLS Lasso as the learner in order to predict in the whole sample the value of $log(\hat{W}age)$.

Table 3: Mean Squared Error (MSE) for empirical lambda of Heckman for Probit Lasso

	Split 1	Split 2	Split 3	Split 4	Split 5	Average
OLS	0.5381882	0.3962795	2.9544220	1.6525162	12.9141052	3.6911022
OLS Lasso	0.2910276	0.2086963	0.3533088	0.2654259	0.3374382	0.2911794
OLS Ridge	0.3359037	0.2487646	0.3998108	0.2820733	0.3992088	0.3331522
OLS Elastic Net	0.3321021	0.2106721	0.3525510	0.2660722	0.3577490	0.3038293
Decision Tree	0.4954084	0.3453911	0.5388070	0.3587684	0.4661048	0.4408959
Random Forest	0.3884593	0.2851422	0.4301835	0.3545152	0.4043242	0.3725249
Neural Networks	0.4404376	0.4065424	0.5671050	0.4581342	0.4830628	0.4710564

In our case different that in Paliouras & Jessen (1999) we do not have the NN learner as the one that outperforms in out-of-sample prediction. Even though, the NN does well, in this case in which we use the empirical lambda of Heckman for the predicted values of the Probit Lasso the learner that best works out-of-sample is the OLS EN $\lambda = 0.5$ (Average MSE=0.3038293) even if in this case OLS Lasso works well in this environment (Average MSE=0.2911794), we will use the version in which $\lambda = 0.5$ to compute $log(\hat{W}age)$ for the whole sample, in order to account for robustness in our results depending on the selection of λ .

After computing this regressions we will use the predicted values of this regression in order to compute the labour supply equation. As stated in Weerasooriya (2018) in its paper Mroz

(1987) suspects that log(Wage) will be endogenous in equation (3), the structural equation. Then, the Reduced Form equation, equation (2) for this porpoise will solve the endogeneity of the structural equation (3) and also we will solve the Sample Selection issue by computing introduction as covariates in equation (3) the empirical Inverse Mill's ratio.

Table 4:

	Dependent variable:	
	(1)	(2)
lwage_hat_1	1,319.401***	
	(92.761)	
lwage_hat_2		1,343.230***
		(97.034)
nwifeinc	−10.035***	−10.657***
	(2.984)	(3.017)
age	−7.014	−7.990*
	(4.429)	(4.471)
educ	−149.149***	−151.618***
	(16.198)	(16.521)
kids	−63.637***	−63.968***
	(24.424)	(24.634)
lambda_logit_lasso	5,992.183	
	(4,460.296)	
lambda_probit_lasso		2,723.176
		(1,951.494)
Constant	2,164.291***	2,218.248***
	(275.167)	(277.495)
Note:	*p<0.1; **p<0.05; ***p<0.01	

The following outputs are the two structural equations that we have computed. We can see that the outcome is sign invariant with respect to the empirical lambda that we use and with the learner used to compute $log(\hat{W}age)$.

References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer.

Mroz, T. A. (1984). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. Stanford University.

Paliouras, G., Jessen, H. C., & Athens, U. K. (1999). Statistical and learning approaches to nonlinear modeling of labour force participation. Neural Network World, 9, 341-364.

Weerasooriya, T. R. (2018). High-dimensional sample selection models: Machine-learning algorithms in the Heckman two-step.