Causal Inference Homework #4

Andrew Bennett

March 15, 2024

March 15, 2024

1 Part 1

Estimate Model (1) implementing the Heckman correction method (two-stage approach) on the selection process described in Equation (4). Include, in the vector zi, education of the spouse and all regressors included in xi. As your out-of-work income variable bi you can use either benefits or its interaction with marital status.

Solution

Table 1: Probit Regression Results

Dep. Variable:	lfp	**No. Observations:**	846295
Model:	Probit	**Df Residuals:**	846288
Method:	MLE	**Df Model:**	6
Date:	Fri, 15 Mar 2024	**Pseudo R-squ.:**	inf
Time:	12:50:32	**Log-Likelihood:**	-2.2378e-10
converged:	False	**LL-Null:**	0.0000
Covariance Type:	nonrobust	**LLR p-value:**	1.000

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
education_spouse	0.0092	2593.423	3.54e-06	1.000	-5083.006	5083.024
const	3.2767	nan	nan	nan	nan	nan
education	0.0083	3217.677	2.58e-06	1.000	-6306.522	6306.539
age	0.0241	1529.945	1.58e-05	1.000	-2998.614	2998.662
region	0.0028	3653.781	7.66e-07	1.000	-7161.277	7161.282
benefits	0.0553	2169.377	2.55e-05	1.000	-4251.845	4251.955
married	3.2767	nan	nan	nan	nan	nan

Complete Separation: The results show that there is complete separation or perfect prediction. In this case, the Maximum Likelihood Estimator does not exist and the parameters are not identified.

Given the results of the probit, there are strongly correlated variables in the model. Specifically, the benefits seem to have the strongest effect on labor force participation. This is not surprising, as the benefits are the only variable that is directly related to the decision to work or not. The other variables are more related to the level of earnings.

The education coefficient is quite significant for the outcome variable of the decision to work also. This is consistent with the literature, as education is a strong predictor of going to work. Interestingly, age is also negatively correlated with earnings. This is consistent with literature as well, considering higher levels of age are likely to age out of the working force.

Table 2: OLS Regression Results

Dep. Variable:	ln_earn	**R-squared:**	0.457
Model:	OLS	**Adj. R-squared:**	0.457
Method:	Least Squares	**F-statistic:**	1.426e+05
Date:	Fri, 15 Mar 2024	**Prob (F-statistic):**	0.00
Time:	12:50:32	**Log-Likelihood:**	-5.4570e+05
No. Observations:	846295	**AIC:**	1.091e+06
Df Residuals:	846289	**BIC:**	1.091e+06
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	3.431e+08	1.52e+08	2.255	0.024	4.49e+07	6.41e+08
education	0.0640	0.000	360.891	0.000	0.064	0.064
age	0.0059	6.02e-05	97.633	0.000	0.006	0.006
region	0.0238	0.000	120.160	0.000	0.023	0.024
benefits	0.5438	0.001	682.186	0.000	0.542	0.545
married	-3.977e+08	1.76e+08	-2.255	0.024	-7.43e+08	-5.21e+07
lambda	1.899e+08	8.42e+07	2.255	0.024	2.49e+07	3.55e+08

Notes:

- Standard Errors assume that the covariance matrix of the errors is correctly specified.
- The smallest eigenvalue is 3.19×10^{-25}. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

The lambda here is highly significant and a great sign, as it shows that there is a selection bias in the model. benefits are again the most significant variable in terms of earnings contribution. Education and region are also significant, which could be attributed to some cities having higher wages than other cities. The age variable is again negatively correlated with the outcome variable which potentially suggests that younger salaries are lower and older salaries are higher. Given the potential multicollinearity in the model, there may also be some overlap between the age and benefits variables.

2 Part 2

Re-estimate the Heckman two-stage model using Matlab/Python without any pre-programmed routines.

Solution

The Heckman two stage model starts by estimating a probit model which gives the likelihood of the model being a part of the correct classification.

Then, the second stage is the estimation of the model using the inverse mills ratio as a regressor.

The inverse mills ratio is calculated as follows:

$$\lambda = \frac{\phi(\rho \frac{z_i}{\sigma})}{1 - \Phi(\rho \frac{z_i}{\sigma})}$$

ϕ is the standard normal density function

Φ is the standard normal distribution function.

ρ is the correlation between the error terms of the two equations.

z_i is the vector of variables that are used to predict the probability of being in the correct classification.

σ is the standard deviation of the error term in the first stage.

3 Part 3

Estimate now the selection model by Maximum Likelihood using Matlab/Python without any pre-programmed routines (except Newton-Rhapson). You can instead use the ml Stata command. Use the results from Point 2.1 as starting values.

Here we do the same approach as before of estimating the probit model and then regressing to find coefficients. However, we are now using a Hesian (2nd order derivative) to find the coefficients and optimizing the likelihood function.

Solution

Here we do the same approach as before of estimating the probit model and then regressing to find coefficients. However, we are now using a Hesian (2nd order derivative) to find the coefficients and optimizing the likelihood function.

Below are the tables from the outcomes of the model. For the related code, please see the notebook.

	params	std_errors	t_stats	p_values
education_spouse	0.028297	0.000214	132.494431	0.000000
const	-0.670091	276107.398318	-0.000002	0.999998
education	0.064862	0.000240	270.283542	0.000000
age	-0.015114	0.000075	-201.029688	0.000000
region	0.003236	0.000264	12.270342	0.000000
benefits	0.275063	0.000283	972.125684	0.000000
married	-0.670091	276107.398318	-0.000002	0.999998

4 Part 4

Estimate Model (1) by OLS on the sample of individuals that work. Compare annual wage predictions with those from the selection model you can chose whether to use the predictions from the two stage approach or those from the ML estimation). Plot your estimates and interpret them making reference to the statistics provided by Panel (b) in Figure 1. What is the role for selection empirically?

Solution

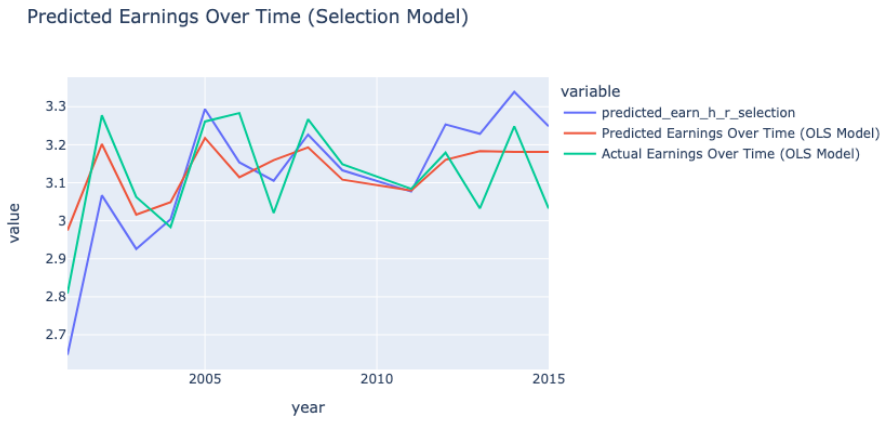
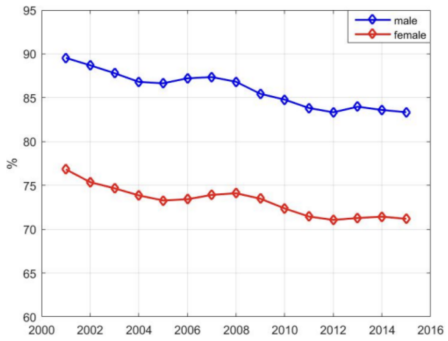


Figure 1: Predicted Earnings for each Year



(b) Labor Force Participation

Figure 2: Figure 2b from the Assignment

In this scenario, the selection model performs slightly better than the OLS model on average, especially towards the more recent years. This is interesting as it shows that as labor force participation increases, the selection bias becomes stronger. This may be due to the high correlation between the benefits and the decision to work. The selection model also seems to follow upward and downward trends better than the OLS model indicating a higher sensitivity.