

Note: See the accompanying do file for the relevant commands and methods.

1. Using the Wooldridge dataset MROZ.RAW, please construct a sample selection model to estimate the effect of education and age on wages. Please code education as a set of dummies for educational attainment (ie HS dropout, HS, some college, BS/BA, more). Use as instruments the family marginal tax rate, the number of kids at home --- both small and big, and the family income without the wife's income (nwifeinc).

- a. Test whether there is selection bias

The 95% confidence interval for rho, the bias parameter, is (-0.35, -0.02), which does not include zero. Thus it appears that there is statistically significant evidence of selection bias.

- b. Test (separately, one at a time) whether the instruments belong in the wage equation

Re-estimating the entire selection model, each time including one of the instruments in the outcome equation, we find that mtr and nwifeinc are significant, while neither kids variable is. Thus, it appears that the kids variables are the only valid instruments.

- c. Construct a test of the hypothesis that education should be entered as dummies rather than linearly as I did in class.

An easy way to test this is to generate dummy variables for each of the 13 levels of education, and include all but one of them in the wage equation. Then the relevant hypothesis is that $\delta_2 = \delta_3 = \dots = \delta_{13} = 0$. If this were true, then the effect of moving from 2 to 3, 3, to 4, ..., 12 to 13 years of education would be the same as moving from 1 to 2 (the omitted category). The p-value associated with this test is essentially zero, so we reject the hypothesis that a linear effect is correct.

- d. Evaluate the derivative of notional wages (ie the wages women would receive were they to work) in age. Estimate and 95% CI.

Here what we want to know is $\partial E[\text{wage}|x]/\partial \text{age} = x\beta_{\text{age}}^H$ where β_{age}^H is the Heckman coefficient on age. The coefficient is 0.025. The 95% CI is (-0.013, 0.064). This is different that the derivative of $E[\text{wage}|x, \text{inlf} = 1]$ with respect to age. The latter expectation is what we would expect wages to be if the woman worked and we observed her selecting into the labor force, which isn't what we want.

Note that I am using the model from part (a), even though it looks like some of the instruments are bad and a linear specification for education is inappropriate. This will be true for the rest of the assignment unless otherwise noted.

- e. Evaluate the derivative of observed wages (ie counting the non-workers as zero wage) in age. Estimate and 95% CI.

Observed wages are

$$wage_i^o = \begin{cases} wage_i & \text{if } inlf_i = 1 \\ 0 & \text{if } inlf_i = 0 \end{cases} .$$

What we want to know is $\partial E[wage^o|x]/\partial age$. The expected observed wage is equal to

$$\begin{aligned} E[wage^o|x] &= E[wage|x, inlf = 1] * \Pr(inlf = 1|z) \\ &= [x\beta^H + \rho\lambda(z\gamma)]\Phi(z\gamma). \end{aligned}$$

The derivative of this with respect to age is

$$\begin{aligned} \frac{\partial E[wage^o|x]}{\partial age} &= \frac{\partial E[wage|x, inlf = 1]}{\partial age} \Pr(inlf = 1|z) \\ &+ \frac{\partial \Pr(inlf = 1|z)}{\partial age} E[wage|x, inlf = 1]. \end{aligned}$$

We can compute all of these terms using the coefficients from the Heckman model. However, we will have a couple of problems. First, we will need to use \bar{x} instead of x , since some of the terms in the above equation will vary over individuals. Second, to compute confidence intervals, we will need to take derivatives of the above equation with respect to both β^H and γ^H (the coefficients from the selection equation), and then compute the confidence intervals using the Delta method. This can be implemented in stata using the command `mf compute, predict(yexpected)`. Using this approach, we find that the estimated effect is -0.029, with CI (-0.063, 0.005).

An easier route is to assume that $E[w^o|x] = x\delta$ and estimate δ by OLS regression of w^o on x . Since we observe w^o for everyone in the sample, we don't have a selection problem here like we do when we try to estimate the parameters of the wage offer equation, $E[w|x] = x\beta$. Using this approach, we find that the OLS coefficient is about 0.002 with CI (-.026, 0.029). Both coefficients are lower than the Heckman coefficient.

- f. Why are the two different?

These coefficients are different because the first (the Heckman coefficient) tells us how age would affect wages if everyone in the sample worked, while the second (OLS) tells us how age affects (i) the wages of those that do work and (ii) the propensity of individuals to work.

This makes sense in the context of our example. Normally, with age, people gain experience and thus earn higher wages. But they also become more likely to attrite from the labor force (i.e. retire). So we expected the effect on observed wages to be smaller than the effect on notional wages.

2. Estimate the model from the previous question using the two-step estimator (manually --- do not use the heckman command). It turns out that the standard errors returned by OLS for the second stage regression are wrong. They are wrong because we can't use the theoretically correct inverse mills ratio (based on the true coefficients from the probit step), rather we use an inverse mills ratio based on estimated betas from the probit step.

- a. Use bootstrapping to obtain correct 95% confidence intervals for the coefficients in the wage equation.

Since using $\hat{\lambda} = \lambda(z\hat{\gamma})$ instead of λ itself introduces extra sampling variability, our bootstrap needs to estimate lambda and then estimate the wage equation in each replication. Doing so, we get

```
Bootstrap statistics                                Number of obs   =    753
                                                    Replications    =   1000
```

Variable	Reps	Observed	Bias	Std. Err.	[95% Conf. Interval]		
b_educ	1000	.4474207	.0009533	.0732358	.303707	.5911345	(N)
					.3086764	.5900798	(P)
					.3097827	.5929164	(BC)
b_age	1000	.0280948	-.0002501	.0214005	-.0139003	.0700899	(N)
					-.0140634	.0707432	(P)
					-.0143145	.0705455	(BC)
b_lambda1	1000	-.9502407	-.0819551	.4415112	-1.816636	-.083845	(N)
					-1.948731	-.2737545	(P)
					-1.823382	-.1782339	(BC)
b_cons	1000	-2.154223	.0415462	1.188219	-4.485915	.1774682	(N)
					-4.327573	.068655	(P)
					-4.483443	.0351784	(BC)

Note: N = normal
P = percentile
BC = bias-corrected

- b. Does it look like there is selection bias?

In either case, the confidence intervals for the coefficient on lambda do not contain zero, so there does appear to be evidence of selection bias.

- c. Use bootstrapping to evaluate the effect of college completion on expected observed (ie including the 0s for non-participation) wages. Estimate and 95% confidence interval.

Using bootstrapped standard errors to estimate the asymptotic t-statistic:

wage	Observed Coef.	Bootstrap Std. Err.	z	P> z	Normal-based [95% Conf. Interval]	
educ	.4532833	.0554388	8.18	0.000	.3446252	.5619413
age	.0015147	.0139917	0.11	0.914	-.0259086	.0289379
_cons	-3.25929	.8798548	-3.70	0.000	-4.983774	-1.534807

Using the percentile method:

wage	Observed Coef.	Bias	Bootstrap Std. Err.	[95% Conf. Interval]		
educ	.45328328	-.0012745	.05543879	.3515864	.5641786	(P)
age	.00151467	.0001264	.0139917	-.0253534	.0300875	(P)
_cons	-3.2592904	.0054968	.87985481	-4.999221	-1.542196	(P)

3. Now estimate the model using Tobit.
a. What effects does this have on your estimates of the value of education?

The resulting coefficient estimate of education is 0.694. This is much larger than in the Heckman models.

- b. Devise and implement a statistical test of which model, Tobit or Heckman's Sample Selection, is the correct one.

Under the Tobit model, individuals select into working according to the same latent index that determines the wage offer. Under the Heckman model, the coefficients on the selection equation can be different, and there are instruments. Thus, if the Tobit model were correct, the ratios of the coefficients β^H / σ would be the same as the coefficients in the selection equation and the coefficients on the instruments would be zero. We can test these joint restrictions in Stata. Doing so, we find that the probability of the test statistic under the null hypothesis is very close to zero. This suggests that the Heckman model is a better specification.

Here is another possible specification test. The latent index in the tobit model is the same latent index from a probit model of $\ln f_i$ on x_i . Thus, if the tobit specification is correct, the ratios $\beta^{\text{Tobit}} / \sigma^{\text{Tobit}}$ should be close to β^{Probit} . A finding that the coefficient from the probit and tobit models differ greatly indicates that the tobit is a misspecification.