PS5: The Roy Model

Alex Jenni

18 April 2019

Problem 1: Generalized Roy Model: The union/non-union wage differential

This exercise is based on an article by Lee (1978). In Lee's model, every worker has two different potential wages, depending on their union membership status

$$\ln w_i^U = x_i'\beta^U + u_i^U, \quad u_i^U \sim Normal(0, \sigma_U^2)$$
$$\ln w_i^N = x_i'\beta^N + u_i^N, \quad u_i^N \sim Normal(0, \sigma_N^2).$$

A worker joins the union if

$$U_i^* = \delta_0 + \delta_1(\ln w_i^U - \ln w_i^N) + x_i'\delta_2 + z_i'\delta_3 - v_i > 0$$

and v_i is assumed to be distributed as $Normal(0, \sigma_v^2)$. For simplicity, we assume that the error terms are **independent**[1] from each other **and from the regressors** (x_i, z_i). For every worker, only $(\ln w_i, x_i, z_i, d_i)$ is observed where d_i is a union membership indicator and $\ln w_i = d_i \ln w_i^U + (1 - d_i) \ln w_i^N$.

a. What is the expected union/non-union log wage differential for a randomly chosen individual with characteristics x_i? For a union worker?

Since the error terms are mean zero, the expected **log** wage differential for a randomly chosen individual with characteristics x_i is

$$E[\ln w_i^U - \ln w_i^N | x_i] = x_i'(\beta^U - \beta^N).$$

Getting the expected log wage differential for a randomly selected union worker with characteristics x_i **and** z_i is more involved because of the selection into union membership. After plugging in the wage equations, the selection equation becomes

$$U_i^* = \delta_0 + \delta_1 x_i'(\beta^U - \beta^N) + x_i'\delta_2 + z_i'\delta_3 - \underbrace{(v_i + \delta_1(u_i^N - u_i^U))}_{\epsilon_i} > 0.$$

Because the error terms are independent and normally distributed, we obtain that

$$\begin{aligned} Pr(U_i^* > 0 | x_i, z_i) &= Pr(\epsilon_i < \delta_0 + \delta_1 x_i'(\beta^U - \beta^N) + x_i'\delta_2 + z_i'\delta_3 | x_i, z_i) \\ &= \Phi\left(\frac{\delta_0 + x_i'\left(\delta_1(\beta^U - \beta^N) + \delta_2\right) + z_i'\delta_3}{\sigma_\epsilon}\right), \end{aligned} \tag{1}$$

where $\sigma_\epsilon^2 = \sigma_v^2 + \delta_1^2(\sigma_U^2 + \sigma_N^2)$.

For a union worker with traits x_i and z_i, the expected **log** wage differential is

[1] Alternatively, joint normality and zero correlation between the error terms would also do the job (since this is the result that we need in the end).

$$E[\ln w_i^U - \ln w_i^N | x_i, z_i, d_i = 1] = x_i'(\beta^U - \beta^N) + E[u_i^U - u_i^N | x_i, z_i, U_i^* > 0]. \tag{2}$$

By virtue of independence of the error terms, $E[u_i^j|\epsilon_i] = E[u_i^j|u_i^j] = u_i^j$ for $j \in \{U, N\}$. We can thus rewrite

$$u_i^j = \gamma_j \epsilon_i + \xi_{ij},$$

where

$$\gamma_N = \frac{Cov(\epsilon_i, u_i^N)}{Var(\epsilon_i)} = \delta_1 \frac{\sigma_N^2}{\sigma_\epsilon^2}, \quad \gamma_U = -\delta_1 \frac{\sigma_U^2}{\sigma_\epsilon^2}, \text{ and } E[\xi_{ij}|\epsilon_i] = 0, \quad j \in \{U, N\}.$$

Hence the expected log wage differential for a union worker is

$$\begin{aligned} E[\ln w_i^U - \ln w_i^N | x_i, z_i, d_i = 1] =& x_i'(\beta^U - \beta^N) + (\gamma_U - \gamma_N) E[\epsilon_i | x_i, z_i, \epsilon_i < \delta_0 + \delta_1 x_i'(\beta^U - \beta^N) + x_i'\delta_2 + z_i'\delta_3] \\ =& x_i'(\beta^U - \beta^N) + (\gamma_U - \gamma_N)\sigma_\epsilon E\left[\frac{\epsilon_i}{\sigma_\epsilon} | x_i, z_i, \frac{\epsilon_i}{\sigma_\epsilon} < c_i\right] \\ =& x_i'(\beta^U - \beta^N) + \delta_1 \frac{\sigma_U^2 + \sigma_N^2}{\sigma_\epsilon} \lambda(-c_i) \end{aligned} \tag{3}$$

where $c_i = (\delta_0 + x_i'\left(\delta_1(\beta^U - \beta^N) + \delta_2\right) + z_i'\delta_3)/\sigma_\epsilon$ and $\lambda(x) = \phi(x)/(1 - \Phi(x))$ is the inverse Mills ratio[2].

b. Is it possible to obtain an unbiased estimate of β^U and β^N by running two linear regressions using the sample of union members and non-members respectively?

We would generally not obtain an unbiased estimate of β^U and β^N if we ignore the selection of workers into union membership. As we have seen in part **a**, due to self-selection into union membership, the wage of union members does not depend on x only through the effect of x on the wage but also through the effect of x on the decision to join a union (i.e. c is a function of x). If x and z are not independent, the naive OLS estimate is also biased by the effect of z on self-selection

c. Sketch how both equations can be estimated jointly by maximum likelihood?

Given our assumptions on the functional form of the structural wage equations and the union status equation and about the distribution of the error terms, we have a fully parameterized model of the conditional distribution of wages and union status. The conditional likelihood of observing a worker with log wage and union status $(\ln w, d)$ given covariates $(x_i, z_i)_{i=1}^n$ and parameters $\theta = \{\beta^U, \beta^N, \delta_0, \delta_1, \delta_2, \delta_3, \sigma_U^2, \sigma_N^2, \sigma_v^2\}$ is given by:

$$\begin{aligned} \mathcal{L}(\ln w, d | x_i, z_i, \theta) =& Pr(d|x_i, z_i, \theta) f(\ln w | d_i, x_i, z_i, \theta) \\ =& \left[\Phi\left(\frac{\delta_0 + x_i'\left(\delta_1(\beta^U - \beta^N) + \delta_2\right) + z_i'\delta_3}{\sqrt{\sigma_v^2 + \delta_1^2(\sigma_U^2 + \sigma_N^2)}}\right)\phi\left(\frac{\ln w_i - x_i'\beta^U}{\sigma_U}\right)\right]^{d_i} \times \\ & \left[\left(1 - \Phi\left(\frac{\delta_0 + x_i'\left(\delta_1(\beta^U - \beta^N) + \delta_2\right) + z_i'\delta_3}{\sqrt{\sigma_v^2 + \delta_1^2(\sigma_U^2 + \sigma_N^2)}}\right)\right)\phi\left(\frac{\ln w_i - x_i'\beta^N}{\sigma_N}\right)\right]^{1-d_i}. \end{aligned}$$

We cannot identify all the parameters of this model. But we could estimate another set of parameters $\tilde{\theta} = \{\beta^U, \beta^N, \gamma_0, \gamma_1, \gamma_2, \sigma_U^2, \sigma_N^2\}$ by maximizing the log likelihood of the sample:

[2]For a standard normal variable z, it is known that $E[z|z < a] = -\lambda(-a)$.

$$\max_{\tilde{\theta}\in\mathbb{R}^5\times\mathbb{R}^2_+} N^{-1}\sum_{i=1}^{N} d_i\left[\ln\Phi\left(\gamma_0+x_i'\gamma_1+z_i'\gamma_2\right)+\ln\phi\left(\frac{\ln w_i-x_i'\beta^U}{\sigma_U}\right)\right]+$$
$$(1-d_i)\left[\ln\Phi\left(\gamma_0+x_i'\gamma_1+z_i'\gamma_2\right)+\ln\phi\left(\frac{\ln w_i-x_i'\beta^N}{\sigma_N}\right)\right].$$

This is probably not easy computationally speaking and I am not sure that it allows to identify the structural parameter of interest δ_1.

d. Describe a two-step method which estimates β^U and β^N consistently by including estimated sample-selection correction variables in the structural wage equations.

i) Run a Probit regression of union status d on a constant and regressors x and z: $Pr(d=1|x,z)=\Phi(\gamma_0+x'\gamma_1+z'\gamma_2)$. Use the estimates from the model to compute $\hat{c}_i=\hat{\gamma}_0+x_i'\hat{\gamma}_1+z_i'\hat{\gamma}_2$.
ii) Run two separate OLS regressions by union status of log wage on a constant and regressors x and the appropriate correction variable, i.e. $\lambda(-\hat{c})$ for union members and $\lambda(\hat{c})$ for non-union members.

e. How can the structural parameter of the union status equation, i.e. δ, be recovered after this estimation procedure?

We can use our estimates from the structural wage equations to predict the potential log wages $\hat{\ln}(w_i^U)$ and $\hat{\ln}(w_i^N)$. We then estimate our selection equation (the probit model) again but include the estimated log potential wages as additional explanatory variables. Because the predicted log wages are a non-linear combination of x and z and we controlled for sample-selection, we identify δ up to a normalization constant[3]. Using this procedure, we are able to back up the marginal effect of the log wage differential on the probability to join a union for different values of the covariates.

f. Is the relative rate of return to education higher in the unionized or the non-unionized sector (Tables 1 and 2)? What about the effects of market experience (ME), female, black and health impediments on wages?

The relative rate of return to education is higher in the non-unionized sector: the predicted wage increases by about 33% going from the lowest to the highest education level in the non-unionized sector and by about 25% in the unionized sector. This is almost entirely driven by the difference in return to going from the second-highest to the highest level of education. Market experience has a larger effect in the unionized sector. Females have a larger relative wage gap in the unionized sector. Blacks have a smaller relative wage gap in the unionized sector. The effect of health impediments on relative wages is larger in the non-unionized sector.

g. Interpret the sign of the selectivity variables (Tables 1 and 2)?

In Table 1, the selectivity variable is $-\lambda(-\hat{c})$. In Table 2, the selectivity variable is $\lambda(\hat{c})$. With the independence assumption on the error terms, the coefficients on the selectivity variables equal $-\delta_1\sigma_U^2/\sigma_\epsilon$ in the union wage equation and $\delta_1\sigma_N^2/\sigma_\epsilon$ in the nonunion wage equation. Hence, we expect the two coefficients to be of opposite signs. In fact, because the variance and standard-deviation terms are always positive, the signs in tables 1 and 2 tell us that δ_1 is positive. In words, the probability to be a union member increases in the relative union/non-union wage differential.

[3]The issue is that, as in any Probit model, we can only identify the parameters up to a constant.

h. Judging from the presented evidence, how important would you say is the wage differential in explaining the probability of union membership?

The log wage differential appears to be the most powerful factor in explaining the probability of union membership (Table 6).

i. Compare Table 6 and Table 7. How can the reduced form estimates be interpreted?

We see that the coefficients for the education variables are lower in Table 7 than in Table 6. The coefficients in Table 7 combine the direct effects of the education level and the indirect effects through the impact of the education level on the wage differential. Since the coefficients are larger in Table 6 when the wage differential is controlled for, we can conclude that the education level is negatively correlated with the log wage differential.

Problem 2: Application: Female labor supply using Altonji-Elder-Taber Bounds

We use the dat set **mroz.dta** (Mroz 1987) to assess the selection into labor force **inlf** and motherhood **kidslt6**.

a. Run two separate probit regression of inlf and kidslt6 on nwifeinc, educ, exper, expersq and age.

For all of this problem, see the Stata code attached.

b. Run a bivariate probit. What is the advantage of estimating the models jointly?

The bivariate probit specification is asymptotically more efficient if the error terms are correlated across the equations. By estimating two separate probit regressions, we are throwing away useful information. This is similar to the motivation behind the seemingly unrelated equations (SUR) model. In addition, we might also be interested in knowing the correlation of the unobservable idiosyncratic factors leading women to select into the labor force and motherhood. In this case, the estimate of this correlation $\hat{\rho}$ equals -0.634. Unobservables that lead women to select into the labor force also tend to lead them not to have two kids or more.

c. Now suppose that selection into the labor force depends on the selection into motherhood.

If selection into the labor force depends on the selection into motherhood, kidslt6 must be included as an explanatory variable in the labor force equation

$$\begin{aligned} Pr(l_i = 1|k_i, x_i) =& Pr(x_i'\beta + k_i\delta > \epsilon_i^l) = \Phi(x_i'\beta + k_i\delta) \\ Pr(k_i = 1|x_i) =& Pr(x_i'\gamma > \epsilon_i^k) = \Phi(x_i'\gamma). \end{aligned}$$

We are making the assumption that ϵ_i^k and ϵ_i^l are jointly normally distributed given x_i. We can actually ignore the simultaneity in the simultaneous bivariate probit model (see Greene (2003), chapter 23). However, a hidden assumption here is that the selection into motherhood does not depend on the selection into the labor force.

The estimated correlation $\hat{\rho}$ between the error terms is now 0.668. The coefficient on motherhood in the labor force participation equation is negative. Hence, once we account for the effects of motherhood on labor force participation, the unobserved idiosyncratic factors ϵ_i^k and ϵ_i^l are actually positively correlated.

d. Based on the reasoning of Altonji, Elder, and Taber (2008), explain briefly how to assess the selection on unobservables.

Altonji, Elder, and Taber (2008) point out that ρ is only parametrically identified because there is one more parameter to estimate than the number of regressors. The coefficient estimates might be highly sensitive to the parametric specification. Hence, it is informative to see which values the estimates would take for alternative reasonable values of ρ.

Altonji et al. suggest to bound the selection on unobservables ρ using the selection on observables and additional assumptions. For example, if we believe that the selection on unobservables is at most as large as the selection on observables, we can use the correlation between the observable component of the latent variable determining motherhood $x'\beta$ and the part of the latent variable determining labor force participation that is determined by the same regressors $x'\gamma$ to upper bound the absolute selection on unobservables: $|\rho| \leq \rho_{max} = cov(x'\beta, x'\gamma)/var(x'\gamma)$. Practically, we need first to estimate the coefficients β and γ to derive ρ_{max} (and could possibly iterate over this procedure, see the code).

e. Find the value of ρ that eliminates the effect of motherhood on labor force participation

This is exactly the value of ρ when we exclude kidslt6 from the list of regressors in the labor force participation equation (-0.634).

f. In light of this article, is this value plausible? To answer this question, relate selection on unobservables to selection on observables.

Implementing the procedure described in **d**, I estimate that if the selection on unobservables is as large as the selection on observables then $\rho = 0.175$. If we are willing to assume that the selection on unobservables goes in the same direction as the selection on observables, then $\rho = -0.634$ is totally implausible.

g. Using your estimates from f., compute the lower bound effect of having more than one child on labor force participation.

Again, if we are willing to assume that the selection on unobservables is bounded between 0 and the selection on observables 0.175, then $\beta_{kidslt6}$ lies in the interval $[-1.787, -1.481]$. I am not sure about the validity of this assumption given the value of $\hat{\rho} = 0.668$ that we find when ρ is directly estimated by joint maximum likelihood (which corresponds to a coefficient $\hat{\beta}_{kidslt6}$ of -2.528).

References

Altonji, Joseph G, Todd E Elder, and Christopher R Taber. 2008. "Using Selection on Observed Variables to Assess Bias from Unobservables When Evaluating Swan-Ganz Catheterization." *American Economic Review* 98 (2): 345–50.

Greene, William H. 2003. *Econometric Analysis*. Pearson Education India.

Lee, Lung-Fei. 1978. "Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables." *International Economic Review*. JSTOR, 415–33.

Mroz, Thomas A. 1987. "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions." *Econometrica*. JSTOR, 765–99.