ECON 211B: Homework 1

Due: Monday, January 23, 2017

Carlos E. Dobkin

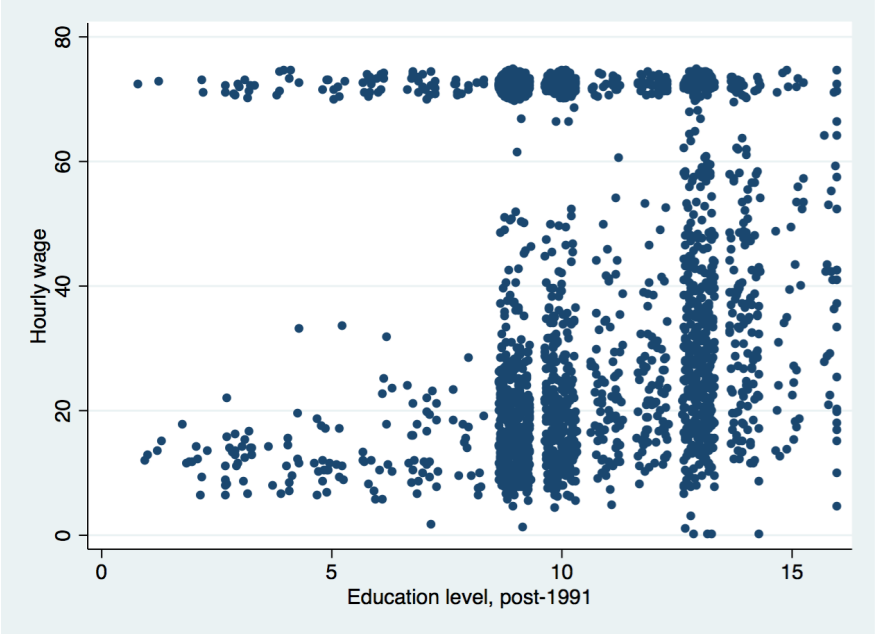
David Sungho Park

Problem 1

$$
\begin{aligned}
E_X[E[Y_i|X_i]] &= E_X\left[\sum_u uP(Y_i = u|X_i)\right] \\
&= \sum_t \left[\sum_u uP(Y_i = u|X_i = t)\right]P(X_i = t) \\
&= \sum_u u\left[\sum_t P(Y_i = u|X_i = t)P(X_i = t)\right] \\
&= \sum_u u\left[\sum_t P(Y_i = u, X_i = t)\right] \\
&= \sum_u uP(Y_i = u) \\
&= E[Y_i].
\end{aligned}
$$

Problem 2

Figure 1: Wage and educational attainment of men in the US between 30 and 40



Source: Current Population Survey (CPS), January 2014.
Note: Hourly wage represents *wage4* from dataset.

Problem 3

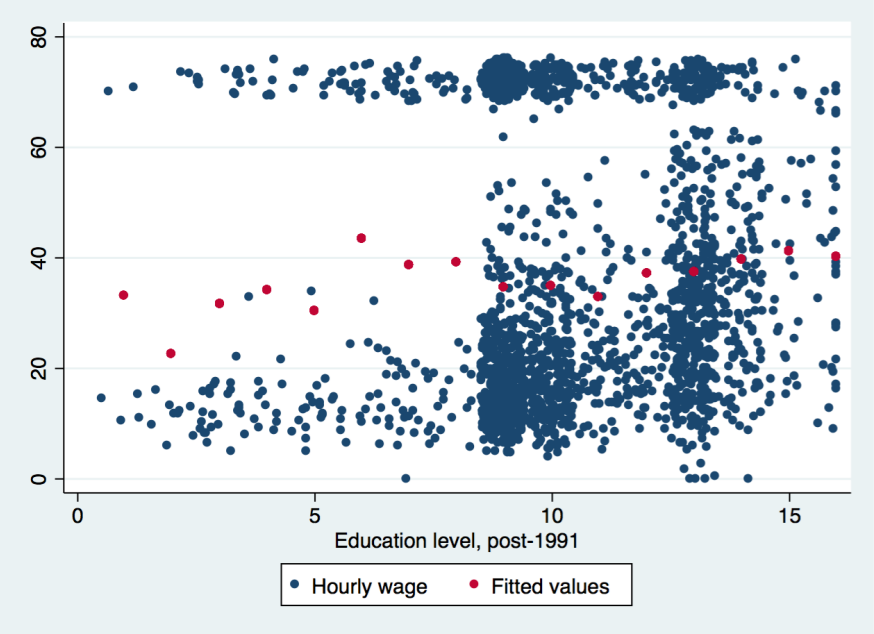
The conditional expected function (CEF) can be estimated by OLS regression on the earnings equation with only the dummy variables on the right hand side.

Problem 3 continued on next page...

VARIABLES	(1) wage4
educ_1	33.06***
	(9.776)
educ_2	22.53***
	(7.220)
educ_3	31.46***
	(4.107)
educ_4	33.94***
	(5.354)
educ_5	30.19***
	(4.608)
educ_6	43.42***
	(4.233)
educ_7	38.50***
	(3.570)
educ_8	39.16***
	(4.696)
educ_9	34.58***
	(0.944)
educ_10	34.85***
	(1.235)
educ_11	32.80***
	(2.395)
educ_12	37.12***
	(2.294)
educ_13	37.37***
	(1.086)
educ_14	39.69***
	(1.949)
educ_15	41.02***
	(4.168)
educ_16	40.12***
	(4.048)
Observations	2,135
R-squared	0.695

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 2: Fitted values for wage and educational attainment of men in the US between 30 and 40



Source: Current Population Survey (CPS), January 2014.
Note: Hourly wage represents *wage4* from dataset.

Problem 4

(a)

There are two general cases in which the CEF is linear. The first is joint normality of Y_i and X_i's, and the second is when we use a saturated model, which is our case here. By definition, saturated regression models have discrete explanatory variables and have one parameter for every possible j in $E[Y_i|s_i = j]$. On the right hand side of our equation, we have 16 dummy variables each representing a level of educational attainment. Thus our case makes the CEF inherently linear.

(b)

We need to show that

$$E[Y_i|X_i] = X_i'\beta \quad \Rightarrow \quad \hat{\beta} = E[X_iX_i']^{-1}E[X_iY_i] = \beta.$$

Proof.

$$\begin{aligned} E[X_iX_i']^{-1}E[X_iY_i] &= E[X_iX_i']^{-1}E\big[E[X_iY_i|X_i]\big] \qquad \text{(by LIE)} \\ &= E[X_iX_i']^{-1}E\big[E[X_i|X_i]E[Y_i|X_i]\big] \\ &= E[X_iX_i']^{-1}E\big[X_iX_i'\beta\big] \qquad \text{(by linearity of CEF)} \\ &= \beta. \end{aligned}$$

□

Problem 5

Using the Rubin's Causal Model, the setting is as follows.

- Treatment:

  $$\text{college attendance vs. high school completion.}$$

  $\Rightarrow$ In our model from Problem 2 to 4, the treatment group includes the groups corresponding to *educ*_10 to *educ*_16 and the control group is the rest of *educ*_1 to *educ*_9.

- Treatment indicator:

  $$D_i = \begin{cases} 1 & \text{if treated (i.e. } educ\_j = 1, \quad \forall j = 10, 11, ..., 16) \\ 0 & \text{if not treated (i.e. } educ\_j = 1, \quad \forall j = 1, 2, ..., 9). \end{cases}$$

- Outcome:

  $$Y_i = \text{hourly wage of individual } i.$$

- Potential outcome:

  $$Y_i(D_i) = \begin{cases} Y_i(1) = \text{hourly wage of individual } i \text{ if he or she received college education} \\ Y_i(0) = \text{hourly wage of individual } i \text{ if he or she received no college education.} \end{cases}$$

- Treatment effect:

  $$\tau_i = Y_i(1) - Y_i(0),$$

  which cannot be observed because only one of the potential outcomes can be realized. Therefore, we measure the *average treatment effect*:

  $$\bar{\tau_i} = E[\tau_i] = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)].$$

  However, we should note that from the actual observed data, we can only get

  $$E[Y_i|D_i = 1] \quad \text{and} \quad E[Y_i|D_i = 0].$$

  To estimate the average treatment effect with the observed date, we need two steps. We start by expressing the outcome as potential outcomes. For each individual i,

  $$Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0).$$

  Hence,

  $$\begin{aligned} E[Y_i|D_i = 1] = E[D_i Y_i(1) + (1 - D_i)Y_i(0)] &= E[Y_i(1)|D_i = 1] \\ E[Y_i|D_i = 0] = E[D_i Y_i(1) + (1 - D_i)Y_i(0)] &= E[Y_i(0)|D_i = 0] \end{aligned}$$

  For the second step, we need an ideal situation where the treatments are assigned randomly. Namely in such a randomized controlled trial (RCT), the potential outcomes are independent of the treatment

Problem 5 continued on next page...

indicator. Therefore,

$$E[Y_i(1)|D_i = 1] = E[Y_i(1)] \quad \text{and} \quad E[Y_i(0)|D_i = 0] = E[Y_i(0)].$$

Now, the average treatment effect can now be measured by the observed difference in hourly wage.

$$\begin{aligned} \bar{\tau}_i &= E[Y_i(1)] - E[Y_i(0)] \\ &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] \quad \text{(by randomization)} \\ &= E[Y_i|D_i = 1] - E[Y_i|D_i = 0]. \end{aligned}$$

However, in our case where the treatment of college education was not randomly assigned but chosen by each individual, the average treatment effect cannot be measured. Instead, the observed difference in hourly wage is

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] \\ &= \underbrace{E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1]}_{average\ treatment\ effect} + \underbrace{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]}_{selection\ bias}. \end{aligned}$$

In words, simply calculating the differences in the observed hourly wages of the people who went to college and those who didn't does not give us accurate information about the causal effect of college education on earnings. This is mainly because the observed difference contains the selection bias, which is the difference between the expected non-college potential incomes of those who received college education and of those who did not. That is, if the people who received college education had not decided to do so but their average income would have been still higher than those who did not go to college, then there is a positive selection bias.

Stata Codes

```
*Opening file
use "/Users/DSP/Dropbox/UCSC (2016- )/1stYear_2Q/211B/Homeworks/cepr_org_2014_hw1.dta",
replace

*Applying restrictions
keep if female==0 & age >=30 & age <=40 & month==1

*Overview of education variable
tab educ92

*A 98% Winsorization
sum wage4, detail
replace wage4 = r(p99) if wage4 > r(p99)
replace wage4 = r(p1) if wage4 < r(p1)

*Plotting earnings against education
scatter wage4 educ92, jitter(7) msize(small)
graph export "/Users/DSP/Dropbox/UCSC (2016- )/1stYear_2Q/211B/Homeworks/211b_hw1_fig1_scatter.png"

*Checking the label values of educ92
label list educ92

*Generating dummy variables
forval i=1/16 {
gen educ_‘i’ = 0
replace educ_‘i’ = 1 if educ92==‘i’
}

*Estimating CEF via OLS
reg wage4 educ_*, noconstant

*Saving the fitted values
predict fitted

*Exporting to tex
outreg2 using reg1_1, tex(fr)

*Plotting the fitted values
twoway (scatter wage4 educ92, jitter(7) msize(small))(scatter fitted educ92, msize(small)
mcolor(cranberry))
graph export "/Users/DSP/Dropbox/UCSC (2016- )/1stYear_2Q/211B/Homeworks/211b_hw1_fig2_fitted.png"
```