Causal Inference - Problem Set 2

Gustavo Baroni

8/26/2020

Question 1

Core concepts:

(a) What is an experiment, and how does it differ from an observational study?

An experiment is a method of data analysis with a control and a treatment group. Those groups are used in order to measure the impact of a particular phenomenon. On the other hand, in an observational study, there is nothing affecting the variables.

(b) What is "unobserved heterogeneity," and what are its consequences for the interpretation of correlations?

An unobserved heterogeneity is a term that suggests a difference between variables that were used (observed) in the model and variables that were not (unobserved). The existence of an unobserved heterogeneity implicates in a not accurate correlation: if there is an independent variable x_i that impacts the dependent y that is not in the model, the correlation in the model is erroneous.

Question 2

You are an economist hired as a consultant by an NGO to perform an impact evaluation agenda. The NGO "FUTURO" offers a training program to individuals that are currently employed or unemployed. The course is free of cost, one year-long and offered to all individuals in Hill Valley, who can choose to participate or not. The enrollment process is in every January, when a baseline questionnaire is applied. There are also surveys one, two and five years after the course.

FUTURO's CEO "Mr Manager", trying to persuade you to take the job, mentions how successful is the training program: "When entering the training program the average monthly wage among the employed students are \$1,000.00 and when they leave FUTURO this number raises to \$1,250.00".

(a) Why is this comparison not a good evaluation of the training program? How would you explain this (in words) to Mr Manager?

The comparison is not a good evaluation of the training program because it is not possible to evaluate how much money the same individual would make if he/she did not get in the program. It, basically, the *fundamental problem of causal inference.*

(b) How would you explain the problem raised on the previous item using the potential outcome notation?

I am defining S as salary, $Y_t(u)$ as potential salary when treated (there is use of training program), $Y_c(u)$ as potential salary when not treated (there is not use of training program), and u the unit of measure. As a consequence, it is possible to provide the potential outcome notation of the previous item as that:

$$S = Y_t(u) - Y_c(u)$$

As I have explained, we can not evaluate the potential salary if the individual would not have been treated ($Y_c(u)$).

(c) Using the potential outcome notation show how randomization solves the problem raised on previous item.

Randomizing the experiment allows us to estimate the treatment effects. Therefore, it is possible to measure the treatment effect by the difference in means between treatment and control groups:

$$S = Y_t(u) - Y_c(u)$$

The NGO decided to implement an randomized experiment. The results are available in the pset2.dta, where nr is the person identifier, educ, women, exper are respectively years of education, women indicator and years of experience at work, all of them constructed based on the baseline information (thus before the treatment). The variable treat is the treatment indicator and wage, lwage, emp are respectively the monthly wage, the log of wage and an indicator of employee status (equals 1 if person is employed). These three are constructed based on the survey realized one year after the course. For the next questions, you must use R and the code should be attached to the list resolution.

(d) Verify if the randomization was well done. In other words, if the control and treatment group are balanced.

It is possible to affirm that the control and treatment group are balanced since almost 50% of the population was treated (47.71 to be more precise).

(e) Regress lwage on educ, exper, expersq, women. Interpret the results.

The model is given as the following:

```r
library(stargazer)
regress <- lm(data = pset2, lwage ~ educ + exper + expersq + women)
stargazer(regress, type="text",
          title = "Regression of lwage",
          df=FALSE,
          digits = 5)
```

```
##
## Regression of lwage
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                lwage
## -----------------------------------------------
## educ                        0.09048***
##                              (0.01610)
##
## exper                       0.21104***
##                              (0.05111)
##
## expersq                     -0.01935***
##                              (0.00633)
##
## women                        -0.03744
##                              (0.04525)
##
## Constant                    5.81064***
##                              (0.23128)
##
## -----------------------------------------------
## Observations                    323
## R2                            0.13304
## Adjusted R2                   0.12214
## Residual Std. Error           0.39453
## F Statistic                 12.19987***
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

As we can see, educ, exper, and expersq have a statistical significance in the regression. Therefore, it is possible to suggest that years of education and experience at work have a positive impact in the log of wage (0.09 and 0.21 respectively). On the other hand, expersq and women have a negative impact in the dependent variable (-0.02 and -0.04 respectively), whereas women indicator is not statistically significant.

(f) Analyze the experiment's results using as outcomes wage and employment status.

```
exp.resul.wage <- lm(data = pset2, wage ~ educ + exper + expersq + women)
exp.resul.emp <- lm(data = pset2, emp ~ educ + exper + expersq + women)
stargazer(exp.resul.wage, exp.resul.emp,
          type="text",
          title = "Regressions of wage and employment status",
          df=FALSE,
          digits = 5)
```

```
##
## Regressions of wage and employment status
## =================================================
##                          Dependent variable:
##                     -----------------------------
##                          wage           emp
##                           (1)           (2)
## -------------------------------------------------
## educ                  137.87150***   0.03382**
##                        (25.89995)    (0.01500)
##
## exper                 321.63410***   0.12041***
##                        (82.23695)    (0.03946)
##
## expersq               -30.12478***   -0.01083**
##                        (10.18426)    (0.00442)
##
## women                   -54.82081     -0.04173
##                        (72.81128)    (0.04336)
##
## Constant               -650.35040*    -0.01433
##                        (372.15640)   (0.20799)
##
## -------------------------------------------------
## Observations               323           545
## R2                       0.12201       0.02783
## Adjusted R2              0.11097       0.02063
## Residual Std. Error    634.83950       0.48669
## F Statistic           11.04788***    3.86470***
## =================================================
## Note:                 *p<0.1; **p<0.05; ***p<0.01
```

As we can see, years of education have a positive - and statistically significance - impact on wage and on employment status (137.8715 and 0.0338 to be more precise). Also, we can affirm that experience of market has a positive - and statistically significance - impact on wage and on employment status (321.6341 and 0.1204 to be more precise). Besides that, experience of market squared has a negative impact on wage and on employment status (-30.1248 and -0.0108 to be more precise). Moreover, women indicator has a negative - and not statistically significance - impact on wage and on employment status (-54.8208 and -0.0417 to be more precise).

(g) Which assumptions must hold to interpret the previous item's results as the average treatment effect?

We must hold the treatment was well randomized. In terms of Average Treatment Effect (ATE), these are the assumptions that must be hold to interpret the previous item‘s results:

$$E[Y_0|D=1] = E[Y_0|D=0]$$
$$E[Y_1|D=1] = E[Y_1|D=0]$$

Question 3

Researcher A intends to evaluate a binary treatment (T_i) effect in outcome Y_i, using the following regression:

$$Y_i = \alpha + \beta T_i + \epsilon_i$$

Assume that $Var(\epsilon_i) = \sigma^2$, the treatment is binary and individuals will be assigned either to treatment or control group.

(a) What is the variance of the estimator $\hat{\beta}$?

Assuming T_i are not random, $Var(\epsilon_i) = \sigma^2$, and $Cov(\epsilon_i, \epsilon_j) = 0$, it is possible to find the variance of $\hat{\beta}$ *as the following*:

$$Var(\hat{\beta}) = Var(\frac{\sum(T_i - \overline{T})Y_i}{\sum(T_i - \overline{T})^2})$$

The denominator is a constant, so:

$$= \frac{1}{(\sum(T_i - \overline{T})^2)^2} Var(\sum(T_i - \overline{T})Y_i)$$

Substituting Y:

$$= \frac{1}{(\sum(T_i - \overline{T})^2)^2} Var(\sum(T_i - \overline{T})(\alpha + \beta T_i + \epsilon_i))$$

$$= \frac{1}{(\sum(T_i - \overline{T})^2)^2} Var(\sum(T_i - \overline{T})(\alpha + \beta T_i) + \sum(T_i - \overline{T})\epsilon_i))$$

Cancelling the constant:

$$= \frac{1}{(\sum(T_i - \overline{T})^2)^2} Var(\sum(T_i - \overline{T})\epsilon_i))$$

$$= \frac{1}{(\sum(T_i - \overline{T})^2)^2} \sum Var((T_i - \overline{T})\epsilon_i)$$

The first term inside the variance is a constant, so:

$$= \frac{1}{(\sum(T_i - \overline{T})^2)^2} \sum(T_i - \overline{T})^2 Var(\epsilon_i)$$

Assuming $Var(\epsilon_i) = \sigma^2$, we have:

$$= \frac{1}{(\sum(T_i - \overline{T})^2)^2} \sum(T_i - \overline{T})^2 \sigma^2$$

Replacing the constant:

$$= \frac{\sigma^2}{(\sum(T_i - \overline{T})^2)^2} \sum(T_i - \overline{T})^2$$

Therefore,

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum(T_i - \overline{T})^2}$$

(b) The researcher is interested in testing if the treatment effect is significant or not, how should he write the alternative and null hypothesis?

In order to test if the treatment effect is significant or not, the null (not significant) and alternative (significant) hypothesis must be written as the following:

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

(c) Now assume the researcher is interested in testing if the treatment effect is positive or not, how should he write the alternative and null hypothesis?

In order to test if the treatment effect is positive or not, the null (not positive) and alternative (positive) hypothesis must be written as the following:

$$H_0 : \beta \leq 0$$
$$H_1 : \beta > 0$$