Exam 4137 – Spring 2021

IMPORTANT: Always motivate your answers. Show knowledge and understanding of the concepts taught in the course. Your answers should be as short as possible and as long as necessary. Each subquestion carries equal weight.

---

1. Economists have long been interested to estimate the return to schooling. The Mincer equation is often used to explain the employment earnings as a function of schooling and labor market experience:

$$\log w_i = \beta_0 + \beta_1 s_i + \beta_2 x_i + \beta_3 x_i^2 + \varepsilon_i \tag{1}$$

where, $\log w_i$ is the log of hourly earnings, s is the years of schooling, and x is the years of potential labor market experience. Suppose that you have access to a cross-section data of 3010 individuals, which contains weekly earnings, weekly number of hours of worked, years of schooling and potential labor market experience, and some other variables for each individual.

   (a) Suppose that you have estimated eq.(1) using OLS on the available data, and obtained $\hat{\beta}_1 = 0.07$ with estimated standard error 0.003. How would you interpret this estimate?

   ANSWER HINT:
   The estimate is significant different from zero.
   Everything else equal, hourly earnings for those individuals with s+1 years of schooling are on average 7% higher compared with those individuals with s years of schooling.
   NOTE: the point here is that i want the students to interpret this as "a comparison" instead of a "cause" like one year of additional schooling increase the hourly earnings by 7%

   (b) One of your fellow students estimated a similar model, but she used weekly earnings instead of hourly earnings. However, her OLS estimate of β_1 is much higher than yours, what could be the reason?

ANSWER HINT:
Hourly earnings is a function of weekly earnings and weekly hours of working.

$$log(w_i) = log(W_i/L_i) = log(W_i) - log(L_i)$$

where W_i weekly earnings and L_i the weekly hours of work**.**
The OLS estimate for β_1 in the regression with log(hourly earnings) as the outcome variable is thus the difference of the following two OLS estimates:

$$log(W_i) = b_0 + b_1 s_i + b_2 x_i + b_3 x_3^2 + \varepsilon_i$$

$$log(L_i) = a_0 + a_1 s_i + a_2 x_i + a_3 x_3^2 + \widetilde{\varepsilon}_i$$

that is $\hat{\beta}_1 = \hat{b_1} - \hat{a}_1$
If those who have higher education tend to work more hours per week ($\hat{a}_1 > 0$), this would happen.

(c) You are worried that unobserved factors, such as individual ability, might impact both the hourly earnings and years of schooling. In particular, those with high ability tend to have higher earnings and obtain also more years of schooling. If this is case, what can you say about the estimate from eq.(1)?

ANSWER HINT: If this is the case, $E(\varepsilon_i|s_i) \neq 0$. We have an endogeneity bias, OLS estimator $\hat{\beta}_1$ will be inconsistent. Since cov(s,a) and cov(s,y) are both positive, then $\hat{\beta}_1$ is biased upward.

(d) To cope with the potential problem in (c), some authors propose an instrumental variables (IV) strategy whereby they instrument the variable schooling with a college-in-the-county dummy variable. Namely, for individual i

$$D_i = \begin{cases} 1 & \text{if between age 14 and 24 lived in a county with local colleges} \\ 0 & \text{otherwise} \end{cases}$$

Specify and discuss the assumptions that need to hold for D_i to be a valid instrument. Are these assumptions likely to hold?

ANSWER HINT:
Random assignment: This assumption assumes whether the individual lives in a county with college is exogenous. This implies that both the reduced form and first stage regression have a casual interpretation. This will be violated if high ability families choose to live in a county with colleges. Here we may expect the students to relate their discussions to the Omitted Variable Bias.
The exclusion restriction: Conditional on the *years of schooling,* where an individual lives in a county with/without college around age 20 does not impact the earnings. In other words, all possible impacts of the instrument on the outcome go through the endogenous variable. (one potential violation is that areas more colleges because more demand for high skilled labor)
The relevant assumption: First step regression can be used to check this.
The Monotonicity assumption: *Living in a county with a college does not reduce the years of schooling for all individuals.* This assumption is likely to hold.

(e) The estimates of both the reduced form regression and first stage regressions are given in the following table (standard error in parentheses). Calculate the IV estimates of β_1.

Dependent Variable	Regressor	Estimate
log (hourly earnings)		
	College in the county	0.042
		(0.018)
Years of schooling		
	College in the county	0.320
		(0.088)
N		3,010

ANSWER HINT:
Just one instrument. 2SLS=ILS. The 2SLS estimate is then 0.042/0.320=0.131. One additional year of schooling will lead to an increase of hourly earnings by 13.1%.

2. In the following, we look at a different strategy to identify the causal effect of additional years schooling. In one study, the author utilized a policy change in the minimal school leaving age in Britain in 1947, in which the minimal school leaving age is changed from 14 to 15. The following figure shows the fraction of British born adults that left full-time education by age 14 and by age 15.

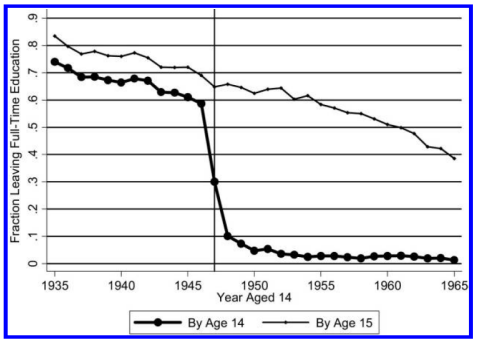


FIGURE 1. FRACTION LEFT FULL-TIME EDUCATION BY YEAR AGED 14 AND 15
(*Great Britain*)

Note: The lower line shows the proportion of British-born adults aged 32 to 64 from the 1983 to 1998 General Household Surveys who report leaving full-time education at or before age 14 from 1935 to 1965. The upper line shows the same, but for age 15. The minimum school-leaving age in Great Britain changed in 1947 from 14 to 15.

(a) The author used a Regression Discontinuity Design (RDD) to identify the causal effect he wants to study. Explain briefly the main idea of this identification strategy based on the setting presented above. What is the key assumption of this identification strategy? What type of causal effect is estimated in this case?

ANSWER HINT:
The causal effect is identified by comparing those who left by age 14 at 1946 vs those who left by age 14 at 1947 (or 1948). Note that since this is a "Fuzzy" design, the outcome discontinuity needed to be adjusted by the treatment discountinity.
The key assumption that continuity of potential outcomes at the cutoff (lecture notes, p 22).
This is a Fuzzy RDD, it actually gives a very specific parameter, the $LATE$.

(b) For simplicity, assume that we only look at the individuals who became 14 years old at the year 1946 and 1948. Let S_i be an indicator for whether individual i leaves at age 14 ($S_i = 1$) or still attends school at age 15 ($S_i = 0$). Let Z_i be a binary variable with $Z_i = 0$ if the minimum school-leaving age for individual i equals 15, and $Z = 1$ if it equals 14. Let y_i be the log hourly earning when they were 45. Suppose that you have the following data:

		Obs	Average log hourly earnings at 45
$Z_i = 0$	$S_i = 0$	4500	1.65
$Z_i = 0$	$S_i = 1$	500	1.45
$Z_i = 1$	$S_i = 0$	2000	1.7
$Z_i = 1$	$S_i = 1$	3000	1.5

Compute an estimate of the RDD estimate of the causal effect of dropping out from school at age *14* and interpret the parameter.

ANSWER HINT:
RDD is the LATE.
E(y|z=0)=0.9*1.65+0.1*1.45=1.63, E(y|z=1)=0.6*1.5+0.4*1.7=1.58,
E(s|z=0)=0.1, E(s|z=1)=0.6, so
LATE=-0.05/0.5=-0.1

(c) Define "compliers", "always takers", and "never takers" in this case. Then compute estimates of the fractions of the population that are "compliers", "always takers", and "never takers".

ANSWER HINT:
Always takers, those who will drop out at 14 no matter what.
Never takers, those who will not drop out at 14 no matter what.
compliers, those who will drop out at 14 if the minimum age is 14, but not drop out at 14 if the minimum age is 15.
p(a)=0.1, p(n)=0.4, p(c)=0.5

(d) Compute estimates of the mean log hourly earnings:

i. for always takers, if they drop out at 14
ii. for never takers, if they don't drop out at 14
iii. for compliers, if they drop out at 14
iv. for compliers, if they don't drop out at 14

ANSWER HINT: .
E(y1|a)=1.45
E(y0|n)=1.7
E(y1|c)*(p(c)/(p(c)+p(a)))+E(y1|a)*(p(a)/(p(c)+p(a)))=1.5
E(y1|c)=1.51
E(y0|c)*(p(c)/(p(c)+p(a)))+E(y0|n)*(p(n)/(p(c)+p(a)))=1.65
E(y0|c)=1.61

3. A group of unemployed individuals are assigned to an online training program which lasts for 60 hours. Suppose that the time that each individual actually spends watching the program x is given by

$$f_X(x|\theta) = \begin{cases} \theta x^{\theta-1}/60^{\theta} & 0 < x \leq 60 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\theta > 1$ is an unknown parameter. It is easy to verify that $E(X) = 60\theta/(\theta+1)$. We have collected information on 100 unemployed individuals and their time spent watching the program. See attached Stata output for detailed information of the data. (You might find the natural logarithm table attached useful.)

(a) Write down the log-likelihood functions, and outline how one can estimate the model using MLE.

ANSWER HINT:
$lnL(x|\theta) = \sum(ln\theta + (\theta-1)ln(x_i) - \theta ln(60)) = n(ln(\theta) + (\theta-1)\overline{ln(x_i)} - \theta ln(60))$

(b) Suppose that we know that the true value of θ is in the set $\{2,3\}$. That is, θ takes either value 2 or 3. What is the $\hat{\theta}_{MLE}$ in this case?

ANSWER HINT:
$lnL(x|\theta) = \sum(ln\theta + (\theta-1)ln(x_i) - \theta ln(60))$
n=100, $\overline{ln(x_i)} \simeq 3.63$, $ln(60) = 4.1$, $ln(1) = 0$, $ln(2) = 0.69$,$ln(3) = 1.1$
$\theta = 2$, $lnL(x|\theta) = -386.55$
$\theta = 3$, $lnL(x|\theta) = -392.44$
so $\hat{\theta}_{MLE} = 2$

(c) Suppose instead that θ can take all real values that is greater than 1, what will be the $\hat{\theta}_{MLE}$ and corresponding log likelihood value $lnL(x|\hat{\theta}_{MLE})$?

ANSWER HINT:
Set the score function to 0

$$\frac{dlnL(x|\theta)}{d(\theta)} = n((1/\theta) + \overline{ln(x_i)} - ln(60)) = 0$$

$\hat{\theta}_{MLE} = 1/(ln60 - \overline{ln(x_i)}) = 1/(4.1 - 3.63) \simeq 2.13$ (ln60 can be found in the attached table)
$lnL(x|\hat{\theta}_{MLE}) = n(ln(\hat{\theta}_{MLE}) + (\hat{\theta}_{MLE} - 1)\overline{ln(x_i)} - \hat{\theta}_{MLE} ln(60)) = -387.5$
NOTE there was an inconsistency on the number of observations in the exam – in the problem it is stated as 500, while in the stata output there is only 100. The above answer is for the case of $n = 100$.
If the student instead uses 500, we should give him full marks, as long as the procedure and calculation is correct (in this case, 5*-387.5=-1937.5)

Suppose that we want study whether θ varies across gender, i.e., ask whether $\theta_f = \theta_m$.

(d) Specify a model where you can test the hypothesis that $\theta_f = \theta_m$. Based on the Stata output, will you reject the null hypothesis? (Hint, use the LR test!)

ANSWER HINT:
This problem might be a little difficult for some students.
As stated in the hint, here one can use the LR test,
under the restriction of the null hypothesis, the log-likelihood function value is what is found in (c).
Without this restriction, we will have to estimate the model separately for male and female.
Similar to (c), we can find that
for male, $\hat{\theta}_{MLE}(male) = 1/(ln60 - \overline{ln(x_i)}) = 1.60$, likelihood function $lnL(x|\hat{\theta}_{MLE}) = -200.4$
for female, $\hat{\theta}_{MLE}(female) = 1/(ln60 - \overline{ln(x_i)}) = 3.18$ likelihood function $lnL(x|\hat{\theta}_{MLE}) = -182.0$
so unrestricted loglikelihood function=$(-200.4 - 182.0) = -382.4$
LR=-2(-387.5+382.4)=10.2
it is chi-squared with degree 1. reject.
Similar to (c), if the student uses $n = 500$, the LR=5*10.2=51. still reject.

Stata Output

```
* data for problem 1;

. gen lnx=ln(x)
. su lnx x gender

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         lnx |        100    3.625334    .4754063   1.830871   4.089396
           x |        100    40.99326    14.26466   6.239318   59.70381
      gender |        100          .5    .5025189          0          1

 . tabstat lnx x, by(gender)

Summary statistics: mean
  by categories of: gender
  gender |       lnx         x
---------+--------------------
       0 |  3.775763  45.48103
       1 |  3.474904  36.5055
---------+--------------------
   Total |  3.625334  40.99326
------------------------------


Critical Values of Chi-square
df       .50      .25      .10      .05      .025     .01      .001
1        0.45     1.32     2.71     3.84     5.02     6.63     10.83
2        1.39     2.77     4.61     5.99     7.38     9.21     13.82
3        2.37     4.11     6.25     7.81     9.35     11.34    16.27
4        3.36     5.39     7.78     9.49     11.14    13.28    18.47
5        4.35     6.63     9.24    11.07     12.83    15.09    20.52
6        5.35     7.84    10.64    12.59     14.45    16.81    22.46
7        6.35     9.04    12.02    14.07     16.01    18.48    24.32
8        7.34    10.22    13.36    15.51     17.53    20.09    26.12
9        8.34    11.39    14.68    16.92     19.02    21.67    27.88
10       9.34    12.55    15.99    18.31     20.48    23.21    29.59
```

The natural logarithm table (Equal to or more than 1.0)

n	$\log_e n$	n	$\log_e n$	n	$\log_e n$	n	$\log_e n$
1.0	0.00000	3.0	1.09861	5.0	1.60944	25.0	3.21887
1.1	0.09531	3.1	1.13140	6.0	1.79176	26.0	3.25809
1.2	0.18232	3.2	1.16315	7.0	1.94591	27.0	3.29583
1.3	0.26236	3.3	1.19392	8.0	2.07944	28.0	3.33220
1.4	0.33647	3.4	1.22377	9.0	2.19722	29.0	3.36729
1.5	0.40546	3.5	1.25276	10.0	2.30258	30.0	3.40119
1.6	0.47000	3.6	1.28093	11.0	2.39789	40.0	3.68888
1.7	0.53063	3.7	1.30833	12.0	2.48491	50.0	3.91202
1.8	0.58779	3.8	1.33500	13.0	2.56495	60.0	4.09434
1.9	0.64185	3.9	1.36097	14.0	2.63905	70.0	4.24849
2.0	0.69314	4.0	1.38629	15.0	2.70805	80.0	4.38202
2.1	0.74193	4.1	1.41099	16.0	2.77259	90.0	4.49981
2.2	0.78845	4.2	1.43508	17.0	2.83321	100.0	4.60517
2.3	0.83291	4.3	1.45861	18.0	2.89037	200.0	5.29832
2.4	0.87547	4.4	1.48160	19.0	2.94444	300.0	5.70378
2.5	0.91629	4.5	1.50408	20.0	2.99573	400.0	5.99146
2.6	0.95551	4.6	1.52605	21,0	3.04452	500.0	6.21461
2.7	0.99325	4.7	1.54756	22,0	3.09104	600.0	6.39693
2.8	1.02962	4.8	1.56861	23.0	3.13549	700.0	6.55108
2.9	1.06471	4.9	1.58923	24.0	3.17805	800.0	6.68461