Econometrics I, Problem Set 1 (Individual)

Jingwan (Amadea) Luo [1]

January 27, 2023

This document answers all questions in the problem set 1 of *Econometrics I* taught by *Prof. Maite Cabeza Gutés*. Content in **blue** is for grading. Sidenote is for self-memo and output labels. For empirical analysis, the software in use is MATLAB. Codes are saved for Section Annex, the `.m` files can be found both accompanied with my submission to *TA. Luis Ignacio Menéndez García* before *27th Jan., 10:15 a.m.* and also my Github Repository [2], in this repository, all the other related files can be found.

[1] IDEA Graduate Programme in Economics, first year student

[2] https://github.com/AmadeaLuo/IDEA_Econometrics-I

Question 1

Consider the following figure:

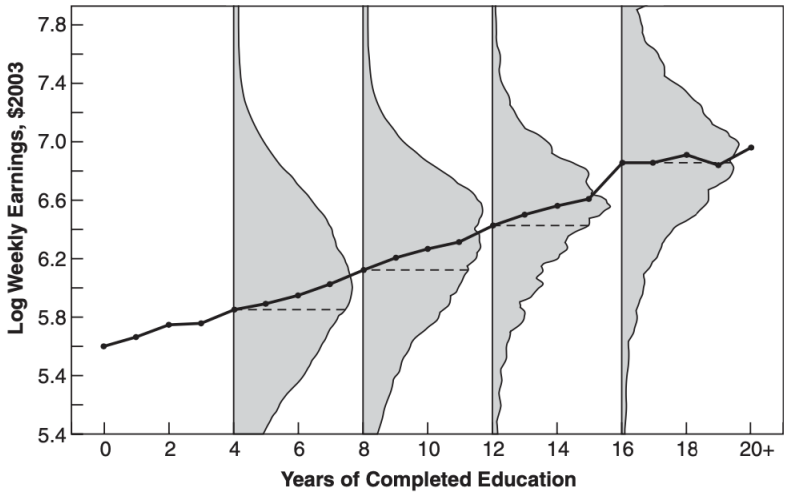

Figure 1: Class Figure 3.1.1 at Slides p.12

Figure 3.1.1 Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40–49 in the 1980 IPUMS 5 percent file.

Figure (1) , reproduced from *Mostly Harmless Econometrics: An Empiricist's Companion*(Angrist, 2009) , page 31, summarises more than 300,000 observations corresponding to a sample of white males in 1980s in the United States, regarding their years of education $(educ)$ and the natural logarithm of their weekly earnings ($lwklywge$).

Datafile `Assig1.csv` includes the observations for variables *educ* and $lwklywge$ (i.e. $lwklywge = \ln(wages)$) similar (not identical) to the ones used to produce the figure above.

Question a)

Provide an estimate of average weekly pay for people with 12 years of education. This statistic would be an estimate of which parameter?

Solution.

$$(\overline{wage}|educ = 12) = 397.9337, \tag{a.1}$$

Numerical value computed in Equation (a.1) is an estimate of

$$E(\overline{wage}|educ = 12).$$

■

Question b)

With the help of MATLAB, reproduce the part of Figure 3.1.1 that includes the conditional sample means and the thick black function joining them. (Please notice you are only asked to reproduce the filled circles and the thick black line joining them, not the 4 Kernel density estimates of the conditional distributions included in the plot). Since the data file provided is slightly different than the one used by the authors, you will not get exactly the same values as those included in Figure 3.1.1. Include plot and MATLAB code as your answer.

Solution. Here we give the average log of weekly wage conditional on the years of education. Figure (2) is produced by firstly compute the conditional means for each sub-group of years of education, then joining them with a line. Everything except for the probability distribution on Figure (1) is reproduced. ■

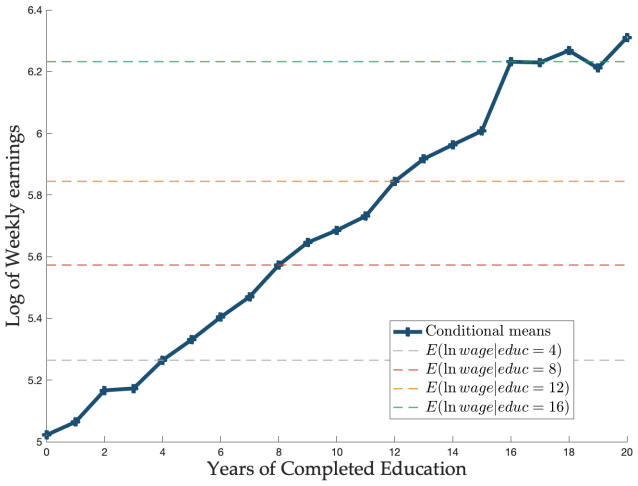

Figure 2: REPLICATION OF FIGURE 3.1.1

Question c)

What does the thick black function you got allow you to say about the relationship between education and earnings? Explain in just a sentence. Try to be as rigorous as possible.

Solution. On **average**, a positive relationship (*not* necessarily a *causal* relationship) between log of weekly earnings and years of education is observed, since logarithmic transformation is monotone, one can posit that wage has a positive (*not* necessarily *linear*) relationship with years of the education **in this sample**. ■

Question d)

Consider that now, instead of focusing on conditional sample means, we focussed on conditional sample medians. With the help of MATLAB calculate the conditional medians, and create a new plot including both, the function joining the conditional sample medians together with the function joining conditional sample means. Include the plot in your answer. Comment.

Solution. COMMENT: The conditional medians are higher than the conditional means for every education group. This observation implies that the distribution of log of weekly earnings is slightly left-skewed (negatively skewed).

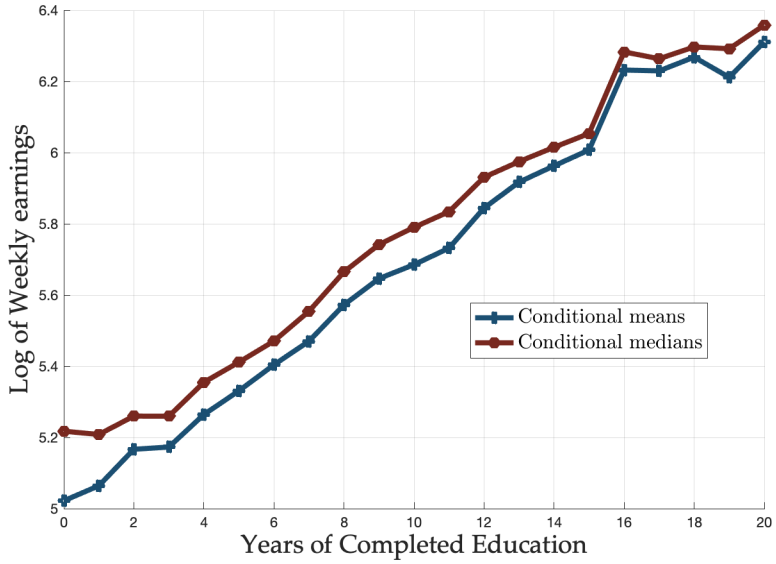

Figure 3: CONDITIONAL MEANS AND MEDIANS FOR LOG OF WEEKLY EARNINGS

We could check this by plotting the density (in Figure (4)). ■

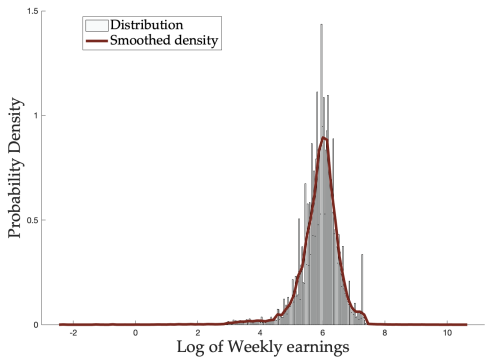

Figure 4: DISTRIBUTION OF LOG OF WEEKLY EARNINGS

Question e)

Let us focus our attention to the thick black function you got in (1b) joining all the conditional sample means (we will ignore now the conditional medians). Your job now is to fit a line through these sample means using MATLAB *polyfit* function.

+ Present a plot including the original sample means and the fitted line.

+ What value did you get for the slope of the line?

Solution. Our goal is to perform a linear fit to the following:

$$\text{lwklywge} = \beta_1 + \beta_2 \text{educ},$$

`p = polyfit(x,y,n)` returns the coefficients for a polynomial $p(x)$ of degree n that is a best fit (in a least-squares sense) for the data in *lwklywge*. The coefficients in p are in descending powers, and the length of p is 2. After the fit, the result is:

$$p = [\beta_2, \beta_1] = [0.0686, 5.0053],$$

Namely,

$$\text{lwklywge} = 5.0053 + 0.0686\text{educ}.$$

So the slope of the line is $\beta_2 = 0.0686$. Please find the fitted line along with the conditional mean at Figure (5). ■

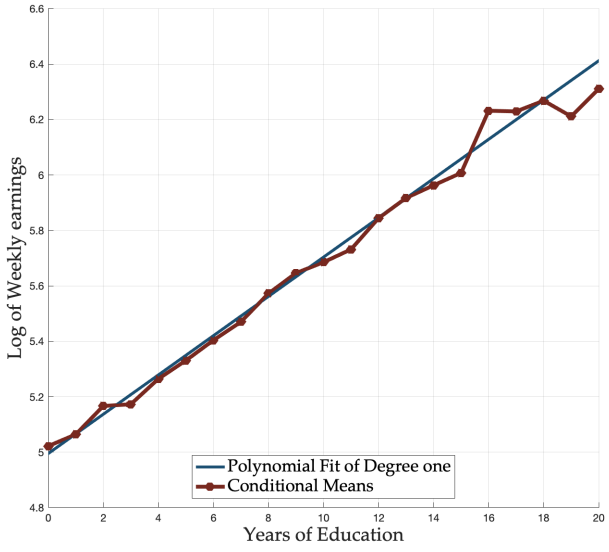

Figure 5: POLYFIT AND CONDITIONAL MEAN

Question f)

Complete this sentence: *"According to the value of the slope of the line fitted in the question above, we can say that..."* Be as specific as possible.

Solution. ...on **average**, an additional year of education means a 7.09 % raise in weekly earnings for this sample. ■

Question 2

Question a)

Consider the following simple linear regression model:

$$y = \beta_1 + \beta_2 x_2 + \epsilon,$$

where $\epsilon \equiv y - (\beta_1 + \beta_2 x_2)$ and

$$\beta_1 = E(y) - \beta_2 E(x_2) \quad \text{and} \quad \beta_2 = \frac{\text{cov}(x_2, y)}{\text{var}(x_2)}$$

Prove that $\beta_1 + \beta_2 x_2$ is the BLP, in MSE sense, of variable y given variable x_2.

Solution. In MSE sense,

$$\arg\min_{y_p(x_2)} E\{y - y_p(x_2) \mid x_2\}^2,$$

Expand gives,

$$\arg\min_{y_p(x_2)} E\left[y^2 \mid x_2\right] - 2E[y \mid x_2]E[y_p(x_2) \mid x] + E\left[y_p(x_2)^2 \mid x_2\right]$$

With F.O.C set to 0 then we reached:

$$y_p(x_2) = E(y \mid x_2) \tag{*}$$

which is essentially **what we want to show.**

Below, the **main technique** in use is that "Expectation of a random variable is a constant, and therefore can be taken out of the Expectation expression".

Consider $E(y \mid x_2)$ and expand y:

$$\begin{aligned} E(\beta_1 + \beta_2 x_2 + \epsilon \mid x_2) &= \beta_1 + \beta_2 x_2 + E(\epsilon \mid x_2) \\ &= \beta_1 + \beta_2 x_2 + E(y - \beta_1 - \beta_2 x_2 \mid x_2) \end{aligned} \tag{2.1}$$

KEY: If substitute $\beta_1 = E(y) - \beta_2 E(x_2)$, $\beta_2 = \frac{cov(x_2, y)}{var(x_2)}$ in Equation (2.1) can give us back $E(y|x_2)$, then it is indeed the predictor

that Equation (*) requires.

$$\beta_1 + \beta_2 x_2 + E(y - \beta_1 - \beta_2 x_2 \mid x_2) = \overbrace{E(y) - \beta_2 E(x_2)}^{\beta_1} + \overbrace{\frac{cov(x_2, y)}{var(x_2)}}^{\beta_2} x_2$$

$$+ E(y \mid x_2) - E\left[\underbrace{E(y) - \beta_2 E(x_2)}_{\beta_1} \mid x_2\right] - \underbrace{\overbrace{E\left(\frac{cov(x_2, y)}{var(x_2)} \mid x_2\right)}^{\beta_2} \cdot x_2}_{\equiv A};$$

$$\begin{aligned}
&= E(y) - \beta_2 E(x_2) + E(y \mid x_2) \\
&- E\left[\underbrace{E(y)}_{\text{a number}} \mid x_2\right] + \beta_2 E\left[\underbrace{E(x_2)}_{\text{a number}} \mid x_2\right]; \\
&= E(y) - \beta_2 E(x_2) + E(y \mid x_2) - E(y) + \beta_2 E(x_2) \\
&= E(y \mid x_2).
\end{aligned}$$

So we reached the requested result.

On a further note, to prove $A = \beta_2 x_2$:

$$\begin{aligned}
E\left(\frac{cov(x_2, y)}{var(x_2)} \mid x_2\right) &= \frac{E\left\{\overbrace{E(x_2 y) - E(x_2)E(y)}^{\text{scalars}} \mid x_2\right\}}{E\left\{\underbrace{E(x_2)^2 - [E(x)]^2}_{\text{scalars}} \mid x_2\right\}} \\
&= \frac{E\{E(x_2 y) \mid x_2\} - E\{E(x_2)E(y) \mid x_2\}}{E\left\{E\left(x_2^2\right) \mid x_2\right\} - E\left\{[E(x)]^2 \mid x_2\right\}} \\
&= \frac{E(x_2 y) - E(x_2)E(y)}{E\left(x_2^2\right) - [E(x)]^2} \\
&= \frac{cov(x_2, y)}{var(x_2)}.
\end{aligned}$$

■

Question b)

Prove the *Decomposition Theorem*, included in slide 1(12), that states that we can always decompose a variable y as:

$$y = E(y|x) + \epsilon$$

with

+ $E(\epsilon|x) = 0$ (ϵ mean independent of x);

+ ϵ uncorrelated with x and with any function of x.

Solution. We know the following equality must holds:

$$y = y + E(y|x) - E(y|x)$$

Write it as

$$y = E(y|x) + \{y - E(y|x)\},$$

and define $\epsilon \equiv y - E(y|x)$.
We check the two properties in turn:

$$[E(\epsilon|x) = 0]: \quad E(y - E(y|x)|x) = E(y|x) - E(y|x) = 0, \quad \checkmark$$
$$[E(x\epsilon) = 0]: \quad E(x\epsilon) = E(x \cdot \underbrace{E(\epsilon|x)}_{\text{L.I.E., and =0}}) = 0 \quad \checkmark$$

Therefore the decomposition theorem holds. ■

Annex

IMPORTANT Codes can also be found at the submission `.zip` to TA's email and Github Repository [3]. Before running the codes, please install the MATLAB add-on *"Professional Plots"* (Aalok, 2023).

[3] https://github.com/AmadeaLuo/IDEA_Econometrics-I

```
%% Setup
clc;clear;

% Set the working directory to the place where the ...
    current file is saved
tmp = matlab.desktop.editor.getActive;
cd(fileparts(tmp.Filename));

dt = readtable("Assig1.csv", ...
    'VariableNamingRule','preserve');
```

To Accompany Question 1a)

```
%% Question 1 a) estimate of average weekly pay with educ=12

% +++ Recover variable `wage` from `lwklywge`: +++
%
% Recover by taking exponetial of `lwklywge`:
wage = exp(dt.lwklywge);
% Make a table of the recovered wage data:
wage1 = array2table(wage);
% Glue to a new dataset:
dt1 = horzcat(dt, wage1);

% +++ Conditional mean E(wage | educ = 12): +++
%
% Create a Sub-sample with educ == 12:
sample_12 = dt1(dt1.educ == 12,:);
% E(wage | educ = 12):
cmean_wage = mean(sample_12.wage);
```

To Accompany Question 1b)

```
%% Question 1 b) plot mean of `lwklywge` conditional on ...
    each educ group

% +++ Conditional mean: +++
%
% Get unique educ values
educ_groups = unique(dt1.educ);
% Initialize vector to store means
cmeans = zeros(length(educ_groups),1);
% Compute mean for each educ value
for i = 1:length(educ_groups)
    cmeans(i) = mean(dt1.lwklywge(dt1.educ == ...
        educ_groups(i)));
end
```

```
% +++ Replicate line chart in Fig 3.1.1: +++
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%           get the add on called               %
%            `professional plot`                %
%                  first                        %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% Before anything, set the graph aesthetics
PS = PLOT_STANDARDS();
```

```
% Plot the means as a line chart with markers
figure(1);
fig1_comps.fig = gcf;
grid off;
hold on;

% Get the conditional mean for educ == 4, 8, 12, 16
cmeans4 = cmeans(5,:);
cmeans8 = cmeans(9,:);
cmeans12 = cmeans(13,:);
cmeans16 = cmeans(17,:);
```

```
% Drawing¬
% Main graph:
fig1_comps.p0 = plot(educ_groups, cmeans);
set(fig1_comps.p0, 'Color', PS.Blue5, ...
    'LineWidth', 4, 'Marker','o');
% cmeans for educ == 4, 8, 12, 16:
fig1_comps.p1 = yline(cmeans4, 'Color',PS.Grey3,...
    'LineStyle','--', LineWidth=1.5);
fig1_comps.p2 = yline(cmeans8, 'Color',PS.Red3,...
    'LineStyle','--', LineWidth=1.5);
fig1_comps.p3 = yline(cmeans12, 'Color',PS.Orange3,...
    'LineStyle','--', LineWidth=1.5);
fig1_comps.p4 = yline(cmeans16, 'Color',PS.Green3,...
    'LineStyle','--', LineWidth=1.5);

xlabel('Years of Completed Education', ...
    'FontSize',20, 'FontName','Palatino');
ylabel('Log of Weekly earnings', ...
    'FontSize',20, 'FontName','Palatino');
legend('Conditional means', ...
    '$E(\ln wage \vert educ = 4)$',...
    '$E(\ln wage \vert educ = 8)$',...
    '$E(\ln wage \vert educ = 12)$',...
    '$E(\ln wage \vert educ = 16)$',...
    'Location','best',...
    'Interpreter', 'latex', ...
    'Fontsize', 16)

hold off;
```

To Accompany Question 1d)

```
%% Question 1 d) plot mean&med of `lwklywge` conditional ...
    on each educ group

% +++ Compute the medians +++:
%
% Initialize vector to store medians
cmeds = zeros(length(educ_groups),1);
% Compute mean for each educ value
for i = 1:length(educ_groups)
    cmeds(i) = median(dt1.lwklywge(dt1.educ == ...
        educ_groups(i)));
end
%
```

```
% +++ Plot conditional means and medians +++:
%
% Drawing¬
figure(2);
fig1_comps.fig = gcf;
grid on;
hold on;

fig2_comps.p0 = plot(educ_groups, cmeans);
fig2_comps.p1 = plot(educ_groups, cmeds);

set(fig2_comps.p0, 'Color', PS.Blue5, ...
    'LineWidth', 4, 'Marker','o');
set(fig2_comps.p1, 'Color', PS.Red5, ...
    'LineWidth', 4, 'Marker','Diamond');

xlabel('Years of Completed Education', ...
    'FontSize',20, 'FontName','Palatino');
ylabel('Log of Weekly earnings', ...
    'FontSize',20, 'FontName','Palatino');
legend('Conditional means', ...
    'Conditional medians',...
    'Location','best',...
    'Interpreter', 'latex', ...
    'Fontsize', 16)

hold off;
%
```

```
%% Additional Checkings for Question 1 d)

% +++ Check distribution of `lwklywge`: +++
%
% Scott's rule for determining optimal bandwidth
bw = 3.5*std(dt1.lwklywge)*length(dt1.lwklywge)^(-1/3);
% Drawing¬
figure(3);
fig1_comps.fig = gcf;
grid off;
hold on;
histogram(dt1.lwklywge, 'Normalization', 'pdf', ...
    'BinWidth', bw, 'FaceColor',PS.Grey1);

[density,xi] = ksdensity(dt1.lwklywge);

% Plot density function
plot(xi,density,'LineWidth',3.5,'Color',PS.Red5);

xlabel('Log of Weekly earnings', 'FontSize',22, ...
    'FontName','Palatino');
ylabel('Probability Density', 'FontSize',22, ...
    'FontName','Palatino');
legend('Distribution', 'Smoothed density',...
    'FontSize',18, 'FontName','Palatino','Location', 'Best');
hold off;
%
```

To Accompany Question 1e)

```
%% Question 1 e) polynomial fitting a line between ...
    `lwklywge` and `educ`
% Fit the polynomial of degree one
p = polyfit(educ_groups, cmeans, 1);

% Get the coefficients of the polynomial
coefs = p;

% Generate the fitted values for the independent variable
y_fit = polyval(p,dt1.educ);

% Plot the data and the fitted line
figure(4);
fig1_comps.fig = gcf;
grid on;
hold on;

fig4_comps.p1 = plot(dt1.educ, y_fit);
fig4_comps.p2 = plot(educ_groups, cmeans);

set(fig4_comps.p1, 'Color', PS.Blue5, ...
    'LineWidth', 3);
set(fig4_comps.p2, 'Color', PS.Red5, ...
    'LineWidth', 3, 'Marker','x');

xlabel('Years of Education', 'FontSize',20, ...
    'FontName','Palatino');
ylabel('Log of Weekly earnings', 'FontSize',20, ...
    'FontName','Palatino');
legend('Polynomial Fit of Degree one', 'Conditional ...
    Means',...
    'FontSize',18, 'FontName','Palatino','Location', 'Best');

hold off;
```

References

Aalok, A. (2023), 'Professional plots', https://www.mathworks.com/matlabcentral/fileexchange/100766-professional_plots. [Online; accessed January 22, 2023].

Angrist, Joshua David;Pischke, J.-S. (2009), *Mostly harmless econometrics: an empiricist's companion*, Princeton University Press.
URL: *libgen.li/file.php?md5=1e30b583b25afaca5fdda278dccc57f4*