Waseda University

School of Political Science and Economics

Homework 2

Daniel Fabio Groth

Econometrics, Fall 2024

Table of contents

Problem 1

Solve excercise 1,3 and 5 in Problem set 2.

Excercise 1: Show the following equalities hold:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i(X_i - \bar{X}_n) \tag{1}$$

Let's start by looking at the left side of the equation:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \to (X_i - \hat{X}_n)^2 \tag{1.1}$$

Then we can expand the equation:

$$(X_i - \bar{X}_n)^2 = X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2 \tag{1.2}$$

thus, we can rewrite the equation as:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2 \tag{1.3}$$

Then we separate the terms:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2\bar{X}_n}{n}\sum_{i=1}^{n} X_i + \frac{\bar{X}_n^2}{n}\sum_{i=1}^{n} 1 \tag{1.4}$$

Then since $\sum_{i=1}^{n} X_i = n\bar{X}_n$, and $\sum_{i=1}^{n} 1 = n$, this just becomes:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\bar{X}_n\bar{X}_n + \bar{X}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 \tag{1.5}$$

Now we can look at the right side of the equation:

$$\frac{1}{n}\sum_{i=1}^{n} X_i(X_i - \bar{X}_n) \to X_i(X_i - \bar{X}_n) \tag{1.6}$$

Then we can expand the equation:

$$X_i(X_i - \bar{X}_n) = X_i^2 - X_i\bar{X}_n \tag{1.7}$$

thus, we can rewrite the equation as:

$$\frac{1}{n}\sum_{i=1}^{n} X_i(X_i - \bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{\bar{X}_n}{n}\sum_{i=1}^{n} X_i \tag{1.8}$$

Then we can use the same logic as before:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 \tag{1.9}$$

Thus, we have shown that the left side of the equation is equal to the right side of the equation.

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i(X_i - \bar{X}_n) \tag{1}$$

Excersice 1 cont: Showing that the second equality holds:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - \bar{Y}_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i(X_i - \bar{X}_n) \tag{2}$$

Let's start by looking at the left side of the equation:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n) \to (X_i - \bar{X}_n)(Y_i - \bar{Y}_n) \tag{2.1}$$

Then we can expand the equation:

$$(X_i - \bar{X}_n)(Y_i - \bar{Y}_n) = X_iY_i - X_i\bar{Y}_n - \bar{X}_nY_i + \bar{X}_n\bar{Y}_n \tag{2.2}$$

thus, we can rewrite the equation as:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n) = \frac{1}{n}\sum_{i=1}^{n} X_iY_i - \frac{1}{n}\sum_{i=1}^{n} X_i\bar{Y}_n - \frac{1}{n}\sum_{i=1}^{n}\bar{X}_nY_i + \frac{1}{n}\sum_{i=1}^{n}\bar{X}_n\bar{Y}_n \tag{2.3}$$

Then we separate the terms:

$$\frac{1}{n}\sum_{i=1}^{n} X_iY_i - \frac{1}{n}\bar{Y}_n\sum_{i=1}^{n} X_i - \frac{1}{n}\bar{X}_n\sum_{i=1}^{n} Y_i + \frac{1}{n}\bar{X}_n\bar{Y}_n\sum_{i=1}^{n} 1 \tag{2.4}$$

Then since $\sum_{i=1}^{n} X_i = n\bar{X}_n$, $\sum_{i=1}^{n} Y_i = n\bar{Y}_n$, and $\sum_{i=1}^{n} 1 = n$, this just becomes:

$$\frac{1}{n}\sum_{i=1}^{n} X_iY_i - \bar{Y}_n\bar{X}_n - \bar{X}_n\bar{Y}_n + \bar{X}_n\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} X_iY_i - \bar{X}_n\bar{Y}_n \tag{2.5}$$

Let's now look at the middle term of the equation:

$$\frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - \bar{Y}_n) \rightarrow X_i(Y_i - \bar{Y}_n) \tag{2.6}$$

Then we can expand the equation:

$$X_i(Y_i - \bar{Y}_n) = X_iY_i - X_i\bar{Y}_n \tag{2.7}$$

thus, we can rewrite the equation as:

$$\frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - \bar{Y}_n) = \frac{1}{n}\sum_{i=1}^{n} X_iY_i - \frac{\bar{Y}_n}{n}\sum_{i=1}^{n} X_i \tag{2.8}$$

Then we can use the same logic as before:

$$\frac{1}{n}\sum_{i=1}^{n} X_iY_i - \frac{\bar{Y}_n}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} X_iY_i - \bar{X}_n\bar{Y}_n \tag{2.9}$$

Finally looking at the right side of the equation:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i(X_i - \bar{X}_n) \rightarrow Y_i(X_i - \bar{X}_n) \tag{2.10}$$

Then we can expand the equation:

$$Y_i(X_i - \bar{X}_n) = X_iY_i - Y_i\bar{X}_n \tag{2.11}$$

thus, we can rewrite the equation as:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i(X_i - \bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} X_iY_i - \frac{\bar{X}_n}{n}\sum_{i=1}^{n} Y_i \tag{2.12}$$

Then using same logic and substituting:

$$\frac{1}{n}\sum_{i=1}^{n} X_iY_i - \frac{\bar{X}_n}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} X_iY_i - \bar{X}_n\bar{Y}_n \tag{2.13}$$

Finally, we can see that equation (2.5), (2.9), and (2.13) are all equal to each other, thus we have shown that:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - \bar{Y}_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i(X_i - \bar{X}_n) \tag{2.14}$$

Excercise 3

Consider a regression model that has no intercept term:

$$Y_i = X_i\beta_1 + \epsilon_i = 1, ..., n.$$

Derive the least squares estimator for β_1:

The least squares estimator for β_1 is given by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

Then we can substitute the equation for Y_i:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i(X_i\beta_1 + \epsilon_i)}{\sum_{i=1}^{n} X_i^2}$$

Then we can expand the equation:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i^2\beta_1 + X_i\epsilon_i}{\sum_{i=1}^{n} X_i^2}$$

Then we can factor out the β_1:

$$\hat{\beta}_1 = \frac{\beta_1 \sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} X_i\epsilon_i}{\sum_{i=1}^{n} X_i^2}$$

Thus, the least squares estimator for β_1 is given by:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} X_i\epsilon_i}{\sum_{i=1}^{n} X_i^2}$$

Excercise 5

Let $(\hat{\beta}_0, \hat{\beta}_1)$ be the ordinary least squares estimator of

$$Y_i = \beta_0 + X_i\beta_1 + \epsilon_i, i = 1, ..., n.$$

The prediction error (i.e, residual) for each i is given by $\hat{e_i} = Y_i - \hat{\beta}_0 n - X_i\hat{\beta}_1 n.$

First show that:

$\sum_{i=1}^{n} \hat{e}_i = 0$
Proof:

$$\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - X_i\hat{\beta}_1)$$

Expanding the equation:

$$\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} X_i\hat{\beta}_1$$

Then since $\hat{\beta}_0$ and $\hat{\beta}_1$ are constants, we can factor them out:

$$\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} X_i$$

Recall that $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimators, thus they satisfy the normal equations:

$$\sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} X_i = 0$$

$$\sum_{i=1}^{n} \hat{e}_i = 0$$

Thus, we have shown that $\sum_{i=1}^{n} \hat{e}_i = 0$

Second, show that:

$$\sum_{i=1}^{n} X_i \hat{e}_i = 0$$

Proof:

$$\sum_{i=1}^{n} X_i \hat{e}_i = \sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_0 - X_i \hat{\beta}_1)$$

Expanding the equation:

$$\sum_{i=1}^{n} X_i \hat{e}_i = \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \hat{\beta}_0 - \sum_{i=1}^{n} X_i X_i \hat{\beta}_1$$

Then since $\hat{\beta}_0$ and $\hat{\beta}_1$ are constants, we can factor them out:

$$\sum_{i=1}^{n} X_i \hat{e}_i = \sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 \sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2$$

Recall that $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimators, thus they satisfy the normal equations:

$$\sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 \sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = 0$$

$$\sum_{i=1}^{n} X_i \hat{e}_i = 0$$

Thus, we have shown that $\sum_{i=1}^{n} X_i \hat{e}_i = 0$

Problem 2

Show that under Assumptions 1-3 in the L.6 slides, the variance of $\hat{\beta}_{n1}$ given $X_1, ..., X_n$ is:

$$\frac{\sigma^2}{n} \frac{1}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

Proof:

Recall that the least squares estimator for β_1 is given by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$

Substituting $Y_i = \beta_0 + X_i\beta_1 + \epsilon_i$ into the equation above, we get:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(\beta_0 + X_i\beta_1 + \epsilon_i) - (\beta_0 + X_i\beta_1 + \bar{\epsilon_i})}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

where $\bar{\epsilon_i} = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i$ and $\sum_{i=1}^{n}(X_i - \bar{X}_n) = 0$

then we can simplify the equation to:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)[\beta 1(X_i - \bar{X}_n)((\epsilon_i - \bar{\epsilon_i})]}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

we then focus on the numerator:

$$\sum_{i=1}^{n}(X_i - \bar{X}_n)[\beta 1(X_i - \bar{X}_n)((\epsilon_i - \bar{\epsilon_i})] = \beta_1\sum_{i=1}^{n}(X_i - \bar{X}_n)^2(\epsilon_i - \bar{\epsilon_i})$$

and since the denominator is a constant, we can factor it out:

$$\hat{\beta}_1 = \beta_1 + \frac{\beta_1\sum_{i=1}^{n}(X_i - \bar{X}_n)^2(\epsilon_i - \bar{\epsilon_i})}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

Recall that the variance of $\hat{\beta}_1$ is given by:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

and since $\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$, we can rewrite the equation as:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{n\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2} = \frac{\sigma^2}{n}\frac{1}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

Thus, we have shown that under Assumptions 1-3 in the L.6 slides, the variance of $\hat{\beta}_{n1}$ given $X_1, ..., X_n$ is:

$$\frac{\sigma^2}{n} \frac{1}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

Problem 3

In this problem, you calculate the OLS estimators using R. Please obtain your own data by using the following code:

```r
set.seed(34)

data <- as.data.frame(state.x77)
data <- data[sample(1:50, 40),]
```

where you need to input the last two digits of your student number for A. Here we use the information of the life expectancy as Y and the illiteracy rate as X. Then answer the following problems.

1) We consider the following two models.

Model 1: $Y_i = \beta_0 + X_i\beta_1 + \epsilon_i$

Model 2: $Y_i = X_i\beta_1 + \epsilon_i$

Obtain the OLS estimators for these two models **without using the lm() function** and compare the results with those given by the lm function.

```r
# Model 1
X <- data$Illiteracy
Y <- data$`Life Exp`
n <- length(X)
X_bar <- mean(X)
Y_bar <- mean(Y)
beta_1_hat <- cov(X, Y)/var(X)
beta_0_hat <- Y_bar - beta_1_hat*X_bar
cat("Model 1: Beta 0 = ", beta_0_hat, " Beta 1 = ", beta_1_hat, "\n")
```

```
Model 1: Beta 0 =  72.44024  Beta 1 =  -1.332373
```

```r
# Double checking with lm function
lm(Y ~ X, data = data)$coefficients
```

```
(Intercept)           X
  72.440238   -1.332373
```

```r
# Model 2

beta_1_hat_model2 <- sum(X*Y)/sum(X^2)

cat("Model 2: beta_1_hat = ", beta_1_hat_model2, "\n")
```

```
Model 2: beta_1_hat =  45.25329
```

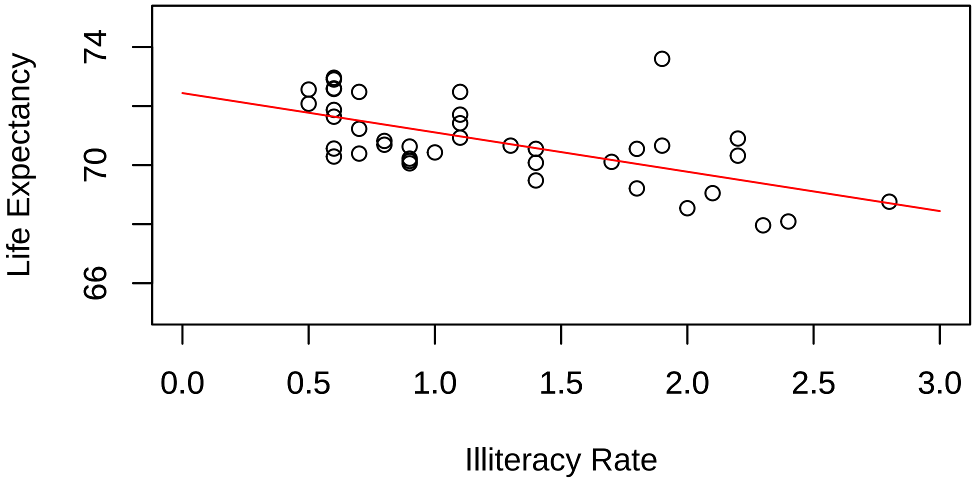
```
# Double checking with lm function
lm(Y ~ 0 + X, data = data)$coefficients
```

```
       X 
45.25329 
```

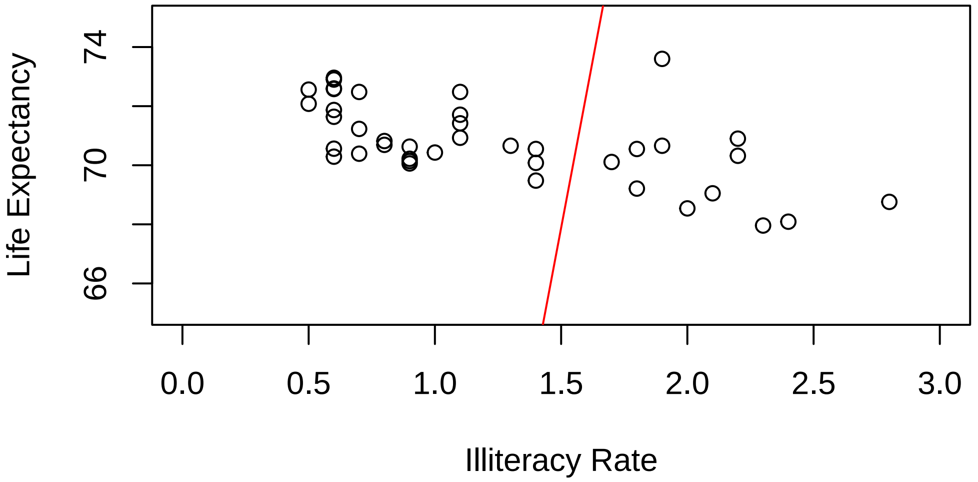
2) For the two models, visually compare the distribution of the data and the lines obtained by OLS as we did in p.16 in the Lecture 6 slides. Discuss which results look more reasonable

```
# Looking at model 1
reg <- function(x) beta_0_hat + beta_1_hat*x
plot(X, Y, ylim = c(65,75), xlim = c(0,3),
     xlab="Illiteracy Rate", ylab="Life Expectancy")

par(new=TRUE)
curve(reg(x), col="red", xlim = c(0,3),
      ylim = c(65,75), xlab = "", ylab = "")
```



```
# Looking at model 2
reg2 <- function(x) beta_1_hat_model2*x
plot(X, Y, ylim = c(65,75), xlim = c(0,3),
     xlab="Illiteracy Rate", ylab="Life Expectancy")
curve(reg2(x), col="red", xlim = c(0,3),
      ylim = c(65,75), add=TRUE)
```



From the plots, it is clear that the first model is more reasonable as it fits the data better, also model 1 shows that the relationship between Life expectancy and illiteracy rate is negative, hence people with better literacy rate are expected to live longer.

For the second model, with no intercept, the line does not fit the data well at all, and it is not reasonable to assume that the life expectancy is 0 when the illiteracy rate is 0, also it shows that there is a positive relationship between the two variables when the graph shows there is not.

3) Based on the “more reasonable” model you chose, explain what the estimated value of β_1 implies about the relationship between the illiteracy rate and the life expectancy.

The estimated value of β_1 in model 1 is -0.52, which implies that for every 1 unit increase in the illiteracy rate, the life expectancy decreases by 0.52 years. This means that people with higher illiteracy rates are expected to live shorter lives compared to people with lower illiteracy rates.

Usage of AI

Copilot

As I am writing this document in Rstudio, there is an integration of copilot which sometimes automatically suggests code snippets. Sometimes it works great and my latex math gets written perfectly, and other times it just gives me a bunch of random unrelevant latex math or code.

Here is a Link to copilot.

ChatGPT

Here is the link to the conversation where I asked questions:

Link to ChatGPT