HOME ASSIGNMENT 1

DANIIL BUCHKO

Question (1)**.** Suppose your calculate estimates for β_0 and β_1 by finding the solution to the following minimization problem.

$$\min_{b_0,b_1} L = \min_{b_0,b_1} \sum_{i=1}^{n} \exp\left\{(y_i - b_0 - b_1 x_i)^2\right\}$$

Write down first-order conditions for the estimates.

Solution.

$$\frac{\partial L}{\partial b_0} = \sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i)\exp\left\{(y_i - b_0 - b_1 x_i)^2\right\} = 0$$

$$\frac{\partial L}{\partial b_1} = \sum_{i=1}^{n} 2x_i(y_i - b_0 - b_1 x_i)\exp\left\{(y_i - b_0 - b_1 x_i)^2\right\} = 0$$

□

Question (2)**.** In the simple linear regression model $y = \beta_0 + \beta_1 x + u$, suppose that $\mathbb{E}(u) \neq 0$. Letting $\alpha_0 = \mathbb{E}(u)$, show that the model can always be rewritten with the same slope, but a new intercept and error, where the new error has a zero expected value.

Solution. Let $\varepsilon = u - \alpha_0$, then model can be rewritten as follows:

$$y = \underbrace{\beta_0 - \alpha_0}_{\beta_0^{'}} + \beta_1 x + u = \beta_0^{'} + \beta_1 x + \varepsilon$$

Where $\mathbb{E}[\varepsilon] = \mathbb{E}[u - \alpha_0] = 0$ □

Question (3)**.** Consider the standard simple linear regression model:

$$y = \beta_0 + \beta_1 x + u$$

When $n = 3$, is it possible that the data point with maximal value of y is located below the OLS regression line? If answer is yes, provide an example, if no, provide a proof.

Solution. Lets assume that there exists pair (x_m, y_m) such that

$$\begin{cases} y_m < \hat{\beta}_0 + \hat{\beta}_1 x_m \\ y_m = \max\{y_1 \ldots y_n\} \end{cases}$$

From the FOCs we know that:

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{cases}$$

Inserting those values to initial statement yields:

$$y_m < \bar{y} - \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \bar{x} + \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} x_m$$

$$y_m \sum_{i=1}^n (x_i - \bar{x})^2 < \bar{y} \sum_{i=1}^n (x_i - \bar{x})^2 + (x_m - \bar{x}) \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$(y_m - \bar{y}) \sum_{i=1}^n (x_i - \bar{x})^2 < (x_m - \bar{x}) \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Next lets assume that $x_m - \bar{x} \neq 0$ and $\sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$ and reformat both sides of inequality above slightly:

$$\frac{y_m - \bar{y}}{x_m - \bar{x}} < \underbrace{\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}}_{\hat{\beta}_1}$$

Thus, we got conditions for which OLS regression of $y = \{y_i\}_{i=1}^n$ on $x = \{x_i\}_{i=1}^n$ with $y_m = \max\{y\}$ lying under regression line. For $n = 3$ points are namely:

$$\begin{cases} x = [1, 2, 3] \\ y = [1, 3, 4] \end{cases}$$

□

Question (4)**.** Consider the following relation:

$$y = \sqrt{x} + u$$

where x is uniformly distributed uniformly on segment $[0, 1]$ and error term u has zero mean conditional on x. A random sample of size n is collected: $\{y_i, x_i\}_{i=1}^n$. Some econometrician performs OLS estimation based on a simple linear model

$$y = \beta_0 + \beta_1 x + e$$

Based on her estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ the econometrician constructs the predictor for the conditional mean of the dependent variable, i.e for $\mathbb{E}(y|x)$ as $\mathbb{E}(\hat{y|}x) = \hat{\beta}_0 + \hat{\beta}_1 x$

(1) Calculate the expected value of y.

(2) Find the expected values of $\hat{\beta}_0$ and $\hat{\beta}_1$ conditional on realized values of x.

Solution. The expected value of y is the following:

$$\mathbb{E}(y) = \mathbb{E}(\mathbb{E}(y|x)) = \mathbb{E}(\sqrt{x}) = \int_0^1 \sqrt{t}dt = \frac{2}{3}$$

Lets now calculate expected values of $\hat{\beta}_0$ and $\hat{\beta}_1$. First of all lets notice that:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + e_i \\ \bar{y} &= \beta_0 + \beta_1 \bar{x} + \bar{e} \end{aligned}$$

then $y_i - \bar{y} = \beta_1(x_i - \bar{x}) + (e_i - \bar{e})$. We will insert this result in the OLS estimation:

$$\begin{aligned} \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} &= \frac{\sum_{i=1}^n (x_i - \bar{x}) \left[\beta_1(x_i - \bar{x}) + (e_i - \bar{e})\right]}{\sum_{i=1}^n (x_i - \bar{x})^2} = \\ &= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) e_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

Now it is helpful to remember, that

$$e_i = y_i - \beta_0 - \beta_1 x_i = \sqrt{x_i} + u_i - \beta_0 - \beta_1 x_i$$

Finally replacing for e_i yields:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})(\sqrt{x_i} + u_i - \beta_0 - \beta_1 x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Calculating conditional expectation for $\hat{\beta}_1$:

$$\begin{aligned} \mathbb{E}(\hat{\beta}_1|X) &= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})\mathbb{E}(\sqrt{x_i} + u_i - \beta_0 - \beta_1 x_i|X)}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})(\sqrt{x_i} - \beta_0 - \beta_1 x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

Now lets move on to calculating the conditional expectation for $\hat{\beta}_0$. From FOCs it follows that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Therefore:

$$\mathbb{E}[\hat{\beta}_0|X] = \mathbb{E}[\bar{y}|X] - \underbrace{\mathbb{E}[\hat{\beta}_1|X]}_{\text{calculated}}$$

What we left to do is to find $\mathbb{E}[\bar{y}|X]$:

$$\mathbb{E}[\bar{y}|X] = \mathbb{E}[\beta_0 + \beta_1 \bar{x} + \bar{e}|X] = \beta_0 + \beta_1 \bar{x} + \mathbb{E}[\bar{e}|X]$$

where

$$\mathbb{E}[\bar{e}|X] = \frac{\mathbb{E}[\sum_{i=1}^{n}\sqrt{x_i} + u_i - \beta_0 - \beta_1 x_i|X]}{n} = \frac{\sum_{i=1}^{n}\left(\sqrt{x_i} - \beta_1 x_i\right)}{n} - \beta_0$$

Inserting back to $\mathbb{E}[\bar{y}|X]$:

$$\mathbb{E}[\bar{y}|X] = \beta_0 + \beta_1\bar{x} + \frac{\sum_{i=1}^{n}\left(\sqrt{x_i} - \beta_1 x_i\right)}{n} - \beta_0 = \frac{\sum_{i=1}^{n}\sqrt{x_i}}{n}$$

Now we have everything to calculate $\mathbb{E}[\hat{\beta}_0|X]$:

$$\mathbb{E}[\hat{\beta}_0|X] = \frac{\sum_{i=1}^{n}\sqrt{x_i}}{n} - \beta_1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\sqrt{x_i} - \beta_0 - \beta_1 x_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

□

Question (5)**.** Using R, compute the sample mean and standard deviation of ahe (average hourly earnings), yrseduc (years of education), and female.

Solution. Standard deviations: ahe = 8.028319, yrseduc = 2.48002, female = 0.4963357

Means:
ahe = 15.19042, yrseduc = 13.46549, female = 0.4385083 □

Question (6)**.** Estimate a regression of ahe on yrseduc.

(1) What is the coefficient on yrseduc? Explain in words what it means. Is the numerical value of your estimate large or small in an economic (real-world) sense?
(2) Graph the data points and the estimated regression line.
(3) Is the slope coefficient statistically significantly different from zero? Show how you reach this conclusion. Use heteroskedasticity robust standard errors to answer this question.
(4) Report the 95% confidence interval for β_1, the slope of the population regression line. Use heteroskedasticity-robust standard errors to answer this question.
(5) What is the R^2 of this regression? What does this mean?
(6) Compute the correlation coefficient between ahe and yrseduc, and compare its square to the R^2. How are the correlation coefficient and the R^2 related?
(7) What is the root mean squared error of the regression? What does this mean?
(8) Based on your graph from (b), does the error term appear to be homoskedastic or heteroskedastic?

Solution. (1) Coefficient is 1.32186. It means that one additional year of education on average increases average hourly earnings for 1.32186 dollars. It means that each additional year of education increases daily earnings by $1.32186 * 8 = 10$ dollars and monthly earnings by $10 * 30 = 300$ dollars. Therefore, starting from 6th class of school, each additional year gives us approximately 20k rubles. By now i have studied for 9 years which would promise me a salary of 180k rubles. For my opinion it is quite a lot in real-world Russian situation.

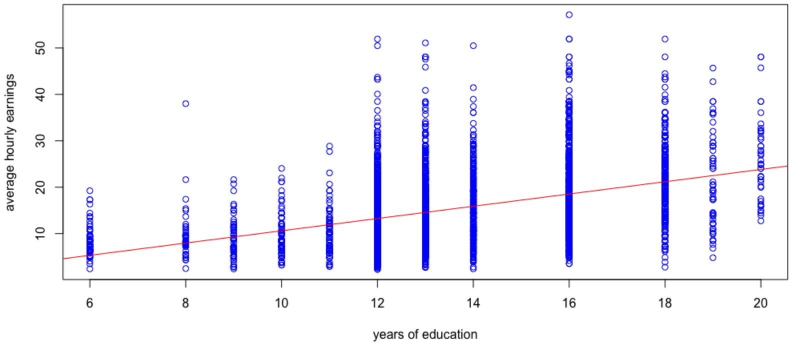

Figure 1. data points and the estimated regression line

(2)

Table 1

	Dependent variable:
yrseduc	1.322^{***}
	(0.050)
Constant	-2.609^{***}
	(0.647)
Note:	*p<0.1; **p<0.05; ***p<0.01

(3) Since P-value is far below 0.05, that means that the probability of obtaining the same value for coefficient yrseduc assuming zero

hypothesis relevance (coeff equals to zero) is very low. Therefore the slope is statistically significantly different from zero.

Table 2. Confidence intervals for β_0 and β_1

	2.5 %	97.5 %
(Intercept)	-3.899	-1.319
yrseduc	1.228	1.416

(4)

(5) Multiple R-squared: 0.1667. It means that years of education explain earnings varinace by fraction of 0.1667. It also means that correlation between years of education and hourly earnings is $0.16^2 = 0.4$.

(6) We see that correlation is very close to the square root of R^2: 0.4082891 vs 0.4083341

(7) RMSE is 7.326572 which means that on average hourly earnings deviate from predicted expectation for 7$. RMSE represents standart deviation of sample error.

(8) Seems like the average hourly earnings distribution becomes more heavy tailed with each additional year of education. It implies that variance of ahe depends on yrseduc and therefore the error term appear to be heteroskedastic.

□

Question (7)**.** Estimate a regression of ahe on female.

(1) What is the coefficient on female? Explain in words what this means. Is the numerical value of your estimate large or small in an economic (real-world) sense?

(2) Test the hypothesis that average hourly earnings are the same for male and female workers, against the alternative that they differ, at the 5% significance level. Make sure to use heteroskedasticity robust standard errors to answer this question.

(3) Compute the sample average of ahe for women, and the sample average of ahe for men; from this compute an estimate of the "gender gap" in earnings; and construct the t-statistic testing the hypothesis that the gender gap is zero. Make sure to use heteroskedasticity-robust standard errors to answer this question. Compare your results to those in parts (a) and (b).

Solution. (1) Coefficient on female is -3.2026. It means that women on average receive hourly by 3.2026 less dollars. It means that women receive (other factors excluded) by $3.2026 * 8 * 30 = 768.6$ dollars less monthly than men. In terms of rubles it would be around 50k per month. Honestly, it is difficult to say, whether women in Russia receive so much less compared to men, because a lot of factors (such as education for example) are not taken into consideration.

(2) Lets look at result of regressing *ahe* on *female* factor, using robust standart errors. One can see that slope coefficient on *female* variable is statistically different from zero, which means that average hourly earnings are different for men and women.

Table 3

	Dependent variable:
female	-3.203^{***}
	(0.251)
Constant	16.595^{***}
	(0.185)
Note:	*p<0.1; **p<0.05; ***p<0.01

(3) Average *ahe* for women is **13.392** and average *ahe* for men is **16.595**. Estimate for gender gap is then: $16.595 - 13.392 = 3.203$. Standart error for gender gap is the following:

$$\sqrt{\frac{s_f^2}{n_f} + \frac{s_m^2}{n_m}} = \sqrt{\frac{s_f^2}{n_f} + \frac{s_m^2}{n_m}} = 0.25$$

and finally constructing the t-statistic:

$$t = \frac{3.203}{0.25} = 12.8 > 1.96$$

which implies that difference in means is significantly different from zero, which coincides with the results from points (a) and (b)

□