Fahri Ulkat   090220756

Question1:

a. Population Regression Line: The line that shows the relationship between variables on avarage within population.

b. Dependent Variable: The dependent variable is a variable that we are trying to predict with changes in independent variable. Also known as outcome variable.

c. Independent Variable: The variable used to explain dependent variable. It is not influenced by the dependent variable. Also known as explanatory variable.

d. Regressor: Another term for an independent variable which helps the the dependent variable.

e. Parameters: They are fixed values we are trying to use to define the relationship between variables.

f. Error Term: The difference between predicted value (E(Y|X)) and observed value of Y.

g. Ordinary Least Squares Estimators: Estimator which minimizes the loss function. It provides to find best parameters.

h. Predicted Value: The value that we obtained from the population regression line. The value that contain error which represents the diffrence between observed and predictel value.

i. Residual: The difference between observed and predicted value of Y.

j. Regression R2: It is a value between 0 and 1 which helps us to measure the how succesful X exlpains Y.

k. Standard Error of the Regression (SER): A value that measures the avarage distance of observe values from the regression line.

l. Least Squares Assumptions: Key assumptions needed for OLS. There are three assumption we covered: Zero-conditional mean, Random sampling, No large outliers assumption.

Question2:

A researcher, using data on class size (CS) and average test scores from 100 third-grade classes wants to estimate the following regression model:

$$TestScore = \beta_0 + \beta_1 \times CS + u.$$

Using R, the researcher obtained the following estimated regression model:

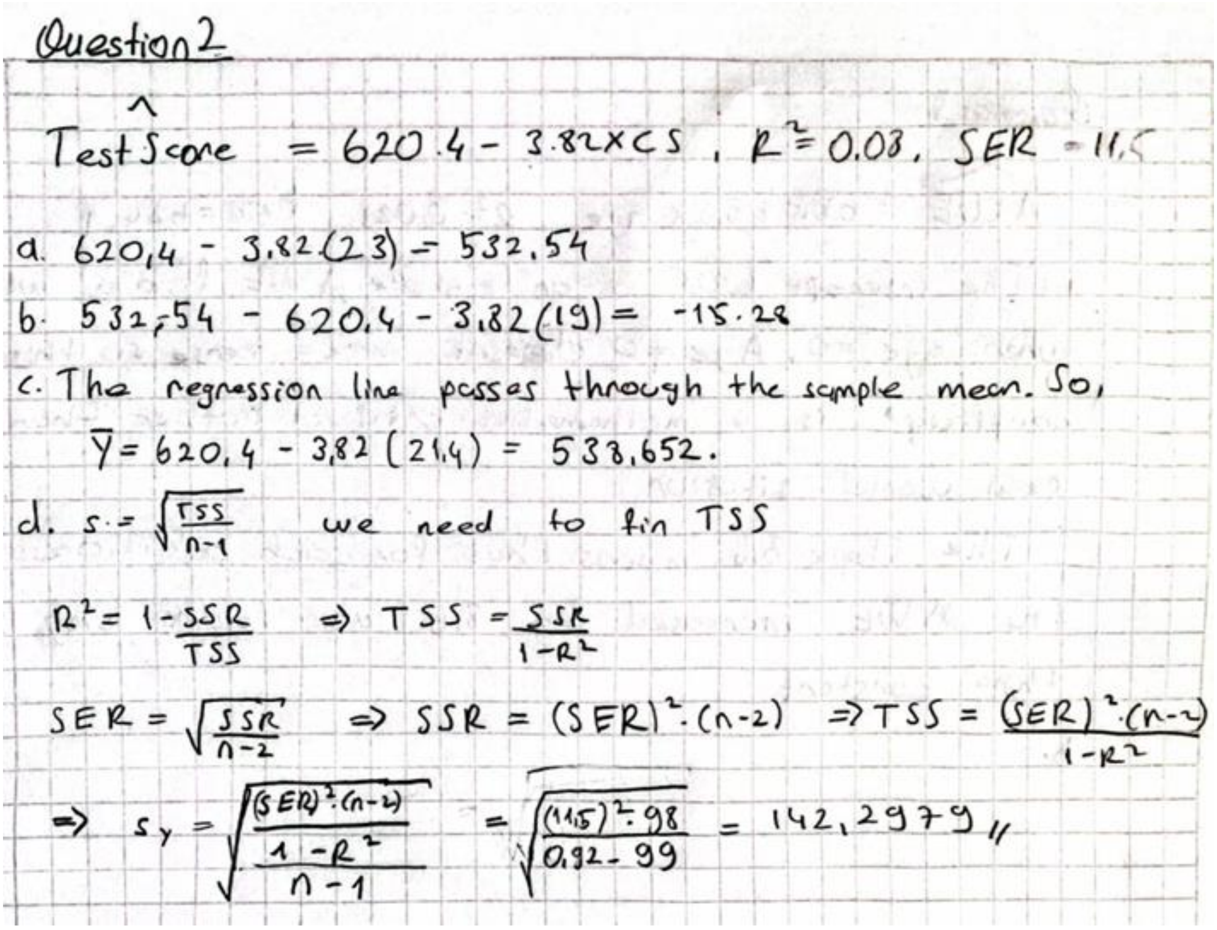
$$\widehat{TestScore} = 620.4 - 3.82 \times CS, \quad R^2 = 0.08, \quad SER = 11.5.$$

a. A classroom has 23 students. What is the regression's prediction for that classroom's average test score?

b. Last year a classroom had 19 students, and this year it has 23 students. What is the regression's prediction for the change in the classroom's average test score?

c. The sample average class size across the 100 classrooms is 21.4. What is the sample average of the test scores across the 100 classrooms? ($Hint$: Review the formulas for the OLS estimators.)

d. What is the sample standard deviation of test scores across the 100 classrooms? ($Hint$: Review the formulas for the R^2 and SER.)



Question 2

$\widehat{TestScore} = 620.4 - 3.82 \times CS$, $R^2 = 0.08$, $SER = 11.5$

a. $620{,}4 - 3.82(23) = 532.54$

b. $532{,}54 - 620.4 - 3.82(19) = -15.28$

c. The regression line passes through the sample mean. So,

$\bar{Y} = 620.4 - 3.82\,(21.4) = 538.652.$

d. $s = \sqrt{\frac{TSS}{n-1}}$ we need to fin TSS

$R^2 = 1 - \frac{SSR}{TSS} \Rightarrow TSS = \frac{SSR}{1-R^2}$

$SER = \sqrt{\frac{SSR}{n-2}} \Rightarrow SSR = (SER)^2 \cdot (n-2) \Rightarrow TSS = \frac{(SER)^2 \cdot (n-2)}{1-R^2}$

$\Rightarrow s_y = \sqrt{\frac{\frac{(SER)^2 \cdot (n-2)}{1-R^2}}{n-1}} = \sqrt{\frac{(11.5)^2 \cdot 98}{0.92 \cdot 99}} = 142{,}2979$ //

Question3:

$$\widehat{AWE} = 630.5 + 8.6 \times Age, \quad R^2 = 0.023, \quad SER = 624.1.$$

a. Interpret the estimated coefficients.

The intercept 630.5 is an estimate AWE for a worker when Age=0 but in this case it does not make sense. So, this coefficient is a mathematical artifact rather than a real world situation.

The slope 8.6 means that, for each additional age AWE incread 8.6 when holding other things constant.

b. Interpret the SER measure. What are the units of measurement for the ? (Dollars? Years? or unit free?)

The Standart Error of the Regression is an estimator of the standart deviation of the regression error terms. In this case SER is a measure of the difference between observed and predicted values of AWEs.

The unit of SER is the same unit with Y so it is dollar.

c. Interpret the R2 measure. What are the units of measurement for the ? (Dollars? Years? or unit free?)

The R2 score is a measure to find out how much X explain Y. In this case R2=0.023 means X explain Y very weakly.

R2 is a value between 0 and 1 in every case, so it is unit free.

d. What is the predicted average weekly earnings for a worker who is ? What is the predicted average weekly earnings for a worker who is ?

For age=23 -> 630.5 + 8.6 x 23 = 828.3

For age=40 -> 630.5 + 8.6 x 40 = 974.5

e. Will the regression give reliable predictions for a worker who is years old? Explain.

It will not give reliable predictions because dataset contains only workers aged 25-65. Even if it included higher ages, SLR may not be the best option to make a prediction because until some ages peoples income increase but after that it will start to decrease because they will get old.

f. Given what you know about the distribution of earnings, do you think it is plausible that the distribution of errors in the regression model is normal? (Do you think that the distribution is symmetric or skewed? What is the smallest value of earnings, and is it consistent with a normal distribution?)

Income data usually have a right-skewed distribution because there is usually a minimum wage but no upper limit, which leads to high outliers. This skewness can cause the error terms in the regression model to be non-normally distributed. Since AWE cannot be negative, a perfect normal distribution is unlikely, since a normal distribution requires symmetry. Therefore the errors in the model are also likely to be skewed.

g. The average age in the sample is 38. What is the average value of AWE in the sample?

Since the regression line must pass through tha sample mean,

630.5 + 8.6 x (38) = 957.3

Question4:

I have prepared a jupyter notebook for question 4 and converted to html file. You can find both in zipped file.