ECON471
Fall 2020
Problem Set 1
Due Wednesday September 16, by 11:59pm, cst.

Name: Yang Wenxiao.

Section: B3.

1. The data in MEAP01 are for the state of Michigan in the year 2001. Use these data to answer the following questions.

(i) Find the largest and the smallest values of *math4*. Does the range make sense? Explain.

Largest : 100 , smallest : 0.

Doesn't make sense : we didn't delete outlier, 0 may be an outlier.

(ii) How many schools have a perfect pass rate on the math test? What percentage is this of the total sample?

38

2.084476%

(iii) How many Schools have math pass rates of exactly 50%?

55

(iv) Compare the average pass rates for the math and reading scores. Which test is harder to pass?

average pass rates of math : 71.909

reading scores : 60.06188

=> reading is harder to pass.

(v) Find the correlation between *math4* and *read4*. What do you conclude?

(vi) The variable *exppp* is expenditure per pupil. Find the average of *exppp* along with its standard deviation. Would you say there is wide variation in per pupil spending?

2. The data in JTRAIN2.TXT come from a job training experiment conducted for low-income men during 1976-1977.

(i) Use the indicator variable *train* to determine the fraction of men receiving job training.

(ii) The variable *re78* is earnings from 1978, measured in thousands of 1982 dollars. Find the averages of *re78* for the sample of men receiving job training and the sample not receiving job training. Is the difference economically large?

(iii) The variable *unem78* is an indicator of whether a man is unemployed or not in 1978. What fraction of the men who received job training are unemployed? What about for men who didn't receive job training? Comment on the difference.

(iv) From parts (ii) and (iii), does it appear that the job training program was effective? What would make our conclusion more convincing?

3. (Please do this problem without using any statistical software.)The following table contains the *ACT* scores and the *GPA* (grade point average) for eight college students. Grade point average is based on a four-point scale and has been rounded to one digit after the decimal.

Students	*GPA*	*ACT*
1	2.8	21
2	3.4	24
3	3.0	26
4	3.5	27
5	3.6	29
6	3.0	25
7	2.7	25
8	3.7	30

(i) Estimate the relationship between *GPA* and *ACT* using OLS; that is, obtain the intercept and slope estimates in the equation

$$\widehat{GPA} = \hat{\beta}_0 + \hat{\beta}_1 ACT.$$

Comment on the direction of the relationship. Does the intercept have a useful interpretation here? Explain. How much higher is the *GPA* predicted to be if the *ACT* score is increased by five points?

(ii) Compute the fitted values and residuals for each observation, and verify that the residuals (approximately) sum to zero.

(iii) What is the predicted value of *GPA* when *ACT* = 20?

(iv) How much of the variation in *GPA* for these 8 students is explained by *ACT*? Explain.

4.

(i) Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the intercept and slope from the regression of y_i on x_i using n observations. Let c_1 and c_2, with $c_2 \neq 0$, be constants. Let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be the intercept and slope from the regression of $c_1 y_i$ on $c_2 x_i$. Show that $\tilde{\beta}_1 = \left(\frac{c_1}{c_2}\right)\hat{\beta}_1$ and $\tilde{\beta}_0 = c_1\hat{\beta}_0$. [Hint: To obtain $\tilde{\beta}_1$, plug the scaled versions of *x* and *y* into $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$. Then use $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ for $\tilde{\beta}_0$, being sure to plug in the scaled *x* and *y* and the correct slope.]

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} (C_2 X_i - C_2\bar{X})(C_1 Y_i - C_1\bar{Y})}{\sum_{i=1}^{n} (C_2 X_i - C_2\bar{X})^2}$$

(ii) Consider the following model:

$$y_i = \beta_0 + u_i$$

Derive the OLS estimator for β_0.

$$\hat{y}_i = \hat{\beta}_0$$

$$\sum (y_i - \hat{\beta}_0)^2 = \sum (y_i^2 + \hat{\beta}_0^2 - 2\hat{\beta}_0 y_i)$$

$$= \sum y_i^2 + n\hat{\beta}_0^2 - 2\beta_0 \sum y_i$$

$$2n\hat{\beta}_0 - 2\sum y_i = 0. \qquad \hat{\beta}_0 = \frac{1}{n}\sum y_i$$

5. The data set in CEOSAL2.TXT contains information on chief executive officers for U.S. corporations. The variable *salary* is annual compensation, in thousands of dollars, and *ceoten* is prior number of years as company CEO.

(i) Find the average salary and the average tenure in the sample.

(ii) How many CEOs are in their first year as CEO (that is, ceoten=0)? What is the longest tenure as a CEO?

(iii) Estimate the simple regression model

$$\log(salary) = \beta_0 + \beta_1 ceoten\text{+u},$$

And report your results in the usual form. What is the (approximate) predicted percentage increase in salary given one more year as a CEO?

6. Use the data in WAGE2.TXT to estimate a simple regression explaining monthly salary (*wage*) in terms of IQ score (*IQ*).

(i) Find the average salary and average IQ in the sample. What is the sample standard deviation of IQ? (IQ scores are standardized so that the average in the population is 100 with a standard deviation of 15.)

(ii) Estimate a simple regression model where a one-point increase in *IQ* changes *wage* by a constant dollar amount. Use this model to find the predicted increase in *wage* for an increase in *IQ* of 15 points. Does *IQ* explain most of the variation in *wage*?

(iii) Now, estimate a model where each one-point increase in *IQ* has the same percentage effect on *wage*. If *IQ* increases by 15 points, what is the approximate percentage increase in predicted *wage*?

7. Use the data in MEAP93.TXT to explore the relationship between the math pass rate (*math10*) and spending per student (*expend*).

(i) Do you think each additional dollar spent has the same effect on the pass rate, or does a diminishing effect seem more appropriate? Explain.

(ii) In the population mode

$$math10 = \beta_0 + \beta_1 \log{(expend)}\text{+u},$$

argue that $\beta_1/10$ is the percentage point change in *math10* given a 10% increase in *expend*.

(iii) Estimate the model in part (ii). Report the estimated equation in the usual way, including the sample size and R-squared.

(iv) How big is the estimated spending effect? Namely, if spending increases by 10%, what is the estimated percentage point increase in *math10*?

(v) One might worry that regression analysis can produce fitted values for *math10* that are greater than 100. Why is this not much of a worry in this data set?

8. To complete this exercise you need to use *R* to generate data from the uniform and normal distributions.

(i) Start by generating 500 observations x_i – the explanatory variable – from the uniform distribution with range [0, 10]. What are the sample mean and sample standard deviation of the x_i?

x = runif (500, 0, 10)

mean(x)   sd(x)

(ii) Randomly generate 500 errors, u_i, from the normal [0, 36] distribution (i.e., a normal random variable with mean zero and variance 36). Is the sample average of the u_i exactly zero? Why or why not? What is the sample standard deviation of the u_i?

this is sample

u = rnorm (500, 0, 6).

mean (u)   sd(u)

(iii) Now generate the y_i as

$$y_i = 1 + 2x_i + u_i = \beta_0 + \beta_1 x_i + u_i;$$

That is, the population intercept is one and the population slope is two. Use the generated data to run the regression of y_i on x_i. What are your estimates of the intercept and slope? Are they equal to the population values in the above equation? Explain.

they are sample   x equal.

(iv) Obtain the OLS residuals, $\hat{u}_i$, and verify that equations $\sum_{i=1}^{n} \hat{u}_i = 0$ and $\sum_{i=1}^{n} x_i \hat{u}_i = 0$ hold (subject to rounding errors).

Sum(resid (A))

sum (X * sesid (A)) .

(v) Compute the same quantities in part (iv) but use the errors u_i in place of the residuals. Now what do you conclude?

Sum (U) may not be very close to zero.

(vi) Repeat parts (i), (ii), and (iii) with a new sample data, starting with generating the x_i. Now what do you obtain for $\hat{\beta}_0$ and $\hat{\beta}_1$? Why are these different from what you obtained in part (iii)?