Albert Wiryawan
ECON 471
673431511
avw2@illinois.edu

Problem Set #1

*NOTE: exact outputs can be seen run through the cells of the attached R markdown file.

Question #1

```{r}
#load in the initial data michigan 01 data set
setwd("/Users/albertwiryawan/Code/Class_Repos/Econometrics/Problem Set #1")
mich_01_data <- read.table("meap01.txt")
print(mich_01_data)
attach(mich_01_data)
```

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

```{r}
#based on the meap01 description given we know that math4 is given by feature V4
#We should also note that math4 represents student's satifactory fourth grade math
math4 = V3
highest_value = max(math4)
lowest_value = min(math4)
print(highest_value)
print(lowest_value)
```

The highest value in this range is 100 and the lowest value in this range is 0. Since each row resembles a school's fourth grade math performance. The number for each row in the table for column math4 corresponds to the average satisfactory or pass rate for a school's fourth grade students in math. In turn, this range does make sense for math4.

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

```{r}
perfect_pass=length(math4[math4 == 100])
total = length(math4)
print(perfect_pass)
print(total)
perfect_pass/total

```
First the number of school's that have a perfect pass rate was found to be 38. While the whole data set was comprised of 1823 observations. The percentage of school's that have a perfect pass rate is thus 2.08%.

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

```{r}
half_pass_rate = length(math4[math4 == 50])
print(half_pass_rate)
```

The number of schools that have exactly 50% pass rate was found to be 17

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

```{r}
read4  = V4
average_reading = mean(read4)
average_math = mean(math4)
print(average_reading)
print(average_math)

```
The average pass rate for fourth grade reading was 71.9, while the average pass rate for fourth grade math was 60.06. This would mean that the reading test is harder to pass.

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

```{r}
cor(math4, read4, method = "pearson")

```

The correlation between these two variables is 0.843. This resembles a high linear association between the two, since pearson correlation is between values of 0 and 1. The value of this pearson correlation makes sense. School's that have students that tend to do well in one test will do well in the other.

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

```{r}
exppp = V8
mean(exppp)
sd(exppp)
summary(exppp)
```

The average exppp is approximately \$5194.87. the standard deviation of the distribution of exppp among schools in Michigan in 2001 is \$1091.89. The minimum is about 1207 and max is 11958 this would show that there is pretty large variation throughout the data.

Question #2

```{r}
# load the data set
jtrain2 = read.table("jtrain2.txt")
print(jtrain2)
attach(jtrain2)
```

(i) Use the indicator variable train to determine fraction of men receiving job training.
```{r}
train = V1
men_getting_job_training = length(train[train == 1 ])
total = length(train)
print(men_getting_job_training)
print(total)
print(men_getting_job_training/total)

```

The total number of men receiving job training is 185 out of 445. As a result, the fraction of men receiving job training will be 41.57%

(ii) Find the averages of re78 for the sample of men receiving job training and the sample of men not receiving job training. Is the difference economically large?

```{r}
re78_with_training = mean(jtrain2$V11[jtrain2$V1==1])
re78_no_training = mean(jtrain2$V11[jtrain2$V1==0])
print(re78_with_training)
print(re78_no_training)
```

The average re78 for the sample of men receiving job training is 6.35 thousands of 1982 dollars. The average re78 for the sample of men that did not receive job training is 4.55 thousands of 1982 dollars. Based on this, there is an economical difference of 1.79 thousands of 1982 dollars. Thus, there is a significant economical difference

(iii) 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.

```{r}
total_trained = nrow(subset(jtrain2, V1==1))
total_untrained = nrow(subset(jtrain2, V1==0))
count_unemployed_with_training = nrow(subset(jtrain2, V1==1 & V14==1))
count_unemployed_no_training = nrow(subset(jtrain2, V1==0 & V14==1))
print(count_unemployed_with_training)
print(count_unemployed_no_training)
fraction_job_train_unemp = count_unemployed_with_training/total_trained
fraction_no_job_train_unemp = count_unemployed_no_training/total_untrained
print(fraction_job_train_unemp)
print(fraction_no_job_train_unemp)
```

The total number of men that received training is 185. Of that, 45 men are still unemployed resulting in a percentage of 24.32% The total number of men that did not receive training is 260. Of that, 92 men are unemployed resulting in a percentage of 35.38%. Thus, given that the person is untrained they will be more likely to be unemployed.

(iv)
Based on the relative ratios created, there is a 24.32% chance that someone that is trained is unemployed and 35.38% change that someone that is NOT trained is unemployed. Due to these percentages, it becomes apparent that the training program was effective in reducing unemployment, thus indicating a successful training program.

Question #3
*Attached scanned pdf of work for this problem below

Question #4
*Attached scanned pdf of work for this problem below

Question #5
```{r}
# load the data set
ceosal = read.table("ceosal2.txt")
print(ceosal)
attach(ceosal)
```

(i) Find the average salary and the average tenure in the sample
```{r}
salary = V1
tenure = V6
mean(salary)
mean(tenure)
```

The average salary is 865.86 thousand dollars and the average prior number of years as company CEO is 7.95 years.

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

```{r}
count_first_year_ceo = length(tenure[tenure == 0])
print(count_first_year_ceo)
longest_tenure = max(tenure)
print(longest_tenure)

```
There are 5 CEOs in their first year as CEO. The longest tenure is 37 years.

(iii) Estimate the simple regression mode
```{r}
regression = lm(log(salary)~tenure, data = ceosal)
summary(regression)
```

Based on the linear predictive model created above the formulated prediction model equation is: log(salary)= 6.505 +0.0097(ceoten)

Question #6

```{r}
# load data
wage2 = read.table("wage2.txt")
print(wage2)
attach(wage2)
```

(i) Find average salary and average IQ in the sample. What is the sample standard deviation of IQ?

```{r}
salary = V1
IQ = V3
mean_salary = mean(salary)
mean_IQ = mean(IQ)
sd_IQ = sd(IQ)
print(mean_salary)
print(mean_IQ)
print(sd_IQ)
```

The average salary is about \$957.95 and the average IQ is about \$101.28. The sample standard deviation of IQ is about 15.05. This is close to the standard deviation of 15 given by the population.

(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?

```{r}
regress_model = lm(wage2$V1~wage2$V3)
summary(regress_model)
```

```{r}
print(8.3*15)
```

The linear regression model generated is 116.9916 + 8.3031*(IQ) + u. Thus, an increase in wage for an increase in IQ of 15 points is 124.5. The r-square is a metric that helps explain the total variation explained by the regression line, since this value is 9.5% this means that the regression line does not cover much of the variation.

(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?

```{r}
regress = lm(log(wage2$V1)~wage2$V3)
summary(regress)
print(100*0.0088072*15)
```

The model that was estimated is given by log(wage) = 5.8869943 + 0.0088072*IQ+u. 15 more points in IQ, will increase wage by 13.2%. and r-square tells us that 9.813% of the variation is explained by this model.

Question #7

```{r}
# load data
meap93 = read.table ("meap93.txt")
print(meap93)
attach(meap93)
```

(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

```{r}
plot(meap93$V4, meap93$V9)
```

Based on the plot produced above graphing the MEAP pass rate with respect to expenditure per student. It can be seen that an increase in expenditure does not increase pass rate significantly which further emphasis the effect of diminishing marginal return on the dollar. Meaning that as more and more money is spent per student there is less of a return -- that is more students passing the Meap exam.

(ii)

*Attached scanned pdf of work for this problem below

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

```{r}
regression = lm(meap93$V9~log(meap93$V4))
summary(regression)
```

The predictive regression model made is defined by math10 = -69.341 + 11.164log(expand) + u. The adjusted R-square is 0.02727 and the sample has n = 408 observations

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

The coefficient of log(expand) is 11.16439 and this coefficient is significant at the 5% significance level as given by Pr(>|t|) being smaller than 5%. Thus, an increase in spending of 10% will increase the estimated percentage point increase in math10 by B1/10 which is 1.116439 percentage points.

(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?

```{r}
expanditure = V4
max(V4)
```

The value of math10 will never surpass 100 as the highest value for expand in the data set is $7419 and this will never result in math10 being larger than 100.

Question #8
(i)What are the sample mean and sample standard deviation of the x?

```{r}
my_data = runif(500, 0, 10)
mean(my_data)
sd(my_data)
```

The sample mean is 5.137937 and the sample standard deviation is 2.983049.

(ii)generate 500 from normal distribution [0,36]. Is the sample average of the ui exactly zero? Why or Why not? What is the sample standard deviation of ui

```{r}

normal_data = rnorm(500, mean = 0, sd=8)
```

```
mean(normal_data)
sd(normal_data)
```

The sample average is not exactly zero. This is because the random points are sampled from a normal distribution that has a mean of zero. Several of these sample distributions combined will generate a population distribution of 0 as given by the central limit theorem. The sample standard deviation is 7.87526 which is close to the square root of the variance of standard deviation of the population distributions sampled from which is 8.

(iii) What are your estimates of the intercept and slope? Are they equal to the population values in the above equation?

```{r}

X = rnorm(500, mean = 0, sd=8)
Y = 1 + 2 * X
regression = lm(Y ~ X)
summary(regression)
```

Although the estimated model is given roughly by Y = 1 + 2 * X (program states that is a close match). The generated model is made by randomly sampled points from a normal distribution following the given linear regression equation. As a result the equations are independent of each other and can be different.

(iv)

```{r}
X = rnorm(500, mean = 0, sd=8)
Y = 1 + 2 * X
regression = lm(Y ~ X)
residuals = resid(regression)
total_resid = sum(residuals)
print(total_resid)
temp = sum(total_resid * X)
print(temp)

```

As verified by the code above summation of the residuals are close to zero but not exactly. When creating the model the residuals between the data points are reduced.

(v)

```{r}
X = rnorm(500, mean = 0, sd=8)
Y = 1 + 2 * X
```

```
regression = lm(Y ~ X)
summary(regression)
error= 1.058e-14
print(error)
temp2 = sum(error * X)
print(temp2)
```

The calculated errors given by the code above is close to zero as the error is also minimized as the model is created. Still there is some error from the given population model which shows that the sampled points from the normal distribution does not exactly match up with Y = 1 + 2 * X

(vi)

```{r}
X = rnorm(500, mean = 0, sd=8)
Y = 1 + 2 * X
regression = lm(Y ~ X)
summary(regression)
```

The obtained values for B1 and B0 is close to 2 and 1 respectively. However, these values are still different because they are taken from a new sample of 500 points randomly generated from a population model that follows the linear equation Y = 1 + 2 * X

Question #3

i $y_i = GPA$; $x_i = ACT$; $n = 8$

$\bar{x} = 207/8 = 25.875$

$\bar{y} = 25.7/8 = 3.2125$

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 5.8125$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 56.875$$

$\hat{B}_1 = 5.8125/56.875 = 0.1022$

$\hat{B}_0 = \bar{y} - \hat{B}\bar{x}$

$\hat{B}_0 = 3.2125 - 0.1022(25.875) \approx 0.5681$

$$\boxed{\hat{GPA} = 0.5681 + 0.1022 ACT}$$

ii

$-0.0857 + -0.3791 + 0.2253 + -0.1725 + -0.0681 +$
$0.1231 + 0.4231 + -0.0659 = \boxed{-.0002}$

A residual is taken for each observational GPA by comparing predicted GPA with the values generated by the model. Summing up all the residuals, the total is approximately 0.

iii $\hat{GPA}_f = 0.5681 + 0.1022(20)$ ACT = 20

$= 2.6121$

$\boxed{\approx 2.61}$

The predicted GPA is approximately 2.61

iv (SSR) Sum of Square Residuals = 0.4347 (data point)

(SST) Total sum of square residuals = 0.10288 (predicted)

$$R\text{-square} = 1 - \frac{SSR}{SST} = 1 - \frac{0.4347}{1.0288} = 0.577 \Rightarrow \boxed{57.7\%}$$

Question #4

i $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad c_1 \hat{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 \cdot c_2 x_i$

show $\tilde{\beta}_1 = \left(\frac{c_1}{c_2}\right)\hat{\beta}_1 \qquad \tilde{\beta}_0 = c_1 \hat{\beta}_0$

$$\rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} (c_2 x_i - c_2 \bar{x})(c_1 y_i - c_1 \bar{y})}{\sum_{i=1}^{n} (c_2 x_i - c_2 \bar{x})^2} = \frac{c_1 c_2 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{c_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \left(\frac{c_1}{c_2}\right)\hat{\beta}_1 = \tilde{\beta}_1$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\tilde{\beta}_0 = c_1 \bar{y} - c_2 \tilde{\beta}_1 \bar{x} = c_1 \bar{y} - c_2 \left(\frac{c_1}{c_2}\right)\hat{\beta}_1 \bar{x} = c_1(\bar{y} - \hat{\beta}_1 \bar{x}) = c_1 \hat{\beta}_0$$

ii $y = \beta_0 + u$

$SSR = \sum (y_i - \hat{y}_i)^2 \qquad \frac{d(SSE)}{d\hat{\beta}_0} = d\left(\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 X_i)\right)$

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

$$\frac{d(SSE)}{d\beta_0} = 0$$

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

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

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

Question #7

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

$$\frac{\partial \text{math10}}{\partial \text{expend}} = \frac{\beta_1}{\text{expend}}$$

$$\partial \text{math10} = \left(\frac{\partial \text{expend}}{\text{expend}}\right) \beta_1$$

$$\frac{\partial \text{expend}}{\text{expend}} = 0.1$$

$$\partial \text{math10} = 0.1 \cdot \beta_1$$

Thus, $\frac{\beta_1}{10}$ is percentage point change in math w/ 10% increase in expend ftare