ECON471
Fall 2020
Problem Set 3
Due Tuesday October 27, by 11:59pm

Name: Albert Wiryawan

Section: B3

1. For the case of the multiple regression problem with two explanatory variables, $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i$, show that minimizing the sum of squared residuals results in three conditions:

$$\sum_{i=1}^{n} \hat{u}_i = 0; \sum_{i=1}^{n} \hat{u}_i X_{i1} = 0; \sum_{i=1}^{n} \hat{u}_i X_{i2} = 0$$

Attached to the end of this document

2. The following model can be used to study whether campaign expenditures affect election outcomes:

$$VoteA = \beta_0 + \beta_1 \log(expendA) + \beta_2 \log(expendB) + \beta_3 prtystrA + u,$$

where $VoteA$ is the percentage of the vote received by Candidate A, $expendA$ and $expendB$ are campaign expenditures by Candidates A and B, and $prtystrA$ is a measure of party strength for Candidate A (the percentage of the most recent presidential vote that went to A's party).

Code:

```{r}
setwd("/Users/albertwiryawan/Code/Class_Repos/Econometrics/Problem Set #3")

#load in Attend txt data set for the exercise
vote1 <- read.table("vote1.txt")

attach(vote1)

voteA = V4
expendA = V5
expendB = V6
prtystrA = V7
model1 = lm(voteA ~ log(expendA) + log(expendB) + prtystrA)
summary(model1)
```

(i) What is the interpretation of β_1?

$\frac{\beta_1}{100}$ (ceteris paribus) – percent point change in percentage of votes received by A when the expenditure by candidate A increases by one percent.

(ii) In terms of the parameters, state the null hypothesis that a 1% increase in A's expenditures is offset by a 1% increase in B's expenditures.

H_0: $\beta_2 = -\beta_1$
H_1: $\beta_2 \neq -\beta_1$

(iii) Estimate the given model using the data in VOTE1.TXT and report the results in usual forms. Do A's expenditures affect the outcome? What about B's expenditures? Can you use these results to test the hypothesis in part (ii)?

```
Call:
lm(formula = voteA ~ log(expendA) + log(expendB) + prtystrA)

Residuals:
     Min       1Q   Median       3Q      Max
-20.3990  -5.4184  -0.8737   4.9563  26.0575

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.08788    3.92680  11.482   <2e-16 ***
log(expendA)  6.08136    0.38211  15.915   <2e-16 ***
log(expendB) -6.61563    0.37889 -17.461   <2e-16 ***
prtystrA      0.15201    0.06203   2.451   0.0153 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.713 on 169 degrees of freedom
Multiple R-squared:  0.7925,    Adjusted R-squared:  0.7888
F-statistic: 215.2 on 3 and 169 DF,  p-value: < 2.2e-16
```

Vote(A) = 45.09 + 6.08log(expendA) - 6.62log(expendB) + 0.15(prtystrA)

From the summary of statistics of the created predicted model, it is seen that the t-value for (expendA) is (15.92) and (-17.46) for (expend) which is significant. Candidate A's expenditure do affect the outcome since a 10% ceteris paribus increase in spending by candidate A increases the predicated share of the vote going to A by 0.61 percentage points. B's expenditure also affects the percentage of votes going to candidate A. With a 10% increase ceteris paribus increase in spending by candidate B reduces perencentage of votes going to candidate A by 0.66 percentage points. A test for the hypothesis test created in part (ii) is not sufficed by the data acquired solely from the data of the model above.

(iv) Estimate a model that directly gives the t statistic for testing the hypothesis in part (ii). What do you conclude? (Use a two-sided test.)

3. The following model is used to study the tradeoff between time spent sleeping and working and to look at other factors affecting sleep:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + u,$$

where *sleep* and *totwrk* (total work) are measured in minutes per week and *educ* and *age* are measured in years.

(i) If adults trade off sleep for work, what is the sign of β_1?

If adults trade off sleep for work, more work would result in less sleep considering all other things are constant. As such, the sign of β_1 is negative.

(ii) What signs do you expect for β_2 and β_3?

The signs for β_2 and β_3 are not exactly defined as the amount of education as well as a person's age can both reduce or increase the amount of sleep. This would explain why data would be necessary to model this relationship.

(iii) The equation was estimated using the data in SLEEP75:

$$\widehat{sleep} = 3638.25 - 0.148\, totwork - 11.13\, educ + 2.20\, age$$

$$(112.28) \quad (.017) \qquad\qquad (5.88) \qquad (1.45)$$

$$n = 706,\ R^2 = 0.113.$$

where the standard errors of the estimates are shown in parentheses. If someone works five more hours per week, by how many minutes is sleep predicted to fall? Is this a large trade off?

If someone works five more hours per week, they will work for 300 minutes. Thus, sleep is predicted to fall by (0.148) * (300) = 44.4 minutes. In the context of a week, a reduction of 45 minutes of sleep per week is not a large trade off.

(iv) Discuss the sign and magnitude of the estimated coefficient on *educ*.

More education results in less predicted time spent sleeping because the predicted model gave a negative coefficient for educ. The magnitude of the coefficient is 11.13 and keep

note that this value is negative. As such, if education increase by one year sleeping time is expected to decrease by the magnitude of 11.13 minutes.

(v) Would you say *totwrk*, *educ*, and *age* explain much of the variation in *sleep*?

The given R^2 is 0.113. As such, this shows that the three explanatory variables given by the predicted model only explains 11.3% of the total variation. Thus, a large percentage of variation remains unexplained.

(vi) Is either *educ* or *age* individually significant at the 5% level against a two-sided alternative? Show your work.

Degrees of freedom (df) =702 and the standard normal critical value from the table is 1.96 at the 5% significance level. The t-statistic can thus be calculated by the expression below for age and educ

-11.13(educ coeff value) / 5.88 (t-statistic for educ) = -1.89
2.20(age coeff)/ 1.45(t-statistic for age) = 1.52

Since 1.89 < 1.96, educ is not significant at the 5% significance level. Also since the absolute value of the calculated t-value for age is 1.52 this is also not statistically significant.

(vii) Dropping *educ* and *age* from the equation produces

$$\widehat{sleep} = 3586.38 - 0.151\ totwork$$

(38.91) (.017)

$n = 706,\ R^2 = 0.103.$

Are *educ* and *age* jointly significant in the original equation at the 5% level of significance? Justify your answer.

The F statistic is calculated in order to assess joint significance.

F-test = (.113-.103)/(1-1.113) * (702/2) = 3.96

The 5% critical value in the F distribution from the table is 3.00. Thus, educ and age are jointly significant at 5% level since 3.96 > 3.00

(viii) Suppose that the sleep equation contains heteroskedasticity. What does this mean about the tests computed in parts (vi) and (vii)?

The t and F statistics that are used assume homoscedasticity. As a result, if there is heteroskedasticity in the equation, the tests are not valid and results are biased.

4. Use the data in WAGE2.TXT for this exercise.

Code

```{r}
setwd("/Users/albertwiryawan/Code/Class_Repos/Econometrics/Problem Set #3")

#load in Attend txt data set for the exercise
wage2 <- read.table("wage2.txt")

attach(wage2)

educ = V5
exper = V6
tenure = V7
wage = V1
mod = lm(log(wage) ~ educ + exper + tenure)
summary(mod)
```

(i) Consider the standard wage equation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 \text{exper} + \beta_3 tenure + u.$$

State the null hypothesis that another year of general workforce experience has the same effect on *log(wage)* as another year of tenure with the current employer.

H_0: $\beta_2 = \beta_3$

(ii) Test the null hypothesis in part (i) against a two-sided alternative, at the 5% significance level, by constructing a 95% confidence interval. State your conclusion.

The 95% confidence interval is given by 0.1523 ± 0.00370. Since the value of 0 does not lie in the confidence interval then θ_2 is statistically different from zero at 5% level. Sp the researcher will fail to accept the null hypothesis given above.

5. The data set 401KSUBS.TXT contains information on net financial wealth (*nettfa*), age of the survey respondent (*age*), annual family income (*inc*), family size (*fsize*), and participation in certain pension plans for people in the United States. The wealth and income variables are both recorded in thousands of dollars. For this question use only the data for single-person households (so *fsize*=1).

(i) How many single-person households are there in the data set?

Code

```{r}
setwd("/Users/albertwiryawan/Code/Class_Repos/Econometrics/Problem Set #3")

#load in Attend txt data set for the exercise
ksubs401 <- read.table("401ksubs.txt")

attach(ksubs401)
nettfa = V7
age = V5
inc =V2
fsize = V6
single_person = length(which(fsize ==1))
single_person
```

There are 2017 single-person house in the dataset.

(ii) Use OLS to estimate the model

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 \text{age} + u,$$

and report the results using the usual format. Be sure to use only the single-person households in the sample. Interpret the slope coefficients. Are there any surprises in the slope estimates?

Code:

```{r}
setwd("/Users/albertwiryawan/Code/Class_Repos/Econometrics/Problem Set #3")

#load in Attend txt data set for the exercise
ksubs401 <- read.table("401ksubs.txt")

attach(ksubs401)

new_data = subset(ksubs401, fsize=='1')
nettfa = V7
age = V5
inc =V2
fsize = V6
mod = lm(new_data$V7 ~ new_data$V2 + new_data$V5)
```

summary(mod)

```
Call:
lm(formula = new_data$V7 ~ new_data$V2 + new_data$V5)

Residuals:
    Min      1Q  Median      3Q     Max
-179.95  -14.16   -3.42    6.03 1113.94

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -43.03981    4.08039 -10.548   <2e-16 ***
new_data$V2   0.79932    0.05973  13.382   <2e-16 ***
new_data$V5   0.84266    0.09202   9.158   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 44.68 on 2014 degrees of freedom
Multiple R-squared:  0.1193,    Adjusted R-squared:  0.1185
F-statistic: 136.5 on 2 and 2014 DF,  p-value: < 2.2e-16
```

The predicted model is found below.

Nettfa = -43.04 + 0.799(inc) + 0.84(age)

The coefficient on inc shows that an increase in the annual income of every 1000 will result in about 799 increase in net total financial assets. The coefficient in age shows that an increase in age by a year will increase net total financial assets by 840 dollars.

(iii) Does the intercept from the regression in part (ii) have an interesting meaning? Explain.

In the context of the predicted model developed the intercept is not that interesting. This is because the intercept is the intial netffa when age and inc is equal to 0. However, net total financial assets cannot be evaluated by someone at birth so this value would be meaningless as it is expected that this value would be zero.

(iv) Find the *p*-value for the test $H_0: \beta_2 = 1$ against $H_1: \beta_2 < 1$. Do you reject H_0 at the 1% significance level? Explain.

Using the output from part (ii) the p-value for β_2 is approximately 0 which is less than the considered level of significance 0.01. as a result, the null hypothesis is rejected at the 1% significance level.

(v) If you do a simple regression of *nettfa* on *inc*, is the estimated coefficient on *inc* much different from the estimate in part (ii)? Why or why not?

Code

```{r}
setwd("/Users/albertwiryawan/Code/Class_Repos/Econometrics/Problem Set #3")

#load in Attend txt data set for the exercise
ksubs401 <- read.table("401ksubs.txt")

attach(ksubs401)

new_data = subset(ksubs401, fsize=='1')
nettfa = V7
age = V5
inc =V2
fsize = V6
mod = lm(new_data$V7 ~ new_data$V2 )
summary(mod)
```

```
Call:
lm(formula = new_data$V7 ~ new_data$V2)

Residuals:
    Min      1Q  Median      3Q     Max
-185.12  -12.85   -4.85    1.78 1112.66

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.5709     2.0607   -5.13 3.18e-07 ***
new_data$V2   0.8207     0.0609   13.48  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45.59 on 2015 degrees of freedom
Multiple R-squared:  0.08267,   Adjusted R-squared:  0.08222
F-statistic: 181.6 on 1 and 2015 DF,  p-value: < 2.2e-16
```

The coefficient on inc in the simple regression output is 0.821 which is not very different from the one obtained when modeled with independent variables of age and inc together which was found to be 0.799.

6. Regression analysis can be used to test whether the market efficiently uses information in valuing stocks. For concreteness, let *return* be the total return from holding a firm's stock over the four-year period from the end of 1990 to the end of 1994. The *efficient market hypothesis* says that these returns should not be systematically related to information known in 1990. If firm characteristic known at the beginning of the period help to predict stock returns, then we could use this information in choosing stocks.

For 1990, let *dkr* be a firm's debt to capital ratio, let *eps* denote the earnings per share, let *netinc* denote net income, and let *salary* denote total compensation for the CEO.

(i) Using the data in RETURN, the following equation was estimated:

$$\widehat{return} = -14.37 + .321\, dkr + .043\, eps - .0051\, netinc + .0035\, salary$$

$$(6.89) \quad (.201) \qquad (.078) \qquad (.0047) \qquad (.0022)$$

$$n = 142, \quad R^2 = .0395.$$

Test whether the explanatory variables are jointly significant at the 5% level. Is any explanatory variable individually significant?

In order to test whether the explanatory variables are jointly significant or not at the 5% significance level we look at the t and F statistic.

The calculated t statistic for dkr is found to be

0.321/ 0.201 = 1.597

Since this value is less than 1.96 this implies that dkr is insignificant at the 5% significance level.

The calculated t statistic for eps is given by

0.043/ 0.078 = 0.551

Since this value is less than 1.96 this implies that eps is insignificant at the 5% significance level.

The calculated t statistic for netinc is given by

-0.0051/0.0047 = 1.08

Since this value is less than 1.96 this means that netinc is insignificant at the 5% significance level

The calculated t statistic for salary is given by

0.0035/ 0.0022 = 1.59

Since this value is less than 1.96 this means that salary is insignificant at the 5% significance level

As a result, all explanatory variablesa re insignificant at the 5% significance level.

The F statistic was found to be

F = ((0.0395/4) / ((1-0.0395)/137)) =1.41

The 5% significance value with 4 numerator and 137 denominator is given by 2.45 which is larger than the computed statistic. As a result, the null hypothesis is not rejected for the hypothesis of H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

(ii) Now, reestimate the model using the log form for *netinc* and *salary*:

$$\widehat{return} = -36.30 + .327\, dkr + .069\, eps - 4.74 \log(netinc) + 7.24 \log(salary)$$

$$(39.37) \quad (.203) \qquad (.080) \qquad (3.39) \qquad\qquad (6.31)$$

$$n = 142, \quad R^2 = .0330.$$

Do any of your conclusion from part (i) change?

The recalculated t statistic for dkr, eps, netinc, salary is 1.61, 0.862, -1.39, 1.15 respectively. Since all of these values are less than 1.96, this implies that the explanatory variables are insignificant.

The calculated F-statistic is given by the expression

F = (0.033/4) / ((1-0.033)/137) = 1.17

The resulting F statistic found is similar to the one calculated to part (i) as a result the null hypothesis is not rejected.

(iii) In this sample, some firms have zero debt and some have negative earnings. Should we try to use log(*dkr*) or log(*eps*) in the model to see if these improve the fit? Explain.

From the sample equation, some estimates of explanatory variables are 0 or negative. The logarithmic function is not defined for zero values. As a result, using logs for these variables might not be a good option to improve the model fit.

(iv) Overall, is the evidence for predictability of stock returns strong or weak?

The evidence is weak since there are no significant results in the above estimations. Both created models show that there is no significance evenly jointly. In addition, the R-squared value explains less than 4% of variation. As a result, the evidence for predictability of the stock market is weak

Problem #1

$y_i = B_0 + B_1 X_{i1} + \beta_2 X_{i2}$

Regress X_1 on X_1 and X_2

$X_1 = \alpha + \alpha_2 X_2 + \alpha_3 X_3 + ee$

$\hat{X}_1 = \hat{\alpha}_0 + \hat{\alpha}_2 X_2 + \alpha_3 X_3$

Regress y on $\hat{r}_1$ and estimate of the slope from this simple regression gives $\hat{B}_1$

$$\hat{B}_1 = \frac{\sum (\hat{r}_{i1} - \hat{r}_1)(y_i - \bar{y})}{\sum (\hat{r}_{i1} - \hat{r}_1)^2}$$

$\sum_{i=1}^{n} \hat{r}_{i1} = 0$ * Property of OLS

$$\hat{B}_1 = \frac{\sum (r_{i1})(y_i - \bar{y})}{\sum_{i=1}^{n} \hat{r}_{i1}^2}$$

$$\hat{\beta}_1 = \frac{\sum \hat{r}_{i1} y_i}{\sum \hat{r}_{i1}^2}$$

$$\hat{B}_2 = \frac{\sum r_{i2} y_i}{\sum \hat{r}_{i2}^2}$$

} since the residual is from the predicted model is 0 the three conditions will be satisfied