Jeffrey Dean

Econometrics 322

Project 1

#1 Generate a new variable given the publication date of the data table is 2013 with years of birth listed.

Given this information I chose to create the variables “age” as 2013- the birth year to find the total years of the student and “age_sq” as the square of the “age” variable. (Output in R Code)

```
text_score_table <- read.table("text_score_econometrics_project.txt", sep="\t", header=TRUE)
text_score_table
birth_years <- text_score_econometrics_project$birth_year
birth_years
age <- 2013 - birth_years
age
age_sq <-(age)^2
age_sq
```

#2 Run an OLS regression on math scores given studying time, cognitive ability, age, and square of age with an explanation of the influences of both study time and cognitive ability.

Below is the linear model regression utilizing each variable to see the effects on math score for students.

```
lm(formula = text_score_econometrics_project$math_score ~ text_score_econometrics_project$studying_time +
    text_score_econometrics_project$cognitive_ability + age +
    age_sq, data = text_score_econometrics_project)
```

The below information is the output found for both the regression and the confidence interval. The model output supports that cognitive ability is significant at any significance level (<0.01) the null hypothesis can be rejected and we can support that the cognitive does affect outcome. Further studying time has a p-value of just higher than 1% so we fail to reject at the 1% level, yet can reject the null hypothesis that studying time does not affect scores at either a 10% or 5% level.

Hypotheses (H0/HA)

Coeff-Studying_Time

H0: Studying time does not affect test score of students

HA: Studying time does affect test score of students

Coeff-Cognitive ability

H0: Cognitive ability does not affect test score of students

HA: Cognitive ability does affect test score of students

Coeff-Age

H0: Age does not affect test score of students

HA: Age does affect test score of students

Coeff-Age ^2

H0: Age ^2 does not affect test score of students

HA: Age ^2 does affect test score of students

```
Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     96.29691    7.65619  12.578   <2e-16 ***
text_score_econometrics_project$studying_time    0.11726    0.04576   2.563   0.0104 *
text_score_econometrics_project$cognitive_ability 4.17226    0.08131  51.312   <2e-16 ***
age                                             -3.78118    1.12759  -3.353   0.0008 ***
age_sq                                           0.13474    0.04127   3.265   0.0011 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.119 on 17718 degrees of freedom
Multiple R-squared:  0.1343,    Adjusted R-squared:  0.1341
F-statistic: 687.1 on 4 and 17718 DF,  p-value: < 2.2e-16
```

```
> confint(OLS_regression_math_scores, level = 0.95)
                                                    2.5 %      97.5 %
(Intercept)                                     81.29002073 111.3037958
text_score_econometrics_project$studying_time    0.02756791   0.2069472
text_score_econometrics_project$cognitive_ability 4.01288130   4.3316382
age                                             -5.99135984  -1.5709985
age_sq                                           0.05385668   0.2156240
```

#3 Find dummy variables for grade_9, high_income, mother_higher_edu, and father_higher_edu.

Below is the output for the question. This work just required developing a subset of the different columns and because the format for each variable was already Boolean (0,1), the list just needed to be converted to a numeric operator. (Output in R Code)

```
grade9 <- as.numeric(text_score_econometrics_project$grade == 9)
grade9

high_income <- as.numeric(text_score_econometrics_project$income_status > 3)
high_income

mother_higher_edu <- as.numeric(text_score_econometrics_project$mother_edu >= 7)
mother_higher_edu

father_higher_edu <- as.numeric(text_score_econometrics_project$father_edu >= 7)
father_higher_edu
```

#4 Find proportion of students in grade 9, proportion of students that are female, proportion of students older than 13, and find proportion of students from a high-income family.

Below is the output for code to find the specific proportions asked based upon the data table grouped as requested in the data code directions. Output for the variables are found within the R code.

```
all_students_grade <- table(text_score_econometrics_project$grade)
students_grade_7_9 <- all_students_grade[names(all_students_grade)==7]+all_students_grade[names(all_students_grade)
                                                                                          ==9]
students_grade_9 <- all_students_grade[names(all_students_grade)==9]

all_students_grade
students_grade_9
students_grade_7_9


prop_grade_9 <- students_grade_9/students_grade_7_9
prop_grade_9

library(MASS)

prop_grade_9_frac <- fractions(prop_grade_9, cycles = 10, max.denominator = 17723)
prop_grade_9_frac

all_students_fem <- table(text_score_econometrics_project$female)
students_gender_fem_0_1 <- all_students_fem[names(all_students_fem)==1]+all_students_fem[names(all_students_fem)==0]
students_gender_fem_1 <- all_students_fem[names(all_students_fem)==1]
```

```
all_students_fem
students_gender_fem_0_1
students_gender_fem_1

prop_gender_fem <- students_gender_fem_1/students_gender_fem_0_1
prop_gender_fem

prop_gender_fem_frac <- fractions(prop_gender_fem, cycles = 10, max.denominator = 17723)
prop_gender_fem_frac

all_students_age <- table(age)
all_students_age
```

```
student_age_11_12 <- all_students_age[names(all_students_age)==11]+all_students_age[names(all_students_age)==12]
student_age_13_17 <- all_students_age[names(all_students_age)==13]+all_students_age[names(all_students_age)==14]+
all_students_age[names(all_students_age)==15]+all_students_age[names(all_students_age)==16]+
all_students_age[names(all_students_age)==17]
student_age_11_12
student_age_13_17
student_age_11_17 <-student_age_11_12 + student_age_13_17
student_age_11_17
student_age_14_17 <-all_students_age[names(all_students_age)==14]+
all_students_age[names(all_students_age)==15]+all_students_age[names(all_students_age)==16]+
all_students_age[names(all_students_age)==17]
student_age_14_17

prop_student_age_14_17 <- student_age_14_17/student_age_11_17
prop_student_age_14_17
```

```
student_income_all_1_2 <-all_students_income_status[names(all_students_income_status)==1]+all_students_income_status[
student_income_all_3_4 <-all_students_income_status[names(all_students_income_status)==3]+all_students_income_status[
student_income_all_5 <-all_students_income_status[names(all_students_income_status)==5]
student_income_all_1_2
student_income_all_3_4
student_income_all_5
```

```
student_income_all <- student_income_all_1_2 + student_income_all_3_4 + student_income_all_5
student_income_all

student_income_all_4 <- all_students_income_status[names(all_students_income_status)==4]
student_income_all_4_5 <- student_income_all_4 + student_income_all_5
student_income_all_4_5

prop_student_income_4_5 <- student_income_all_4_5/student_income_all
prop_student_income_4_5

prop_student_income_4_5_frac <- fractions(prop_student_income_4_5, cycles = 10, max.denominator = 17723)
prop_student_income_4_5_frac
```

```
> prop_grade_9_frac
        9
2189/4629
```

```
> prop_gender_fem_frac
         1
8718/17723
```

```
> prop_student_age_14_17_frac
        14
8934/17723
```

```
> prop_student_income_4_5_frac
         4
1083/17723
```

#5 Build OLS regression using Q2, gender variables, and dummy variables generated. Use robust standard error for the regression.

This is a second OLS regression for the output with new variables included to expand the conclusions drawn to include factors that may impact the resulting math scores within robust standard errors included to account for variation in the model. Below is output and hypotheses.

```
Call:
lm(formula = text_score_econometrics_project$math_score ~ text_score_econometrics_project$studying_time +
    text_score_econometrics_project$cognitive_ability + age +
    age_sq + text_score_econometrics_project$female + text_score_econometrics_project$male +
    grade9 + high_income + mother_higher_edu + father_higher_edu,
    data = text_score_econometrics_project)
```

Hypotheses (H0/HA)

Coeff-Studying_Time

H0: Studying time does not affect test score of students

HA: Studying time does affect test score of students

Coeff-Cognitive ability

H0: Cognitive ability does not affect test score of students

HA: Cognitive ability does affect test score of students

Coeff-Age

H0: Age does not affect test score of students

HA: Age does affect test score of students

Coeff-Age ^2

H0: Age ^2 does not affect test score of students

HA: Age ^2 does affect test score of students

Coeff-Female

H0: Female gender does not affect test score of students

HA: Female gender does affect test score of students

Coeff- Male gender

H0: Male gender does not affect test score of students

HA: Male gender does affect test score of students

Coeff-grade9

H0: Being in grade 9 does not affect test score of students

HA: Being in grade 9 does affect test score of students

Coeff-High Income

H0: Being high income does not affect test score of students

HA: Being high income does affect test score of students

Coeff-Mother Education

H0: Mother's education does not affect test score of students

HA: Mother's education does affect test score of students

Coeff- Father Education

H0: Father's education does not affect test score of students

HA: Father's education does affect test score of students

Below is the result of the code output supporting which t test coefficents are considered significant at different levels given robust standard errors in place. Studying time appears is significant at a 5% level, cognitive ability is significant is at a 1% level, and gender for both males and females are significant at a 1% level. Being from a higher income family is significant at any level with a p-value below 1%, while both mother and father education levels are not significant at any levels. Below is also the confidence intervals for the variables as well.

```
t test of coefficients:

                                                Estimate Std. Error t value  Pr(>|t|)
(Intercept)                                   97.612342   7.963718 12.2571 < 2.2e-16 ***
text_score_econometrics_project$studying_time  0.102723   0.046084  2.2290 0.0258237 *
text_score_econometrics_project$cognitive_ability  4.175894   0.084669 49.3200 < 2.2e-16 ***
age                                           -3.947683   1.160555 -3.4015 0.0006715 ***
age_sq                                         0.136393   0.042047  3.2438 0.0011815 **
text_score_econometrics_project$female         1.126230   0.137507  8.1904 2.780e-16 ***
grade9                                         0.396950   0.250873  1.5823 0.1136049
high_income                                   -1.479507   0.311174 -4.7546 2.004e-06 ***
mother_higher_edu                             -0.260771   0.249942 -1.0433 0.2968116
father_higher_edu                              0.161624   0.235793  0.6854 0.4930706
---
```

```
                                                    2.5 %      97.5 %
(Intercept)                                    82.01705182 113.2076324
text_score_econometrics_project$studying_time   0.01257878   0.1928663
text_score_econometrics_project$cognitive_ability  4.01126383   4.3405232
age                                           -6.21300978  -1.6823564
age_sq                                         0.05459946   0.2181870
text_score_econometrics_project$female         0.85677230   1.3956874
text_score_econometrics_project$male                   NA          NA
grade9                                        -0.07789630   0.8717954
high_income                                   -2.04455388  -0.9144602
mother_higher_edu                             -0.77917281   0.2576312
father_higher_edu                             -0.32242075   0.6456684
```

#6 Develop two OLS regressions with studying time, cognitive ability, age and the square of age as explanatory variables on English scores for females and for males.

For this portion I developed two new regressions one for the female group and one for the male group based on the parameters given. Code and output are below for the two group regressions.

Female Group Regression

```
regression_1_f <- lm(numeric_english_score_f ~ numeric_studying_time_f + numeric_cognitive_ability_f,
                     data=text_score_econometrics_project)
regression_1_f

t_test_regression_1_f <- t.test(regression_1_f)

summary(regression_1_f)
```

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               72.93687    0.10250  711.58   <2e-16 ***
numeric_studying_time_f    0.07821    0.05626    1.39    0.165
numeric_cognitive_ability_f  2.72309    0.10376   26.24   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.063 on 8715 degrees of freedom
Multiple R-squared:  0.07391,   Adjusted R-squared:  0.0737
F-statistic: 347.8 on 2 and 8715 DF,  p-value: < 2.2e-16
```

```
> confint(regression_1_f, level = 0.95)
                                2.5 %     97.5 %
(Intercept)                72.73594751 73.1377939
numeric_studying_time_f    -0.03208283  0.1884951
numeric_cognitive_ability_f  2.51968964  2.9264879
```

Male Group Regression

```
regression_1_m <- lm(numeric_english_score_m ~ numeric_studying_time_m + numeric_cognitive_ability_m,
                     data=text_score_econometrics_project)
regression_1_m

t_test_regression_1_m <- t.test(regression_1_m)

summary(regression_1_m)
```

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               67.17889    0.11714 573.485  < 2e-16 ***
numeric_studying_time_m    0.22889    0.06934   3.301 0.000968 ***
numeric_cognitive_ability_m  3.87297    0.11648  33.250  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.63 on 9002 degrees of freedom
Multiple R-squared:  0.1119,    Adjusted R-squared:  0.1117
F-statistic: 566.9 on 2 and 9002 DF,  p-value: < 2.2e-16
```

```
> confint(regression_1_m, level = 0.95)
                                2.5 %     97.5 %
(Intercept)                66.94926364 67.4085114
numeric_studying_time_m     0.09296196  0.3648119
numeric_cognitive_ability_m  3.64463781  4.1012980
```

As shown above, for the female regression, studying time appeared not to be statistically significant at any level (p >10%), while for males the finding was significant at a 1% level. This shows a stark contrast that can be examined further. It could be studying time could be a more important consideration for males in the study areas where males yield higher test scores. Regarding cognitive ability, both regressions found statistical significance at a 1% level, supporting that this factor is an important component of test scores for both genders.

#7 Adding # of siblings and dummy variables to the prior model, find influences on male vs female

For this portion I needed to set up a regression just with previously built variables in the Q6 model to understand how these factors impact English scores using robust standard errors. Below is the code and the output from testing.

Female Group

```r
regression_2_f <- lm(female_group$english_score ~ female_group$cognitive_ability + female_group$studying_time +
                       female_group$sibling_number, data=text_score_econometrics_project)
regression_2_f

t_test_regression_2_f <- t.test(regression_2_f)

summary(regression_2_f)

library(lmtest)
library(sandwich)

coeftest(regression_2_f, vcov = vcovHC(regression_2_f, type="HC1"))

confint(regression_2_f, level = 0.95)
```

```
t test of coefficients:

                                Estimate Std. Error  t value  Pr(>|t|)
(Intercept)                    72.462767   0.137693 526.2622 < 2.2e-16 ***
female_group$cognitive_ability  2.865969   0.108904  26.3165 < 2.2e-16 ***
female_group$studying_time      0.096723   0.054924   1.7610   0.07827 .
female_group$sibling_number     0.574686   0.119480   4.8099 1.535e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> confint(regression_2_f, level = 0.95)
                                    2.5 %     97.5 %
(Intercept)                    72.19639617 72.7291370
female_group$cognitive_ability  2.65613131  3.0758067
female_group$studying_time     -0.01360689  0.2070535
female_group$sibling_number     0.36227132  0.7871012
```

Male Group

```r
regression_2_m <- lm(male_group$english_score ~ male_group$cognitive_ability + male_group$studying_time +
                       male_group$sibling_number, data=text_score_econometrics_project)
regression_2_m

t_test_regression_2_m <- t.test(regression_2_m)

summary(regression_2_m)

library(lmtest)
library(sandwich)

coeftest(regression_2_m, vcov = vcovHC(regression_2_m, type="HC1"))

confint(regression_2_m, level = 0.95)
```

```
t test of coefficients:

                              Estimate Std. Error  t value  Pr(>|t|)
(Intercept)                  67.226103   0.149775 448.8475 < 2.2e-16 ***
male_group$cognitive_ability  3.860937   0.116551  33.1265 < 2.2e-16 ***
male_group$studying_time      0.227283   0.069268   3.2812  0.001037 **
male_group$sibling_number    -0.068350   0.132784  -0.5147  0.606745
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> confint(regression_2_m, level = 0.95)
                                  2.5 %     97.5 %
(Intercept)                  66.93842115 67.5137845
male_group$cognitive_ability  3.62836803  4.0935065
male_group$studying_time      0.09122524  0.3633408
male_group$sibling_number    -0.31920758  0.1825081
```

The output values remain the same in practice except for the addition of sibling number in the output. For the female group the number of siblings is significant at a 1% level, while the variable does not hold significance at any level for the male group. This could be explored further. Women from larger families may expect more competitive/ higher academic results than smaller family children. The result could also be confounding from other unseen factors in the model.

#8 Find correlation between cognitive ability and math scores. Find correlation between cognitive ability and English scores.

Below is the output for the correlation function relating cognitive ability and math scores as well as cognitive ability and English scores. (output in R Code) For correlation coefficient, the expected range lies between -1 and 1 so the below supports a small positive correlation. Anything positive, but below 0.5 correlation means as cognitive ability is higher, both math and English scores will most likely improve to some degree.

```
> math_score_corr
[1] 0.3651358
> english_score_corr
[1] 0.2951107
```

#9 Find if the mean of math scores differ between male and female students.

Below is output for difference of means between the male group and the female group for math scores. Output was discovered through subsetting both male and female math scores from the data table. The result was delisted and converted to numeric operators. From there I developed a two sample mean t.test for the groups. The output shows a confidence interval over a 95% interval with a p-value that is

below 1% supporting that the H0 can be rejected. This leads to the conclusion that there is a statistically significant difference between males and females for math scores.

Ho: There is no true difference in means between males and females in terms of math scores

HA: There is a true difference in means between males and females in terms of math scores

```
        Two Sample t-test

data:  numeric_math_score_female and numeric_math_score_male
t = 8.5341, df = 17721, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9660396 1.5421035
sample estimates:
mean of x mean of y
 70.88538  69.63131
```

#10 Find if the mean of English scores differ between male and female students.

Below is output for difference of means between the male group and the female group for English scores. Output was discovered through subsetting both male and female math scores from the data table. The result was delisted and converted to numeric operators. From there I developed a two sample mean t.test for the groups. The output shows a confidence interval over a 95% interval with a p-value that is below 1% supporting that the H0 can be rejected. This leads to the conclusion that there is a statistically significant difference between males and females for English scores.

Ho: There is no true difference in means between males and females in terms of math scores

HA: There is a true difference in means between males and females in terms of math scores

```
        Two Sample t-test

data:  numeric_english_score_female and numeric_english_score_male
t = 40.241, df = 17721, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 5.382512 5.933707
sample estimates:
mean of x mean of y
 73.11831  67.46020
```