Assignment 1

Moh Jaiswal, 500916860, Assignment #1, Prof. Turner

'Question 1'

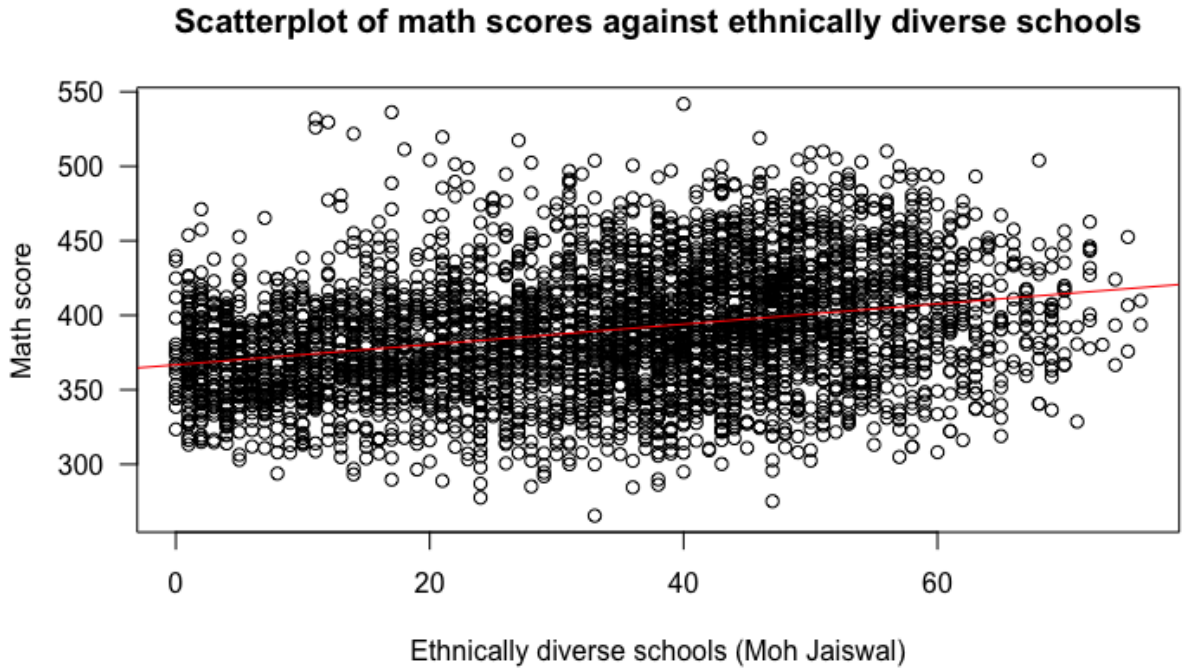
```
1  #Moh Jaiswal, 500916860, Assignment #1, Prof. Turner
2  library(haven)
3  STAR_big <- read_dta("STAR_big.dta")
4  View(STAR_big)
5  'Question 1'
6  #Ryerson No: 500"9""1"6860
7  # VARIABLES ARE x = edi_s, y = math_score, Z = sex_frac_female_z; x^2 = edi_s2
8  question1a <- plot(STAR_big$edi_s,STAR_big$math_score, main="Scatterplot of math scores against ethnically diverse schools",
9       xlab="Ethnically diverse schools (Moh Jaiswal)",
10      ylab="Math score",
11      las=1)
12 question1b <- abline(lm(math_score ~ edi_s, data=STAR_big), col='red')
```

10:24 (Top Level) R Script

Console Terminal Jobs

~/Downloads/

```
> #Ryerson No: 500"9""1"6860
> # VARIABLES ARE x = edi_s, y = math_score, Z = sex_frac_female_z; x^2 = edi_s2
> question1a <- plot(STAR_big$edi_s,STAR_big$math_score, main="Scatterplot of math scores against ethnically diverse schools",
+      xlab="Ethnically diverse schools (Moh Jaiswal)",
+      ylab="Math score",
+      las=1)
> question1b <- abline(lm(math_score ~ edi_s, data=STAR_big), col='red')
```



'Question 2'

```r
'Question 2'
#new variable for edi_s squared.
edi_s2 = STAR_big$edi_s^2
install.packages("sandwich")
install.packages("lmtest")
install.packages("jtools")
install.packages("car")
library(sandwich)
library(lmtest)
library(jtools)
library(car)
question2 <- lm(math_score ~ edi_s+edi_s2, data=STAR_big)
summ(question2, digits=3, robust = "HC1")
cor(STAR_big$edi_s, STAR_big$math_score)
coefci(question2, vcov = vcovHC, type="HC1", level=0.95)
coefci(question2, vcov = vcovHC, type="HC1", level=0.90)
linearHypothesis(question2, c("edi_s=0", "edi_s2=0"), white.adjust = "hc1")
#This is the "model F test" which also is reported using the "summary" command,
#option "hc1" uses the correct standard errors.The results of the test implies
#The correlation (cor.estimate) value between x and y is 0.295, which implies t
#If  we compare F values, there is a slight favoring towards a quadratic relati
#however it is not significant and can be omitted. My recommendation to best pr
```

31:2 (Top Level)

Console | Terminal | Jobs

~/Downloads/

```
> question2 <- lm(math_score ~ edi_s+edi_s2, data=STAR_big)
> summ(question2, digits=3, robust = "HC1")
MODEL INFO:
Observations: 3932
Dependent Variable: math_score
Type: OLS linear regression

MODEL FIT:
F(2,3929) = 188.533, p = 0.000
R² = 0.088
Adj. R² = 0.087

Standard errors: Robust, type = HC1
-----------------------------------------------------
                      Est.    S.E.    t val.       p
----------------- --------- ------- --------- -------
(Intercept)         364.863   1.545   236.183   0.000
edi_s                 0.852   0.114     7.476   0.000
edi_s2               -0.003   0.002    -1.491   0.136
-----------------------------------------------------
```

```
> cor(STAR_big$edi_s, STAR_big$math_score)
[1] 0.2951654
> coefci(question2, vcov = vcovHC, type="HC1", level=0.95)
                    2.5 %       97.5 %
(Intercept) 361.834559775 3.678921e+02
edi_s         0.628496382 1.075309e+00
edi_s2       -0.006094737 8.295896e-04
> coefci(question2, vcov = vcovHC, type="HC1", level=0.90)
                     5 %         95 %
(Intercept) 362.321685858 3.674049e+02
edi_s         0.664427781 1.039378e+00
edi_s2       -0.005537902 2.727553e-04
> linearHypothesis(question2, c("edi_s=0", "edi_s2=0"), white.adjust = "hc1")
Linear hypothesis test

Hypothesis:
edi_s = 0
edi_s2 = 0

Model 1: restricted model
Model 2: math_score ~ edi_s + edi_s2

Note: Coefficient covariance matrix supplied.

  Res.Df Df     F    Pr(>F)
1   3931
2   3929  2 226.1 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

#This is the "model F test" which also is reported using the "summary" command, but "linearHypothesis" with option "hc1" uses the correct standard errors. The results of the test imply that the null hypothesis is rejected due to the high F stat value.
#The correlation (cor.estimate) value between x and y is 0.295, which implies that a linear relationship is not supported.
#If we compare F values, there is a slight favoring towards a quadratic relation over a linear relation, however, it is not significant and can be omitted. My recommendation to best predicts the math_score entails a linear relation between x and y.

'Question 3'

```r
'Question 3'
question3 <- lm(math_score ~ edi_s+edi_s2+sex_frac_female_z, data=STAR_big)
summ(question3, digits=3, robust = "HC1")
coefci(question3, vcov = vcovHC, type="HC1", level=0.95)
coefci(question3, vcov = vcovHC, type="HC1", level=0.90)
linearHypothesis(question3, c("edi_s=0", "edi_s2=0"), white.adjust = "hc1")
#Regression Equation -> math_score = 336.69 + 0.82(edi_s) - 0.002294(edi_s2) + 57.31(sex_frac_female_z)
#As suspected, our Adj. R sq value is marginally higher than the model in question 2. The hypothesis of adding this variable was to test the
#theory that women have higher math scores in general.
#This hypothesis can be validated with the data, we can see the addition of var. Z has a coefficient value of 57, which signifies its
#importance in estimating math_score with greater accuracy.
#However the addition of variable z is rendered moot, hence not a signifant addition to the model and can be omitted.
#Gender has a very small effect in predicting the math score.
```

47:2 (Top Level)

Console Terminal Jobs

~/Downloads/

```
> question3 <- lm(math_score ~ edi_s+edi_s2+sex_frac_female_z, data=STAR_big)
> summ(question3, digits=3, robust = "HC1")
MODEL INFO:
Observations: 3932
Dependent Variable: math_score
Type: OLS linear regression

MODEL FIT:
F(3,3928) = 127.260, p = 0.000
R² = 0.089
Adj. R² = 0.088

Standard errors: Robust, type = HC1
------------------------------------------------------------
                          Est.     S.E.   t val.       p
---------------------- --------- -------- -------- -------
(Intercept)            336.687   13.271   25.371   0.000
edi_s                    0.817    0.115    7.098   0.000
edi_s2                  -0.002    0.002   -1.294   0.196
sex_frac_female_z       57.306   26.883    2.132   0.033
------------------------------------------------------------
> coefci(question3, vcov = vcovHC, type="HC1", level=0.95)
                        2.5 %       97.5 %
(Intercept)       310.668913609 3.627048e+02
edi_s               0.591111526 1.042318e+00
edi_s2             -0.005771742 1.183028e-03
sex_frac_female_z   4.600629189 1.100108e+02
```

```
> coefci(question3, vcov = vcovHC, type="HC1", level=0.90)
                          5 %         95 %
(Intercept)       314.85349041 3.585202e+02
edi_s               0.62739621 1.006033e+00
edi_s2             -0.00521246 6.237457e-04
sex_frac_female_z  13.07740984 1.015340e+02
> linearHypothesis(question3, c("edi_s=0", "edi_s2=0", "sex_frac_female_z=0"), white.adjust = "hc1")
Linear hypothesis test

Hypothesis:
edi_s = 0
edi_s2 = 0
sex_frac_female_z = 0

Model 1: restricted model
Model 2: math_score ~ edi_s + edi_s2 + sex_frac_female_z

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1   3931
2   3928  3 151.89 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

#Regression Equation -> math_score = 336.69 + 0.82(edi_s) - 0.002294(edi_s2) + 57.31(sex_frac_female_z)
#As suspected, our Adj. R sq value is marginally higher than the model in question 2. The hypothesis of adding this variable was to test the theory that women have high math scores. ‘Sex_frac_female_z’ can't really be interpreted as measuring women's performance on tests since this variable measures the female share in the zipcode of the school, not the student body of the school (z means zipcode).
#The coefficient on z has the interpretation that a 1 percentage point increase in share female leads to a 0.57 increase in test scores (since the variable ranges from 0 to 1), conditional on x and its square. This is not a very big effect (relative to a mean test score of about 350), but it is marginally statistically significant, at the 5% level but not at the 1% level, so should probably remain in the model.
#However the addition of variable z is rendered moot, hence not a significant addition to the model and can be omitted.
#Gender has a very small effect in predicting the math score. β1-hat and β2-hat can only be analyzed together as they describe a non-linear relationship between x and y conditional on z.

'Question 4'
4a

```
49  'Question 4'
50  question4a <- lm(math_score ~ edi_s, data=STAR_big)
51  summ(question4a, digits=3, robust = "HC1")
52  coefci(question4a, vcov = vcovHC, type="HC1", level=0.95)
53  coefci(question4a, vcov = vcovHC, type="HC1", level=0.90)
54  linearHypothesis(question4a, c("edi_s=0"), white.adjust = "hc1")
55
```

55:1 (Top Level)

Console | Terminal | Jobs

~/Downloads/

```
> question4a <- lm(math_score ~ edi_s, data=STAR_big)
> summ(question4a, digits=3, robust = "HC1")
MODEL INFO:
Observations: 3932
Dependent Variable: math_score
Type: OLS linear regression

MODEL FIT:
F(1,3930) = 375.069, p = 0.000
R² = 0.087
Adj. R² = 0.087

Standard errors: Robust, type = HC1
------------------------------------------------------
                      Est.    S.E.    t val.       p
---------------- --------- ------- --------- -------
(Intercept)        366.752   1.167   314.394   0.000
edi_s                0.682   0.033    20.753   0.000
------------------------------------------------------
> coefci(question4a, vcov = vcovHC, type="HC1", level=0.95)
                  2.5 %      97.5 %
(Intercept) 364.4645670 369.0387142
edi_s         0.6177733   0.7466739
> coefci(question4a, vcov = vcovHC, type="HC1", level=0.90)
                    5 %        95 %
(Intercept) 364.8324066 368.6708745
edi_s         0.6281391   0.7363081
```

```
> linearHypothesis(question4a, c("edi_s=0"), white.adjust = "hc1")
Linear hypothesis test

Hypothesis:
edi_s = 0

Model 1: restricted model
Model 2: math_score ~ edi_s

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1   3931
2   3930  1 430.69 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

4b

```
56  question4b <- lm(math_score ~ edi_s + sex_frac_female_z, data=STAR_big)
57  summ(question4b, digits=3, robust = "HC1")
58  coefci(question4b, vcov = vcovHC, type="HC1", level=0.95)
59  coefci(question4b, vcov = vcovHC, type="HC1", level=0.90)
60  linearHypothesis(question4b, c("edi_s=0", "sex_frac_female_z=0"), white.adjust = "hc1")
61
```

60:88 (Top Level)

Console | Terminal | Jobs

~/Downloads/

```
> question4b <- lm(math_score ~ edi_s + sex_frac_female_z, data=STAR_big)
> summ(question4b, digits=3, robust = "HC1")
MODEL INFO:
Observations: 3932
Dependent Variable: math_score
Type: OLS linear regression

MODEL FIT:
F(2,3929) = 190.147, p = 0.000
R² = 0.088
Adj. R² = 0.088

Standard errors: Robust, type = HC1
------------------------------------------------------------
                            Est.     S.E.   t val.       p
----------------------- --------- -------- -------- -------
(Intercept)              336.955   13.250   25.430   0.000
edi_s                      0.668    0.033   19.961   0.000
sex_frac_female_z         60.083   26.746    2.246   0.025
------------------------------------------------------------
> coefci(question4b, vcov = vcovHC, type="HC1", level=0.95)
                         2.5 %      97.5 %
(Intercept)       310.9769825 362.9332728
edi_s               0.6025578   0.7338135
sex_frac_female_z   7.6456860 112.5209304
> coefci(question4b, vcov = vcovHC, type="HC1", level=0.90)
                          5 %        95 %
(Intercept)       315.155157 358.7550982
edi_s               0.613113   0.7232583
sex_frac_female_z  16.079450 104.0871666
```

```
> linearHypothesis(question4b, c("edi_s=0", "sex_frac_female_z=0"), white.adjust = "hc1")
Linear hypothesis test

Hypothesis:
edi_s = 0
sex_frac_female_z = 0

Model 1: restricted model
Model 2: math_score ~ edi_s + sex_frac_female_z

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1   3931
2   3929  2 217.98 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

question4a <- lm(math_score ~ edi_s, data=STAR_big)
summary(question4a)
question4b <- lm(math_score ~ edi_s + sex_frac_female_z, data=STAR_big)
summary(question4b)
# Here, the coefficient on x barely changes since z does not have a strong correlation with u from regression #1, as can be seen from regression #2. So even though x and z are positively correlated (this is easy to check in R although the reason for it isn't obvious, at least it's not obvious to me why more female neighborhoods should have more diverse schools), only one condition for omitted variable bias is satisfied and z is not an omitted variable we need to worry about.