Econometrics Analysis HW01 (R Empirical)

Breakout Room 6

Chai Mu Xuan (01418350) Ge Xiao Min (01107669)

Song Yangao (01436075) Wilson Quah (01425826)

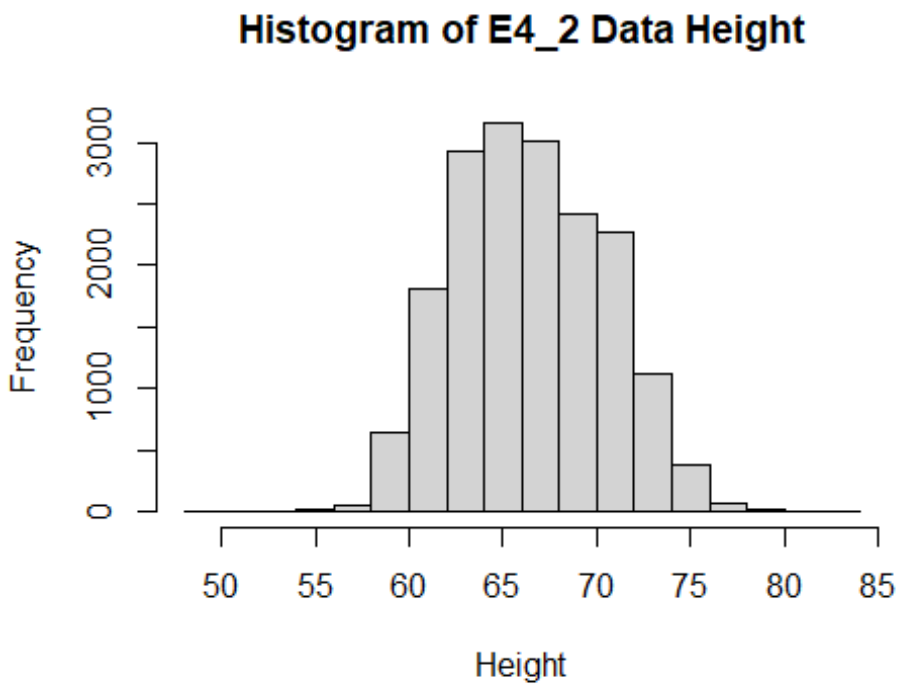
Q1.Empirical Exercise 4.2

Import dataset E4.2_Earnings_and_Height.xlsx

```
library(tinytex)
library(readxl)
E4_2_data <- read_excel("E4.2_Earnings_and_Height.xlsx")
```

Plot Histogram for Heights

```
hist(E4_2_data$height,
     main = "Histogram of E4_2 Data Height",
     xlab = "Height")
```



```
#Compute Median
getMedian <- median(E4_2_data$height)
```

a)

The median is: 67

Test for Normality (Jarque-Bera Test)

```
library(moments)
jarque.test(E4_2_data$height)

##
##  Jarque-Bera Normality Test
##
## data:  E4_2_data$height
## JB = 240.21, p-value < 2.2e-16
## alternative hypothesis: greater

JB_p_value <- jarque.test(E4_2_data$height)$p.value
```

The p-value is: 0

```
JB_test_statistic <- as.numeric(jarque.test(E4_2_data$height)$statistic)
JB_test_statistic <- round(JB_test_statistic, 3)
```

The JB-statistic is: 240.208

Create Dummy Variables for Height (DHeight)

```
E4_2_data$DHeight <- ifelse(E4_2_data$height > 67, 1 , 0)
```

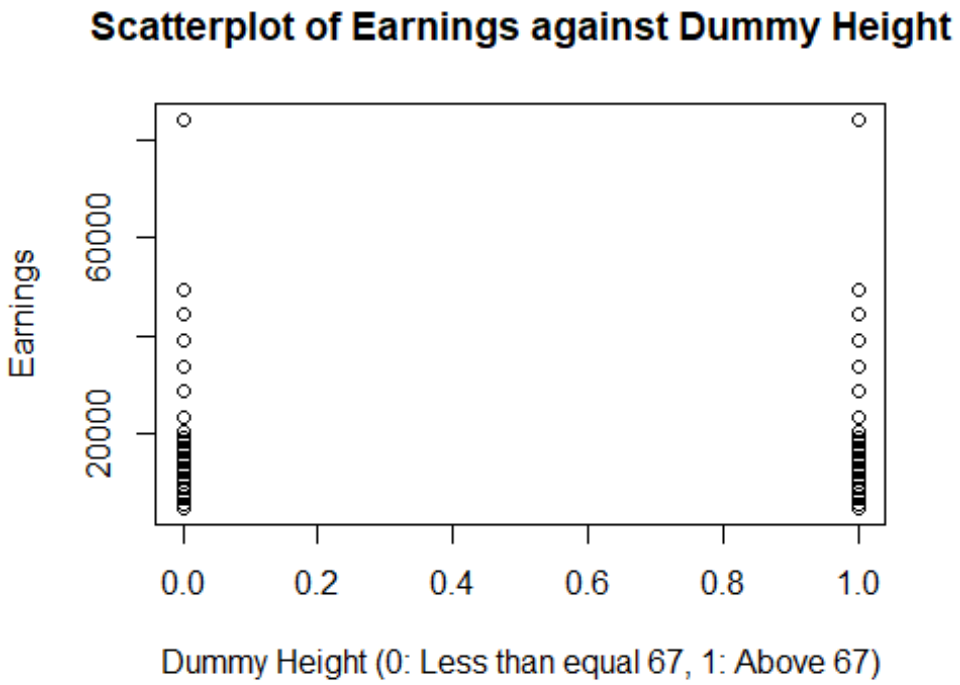
Estimate Model (with DHeight)

```
getModel <- lm(E4_2_data$earnings ~ E4_2_data$DHeight)
summary(getModel)

##
## Call:
## lm(formula = E4_2_data$earnings ~ E4_2_data$DHeight)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -45261 -21427  -5836  34067  39566
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        44488.4      266.3   167.0   <2e-16 ***
## E4_2_data$DHeight   5499.4      404.3    13.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26790 on 17868 degrees of freedom
```

```
## Multiple R-squared:  0.01025,    Adjusted R-squared:  0.01019
## F-statistic:   185 on 1 and 17868 DF,  p-value: < 2.2e-16

plot.default(x = E4_2_data$DHeight,
             y = E4_2_data$earnings,
             main = "Scatterplot of Earnings against Dummy Height",
             xlab = "Dummy Height (0: Less than equal 67, 1: Above 67)",
             ylab = "Earnings")
```




```
earnings_DHeight_coeff <- summary(getModel)$coefficients
earnings_estimated <- predict(getModel)
earnings_DHeight_intercept <- round(earnings_DHeight_coeff[1,1], 3)
earnings_DHeight_slope <- round(earnings_DHeight_coeff[2,1], 3)
```

Earnings = 44488.44 + 5499.44 * DHeight

b)i)

```
earnings_67_coeff <- summary(getModel)$coefficients
earnings_67_slope <- round(earnings_67_coeff[2,1],3)
earnings_67_intercept <- round(earnings_67_coeff[1,1],3)
```

Estimated Avg. Earnings for workers with Height at most 67 Inches: $44488.44

b)ii)

```
earnings_more67 <- round(earnings_67_coeff[1] + earnings_67_coeff[2],3)
```

Estimated Avg. Earnings for workers with Height greater than 67 Inches: $49987.88

b)iii) Do taller workers earn more than shorter workers?

Test if DHeight Coefficient = 0

H0: There is No difference in earnings between Tall workers and Short workers

beta1 = 0

H1: There is a difference in earnings between Tall workers and Short workers

beta1 != 0

```
getCoefficients      <- summary(getModel)$coefficients
get_DHeight_TestStat <- round(getCoefficients[2,3], 3)
get_DHeight_PValue   <- getCoefficients[2,4]
```

DHeight Test Statistic: 13.603

Since Test Statistic is greater than z = 1.96, we reject H0.

DHeight p-value : 6.220183e-42

Since p-value is very small, we reject H0. Hence, there is significant evidence to reject H0 that there is no difference in earnings between Tall and Short workers

How much more?

What is 95% Confidence Interval for difference in earnings

```
confint(getModel, level = 0.95)

##                         2.5 %    97.5 %
## (Intercept)       43966.378 45010.494
## E4_2_data$DHeight  4707.007  6291.873

getCI <- confint(getModel, level = 0.95)
getCI_DHeight_lower <- round(getCI[2,1], 2)
getCI_DHeight_Upper <- round(getCI[2,2], 2)
```

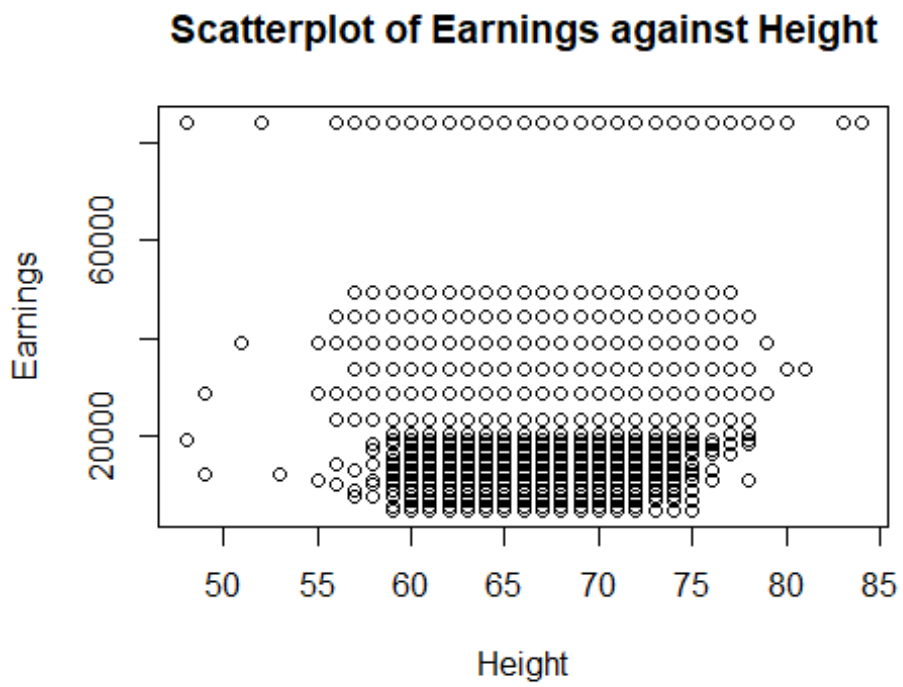
The difference in earnings for Taller workers compared to Short Workers is between

[$4707.01 , $6291.87]

We note that the confidence interval lies in the positive region, suggesting that the the population difference in earnings between Tall and Short workers is a positive value.

Scatterplot of Earnings against Height

```
plot.default(x = E4_2_data$height,
             y = E4_2_data$earnings,
             main = "Scatterplot of Earnings against Height",
             xlab = "Height",
             ylab = "Earnings")
```



The Height is computed to the nearest inches. Hence it can be treated as a discrete independent variable. Thus the height data can only take specific integers of inches. If the height data is allowed to take continuous form, then the data will be spread out in between integer values.

d) Regression of Earnings on Height

```
model_earnings_height <- lm(E4_2_data$earnings ~ E4_2_data$height)
summary(model_earnings_height)

##
## Call:
## lm(formula = E4_2_data$earnings ~ E4_2_data$height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47836 -21879  -7976  34323  50599
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -512.73    3386.86  -0.151     0.88
## E4_2_data$height   707.67      50.49  14.016   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26780 on 17868 degrees of freedom
```

```
## Multiple R-squared:  0.01088,    Adjusted R-squared:  0.01082
## F-statistic: 196.5 on 1 and 17868 DF,  p-value: < 2.2e-16

getCoefficients <- summary(model_earnings_height)$coefficients
getInterceptCoefficient <- round(getCoefficients[1,1], 3)
getHeightCoefficient <- round(getCoefficients[2,1], 3)
```

i)The estimated slope : 707.672

ii)The estimated Intercept: -512.734

Estimated Earnings = -512.734 + 707.672 * Height

```
#Compute Estimated Earnings based on different heights
getCoefficients <- summary(model_earnings_height)$coefficients
getInterceptCoeff <- round(getCoefficients[1,1], 3)
getHeightCoeff <- round(getCoefficients[2,1], 3)
earnings_height67 <- round(getInterceptCoeff + (getHeightCoeff * 67), 3)
earnings_height70 <- round(getInterceptCoeff + (getHeightCoeff * 70), 3)
earnings_height65 <- round(getInterceptCoeff + (getHeightCoeff * 65), 3)
```

At Height: 67 Estimated Earnings: $46901.29

At Height: 70 Estimated Earnings: $49024.31

At Height: 65 Estimated Earnings: $45485.95

e) Suppose height measured in cm instead of inches (1 inch == 2.54 cm)

```
E4_2_data$cHeight = E4_2_data$height * 2.54
model_earnings_cHeight = lm(E4_2_data$earnings ~ E4_2_data$cHeight)
summary(model_earnings_cHeight)

##
## Call:
## lm(formula = E4_2_data$earnings ~ E4_2_data$cHeight)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47836 -21879  -7976  34323  50599
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -512.73    3386.86  -0.151     0.88
## E4_2_data$cHeight   278.61      19.88  14.016   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26780 on 17868 degrees of freedom
## Multiple R-squared:  0.01088,    Adjusted R-squared:  0.01082
## F-statistic: 196.5 on 1 and 17868 DF,  p-value: < 2.2e-16
```

```
get_cHeight_coeff <- summary(model_earnings_cHeight)$coefficients
intercept_cm <- round(get_cHeight_coeff[1,1], 3)
slope_cm <- round(get_cHeight_coeff[2,1], 3)
getRSquared <- summary(model_earnings_cHeight)$r.squared
```

Earnings = -512.734 + 278.611 * cHeight

i)**Slope Decreases (by a factor 2.54): 278.611**

ii)**When Height = 0, no change to Earnings Intercept: $ -512.734**

iii)**Multiple R-squared: 0.0108753**

iv)**Standard Error of Regression: 26780**

f)Regression of Earnings on Height for Female workers only

```
E4_2_data_females <- subset(E4_2_data, sex == 0)
get_model_Height_Females <- lm(E4_2_data_females$earnings ~
E4_2_data_females$height)
summary(get_model_Height_Females)

##
## Call:
## lm(formula = E4_2_data_females$earnings ~ E4_2_data_females$height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -42748 -22006  -7466  36641  46865
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 12650.9     6383.7   1.982   0.0475 *
## E4_2_data_females$height      511.2       98.9   5.169  2.4e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26800 on 9972 degrees of freedom
## Multiple R-squared:  0.002672,   Adjusted R-squared:  0.002572
## F-statistic: 26.72 on 1 and 9972 DF,  p-value: 2.396e-07

get_model_Height_Females_coeff <-
summary(get_model_Height_Females)$coefficients
female_slope <- round(get_model_Height_Females_coeff[2,1], 3)
```

The estimated slope for females is: 511.222

Compute Female Earnings change if height delta is +1

```
delta_earnings_female <- get_model_Height_Females_coeff[2,1]*1
delta_earnings_female <- round(delta_earnings_female,2)
```

The estimated increase in earnings for females when height increases by 1 inch is: +$511.22

g)Regression of Earnings on Height for Male workers only

```
E4_2_data_males <- subset(E4_2_data, sex == 1)
get_model_Height_Males <- lm(E4_2_data_males$earnings ~
E4_2_data_males$height)
summary(get_model_Height_Males)

##
## Call:
## lm(formula = E4_2_data_males$earnings ~ E4_2_data_males$height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50158 -22373  -8118  33091  59228
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -43130.3     7068.5  -6.102  1.1e-09 ***
## E4_2_data_males$height   1306.9      100.8  12.969  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26670 on 7894 degrees of freedom
## Multiple R-squared:  0.02086,    Adjusted R-squared:  0.02074
## F-statistic: 168.2 on 1 and 7894 DF,  p-value: < 2.2e-16

get_model_Height_Males_coeff <- summary(get_model_Height_Males)$coefficients
male_slope <- round(get_model_Height_Males_coeff[2,1], 3)
```

The estimated slope for Males is: 1306.86

Compute Male Earnings change if height delta is +1

```
delta_earnings_male <- get_model_Height_Males_coeff[2,1]*1
delta_earnings_male <- round(delta_earnings_male, 2)
```

The estimated increase in earnings for males when height increases by 1 inch is: +$1306.86

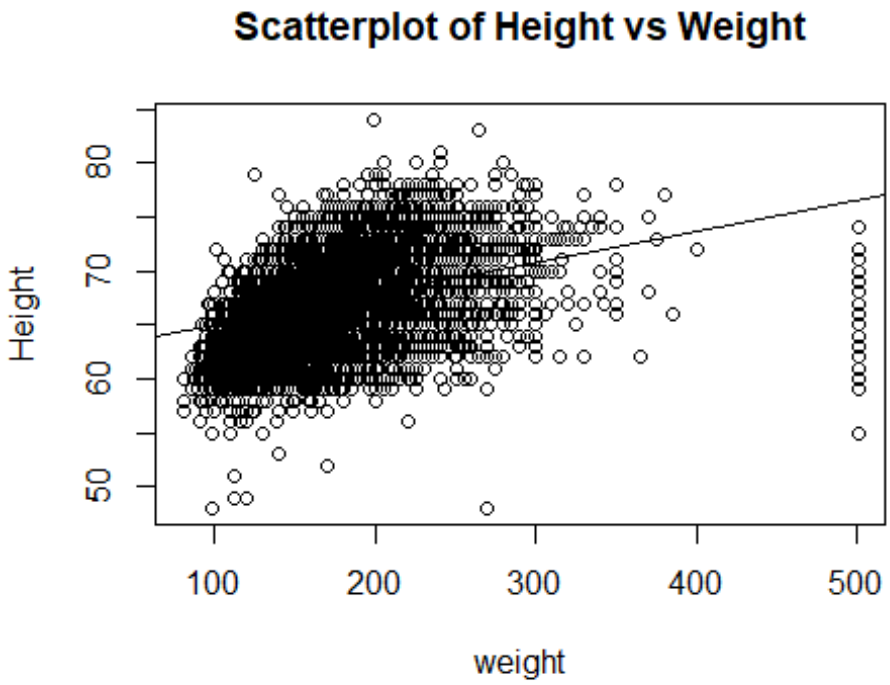
h)Do you think that height is uncorrelated with other factors that cause earnings

No, height is correlated with other factors

Scatterplot of Height on Weight

```
plot.default(y = E4_2_data$height,
             x = E4_2_data$weight,
             main = "Scatterplot of Height vs Weight",
```

```
             ylab = "Height",
             xlab = "weight")
model_height_weight <- lm(E4_2_data$height ~ E4_2_data$weight)
abline(model_height_weight)
```



There is a positive correlation between Height and weighted. For Simple Linear Regression, the Weight is captured by the error terms. Therefore, the conditional mean of error terms given Height is not 0. We need to further extend the Simple Linear Regression with Multiple regression including Weight as a control variable to model its dependencies on earnings.

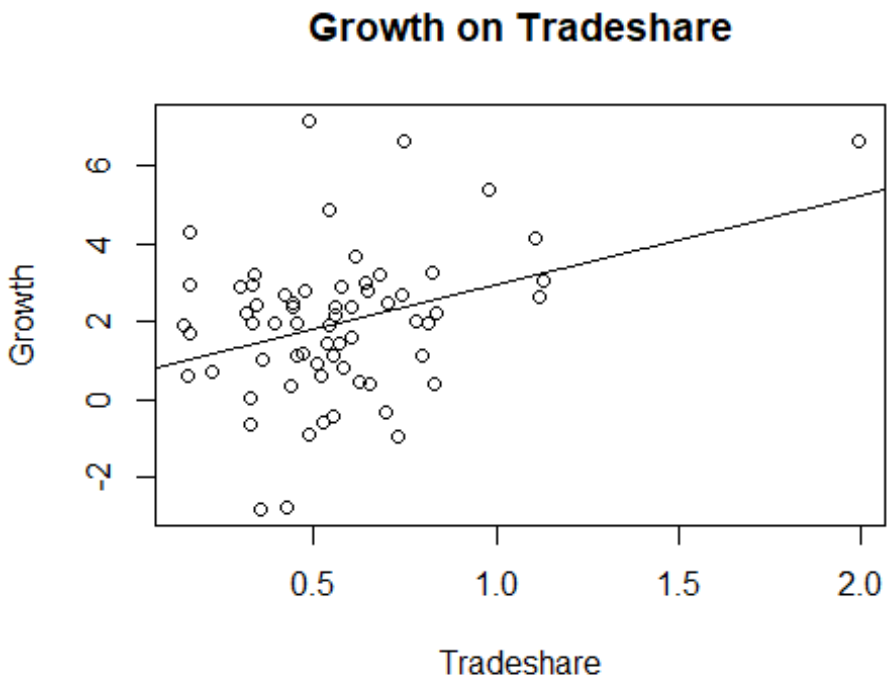
Q2.Empirical Exericses 4.1

Import dataset E4.1_Growth.xlsx

```
library(readr)
Growth <- read_csv("E4.1_Growth.csv", show_col_types = FALSE)
```

Q2a - E4.1a

```
model <- lm(Growth$growth~Growth$tradeshare)
plot.default(x=Growth$tradeshare, y=Growth$growth,
             main = "Growth on Tradeshare", type = "p",
             xlab = "Tradeshare", ylab = "Growth")
# x and y labels
abline(model)
```



Qn 2b - E4.1b

Yes, based on Fig 4b, Malta is indeed an outlier as it is far from the regression function line.

Qn 2c - E4.1c

$$Growth = \beta_0 + \beta_1 * Tradeshare + u$$

```
model <- lm(Growth$growth~Growth$tradeshare)
summary(model)

##
## Call:
## lm(formula = Growth$growth ~ Growth$tradeshare)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3739 -0.8864  0.2329  0.9248  5.3889
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.6403     0.4900   1.307  0.19606
## Growth$tradeshare   2.3064     0.7735   2.982  0.00407 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.79 on 63 degrees of freedom
## Multiple R-squared:  0.1237, Adjusted R-squared:  0.1098
## F-statistic: 8.892 on 1 and 63 DF,  p-value: 0.00407
```

$$Estimatedslope, \beta_1 = 2.306434$$

$$Estimatedintercept, \beta_0 = 0.640265$$

$$(Growth\hat{)} = 0.640265 + 2.306434 TradeShare$$

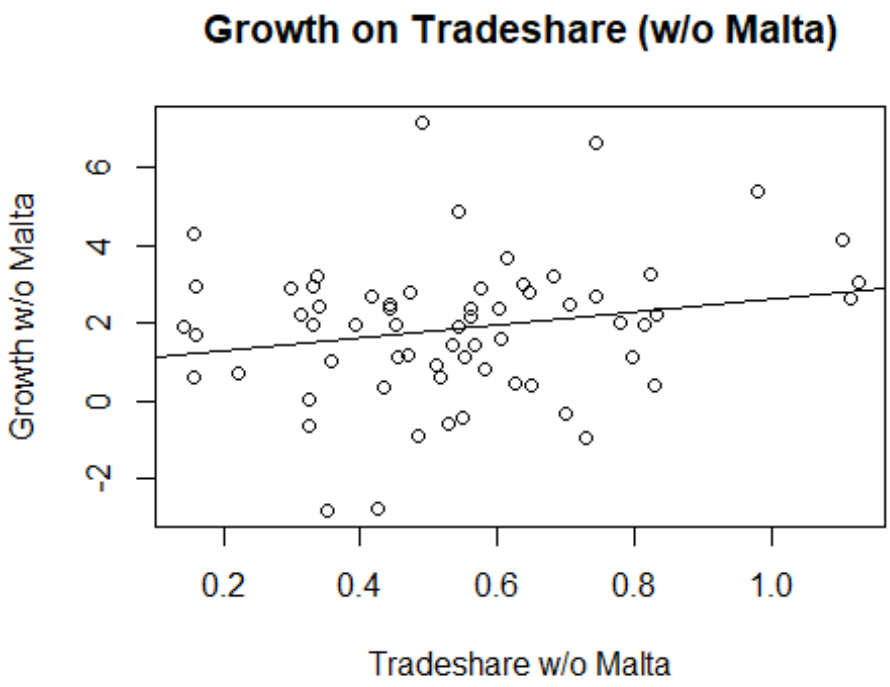
$$if TradeShare = 0.5, (Growth\hat{)} = 1.793482$$

$$TradeShare = 1.0, (Growth\hat{)} = 2.946699$$

Qn 2d

```
library(readr)
Growth_no_M <- read_csv("E4.1_Growth_exclude_Malta.csv", show_col_types =
FALSE)

model <- lm(Growth_no_M$growth~Growth_no_M$tradeshare)
plot.default(x=Growth_no_M$tradeshare, y=Growth_no_M$growth,
             main = "Growth on Tradeshare (w/o Malta)", type = "p",
             xlab = "Tradeshare w/o Malta", ylab = "Growth w/o Malta")
# x and y labels
abline(model)
```




```
model <- lm(Growth_no_M$growth~Growth_no_M$tradeshare)

summary(model)

##
## Call:
## lm(formula = Growth_no_M$growth ~ Growth_no_M$tradeshare)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4247 -0.9383  0.2091  0.9265  5.3776
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.9574     0.5804   1.650   0.1041
## Growth_no_M$tradeshare     1.6809     0.9874   1.702   0.0937 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.789 on 62 degrees of freedom
## Multiple R-squared:  0.04466,    Adjusted R-squared:  0.02925
## F-statistic: 2.898 on 1 and 62 DF,  p-value: 0.09369
```

$$Estimated\ slope, \beta_1 = 1.680905$$

$$Estimated\ intercept, \beta_0 = 0.957411$$

$$Growth = 0.957411 + 1.680905 TradeShare$$

$$if\ TradeShare = 0.5, Growth = 1.7978635$$

$$if\ TradeShare = 1.0, Growth = 2.638316$$

Qn 2e part f in E4.1

Malta is a Southern European island country in the Mediterranean Sea and the world's tenth smallest country in terms of land area. Being a coastal country with deep port, it is a popular freight transport site, receiving imports and exports enroute from other countries travelling from the northern to southern hemisphere via the Suez Canal, hence, explaining its massive shipping transaction volume and hight tradeshare.

Malta should not be included in the analysis as its large shipping transaction volume is not representative of the country's actual annual export or import. The shipping transactions are not intermediate and do not receive further processing or value-added production in Malta itself. Instead, they are passing through Malta as part of a logistic route. Thus, the high tradeshare is not indicative of the country's actual trade volume as it does not contribute to Malta's organic economic growth.

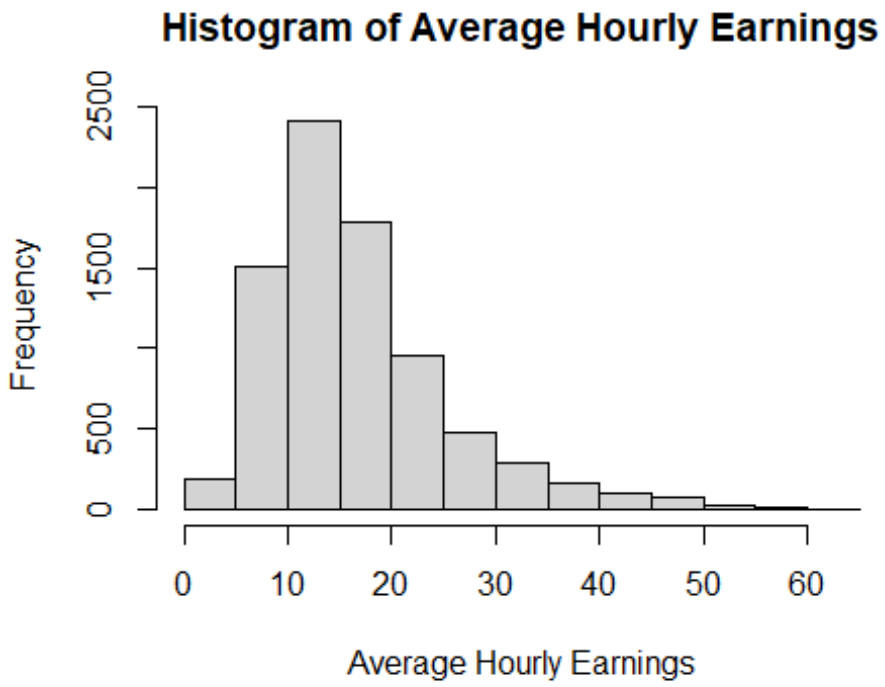
Question 3: CPS04.xls

Import in CPS04 dataset

```
library(readxl)
cps04_data <- read_excel("CPS04.xls")
```

a) Plot Histogram of Average Hourly Earnings

```
hist(cps04_data$ahe,
     main = "Histogram of Average Hourly Earnings",
     xlab = "Average Hourly Earnings")
```



Do you think that ahe is Normally Distributed?

Use Jarque-bera Test for Normality

```
library(tseries)

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

library(moments)
jarque.bera.test(cps04_data$ahe)

##
##  Jarque Bera Test
##
```

```
## data:  cps04_data$ahe
## X-squared = 4991.6, df = 2, p-value < 2.2e-16

jarque.test(cps04_data$ahe)

##
##  Jarque-Bera Normality Test
##
## data:  cps04_data$ahe
## JB = 4991.6, p-value < 2.2e-16
## alternative hypothesis: greater

jb_statistic <- jarque.bera.test(cps04_data$ahe)[1]
jb_p_value <- jarque.bera.test(cps04_data$ahe)[3]
```

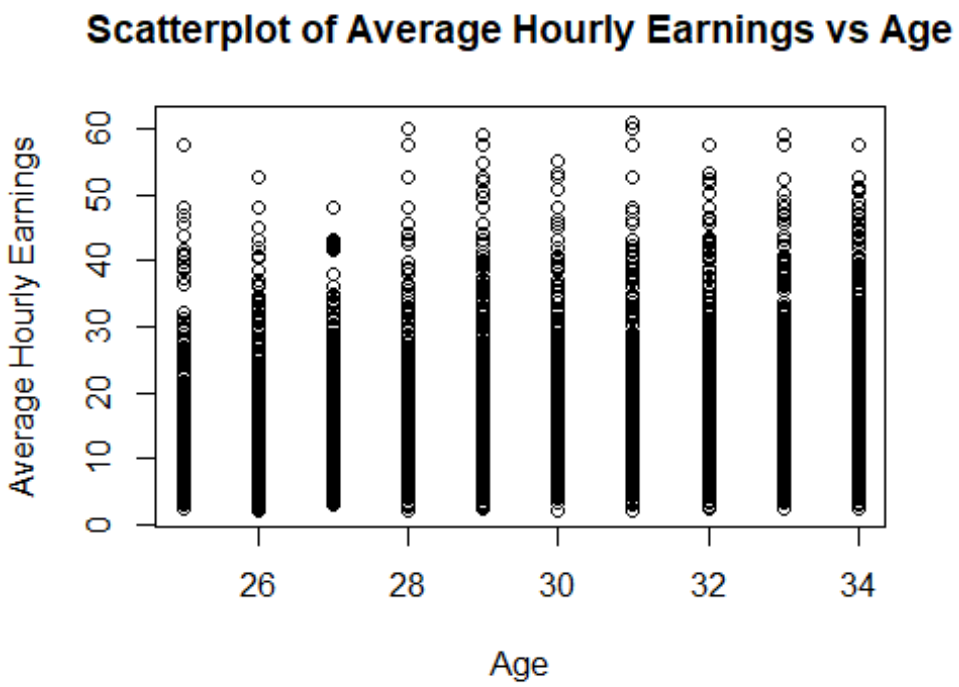
The result of Jarque-Bera Test: 4991.603

p-value: 0

Since p-value is low, we reject the null that the average hourly earnings is normal. We conclude that there is significant evidence that the dataset is not normal.

b)Scatterplot of Average Hourly Earnings on Age

```
plot.default(x = cps04_data$age,
             y = cps04_data$ahe,
             main = "Scatterplot of Average Hourly Earnings vs Age",
             xlab = "Age",
             ylab = "Average Hourly Earnings")
```



Visually, there is no heteroskedasticity (variance of error terms do not increase as independent variable age changes)

Run Regression of ahe on age with White's Standard Errors

c)

```
library(estimatr)
model_robust_ahe_age <- lm_robust(cps04_data$ahe ~ cps04_data$age,
                                  se_type = "HC1")
summary(model_robust_ahe_age)

##
## Call:
## lm_robust(formula = cps04_data$ahe ~ cps04_data$age, se_type = "HC1")
##
## Standard error type:  HC1
##
## Coefficients:
##                 Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper
DF
## (Intercept)      3.3242    0.96570   3.442 5.798e-04   1.4312   5.2172
7984
## cps04_data$age   0.4519    0.03297  13.708 2.715e-42   0.3873   0.5166
7984
##
```

```
## Multiple R-squared:  0.02225 ,   Adjusted R-squared:  0.02213
## F-statistic: 187.9 on 1 and 7984 DF,  p-value: < 2.2e-16

robust_ahe_age_coeff <- summary(model_robust_ahe_age)$coeff

robust_ahe_age_intercept <- round(robust_ahe_age_coeff[1,1], 3)
#robust_ahe_age_intercept <- format(robust_ahe_age_intercept, digits = 2,
nsmall = 3)

robust_ahe_age_slope <- round(robust_ahe_age_coeff[2,1], 3)
#robust_ahe_age_slope <- format(robust_ahe_age_slope, digits = 2, nsmall = 3)
```

Intercept Term: 3.324

Slope Term : 0.452

d)Bob (age = 26 years old), Alexis (age = 30 years old) Predict their earnings

```
earnings_bob    <- robust_ahe_age_coeff[1,1] + (robust_ahe_age_coeff[2,1] *
26)
earnings_alexis <- robust_ahe_age_coeff[1,1] + (robust_ahe_age_coeff[2,1] *
30)

earnings_bob <- format(earnings_bob, digits = 4)
earnings_alexis <- format(earnings_alexis, digits = 4)
```

Bob's Estimated Earnings : $15.07

Alexis's Estimated Earnings: $16.88

e)Test the hypothesis that the slope is 0

H0: The Slope(beta1) is 0 H1: The Slope(beta1) is != 0

```
slope_coeff_p_value  <- robust_ahe_age_coeff[2,4]
slope_coeff_testStat <- round(robust_ahe_age_coeff[2,3], 3)
slope_coeff_SE       <- robust_ahe_age_coeff[2,2]
slope_coeff_CI_Lower <- round(robust_ahe_age_coeff[2,5], 3)
slope_coeff_CI_Upper <- round(robust_ahe_age_coeff[2,6], 3)
```

At alpha = 5% Significance level

Slope p-value: 2.7153466^{-42} is small, we reject H0. We conclude that there is sufficent evidence that the slope is not 0

Slope Test Statistic: 13.708. Since the test statistic is greater than 1.960. We reject H0

Confidence Interval for Slope: [0.387 , 0.517] We note that the Confidence interval is in the positive region. Hence the slope is not 0

f) Inteprete RSquare

The Regression R2 is a measure of goodness of fit of the regression model on the sample data, it shows the fraction of the sample variance of Y predicted by X. R2 is the ratio of ESS (Explained Sum of Squares) to TSS (Total Sum of Squares). In this study, R2 it is 0.0222 (the model only explains 2.22% of the variation of the average hourly earnings). In summary, this regression model of single regressor age does not predict the average hourly earnings well. This suggests that there may be other relevant factors which may influence the earnings.

g) Run Regression with White Standard Errors (lm_robust)

```
model_robust_ahe_bachelor <- lm_robust(cps04_data$ahe ~ cps04_data$bachelor,
                                       se_type = "HC1")
summary(model_robust_ahe_bachelor)

##
## Call:
## lm_robust(formula = cps04_data$ahe ~ cps04_data$bachelor, se_type = "HC1")
##
## Standard error type:  HC1
##
## Coefficients:
##                     Estimate Std. Error t value   Pr(>|t|) CI Lower CI
Upper
## (Intercept)           13.810     0.1021  135.29  0.000e+00   13.610
14.010
## cps04_data$bachelor    6.497     0.1884   34.49 4.528e-243    6.128
6.867
##                       DF
## (Intercept)         7984
## cps04_data$bachelor 7984
##
## Multiple R-squared:  0.1365 ,    Adjusted R-squared:  0.1364
## F-statistic:  1189 on 1 and 7984 DF,  p-value: < 2.2e-16

robust_ahe_bachelor_coeff <- summary(model_robust_ahe_bachelor)$coeff
robust_ahe_bachelor_intercept <- round(robust_ahe_bachelor_coeff[1,1], 2)
robust_ahe_bachelor_slope <- round(robust_ahe_bachelor_coeff[2,1], 2)
robust_ahe_bachelor_intercept_slope <- round(robust_ahe_bachelor_coeff[2,1] +
robust_ahe_bachelor_coeff[1,1], 2)
```

A binary variable is also an indicator variable (aka Dummy variable). The textbook mentions that the slope for a binary variable regressor does not make sense. Given that the worker has no Bachelor (Bachelor = 0), Average hourly earnings will be $13.81/hour. Given that the worker has a Bachelor (Bachelor = 1), Average hourly earnings will be 20.31/hour. A worker with a Bachelor commands a premium average hourly earnings of $6.5/hour.

h) Run Regression with White Standard Errors (lm_robust)

```
model_robust_ahe_gender <- lm_robust(cps04_data$ahe ~ cps04_data$female,
                                     se_type = "HC1")

summary(model_robust_ahe_gender)

##
## Call:
## lm_robust(formula = cps04_data$ahe ~ cps04_data$female, se_type = "HC1")
##
## Standard error type:  HC1
##
## Coefficients:
##                   Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper
DF
## (Intercept)         17.773     0.1361  130.58 0.000e+00   17.506    18.04
7984
## cps04_data$female   -2.414     0.1910  -12.64 2.761e-36   -2.788    -2.04
7984
##
## Multiple R-squared:  0.01844 ,   Adjusted R-squared:  0.01832
## F-statistic: 159.8 on 1 and 7984 DF,  p-value: < 2.2e-16

robust_ahe_gender_coeff <- summary(model_robust_ahe_gender)$coeff
robust_ahe_gender_intercept <- round(robust_ahe_gender_coeff[1,1], 2)
robust_ahe_gender_slope <- round(robust_ahe_gender_coeff[2,1], 2)
robust_ahe_gender_slope_intercept <- round(robust_ahe_gender_intercept +
robust_ahe_gender_slope, 2)
```

Similar to the previous regressor on Bachelor, but the coefficient of Female regressor is negative value. Given that a worker is a male (Female = 0), he will be predicted to have an average hourly earning of $17.77/hour. Given that a worker is a female (Female = 1), she will be predicted to have an average hourly earning of $15.36/hour.

Question 4: Empirical Exercise 5.3

Q E.5.3

```
BS <- read_excel("BS.xlsx")
```

(a) i) What is average birth weight for infants for all mothers?

```
birthweight_avg <- mean(BS$birthweight)
birthweight_avg <- round(birthweight_avg, 2)
```

The avg. birth weight of infants for all mothers: 3382.93 grams

(a) ii) What is average birth weight for infants for mothers who smoked?

```
## filter smoker from the data set
smoker.weight <- subset(BS, smoker == 1)
smoker.weight_avg <- mean(smoker.weight$birthweight)
smoker.weight_avg <- round(smoker.weight_avg, 2)
```

The avg. birth weight of infants for mothers who smoke: 3178.83 grams

(a) iii) What is average birth weight for infants for mothers who do not smoke?

```
## filter non-smoker from the data set
nonsmoker.weight <- subset(BS, smoker == 0)
nonsmoker.weight_avg <- mean(nonsmoker.weight$birthweight)
nonsmoker.weight_avg <- round(nonsmoker.weight_avg, 2)
```

The avg. birth weight of infants for mothers who do not smoke: 3432.06 grams

(b)i) Estimate the differerence in birth weight for Smoking & Non-Smoking Mothers

```
model_birthweight_smoker <- lm(BS$birthweight ~ BS$smoker)
summary(model_birthweight_smoker)
```

```
##
## Call:
## lm(formula = BS$birthweight ~ BS$smoker)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3007.06  -313.06    26.94   366.94  2322.94
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3432.06      11.87 289.115   <2e-16 ***
## BS$smoker    -253.23      26.95  -9.396   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 583.7 on 2998 degrees of freedom
## Multiple R-squared:  0.0286, Adjusted R-squared:  0.02828
## F-statistic: 88.28 on 1 and 2998 DF,  p-value: < 2.2e-16
```

```
birthweight_smoker_coeff <- summary(model_birthweight_smoker)$coefficients

birthweight_smoker_intercept <- birthweight_smoker_coeff[1,1]
birthweight_smoker_intercept <- round(birthweight_smoker_intercept, 2)

birthweight_smoker_slope <- birthweight_smoker_coeff[2,1]
birthweight_smoker_slope <- round(birthweight_smoker_slope, 2)
```

Birthweight = 3432.06 + -253.23 * smoker

Birthweight (non-smoking mother) = (3432.06 + -253.23 * 0) grams

Birthweight (mother who smokes) = (3432.06 + -253.23 * 1) grams

Since the smoker regressor is a dummy variable, the difference in average birth weight of infants for mothers who smoke vs mothers who do not smoke is just the slope (-253.23 grams)

b)ii) What is the Standard Error for the estimated difference?

```
birthweight_smoker_slope_SE <- birthweight_smoker_coeff[2,2]
birthweight_smoker_slope_SE <- round(birthweight_smoker_slope_SE, 2)
```

The Standard Error for the slope coefficient is 26.95

Alternatively, we can compute the Standard Error of the slope for Smokers

```
sd(BS$birthweight)/sqrt(length(BS$birthweight))

## [1] 10.81137

## To caculate the standard error of  birthweight of smoker mothers
smoker.weight_sd <- sd(smoker.weight$birthweight)

## To caculate the standard error of  birthweight of smoker mothers
nonsmoker.weight_sd <- sd(nonsmoker.weight$birthweight)

nonsmoker.weight_sd <- sd(nonsmoker.weight$birthweight) /
                          sqrt(length(nonsmoker.weight$birthweight))

## The standard error of the difference between smoker and nonsmoker
birthweight
std.s.non <-
sqrt((sd(nonsmoker.weight$birthweight)/sqrt(length(nonsmoker.weight$birthweig
ht)))^2+(sd(smoker.weight$birthweight)/sqrt(length(smoker.weight$birthweight)
))^2)
std.s.non <- format(round(std.s.non, 2), nsmall = 3)
std.s.non <- as.numeric(std.s.non)
```

The Standard Error for the difference in birth weight: 26.82

b) iii) Construct 9% Confidence Interval for the Difference in birth weight

```
CI_error <- round((qnorm(0.975)* std.s.non),2)
CI_left <- round((birthweight_smoker_slope - CI_error), 2)
CI_right <- round((birthweight_smoker_slope + CI_error), 2)
```

Therefore, the 95% confidence interval is [-305.8 , -200.66]

c) Run Regression of Infant Birth Weight on Smoker

The intercept is the average infant birth weight for non-smokers (Smoker = 0). The slope is the difference between average infant birth weights for smokers (Smoker = 1) and non-smokers (Smoker = 0)

c)ii)

They are roughly the same.

c)iii)

```
CI_smoker_slope <- confint(model_birthweight_smoker, level = 0.95)
```

The Confidence Interval is [-306.0736375 , -200.383066]. This the same as the confidence interval in (b). We note that the Confidence Interval lie in the negative region and that we have 95% confidence that the difference in infant birth weight lies in the negative region (Mothers who smoke are correlated with a decrease in infant birth weight)

d)

No, smoking is not uncorrelated with other factors. Just solely determining the birth weight of infants based on whether a mother smokes is not a good gauge. The simple linear regression model RSquared gives 0.0286 (which allows it to estimate only 2.8% of the infant's weight). Additionally, we know that there are other factors that are correlated with whether a mother smokes or not, these variables may include education level, married or unmarried, alcohol consumption, number of drinks per week.