DARIEN OLIVA
ECON 453 - PSET 4/5

Question #1

A) Estimate a Regression Model for Medical Expenses

Showing R code and output for the Regression, see script for more detail

model1<-lm(medicalexpn~income+education+D_RURAL+D_USA+D_CANADA, data=pset4_data)

PREDICTED MEDICAL EXPENSE = 8.84032+.33494(income)+.50136(education)-12.64338(D_RURAL)

(Since D_USA and D_CANADA have insignificant p-values, they are removed from the model.)

#Residual standard error: 6.891 on 79 degrees of freedom
#Multiple R-squared: 0.7787, Adjusted R-squared: 0.7646
#F-statistic: 55.58 on 5 and 79 DF, p-value:<2.2e-16

The p-value is less than the 1% level of significance, telling us that the data is significant.

```
> print(vcov(model1),digits=4)

            (Intercept)    income education  D_RURAL    D_USA D_CANADA
(Intercept)      6.9387 -0.022097 -0.322000 -2.80130 -1.18160 -1.06601
income          -0.0221  0.002913 -0.007561  0.01437 -0.02763 -0.02222
education       -0.3220 -0.007561  0.052133  0.10664  0.03339  0.01749
D_RURAL         -2.8013  0.014370  0.106644  3.01003 -0.44412 -0.73254
D_USA           -1.1816 -0.027634  0.033387 -0.44412  3.79750  2.18407
D_CANADA        -1.0660 -0.022221  0.017486 -0.73254  2.18407  3.78708
```

NOW FOR THE MODEL FOR MEXICO. AVOID DUMMY VARIABLE TRAP

> model2 <- lm(medicalexpn~income+education+D_RURAL+D_USA+D_MEXICO, data=pset4_data)

Residual standard error: 6.891 on 79 degrees of freedom

Multiple R-squared: 0.7787, Adjusted R-squared: 0.7646

F-statistic: 55.58 on 5 and 79 DF, p-value: < 2.2e-16

> print(vcov(model2),digits=4)

	(Intercept)	income	education	D_RURAL	D_USA	D_MEXICO
(Intercept)	8.59374	-0.044318	-0.304514	-3.53384	-1.718607	-2.72107
income	-0.04432	0.002913	-0.007561	0.01437	-0.005413	0.02222
education	-0.30451	-0.007561	0.052133	0.10664	0.015901	-0.01749
D_RURAL	-3.53384	0.014370	0.106644	3.01003	0.288416	0.73254
D_USA	-1.71861	-0.005413	0.015901	0.28842	3.216439	1.60301
D_MEXICO	-2.72107	0.022221	-0.017486	0.73254	1.603012	3.78708

B) Test the null hypothesis that country and location jointly have no effect on medical expenses.

SEE ALL REGRESSIONS FOR Q1B IN ADDENDUM

For this question I ran regressions for medical scores with all countries and locations. For each regression, the result was that the p-value was less than alpha at at least a 5% confidence level. From these results, we can conclude that country and location have a joint effect on medical expenses. Since we could confirm this for each country and location at at least a 5% significance level, we know that the data/relationship is significant, and we must reject the null hypothesis.

C) Construct a 95% Confidence Interval for the Two Households

SEE CODE IN ADDENDUM

The 95% confidence interval for medical expenses for a rural Mexican household with an income of $50,000 and 12 years of education is between 18.96027 and 33.25869. The 95% confidence interval for medical expenses for an urban US household with the same level of income and education is between 30.03268 and 44.06091. From these intervals we can see that these overlap. The difference in costs for these models is approximately $11 higher for the USA urban model. The distribution is right shifting.

Question #2

A) Do both schools have increases in scores from 2014-2016?

score+ = β β *d*+2016 ε *i*01*ii*

SEE CODE/FULL REGRESSION RESULTS IN ADDENDUM

$H0 : \beta 1 = 0$, $H1 : \beta 1 > 0$

The first regression for the Kennedy School had a p-values less than alpha at at least 5% significance levels, meaning we could reject the null and conclude there was an increase in scores for the Kennedy School between 2014 and 2016. Next I repeated these steps for the Lincoln School, which also had a p-values less than alpha at at least 5% significance levels. This allows us to conclude that the claim is correct that both schools experienced increases in scores between 2014 and 2016.

B) Does performance of fourth graders differ between schools?

SEE CODE/FULL REGRESSION RESULTS IN ADDENDUM

$score i =+\beta 0\ \beta 1 d+keni\ \varepsilon i$

$H0 : \beta 1 = 0$, $H1 : \beta 1 \neq 0$

The regression for testing the differences in scores between schools resulted in a p-value less than alpha at a 5% significance level. It showed that the difference in the data between schools is significant. This allows us to reject the null and conclude that the performance of fourth graders did differ between the Kennedy School and the Lincoln School.

C) Does performance of fourth graders differ between schools each year?

SEE CODE/FULL REGRESSION RESULTS IN ADDENDUM

$score i =+\beta 0\ \beta 1 d+keni\ \beta 2 d2+016\ \varepsilon i.$

$H0 :\beta 1 =0,\beta 2\ 0,\ H=1$:Atleastoneof$\beta 1$ *and*$\beta 2 \neq 0$

This regression also resulted in p-values at at least 5% of significance level. This shows us that the difference in data between scores and schools are significantly different in bith years 2014 and 2016. Therefore we can reject the null hypothesis and conclude the average test scores for both schools are not the same for either year.

Question #3

A) Scatterplot for the Data (Generated in R)

From the scatterplot below, we can see that the living wage increases with age until approximately the 50 year mark. After the 50 year age mark the living age begins to decrease as age continues to rise.

B) Data Summary Tables (EXCEL)

TRAINING SET (OBSERVATIONS 1:140)	MODEL 1
ESTIMTATED COEFFICENTS	-
INTERCEPT	40.14946
GRADUATE	6.1061
AGE	0.07249
STANDARD ERRORS	-
INTERCEPT	1.36209
GRADUATE	0.78197
AGE	0.02612
P-VALUES	-
INTERCEPT	2.00E-16
GRADUATE	1.35E-12
AGE	0.00628
R-SQUARED	0.3348
ADJUSTED R-SQUARED	0.325

VALIDATION SET (OBSERVATIONS 140:160)	MODEL 2
ESTIMTATED COEFFICENTS	-
INTERCEPT	-1.018002
GRADUATE	6.078027
AGE	1.856074
I(AGE^2)	-0.018106
STANDARD ERRORS	-
INTERCEPT	5.105533
GRADUATE	1.05605
AGE	0.240329
I(AGE^2)	0.002702
P-VALUES	-
INTERCEPT	8.44E-01
GRADUATE	2.95E-05
AGE	8.73E+00
I(AGE^2)	5.09E-06
R-SQUARED	0.8969
ADJUSTED R-SQUARED	0.8775

C) Predict Expected Wage from Both Models.

FOR MODEL 1 : predict.lm(model3_1, data.frame(Graduate=1, Age=30))
For model 1, the estimated wage for a 30-year-old individual with a graduate degree is 48.43022.
FOR MODEL 2 : predict.lm(model3_2, data.frame(Graduate=1, Age=30))
For model 2, the estimated wage for a 30-year-old individual with a graduate degree is 44.44729.

D) At What Age Are Wages Maximum?

R CALCULATIONS:

```
> res3_2=summary(model3_2)

> coef(res3_2)
              Estimate  Std. Error    t value     Pr(>|t|)
(Intercept) -1.01800194 5.105533196 -0.1993919 8.444681e-01
Graduate     6.07802700 1.056049991  5.7554349 2.952426e-05
Age          1.85607421 0.240328578  7.7230691 8.731945e-07
I(Age^2)    -0.01810551 0.002701852 -6.7011464 5.090307e-06
> -coef(res3_2)[3,1]/(2*coef(res3_2)[4,1])

[1] 51.25717
```

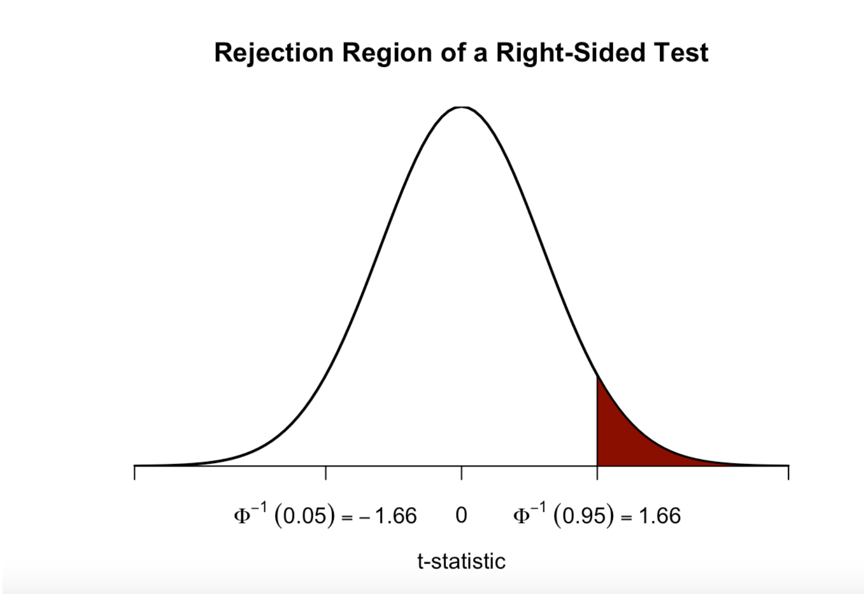
RESULTS: According to our model, the age in which the living wages are maximized is approximately 51 years old.

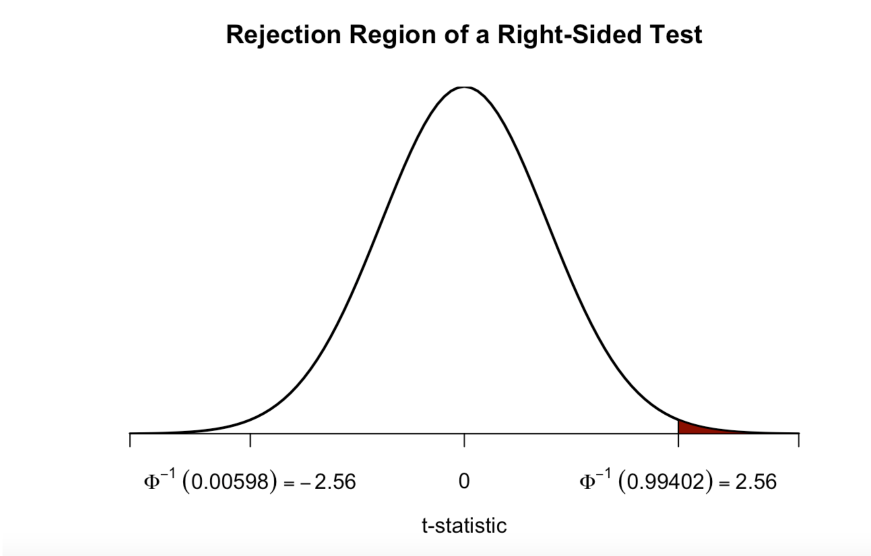
E) Cross Validation Results

DARIEN OLIVA - ECON 453 - PSET4/5		
Cross Validation Results for Wage Data		
TRAINING SET (OBSERVATIONS 1-140)	MODEL 1	MODEL 2
R-SQUARED	0.3348	
ADJUSTED R-SQUARED	0.325	
[correlation (wages, wages)^2]	0.3347565	
MSE	20.10457	
RMSE	4.483812	
MAE	3.664206	
MAPE	8.092382	
RSE	4.532639	
VALIDATION SET (OBSERVATIONS 141-160)		
R-SQUARED		0.8969
ADJUSTED R-SQUARED		0.8775
[correlation (wages, wages)^2]		0.8968757
MSE		2.830505
RMSE		1.682411
MAE		1.385413
MAPE		2.924267
RSE		1.880992

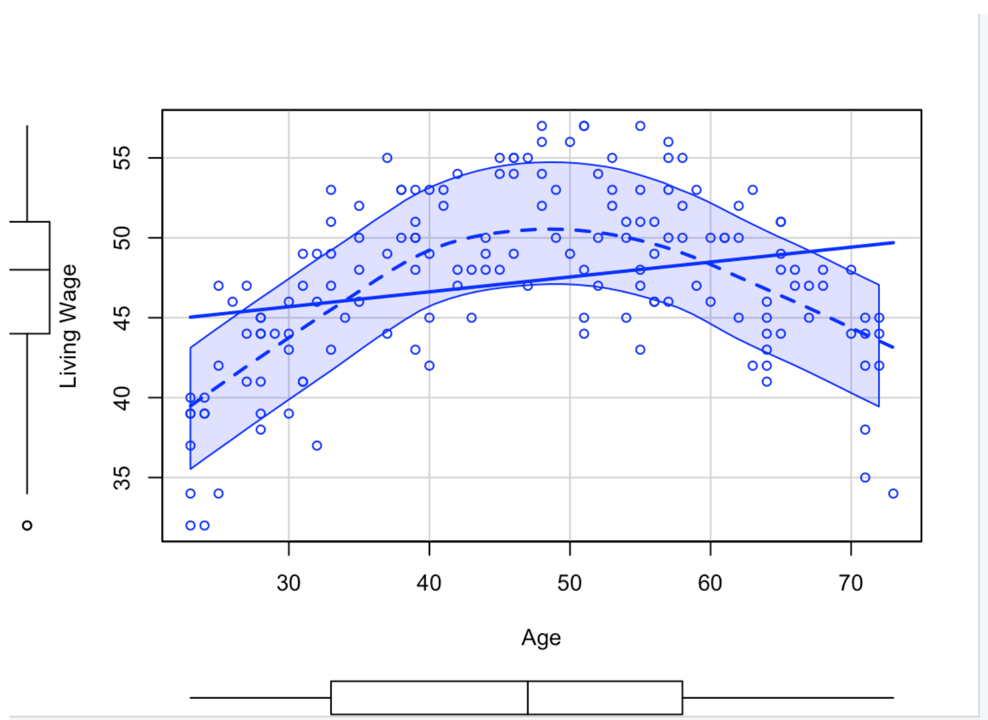
We know that the lower the MAE, MSE, and RMSE, the more accurate the regression model is. Our cross validation shows that all three of the MAE, MSE, and RMSE are lower for the validation model/model 2. Additionally, the model with the higher R-Squared value is typically preferred and that model is also model 2.

ADDENDUM - R GENERATED PLOTS BY DARIEN OLIVA






Living Wage Pertaining to Age
Living Wage
55
50
45
40
35
30
40
50
60
70
Age


Living Wage
55
50
45
40
35
30
40
50
60
70
Age

DARIEN OLIVA - ECON 453 - PSET4/5 - EXCEL TABLES

TRAINING SET (OBSERVATIONS 1:140)	MODEL 1
ESTIMTATED COEFFICENTS	-
INTERCEPT	40.14946
GRADUATE	6.1061
AGE	0.07249
STANDARD ERRORS	-
INTERCEPT	1.36209
GRADUATE	0.78197
AGE	0.02612
P-VALUES	-
INTERCEPT	2.00E-16
GRADUATE	1.35E-12
AGE	0.00628
R-SQUARED	0.3348
ADJUSTED R-SQUARED	0.325

VALIDATION SET (OBSERVATIONS 140:160)	MODEL 2
ESTIMTATED COEFFICENTS	-
INTERCEPT	-1.018002
GRADUATE	6.078027
AGE	1.856074
I(AGE^2)	-0.018106
STANDARD ERRORS	-
INTERCEPT	5.105533
GRADUATE	1.05605
AGE	0.240329
I(AGE^2)	0.002702
P-VALUES	-
INTERCEPT	8.44E-01
GRADUATE	2.95E-05
AGE	8.73E+00
I(AGE^2)	5.09E-06
R-SQUARED	0.8969
ADJUSTED R-SQUARED	0.8775

Cross Validation Results for Wage Data		
TRAINING SET (OBSERVATIONS 1-140)	**MODEL 1**	**MODEL 2**
R-SQUARED	0.3348	
ADJUSTED R-SQUARED	0.325	
[correlation (wages, wages)^2]	0.3347565	
MSE	20.10457	
RMSE	4.483812	
MAE	3.664206	
MAPE	8.092382	
RSE	4.532639	
VALIDATION SET (OBSERVATIONS 141-160)		
R-SQUARED		0.8969
ADJUSTED R-SQUARED		0.8775
[correlation (wages, wages)^2]		0.8968757
MSE		2.830505
RMSE		1.682411
MAE		1.385413
MAPE		2.924267
RSE		1.880992

ADDENDUM / R- SCRIPT BY DARIEN OLIVA

```
> # DARIEN OLIVA - ECON 453 - PSET4&5
> # DATA FROM PSET1
>
> rm(list = ls())
> library(car)
> library(carData)
> library(readxl)
> pset4_data <- read_excel("Downloads/pset1(4&5)_data.xlsx",
+ sheet = "medical_expenses")
> View(pset4_data)
>
> #VIEW DATA STRUCTURE
```

```
> str(pset4_data)
tibble [85 × 5] (S3: tbl_df/tbl/data.frame)
 $ country    : chr [1:85] "USA" "USA" "USA" "USA" ...
 $ location   : chr [1:85] "RURAL" "RURAL" "RURAL" "RURAL" ...
 $ medicalexpn: num [1:85] 22.27 8.98 8.06 15.6 7.99 ...
 $ income     : num [1:85] 19 42 54 14 36 39 21 75 22 16 ...
 $ education  : num [1:85] 12 10 11 5 11 10 5 14 6 13 ...
>
> # QUESTION 1 A
>
>
>
> #ASSIGN DUMMY VARIABLES, I DISCLUDED MEXICO.
> pset4_data$D_RURAL=ifelse((pset4_data$location == "RURAL"), 1,0)
> pset4_data$D_USA=ifelse((pset4_data$country == "USA"), 1,0)
> pset4_data$D_CANADA=ifelse((pset4_data$country == "CANADA"), 1,0)
> pset4_data$D_MEXICO=ifelse((pset4_data$country == "MEXICO"), 1,0)
>
> #CHECK FOR 1S AND 2S
> pset4_data$Check<-
pset4_data$D_RURAL+pset4_data$D_USA+pset4_data$D_CANADA+pset4_data$D_MEXICO
> head(pset4_data,10)
# A tibble: 10 × 10
   country location medicalexpn income education D_RURAL D_USA D_CANADA D_MEXICO
Check
   <chr>   <chr>          <dbl>  <dbl>     <dbl>   <dbl> <dbl>    <dbl>    <dbl> <dbl>
 1 USA     RURAL          22.3      19        12       1     1        0        0     2
 2 USA     RURAL           8.98     42        10       1     1        0        0     2
 3 USA     RURAL           8.06     54        11       1     1        0        0     2
 4 USA     RURAL          15.6      14         5       1     1        0        0     2
 5 USA     RURAL           7.99     36        11       1     1        0        0     2
 6 USA     RURAL           5.57     39        10       1     1        0        0     2
 7 USA     RURAL           1        21         5       1     1        0        0     2
 8 USA     RURAL          25.4      75        14       1     1        0        0     2
 9 USA     RURAL           3.43     22         6       1     1        0        0     2
10 USA     RURAL           4.87     16        13       1     1        0        0     2
> tail(pset4_data,5)
# A tibble: 5 × 10
  country location medicalexpn income education D_RURAL D_USA D_CANADA D_MEXICO
Check
  <chr>   <chr>          <dbl>  <dbl>     <dbl>   <dbl> <dbl>    <dbl>    <dbl> <dbl>
1 MEXICO  URBAN          39.7      59        14       0     0        0        1     1
2 MEXICO  URBAN          17.9      11         5       0     0        0        1     1
3 MEXICO  URBAN          29.6      42        13       0     0        0        1     1
```

```
4 MEXICO  URBAN          3.31     13        3       0     0        0        1     1
5 MEXICO  URBAN         41.3      39       18       0     0        0        1     1
>
> #NEW DATA SUMMARY
> summary(pset4_data)
   country            location          medicalexpn         income         education
 Length:85          Length:85          Min.   : 1.000   Min.   : 4.00   Min.   : 0.00
 Class :character   Class :character   1st Qu.: 8.208   1st Qu.:21.00   1st Qu.: 6.00
 Mode  :character   Mode  :character   Median :16.351   Median :34.00   Median :11.00
                                       Mean   :19.188   Mean   :37.42   Mean   :10.18
                                       3rd Qu.:26.822   3rd Qu.:48.00   3rd Qu.:13.00
                                       Max.   :62.231   Max.   :99.00   Max.   :18.00
    D_RURAL           D_USA           D_CANADA         D_MEXICO          Check
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :1.000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.000
 Median :1.0000   Median :0.0000   Median :0.0000   Median :0.0000   Median :2.000
 Mean   :0.5294   Mean   :0.3529   Mean   :0.3529   Mean   :0.2941   Mean   :1.529
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:2.000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :2.000
>
> #FIRST LINEAR REGRESSION MODEL:
>
> model1<-lm(medicalexpn~income+education+D_RURAL+D_USA+D_CANADA,
data=pset4_data)
> summary(model1)

Call:
lm(formula = medicalexpn ~ income + education + D_RURAL + D_USA +
    D_CANADA, data = pset4_data)

Residuals:
    Min      1Q  Median      3Q     Max
-11.392  -4.553  -1.285   4.045  15.259

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.84034    2.63414   3.356  0.00122 **
income        0.33494    0.05397   6.205 2.36e-08 ***
education     0.50136    0.22833   2.196  0.03104 *
D_RURAL     -12.64338    1.73494  -7.287 2.10e-10 ***
D_USA        -1.57098    1.94872  -0.806  0.42257
D_CANADA     -0.11764    1.94604  -0.060  0.95195
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 6.891 on 79 degrees of freedom
Multiple R-squared:  0.7787,  Adjusted R-squared:  0.7646
F-statistic: 55.58 on 5 and 79 DF,  p-value: < 2.2e-16

>
> # REGRESSION MODEL/ EQUATION FOR Q#1A
>
> # PREDICTED MEDICAL EXPENSE =
8.84032+.33494(income)+.50136(education)-12.64338(D_RURAL)
> # SINCE D_USA AND D_CANADA HAVE INSIGNIFICANT P-VALUES, THEY ARE
REMOVED FROM THE MODEL.
>
> #Residual standard error: 6.891 on 79 degrees of freedom
> #Multiple R-squared: 0.7787, Adjusted R-squared: 0.7646
> #F-statistic: 55.58 on 5 and 79 DF, p-value:<2.2e-16
> # The p-value is less than the 1% level of significance, telling us that the data is significant.
>
> print(vcov(model1),digits=4)
            (Intercept)    income education  D_RURAL    D_USA D_CANADA
(Intercept)      6.9387 -0.022097 -0.322000 -2.80130 -1.18160 -1.06601
income          -0.0221  0.002913 -0.007561  0.01437 -0.02763 -0.02222
education       -0.3220 -0.007561  0.052133  0.10664  0.03339  0.01749
D_RURAL         -2.8013  0.014370  0.106644  3.01003 -0.44412 -0.73254
D_USA           -1.1816 -0.027634  0.033387 -0.44412  3.79750  2.18407
D_CANADA        -1.0660 -0.022221  0.017486 -0.73254  2.18407  3.78708
>
> # NOW FOR THE MODEL FOR MEXICO. AVOID DUMMY VARIABLE TRAP
> model2 <- lm(medicalexpn~income+education+D_RURAL+D_USA+D_MEXICO,
data=pset4_data)
> summary(model2)

Call:
lm(formula = medicalexpn ~ income + education + D_RURAL + D_USA +
   D_MEXICO, data = pset4_data)

Residuals:
    Min      1Q  Median      3Q     Max
-11.392  -4.553  -1.285   4.045  15.259

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.72271    2.93151   2.976  0.00388 **
income       0.33494    0.05397   6.205 2.36e-08 ***
```

```
education     0.50136    0.22833   2.196  0.03104 *
D_RURAL     -12.64338    1.73494  -7.287 2.10e-10 ***
D_USA        -1.45334    1.79344  -0.810  0.42017
D_MEXICO      0.11764    1.94604   0.060  0.95195
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.891 on 79 degrees of freedom
Multiple R-squared:  0.7787,  Adjusted R-squared:  0.7646
F-statistic: 55.58 on 5 and 79 DF,  p-value: < 2.2e-16

>
> print(vcov(model2),digits=4)
            (Intercept)    income education  D_RURAL     D_USA D_MEXICO
(Intercept)     8.59374 -0.044318 -0.304514 -3.53384 -1.718607 -2.72107
income         -0.04432  0.002913 -0.007561  0.01437 -0.005413  0.02222
education      -0.30451 -0.007561  0.052133  0.10664  0.015901 -0.01749
D_RURAL        -3.53384  0.014370  0.106644  3.01003  0.288416  0.73254
D_USA          -1.71861 -0.005413  0.015901  0.28842  3.216439  1.60301
D_MEXICO       -2.72107  0.022221 -0.017486  0.73254  1.603012  3.78708
>
> # QUESTION 1 B
> #LINEAR HYPOTHESIS FOR MODEL 2
> linearHypothesis(model2, c("D_RURAL = 0", "D_USA=0","D_MEXICO=0"))
Linear hypothesis test

Hypothesis:
D_RURAL = 0
D_USA = 0
D_MEXICO = 0

Model 1: restricted model
Model 2: medicalexpn ~ income + education + D_RURAL + D_USA + D_MEXICO

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     82 6454.6
2     79 3750.9  3    2703.7 18.981 2.301e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # P-VALUE IS 0.000000002301 AND LESS THAN THE ALPHA. REJECT THE NULL
HYPOTHESIS AT 5% SIGNIFICANCE LEVEL.
>
> # LINEAR HYPOTHESIS FOR MODEL 1
```

```
> linearHypothesis(model1, c("D_RURAL = 0", "D_USA=0","D_CANADA=0"))
Linear hypothesis test

Hypothesis:
D_RURAL = 0
D_USA = 0
D_CANADA = 0

Model 1: restricted model
Model 2: medicalexpn ~ income + education + D_RURAL + D_USA + D_CANADA

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     82 6454.6
2     79 3750.9  3    2703.7 18.981 2.301e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # P-VALUE IS 0.000000002301 AND LESS THAN ALPHA SO WE CAN REJECT THE NULL HYPOTHEIS
> # AT EITHER 1 OR 5 PERCENT SIGNIFICANCE LEVEL.
> # RESULTS: LIVING IN MEXICO OR IN AN URBAN LOCATON DOES HAVE AN EFFECT ON MEDICAL EXPENSES.
>
> # LINEAR HYPOTHESIS FOR REFERENCE GROUP AND RURAL 1
>
> linearHypothesis(model1, c("D_RURAL = 1", "D_USA=0","D_CANADA=0"))
Linear hypothesis test

Hypothesis:
D_RURAL = 1
D_USA = 0
D_CANADA = 0

Model 1: restricted model
Model 2: medicalexpn ~ income + education + D_RURAL + D_USA + D_CANADA

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     82 6890.9
2     79 3750.9  3      3140 22.044 1.807e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # P-VALUE IS .0000000001807 WHICH IS LESS THAN ALPHA AT 1% SIGNIFICANCE LEVEL.
```

```
> # REJECT THE NULL
> # RESULTS: LIVING IN MEXICO OR IN AN URBAN LOCATION DOES HAVE AN EFFECT
ON MEDICAL EXPENSES.
>
> #LINEAR HYPOTHESIS FOR USA AND RURAL
> linearHypothesis(model1, c("D_RURAL = 1", "D_USA=1","D_CANADA=0"))
Linear hypothesis test

Hypothesis:
D_RURAL = 1
D_USA = 1
D_CANADA = 0

Model 1: restricted model
Model 2: medicalexpn ~ income + education + D_RURAL + D_USA + D_CANADA

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     82 6969.8
2     79 3750.9  3    3218.8 22.598 1.161e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # P_VALUE IS LESS THAN 1% SIGNIFICANCE LEVEL. REJECT THE NULL.
> # RESULTS; USA AND RURAL DOES HAVE AN EFFECT ON MDEICAL EXPENSES.
>
> # LINEAR HYPOTHESIS FOR USA AND URBAN
>
> linearHypothesis(model1, c("D_RURAL = 0", "D_USA=1","D_CANADA=0"))
Linear hypothesis test

Hypothesis:
D_RURAL = 0
D_USA = 1
D_CANADA = 0

Model 1: restricted model
Model 2: medicalexpn ~ income + education + D_RURAL + D_USA + D_CANADA

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     82 6533.1
2     79 3750.9  3    2782.2 19.532 1.438e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

```
> # P-VALUE IS LESS THAN APLHA AT 1% SIGNIFICANCE LEVEL, REJECT THE NULL.
> # USA AND URBAN DOES HAVE AN EFFECT ON MEDICAL EXEPENSES
>
> # LINEAR HYPOTHESIS FOR CANADA AND RURAL
>
> linearHypothesis(model1, c("D_RURAL = 1", "D_USA=0","D_CANADA=1"))
Linear hypothesis test

Hypothesis:
D_RURAL = 1
D_USA = 0
D_CANADA = 1

Model 1: restricted model
Model 2: medicalexpn ~ income + education + D_RURAL + D_USA + D_CANADA

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     82 6966.2
2     79 3750.9  3    3215.2 22.572 1.184e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # P-VALUE IS LESS THAN APLHA AT 1% SIGNIFICANCE LEVEL, REJECT THE NULL.
> # CANADA AND RURAL DOES HAVE AN EFFECT ON MEDICAL EXEPENSES
>
> # LINEAR HYPOTHESIS FOR CANADA AND URBAN
>
> linearHypothesis(model1, c("D_RURAL = 0", "D_USA=0","D_CANADA=1"))
Linear hypothesis test

Hypothesis:
D_RURAL = 0
D_USA = 0
D_CANADA = 1

Model 1: restricted model
Model 2: medicalexpn ~ income + education + D_RURAL + D_USA + D_CANADA

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     82 6523.6
2     79 3750.9  3    2772.7 19.465 1.522e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # P-VALUE IS LESS THAN APLHA AT 1% SIGNIFICANCE LEVEL, REJECT THE NULL.
> # CANADA AND URBAN DOES HAVE AN EFFECT ON MEDICAL EXEPENSES
```

```
>
> #DOUBLE CHECK AGAINST REFERENCE GROUP
> # (MEXICO AND URBAN AREA WITH MODEL #2)
> linearHypothesis(model2, c("D_RURAL = 0", "D_USA=0","D_MEXICO=1"))
Linear hypothesis test

Hypothesis:
D_RURAL = 0
D_USA = 0
D_MEXICO = 1

Model 1: restricted model
Model 2: medicalexpn ~ income + education + D_RURAL + D_USA + D_MEXICO

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     82 6361.7
2     79 3750.9  3    2610.7 18.328 4.041e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> #DOUBLE CHECK AGAINST REFERENCE GROUP
> # (MEXICO AND RURAL AREA WITH MODEL #2)
>
> linearHypothesis(model2, c("D_RURAL=1", "D_USA=0","D_MEXICO=1"))
Linear hypothesis test

Hypothesis:
D_RURAL = 1
D_USA = 0
D_MEXICO = 1

Model 1: restricted model
Model 2: medicalexpn ~ income + education + D_RURAL + D_USA + D_MEXICO

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     82 6791.5
2     79 3750.9  3    3040.5 21.346 3.183e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # QUESTION 1 C
> # CONSTRUCT A 95% CONFIDENCE INTERVAL
>
```

```
> predict.lm(model1, data.frame(income=50, education=12, D_RURAL=1, D_USA=0,
D_CANADA=0),
+           interval="predict", level = 0.95)
      fit      lwr      upr
1 18.96027 4.661852 33.25869
>
> # CHECK THIS WITH MODEL 2 FOR MEXICO AND RURAL
>
> predict.lm(model2,data.frame(income=50, education=12, D_RURAL=1, D_USA=0,
D_MEXICO=1))
      1
18.96027
>
> # IDENTICAL TO MODEL 1
>
> predict.lm(model1, data.frame(income=50, education=12, D_RURAL=0, D_USA=1,
D_CANADA=0),
+           interval="predict", level = 0.95)
      fit      lwr      upr
1 30.03268 16.00445 44.06091
>
> # CHECK THIS WITH MODEL 2 FOR USA AND URBAN
> predict.lm(model2, data.frame(income=50, education=12, D_RURAL=0, D_USA=1,
D_MEXICO=0))
      1
30.03268
>
> # IDENTICAL TO MODEL 1
>
> # COMPARE THE PREDICTED VALUES
>
> PREDICTEDMEXICORURAL = 18.96027
> PREDICTEDUSAURBAN = 30.03268
>
> (PREDICTEDMEXICORURAL)-(PREDICTEDUSAURBAN)
[1] -11.07241
>
> # THE DIFFERENCE IN COSTS FOR THESE MODELS IS APPROXIMATELY $11 FOR THE
USA URBAN MODEL.
> # OUR 95% CONFIDENCE INTERVAL SHOWS THAT THE INTERVALS FOR THE MODELS
DO OVERLAP. RIGHT SHIFTING DISTRIBUTION
>
>
> # QUESTION 2 A
```

```
>
> View(pset4_data)
> library(readxl)
> pset4_data <- read_excel("Downloads/pset1(4&5)_data.xlsx", sheet = "scores")
> View(pset4_data)
> str(pset4_data)
tibble [200 × 3] (S3: tbl_df/tbl/data.frame)
 $ year  : num [1:200] 2014 2014 2014 2014 2014 ...
 $ school: chr [1:200] "Lincoln" "Lincoln" "Lincoln" "Lincoln" ...
 $ score : num [1:200] 40 12 88 64 94 12 84 40 94 73 ...
>
>
> # ASSIGN DUMMY VARIABLES
> # 0 FOR LINCOLN
> # 1 FOR KENNEDY
>
> pset4_data$d_ken=ifelse((pset4_data$school =="Kennedy"), 1, 0)
> pset4_data$d_2016=ifelse((pset4_data$year ==2016), 1, 0)
> summary(pset4_data)
     year         school              score            d_ken         d_2016
 Min.   :2014   Length:200         Min.   :  3.00   Min.   :0.0   Min.   :0.0
 1st Qu.:2014   Class :character   1st Qu.: 47.75   1st Qu.:0.0   1st Qu.:0.0
 Median :2015   Mode  :character   Median : 63.00   Median :0.5   Median :0.5
 Mean   :2015                      Mean   : 62.37   Mean   :0.5   Mean   :0.5
 3rd Qu.:2016                      3rd Qu.: 83.00   3rd Qu.:1.0   3rd Qu.:1.0
 Max.   :2016                      Max.   :100.00   Max.   :1.0   Max.   :1.0
>
> # FIRST LINEAR MODEL FOR KENNEDY
>
> mken16 = lm(data=pset4_data[pset4_data$d_ken==1,],score~ d_2016)
> summary(mken16)

Call:
lm(formula = score ~ d_2016, data = pset4_data[pset4_data$d_ken ==
    1, ])

Residuals:
   Min     1Q Median     3Q    Max
-31.58 -15.37  -0.27  17.04  37.42

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.580      2.590  24.161   <2e-16 ***
d_2016         9.380      3.663   2.561    0.012 *
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.31 on 98 degrees of freedom
Multiple R-squared:  0.06272,	Adjusted R-squared:  0.05315 
F-statistic: 6.558 on 1 and 98 DF,  p-value: 0.01197

> 
> # RESULTS; THE P-VALUE IS LESS THAN THE ALPHA AT 5% SIGNIFICANCE LEVEL.
> # REJECT THE NULL HYPOTHESIS. THE DATA IN THIS MODEL IS SIGNIFCANT.
> 
> res<-summary(mken16)
> coef(res)
            Estimate Std. Error   t value     Pr(>|t|)
(Intercept)    62.58   2.590095 24.161276 4.571907e-43
d_2016          9.38   3.662947  2.560779 1.196947e-02
> pt(coef(res)[2,3], 98)
[1] 0.9940153
> 1-pt(coef(res)[2,3], 98)
[1] 0.005984733
> 
> m16 <- lm(score~d_ken+d_2016, data=pset4_data)
> summary(m16)

Call:
lm(formula = score ~ d_ken + d_2016, data = pset4_data)

Residuals:
   Min     1Q Median     3Q    Max 
-49.05 -16.10  -0.37  18.04  46.95 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   52.050      2.731  19.056  < 2e-16 ***
d_ken          9.800      3.154   3.107 0.002168 ** 
d_2016        10.840      3.154   3.437 0.000718 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.3 on 197 degrees of freedom
Multiple R-squared:  0.09826,	Adjusted R-squared:  0.08911 
F-statistic: 10.73 on 2 and 197 DF,  p-value: 3.762e-05

> 
```

```
> # RESULTS; THE P-VALUE IS LESS THAN THE ALPHA AT 1% SIGNIFICANCE LEVEL.
> linearHypothesis(m16, c("d_ken=1", "d_2016=1"))
Linear hypothesis test

Hypothesis:
d_ken = 1
d_2016 = 1

Model 1: restricted model
Model 2: score ~ d_ken + d_2016

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    199 106697
2    197  97983  2    8713.3 8.7592 0.0002268 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # PLOT THE T DISTRIBUTION DENSITY OVER THE DOMAIN [-4,4]
> curve(dt(x, df=98),
+       xlim = c(-4, 4),
+       main = "Rejection Region of a Right-Sided Test", yaxs = "i",
+       xlab = "t-statistic",
+       ylab = "",
+       lwd = 2,
+       axes = "F")
> axis(1,
+      at = c(-4,-qt(0.95, 98), 0, qt(0.95, 98), 4),
+      padj = 0.5,
+      labels = c("",expression(Phi^-1~(.05)==-1.66), 0, expression(Phi^-1~(.95)==1.66), ""))
> polygon(x = c(1.66, seq(1.66, 4, 0.01), 4),
+           y = c(0, dt(seq(1.66, 4, 0.01),df=98), 0), col = "darkred")
>
> pt(coef(res)[2,3], 98)
[1] 0.9940153
> 1-pt(coef(res)[2,3], 98)
[1] 0.005984733
> 2*(1-pt(coef(res)[2,3], 98))
[1] 0.01196947
>
> curve(dt(x, df=98),
+       xlim = c(-4, 4),
+       main = "Rejection Region of a Right-Sided Test", yaxs = "i",
+       xlab = "t-statistic",
+       ylab = "",
```

```
+       lwd = 2,
+       axes = "F")
> axis(1,
+      at = c(-4,-2.560779, 0, 2.560779, 4),
+      padj = 0.5,
+      labels = c("",expression(Phi^-1~(0.00598)==-2.56), 0,
expression(Phi^-1~(0.99402)==2.56),
+                  ""))
> polygon(x = c(2.56, seq(2.56, 4, 0.01), 4),
+         y = c(0, dt(seq(2.56, 4, 0.01), df=98),0),
+         col = "darkred")
>
> # FOR LINCOLN 2016
>
> ml16=lm(data=pset4_data[pset4_data$d_ken==0,],score~ d_2016)
> summary(ml16)

Call:
lm(formula = score ~ d_2016, data = pset4_data[pset4_data$d_ken ==
    0, ])

Residuals:
   Min     1Q Median     3Q    Max
-48.32 -17.62  -0.62  19.38  47.68

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   51.320      3.642  14.090   <2e-16 ***
d_2016        12.300      5.151   2.388   0.0189 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.75 on 98 degrees of freedom
Multiple R-squared:  0.05499,           Adjusted R-squared:  0.04534
F-statistic: 5.702 on 1 and 98 DF,  p-value: 0.01886


>
> # P-VALUE IS LESS THAN ALPHA AT SIGNIIFICANCE LEVEL 5%.
> # REJECT THE NULL HYPOTHESIS
>
> # FOR LINCOLN 2014
>
> linearHypothesis(ml16, c("d_2016 = 0"))
Linear hypothesis test
```

```
Hypothesis:
d_2016 = 0

Model 1: restricted model
Model 2: score ~ d_2016

  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     99 68787
2     98 65005  1    3782.3 5.7021 0.01886 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # P-VALUE IS LESS THAN ALPHA AT 5% SIGNIFICANCE LEVEL
> # REJECT THE NULL. RESULTS: THERE WAS AN EFFECT ON SCORES IN YEAR 2016.
>
> linearHypothesis(ml16, c("d_2016 = 1"))
Linear hypothesis test

Hypothesis:
d_2016 = 1

Model 1: restricted model
Model 2: score ~ d_2016

  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     99 68197
2     98 65005  1    3192.3 4.8126 0.03062 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # P-VALUE IS LESS THAN ALPHA AT 5% SIGNIFICANCE LEVEL
> # REJECT THE NULL
>
> m2_3=lm(data=pset4_data[pset4_data$d_2016==0,],score~ d_ken)
> summary(m2_3)

Call:
lm(formula = score ~ d_ken, data = pset4_data[pset4_data$d_2016 ==
    0, ])

Residuals:
   Min     1Q Median     3Q    Max
-48.32 -18.39   1.42  18.48  47.68
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   51.320      3.596  14.270   <2e-16 ***
d_ken         11.260      5.086   2.214   0.0292 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.43 on 98 degrees of freedom
Multiple R-squared:  0.04763,	Adjusted R-squared:  0.03791
F-statistic: 4.901 on 1 and 98 DF,  p-value: 0.02915

>
> # P-VALUE IS LESS THAN ALPHA AT 5% SIGNIFICANCE LEVEL
> # REJECT THE NULL
>
> linearHypothesis(m2_3, c("d_ken = 0"))
Linear hypothesis test

Hypothesis:
d_ken = 0

Model 1: restricted model
Model 2: score ~ d_ken

  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     99 66545
2     98 63375  1    3169.7 4.9014 0.02915 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # P-VALUE IS LESS THAN ALPHA AT 5% SIGNIFICANCE LEVEL
> # REJECT THE NULL
>
> linearHypothesis(m2_3, c("d_ken = 1"))
Linear hypothesis test

Hypothesis:
d_ken = 1

Model 1: restricted model
Model 2: score ~ d_ken

  Res.Df   RSS Df Sum of Sq      F Pr(>F)
```

```
1     99 66007
2     98 63375  1    2631.7 4.0695 0.0464 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # P-VALUE IS LESS THAN ALPHA AT 5% SIGNIFICANCE LEVEL
> # REJECT THE NULL. RESULTS: THE SCHOOL HAS A SIGNIFICANT EFFECT.
>
> m2_4=lm(data=pset4_data[pset4_data$d_2016==1,],score~ d_ken)
> summary(m2_4)

Call:
lm(formula = score ~ d_ken, data = pset4_data[pset4_data$d_2016 ==
    1, ])

Residuals:
   Min     1Q Median     3Q    Max
-40.62 -14.71  -2.29  18.12  34.38

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   63.620      2.654  23.976   <2e-16 ***
d_ken          8.340      3.753   2.222   0.0286 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.76 on 98 degrees of freedom
Multiple R-squared:  0.04798,           Adjusted R-squared:  0.03827
F-statistic: 4.939 on 1 and 98 DF,  p-value: 0.02855

>
> # THE P-VALUE IS 0.02855. THIS IS LESS THAN THE ALPHA AT 5% SIGNIFICANCE
LEVEL.
> # REJECT THE NULL.
>
> m2_5=lm(data=pset4_data,score~ d_ken+d_2016)
> summary(m2_5)

Call:
lm(formula = score ~ d_ken + d_2016, data = pset4_data)

Residuals:
   Min     1Q Median     3Q    Max
-49.05 -16.10  -0.37  18.04  46.95
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   52.050      2.731  19.056  < 2e-16 ***
d_ken          9.800      3.154   3.107 0.002168 **
d_2016        10.840      3.154   3.437 0.000718 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.3 on 197 degrees of freedom
Multiple R-squared:  0.09826,	Adjusted R-squared:  0.08911
F-statistic: 10.73 on 2 and 197 DF,  p-value: 3.762e-05

>
> linearHypothesis(m2_5, c("d_ken = 0", "d_2016=0"))
Linear hypothesis test

Hypothesis:
d_ken = 0
d_2016 = 0

Model 1: restricted model
Model 2: score ~ d_ken + d_2016

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    199 108661
2    197  97983  2     10677 10.734 3.762e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # THE P-VALUE IS 0.00003762. THIS IS LESS THAN THE ALPHA AT 5% SIGNIFICANCE
LEVEL.
> # REJECT THE NULL.
>
> linearHypothesis(m2_5, c("d_ken = 1", "d_2016=1"))
Linear hypothesis test

Hypothesis:
d_ken = 1
d_2016 = 1

Model 1: restricted model
Model 2: score ~ d_ken + d_2016
```

```
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1    199 106697                                  
2    197  97983  2    8713.3 8.7592 0.0002268 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
> # THE P-VALUE IS 0.0002268. THIS IS LESS THAN THE ALPHA AT 5% SIGNIFICANCE LEVEL.
> # REJECT THE NULL.
> 
> 
> #QUESTION 3 A
> 
> # DRAW SCATTERPLOT
> 
> library(readxl)
> chpt08wagedata <- read_excel("Downloads/jaggia_ba_2e_ch08_data.xlsx", 
+        sheet = "Wages")
> View(chpt08wagedata)
> str(chpt08wagedata)
tibble [160 × 3] (S3: tbl_df/tbl/data.frame)
 $ Wage    : num [1:160] 53 47 32 43 42 47 41 55 48 46 ...
 $ Graduate: num [1:160] 1 0 0 0 1 0 1 1 1 1 ...
 $ Age     : num [1:160] 57 42 24 39 72 52 27 46 68 35 ...
> 
> x<- chpt08wagedata$Age
> y<- chpt08wagedata$Wage
> plot(x, y, main = "Living Wage Pertaining to Age",
+     xlab = "Age", ylab = "Living Wage", pch = 20, frame = FALSE)
> 
> # ADD FIT LINES AND REGRESSION
> molog <- lm(y ~ log(x))
> summary(molog)

Call:
lm(formula = y ~ log(x))

Residuals:
     Min       1Q   Median       3Q      Max 
-16.0940  -3.5116   0.5128   3.6541   9.2964 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   25.632      4.632   5.533 1.28e-07 ***
```

```
log(x)        5.702      1.217   4.685 6.00e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.155 on 158 degrees of freedom
Multiple R-squared:  0.122,    Adjusted R-squared:  0.1164
F-statistic: 21.95 on 1 and 158 DF,  p-value: 6e-06

> scatterplot(chpt08wagedata$Wage~chpt08wagedata$Age,
+             xlab = "Age" , ylab = "Living Wage")
>
> # THE LIVING WAGE INCREASES WITH AGE UNTIL APPROXIMATELY 50 YAERS OLD.
> # AFTER THE 50 YEAR AGE MARK, THE LIVING WAGE DECREASES AS AGE
INCREASES.
>
> # QUESTION 3 B
>
> # SEPERATE THE DATA FROM 1-140 AND 141-160
>
> chpt08wagedata_T = chpt08wagedata[1:140,];
> chpt08wagedata_v = chpt08wagedata [141:160,];
> model3_1=lm( Wage ~ Graduate+Age,chpt08wagedata_T)
> summary(model3_1)

Call:
lm(formula = Wage ~ Graduate + Age, data = chpt08wagedata_T)

Residuals:
    Min       1Q   Median       3Q      Max
-11.4411  -2.9854   0.4416   3.8157   7.2650

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.14946    1.36209  29.476  < 2e-16 ***
Graduate     6.10610    0.78197   7.809 1.35e-12 ***
Age          0.07249    0.02612   2.776  0.00628 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.533 on 137 degrees of freedom
Multiple R-squared:  0.3348,  Adjusted R-squared:  0.325
F-statistic: 34.47 on 2 and 137 DF,  p-value: 7.485e-13

>
```

```
> # P-VALUE IS LESS THAN THE ALPHA AT 5% SIGNIFICANCE LEVEL.
> # REJECT THE NULL HYPOTHEIS
>
> # ADD ANOTHER COLUMN FOR AGE SQUARED AND RUN NEW REGRESSION
> model3_2=lm(data = chpt08wagedata_v, Wage~Graduate+Age+ I(Age^2))
> summary(model3_2)

Call:
lm(formula = Wage ~ Graduate + Age + I(Age^2), data = chpt08wagedata_v)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3941 -1.0262  0.0088  1.1640  2.9685

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.018002   5.105533  -0.199    0.844
Graduate     6.078027   1.056050   5.755 2.95e-05 ***
Age          1.856074   0.240329   7.723 8.73e-07 ***
I(Age^2)    -0.018106   0.002702  -6.701 5.09e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.881 on 16 degrees of freedom
Multiple R-squared:  0.8969,  Adjusted R-squared:  0.8775
F-statistic: 46.38 on 3 and 16 DF,  p-value: 4.07e-08

>
> # QUESTION 3 C
> # PREDICT THE WAGE FOR A 30 YEAR OLD GRADUATE STUDENT.
> #USING MODEL 3_1 AND MODEL 3_2
>
> # FOR MODEL 3_1
> predict.lm(model3_1, data.frame(Graduate=1, Age=30))
       1
48.43022
> # USING MODEL3_1 THE PREDICTED WAGE FOR A 30 YEAR OLD GRADUATE
STUDENT IS 48.43
>
>
> # FOR MODEL 3_2
> predict.lm(model3_2, data.frame(Graduate=1, Age=30))
       1
44.44729
```

```
> # USING MODEL3_1 THE PREDICTED WAGE FOR A 30 YEAR OLD GRADUATE
STUDENT IS 44.45
>
>
> # QUESTION 3 D
> # AT WHAT AGE ARE WAGES THE MAXIMUM AMOUNT?
>
> res3_2=summary(model3_2)
> coef(res3_2)
              Estimate  Std. Error    t value     Pr(>|t|)
(Intercept) -1.01800194 5.105533196 -0.1993919 8.444681e-01
Graduate     6.07802700 1.056049991  5.7554349 2.952426e-05
Age          1.85607421 0.240328578  7.7230691 8.731945e-07
I(Age^2)    -0.01810551 0.002701852 -6.7011464 5.090307e-06
> -coef(res3_2)[3,1]/(2*coef(res3_2)[4,1])
[1] 51.25717
>
> # THE AGE IN WHICH WAGES ARE THE MAXIMUM IS APPROXIMATELY 51 YEARS OLD.
>
>
> # QUESTION 3 E
>
> # Cross validation. Calculate, MSE, RMSE, MAPE, MAE, MAPE
> # Do for both models over the training set and over the validation set.
>
> # FIRST FOR TRAINING SET
>
> res3_1=summary(model3_1)
> coef(res3_1)
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 40.14945877 1.36209119 29.476337 3.595770e-61
Graduate     6.10610322 0.78196849  7.808631 1.346943e-12
Age          0.07248865 0.02611639  2.775600 6.280670e-03
>
> chpt08wagedata_T$yhat=predict.lm(model3_1,chpt08wagedata_T)
> chpt08wagedata_T$res3_1=resid(model3_1)
>
> (cor(chpt08wagedata_T$Wage, chpt08wagedata_T$yhat))^2
[1] 0.3347565
>
> MSE3_1=(sum((chpt08wagedata_T$Wage-chpt08wagedata_T$yhat)^2)/(140))
> MSE3_1
[1] 20.10457
>
```

```
> RMSE3_1=MSE3_1^0.5
> RMSE3_1
[1] 4.483812
>
> MAE3_1=mean(abs(chpt08wagedata_T$Wage-chpt08wagedata_T$yhat))
> MAE3_1
[1] 3.664206
>
> MAPE3_1=mean(abs(chpt08wagedata_T$Wage-
chpt08wagedata_T$yhat)/chpt08wagedata_T$Wage*100)
> MAPE3_1
[1] 8.092382
>
> RSE3_1=(sum((chpt08wagedata_T$Wage-chpt08wagedata_T$yhat)^2)/(140-3))^0.5
> RSE3_1
[1] 4.532639
>
> # NOW VALIDATIION SET
>
> res3_2=summary(model3_2)
> coef(res3_2)
              Estimate  Std. Error    t value     Pr(>|t|)
(Intercept) -1.01800194 5.105533196 -0.1993919 8.444681e-01
Graduate     6.07802700 1.056049991  5.7554349 2.952426e-05
Age          1.85607421 0.240328578  7.7230691 8.731945e-07
I(Age^2)    -0.01810551 0.002701852 -6.7011464 5.090307e-06
> chpt08wagedata_v$yhat=predict.lm(model3_2,chpt08wagedata_v)
> chpt08wagedata_v$res3_2=resid(model3_2)
> (cor(chpt08wagedata_v$Wage,chpt08wagedata_v$yhat))^2
[1] 0.8968757
>
> MSE3_2=(sum((chpt08wagedata_v$Wage-chpt08wagedata_v$yhat)^2)/(20))
> MSE3_2
[1] 2.830505
>
> RMSE3_2=MSE3_2^0.5
> RMSE3_2
[1] 1.682411
>
> MAE3_2=mean(abs(chpt08wagedata_v$Wage-chpt08wagedata_v$yhat))
> MAE3_2
[1] 1.385413
>
```

```
> MAPE3_2=mean(abs(chpt08wagedata_v$Wage-
chpt08wagedata_v$yhat)/chpt08wagedata_v$Wage*100)
> MAPE3_2
[1] 2.924267
>
> RSE3_2=(sum((chpt08wagedata_v$Wage-chpt08wagedata_v$yhat)^2)/(20-4))^0.5
> RSE3_2
[1] 1.880992
>
> # CREATE TABLE IN EXCEL TO DISPLAY THESE RESULTS.
```