R Notebook

Code

Hide

```
setwd('C:/Users/Samay Panwar/Documents/GitHub/MH3511-DataAnalysis')
#install.packages("wooldridge")
#install.packages("lmtest")
#install.packages('regclass')
#install.packages('gclus')

library(wooldridge)
```

```
package 'wooldridge' was built under R version 4.0.4
```

Hide

```
library(dplyr)
```

```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

Hide

```
library(lmtest)
```

```
package 'lmtest' was built under R version 4.0.4Loading required package: zoo
package 'zoo' was built under R version 4.0.4
Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

Hide

```
library(regclass)
```

```
package 愀牶regclass愀牛 was built under R version 4.0.4Loading required package: bestglm
package 愀牶bestglm愀牛 was built under R version 4.0.4Loading required package: leaps
package 愀牶leaps愀牛 was built under R version 4.0.4Loading required package: VGAM
package 愀牶VGAM愀牛 was built under R version 4.0.4Loading required package: stats4
Loading required package: splines

Attaching package: 愀牶VGAM愀牛

The following object is masked from 愀牶package:lmtest愀牛:

    lrtest

The following object is masked from 愀牶package:wooldridge愀牛:

    wine

Loading required package: rpart
Loading required package: randomForest
package 愀牶randomForest愀牛 was built under R version 4.0.4randomForest 4.6-14
Type rfNews() to see new features/changes/bug fixes.

Attaching package: 愀牶randomForest愀牛

The following object is masked from 愀牶package:dplyr愀牛:

    combine

Important regclass change from 1.3:
All functions that had a . in the name now have an _
all.correlations -> all_correlations, cor.demo -> cor_demo, etc.
```

Hide

```
library(tidyverse)
```

```
package 愀牶tidyverse愀牛 was built under R version 4.0.4Registered S3 methods overwritten by
'dbplyr':
  method         from
  print.tbl_lazy
  print.tbl_sql
-- Attaching packages --------------------------------------------------- tidyverse 1.3.0 --
v ggplot2 3.3.3     v purrr   0.3.4
v tibble  3.0.4     v stringr 1.4.0
v tidyr   1.1.2     v forcats 0.5.0
v readr   1.4.0
package 愀牶ggplot2愀牛 was built under R version 4.0.4-- Conflicts -------------------------
------------------------------ tidyverse_conflicts() --
x randomForest::combine() masks dplyr::combine()
x tidyr::fill()           masks VGAM::fill()
x dplyr::filter()         masks stats::filter()
x dplyr::lag()            masks stats::lag()
x ggplot2::margin()       masks randomForest::margin()
```

Hide

```
library(gclus)
```

```
package 愀牶gclus愀牷 was built under R version 4.0.4Loading required package: cluster
```

Hide

```
data('bwght')

write.csv(bwght, "C:/Users/Samay Panwar/Documents/GitHub/MH3511-DataAnalysis/bwght.csv")
bw <- na.omit(bwght)
```

Hide

```
bw
```

	faminc <dbl>	cigtax <dbl>	cigprice <dbl>	bwght <int>	fatheduc <int>	motheduc <int>	parity <int>	male <int>	white <int>
1	13.5	16.5	122.3	109	12	12	1	1	1
2	7.5	16.5	122.3	133	6	12	2	1	0
4	15.5	16.5	122.3	126	12	12	2	1	0
5	27.5	16.5	122.3	134	14	12	2	1	1
6	7.5	16.5	122.3	118	12	14	6	1	0
7	65.0	16.5	122.3	140	16	14	2	0	1
8	27.5	16.5	122.3	86	12	14	2	0	0
9	27.5	16.5	122.3	121	12	17	2	0	1
10	37.5	16.5	122.3	129	16	18	2	0	1
11	27.5	16.5	122.3	101	12	16	2	1	0

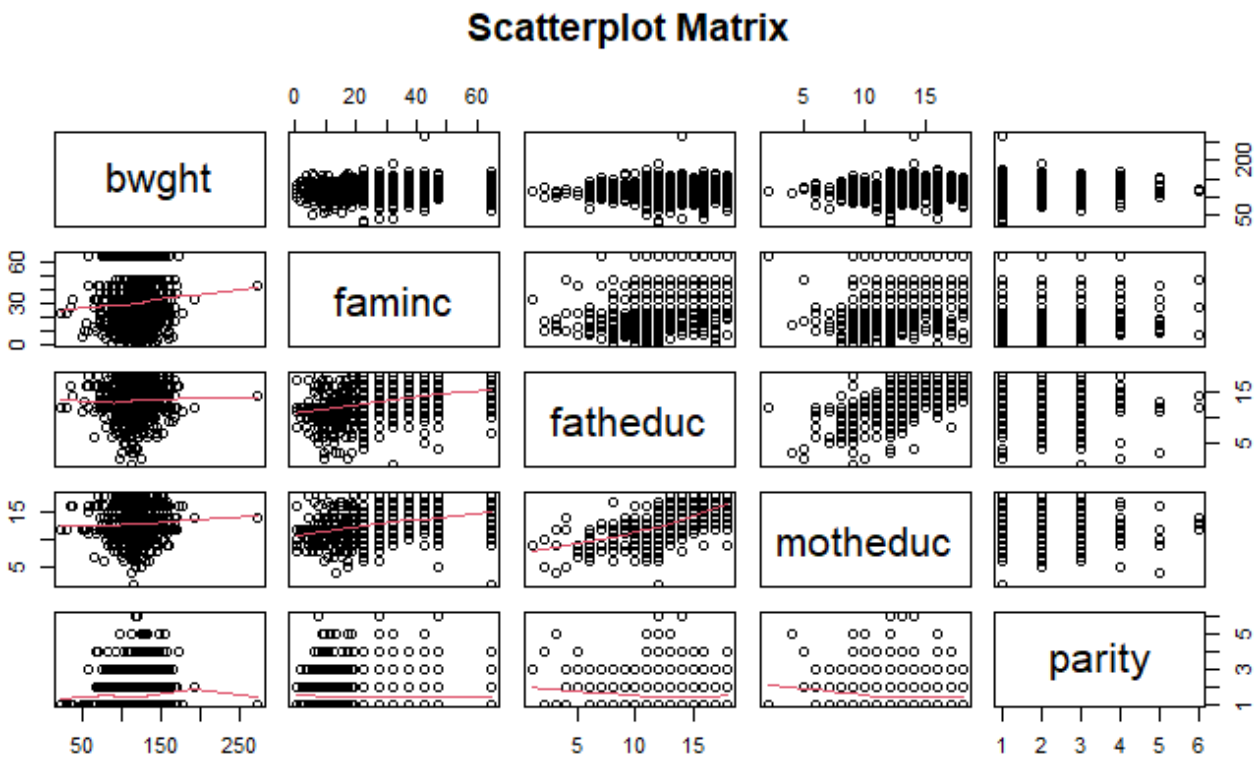
1-10 of 1,191 rows | 1-10 of 14 columns Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
cleanedData <- select(bw, -c("packs", "bwghtlbs", "lbwght", "lfaminc", "cigtax", "cigprice"))
# removing all the variables that we will not be using in our data analysis
cleanedData$male = as.factor(cleanedData$male)          # Male is a categorical variable that
takes on the value 1 when the child born is Male
cleanedData$white = as.factor(cleanedData$white)        # White is a categorical variable tha
t takes on the value of 1 when the child born is White
cleanedData$bwght = as.numeric(cleanedData$bwght)       # Changing the birth weight of the ba
by(in grams) into a numerical data type
cleanedData$faminc = as.numeric(cleanedData$faminc)     # Changing the family income of the b
aby's family into numerical data type
cleanedData = na.omit(cleanedData)                      # Omitting all the NA values present
 in our data

str(cleanedData)
head(cleanedData)
```

Hide

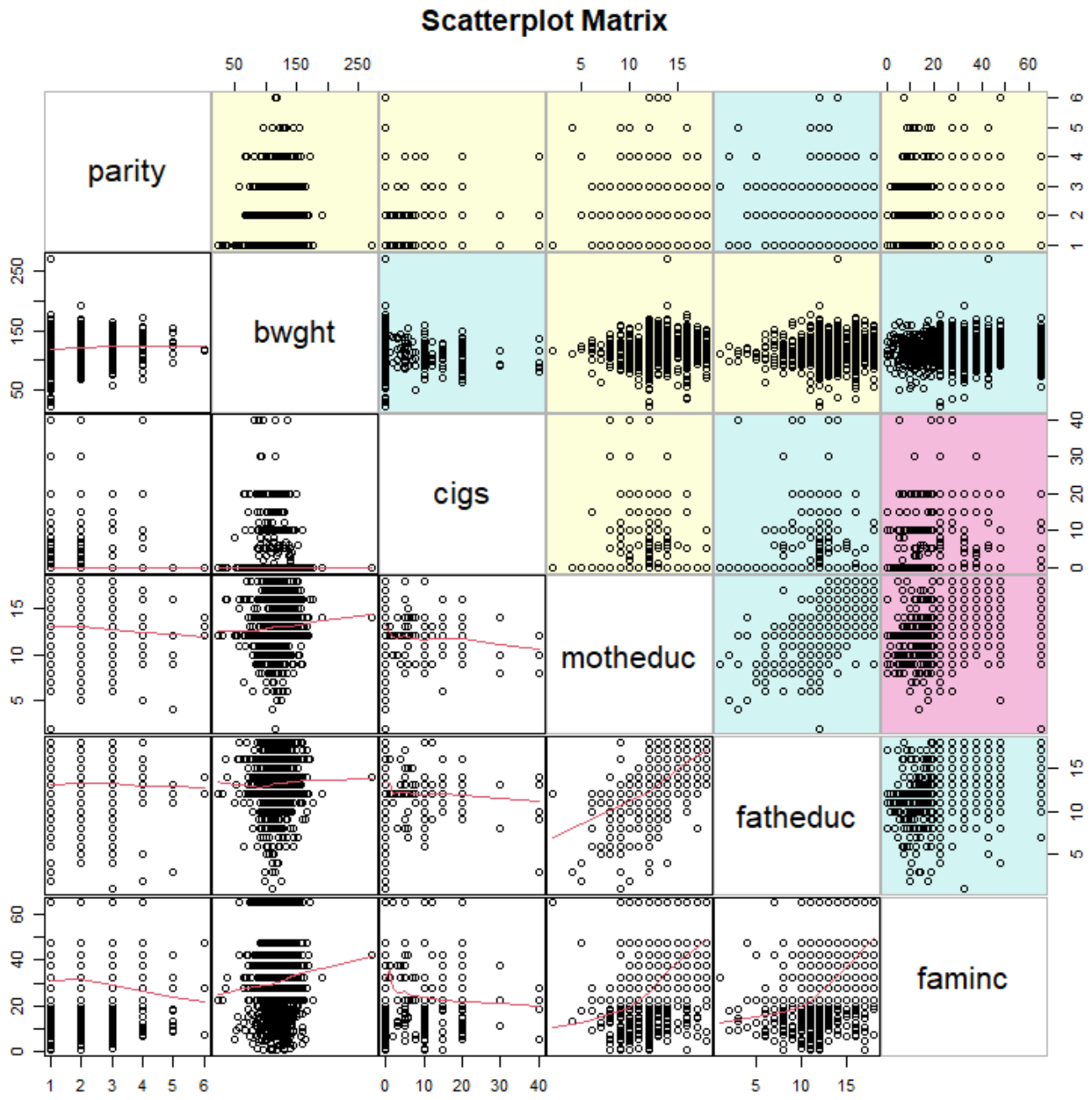
```
# Make a correlation scatterplot with lowess lines
categoricalVariables = c('male', 'white')
correlationData = select(cleanedData, -all_of(categoricalVariables))
pairs(~bwght + faminc + fatheduc + motheduc + parity ,data=correlationData,
      main="Scatterplot Matrix",
  lower.panel = panel.smooth)
```



Testing for Correlations with the Numerical Variables

Hide

```
correlationData.r <- abs(cor(correlationData))             # Get correlations of all the dat
a values
correlationData.colors <- dmat.color(correlationData.r)     # Get colors
# reorder variables so those with highest correlation
# are closest to the diagonal
correlationData.ordered <- order.single(correlationData.r)
cpairs(correlationData, correlationData.ordered, panel.colors=correlationData.colors, gap=.1,
       main="Scatterplot Matrix", lower.panel = panel.smooth )
```



ANOVA for all Categorical Variables

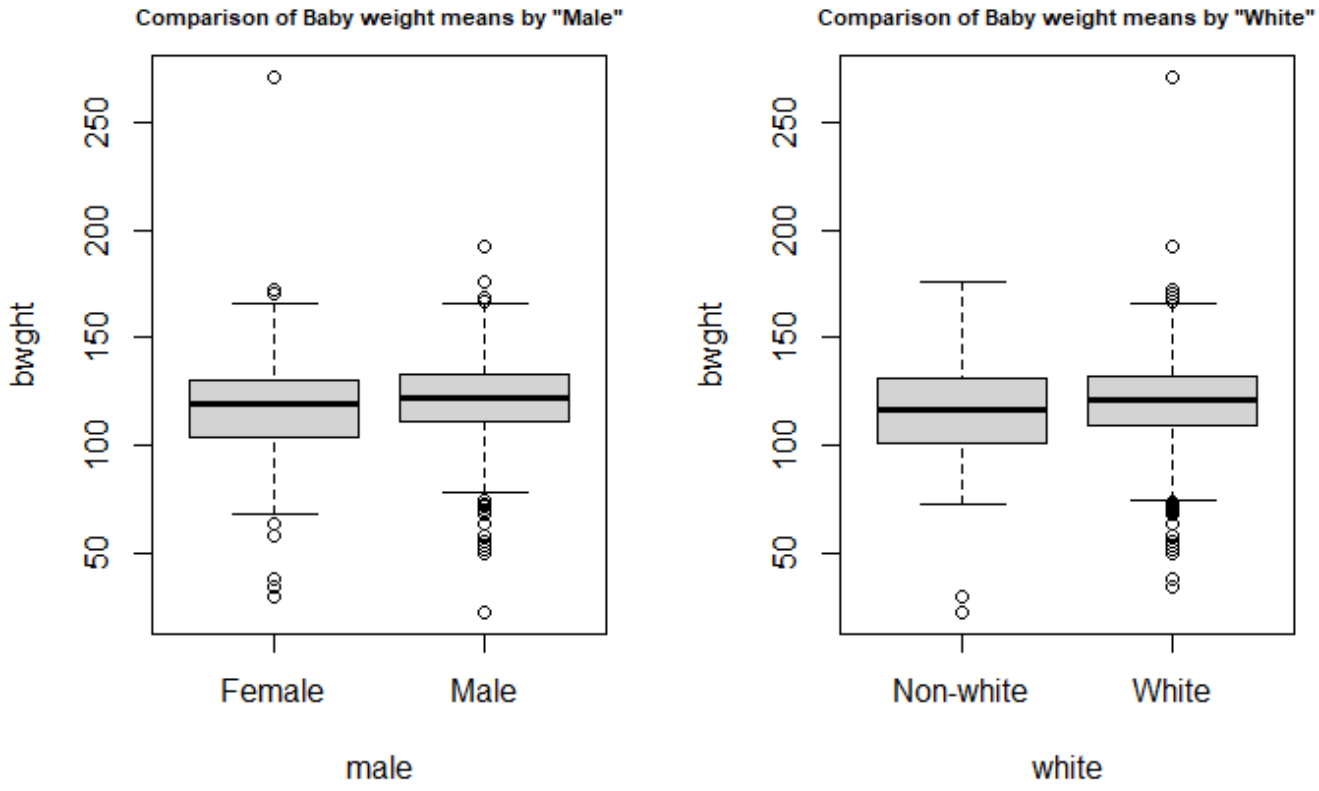
Hide

```
lapply(select(cleanedData, all_of(categoricalVariables)), function(categoricalVariable) t.tes
t(cleanedData$bwght ~ categoricalVariable, mu=0, pair=FALSE))
```

```
$male

    Welch Two Sample t-test

data:  cleanedData$bwght by categoricalVariable
t = -3.2233, df = 1167.4, p-value = 0.001302
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.046747 -1.470816
sample estimates:
mean in group 0 mean in group 1
       117.5794        121.3382


$white

    Welch Two Sample t-test

data:  cleanedData$bwght by categoricalVariable
t = -2.9388, df = 251.2, p-value = 0.003602
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.125048 -1.604594
sample estimates:
mean in group 0 mean in group 1
       115.4247        120.2896
```

Hide

```
# Conduct ANOVA for all categorical variables
par(mfrow=c(1,2))
boxplot(bwght~male, data = cleanedData, main='Comparison of Baby weight means by "Male"', ce
x.main=0.7, names=c('Female', 'Male'))
boxplot(bwght~white, data = cleanedData, main='Comparison of Baby weight means by "White"', c
ex.main=0.7, names=c('Non-white', 'White'))
```



Hide

```
male.aov = aov(bwght ~ male, data=cleanedData)
summary(male.aov)
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
male          1   4201    4201   10.44 0.00127 **
Residuals   1189 478546     402
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

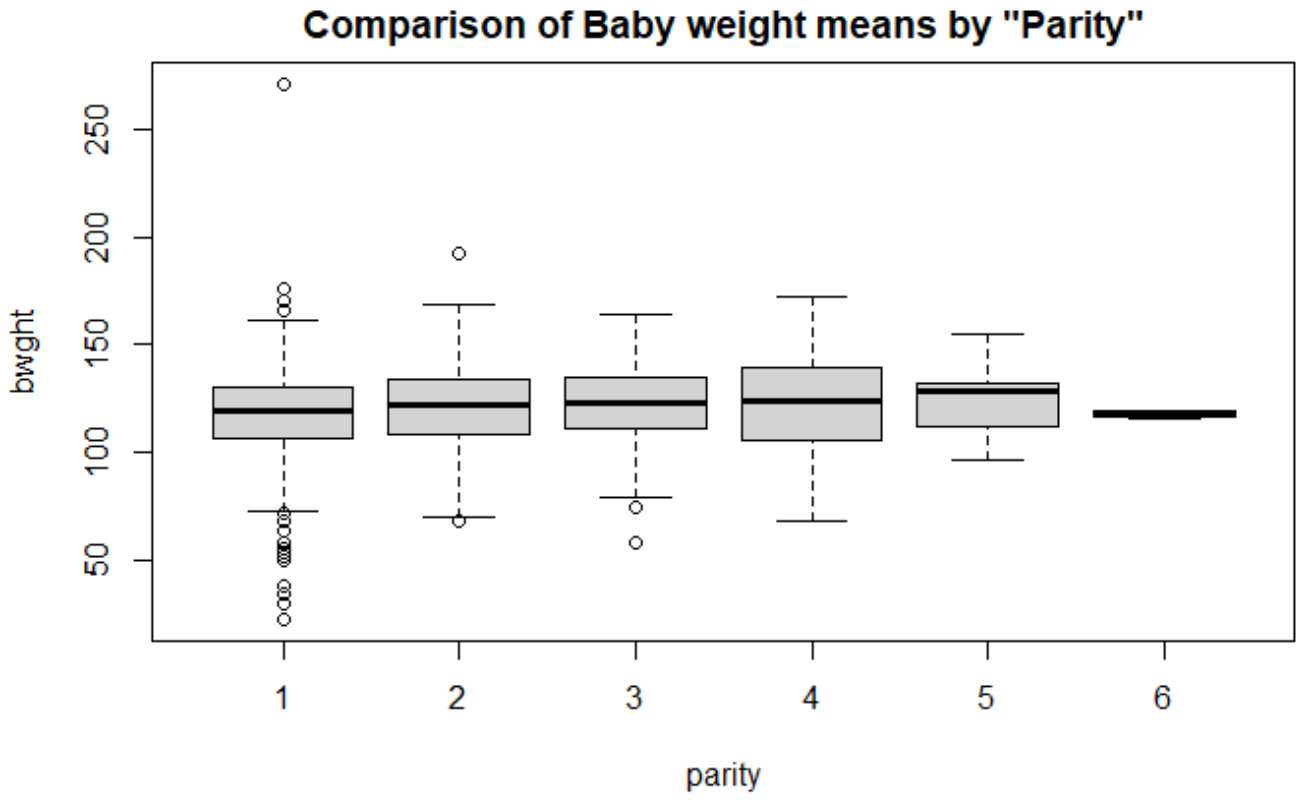
Hide

```
white.aov = aov(bwght ~ white, data=cleanedData)
summary(white.aov)
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
white         1   3715    3715    9.22 0.00245 **
Residuals   1189 479032     403
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hide

```
boxplot(bwght~parity, data = cleanedData, main='Comparison of Baby weight means by "Parity"')
```



Hide

```
parity.aov = aov(bwght ~ parity, data=cleanedData)
summary(parity.aov)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
parity        1   2334    2334   5.777 0.0164 *
Residuals  1189 480412     404
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conducting Simple Linear Regression for all the Independent Variables

Hide

```
# Conduct simple regression models for all remaining variables

SLRObjects = lapply( select(cleanedData, -c('bwght')), function(x) lm(cleanedData$bwght ~ x))
```

Conducting Multiple Linear Regression

Hide

```
# Conduct Multiple Linear Regression with all variables

linearRegression <- lm(bwght~., data=cleanedData)
summary(linearRegression)
```

```
Call:
lm(formula = bwght ~ ., data = cleanedData)

Residuals:
    Min      1Q  Median      3Q     Max
-93.715 -11.741   0.369  12.633 152.326

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.02382    3.93758  27.688  < 2e-16 ***
faminc        0.04376    0.03698   1.184 0.236822
fatheduc      0.41127    0.28101   1.464 0.143585
motheduc     -0.32836    0.31786  -1.033 0.301796
parity        1.91589    0.65539   2.923 0.003530 **
male1         3.79554    1.14268   3.322 0.000922 ***
white1        4.71347    1.60772   2.932 0.003435 **
cigs         -0.59811    0.10973  -5.451 6.11e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.65 on 1183 degrees of freedom
Multiple R-squared:  0.05413,   Adjusted R-squared:  0.04854
F-statistic: 9.672 on 7 and 1183 DF,  p-value: 9.562e-12
```

Checking for Variance Inflation Factor and Heteroskedasticity

Hide

```
# Check for VIF and Heteroskedasticity

bptest(linearRegression)
```

```

    studentized Breusch-Pagan test

data:  linearRegression
BP = 1.6114, df = 7, p-value = 0.9782
```

Hide

```
VIF(linearRegression)
```

```
  faminc fatheduc motheduc   parity     male    white     cigs
1.359176 1.829469 1.820396 1.013066 1.005806 1.051058 1.060133
```

QQ-plots of all the Regressions run

Hide

```
# Make qq plots of all residuals for all the regressions conducted
par(mfrow=c(7,1))

qqnorm(resid(SLRObjects$faminc), main = 'Residual QQ-plot for Family Income')
qqline(resid(SLRObjects$faminc), col='red')
```

Hide

```
qqnorm(resid(SLRObjects$fatheduc), main = 'Residual QQ-plot for Father Education')
qqline(resid(SLRObjects$fatheduc), col='red')
```

Hide

```
qqnorm(resid(SLRObjects$motheduc), main = 'Residual QQ-plot for Mother Education')
qqline(resid(SLRObjects$motheduc), col='red')
```

Hide

```
qqnorm(resid(SLRObjects$parity), main = 'Residual QQ-plot for Parity')
qqline(resid(SLRObjects$parity), col='red')
```

Hide

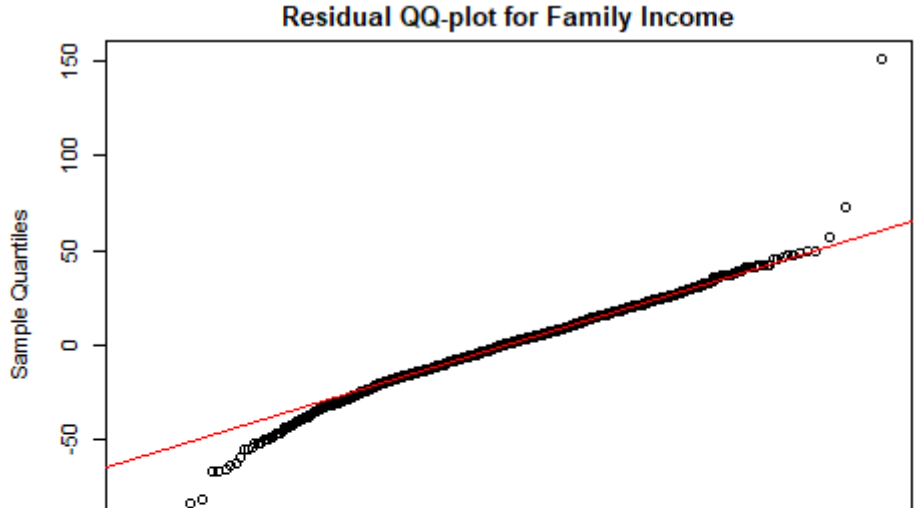
```
qqnorm(resid(SLRObjects$male), main = 'Residual QQ-plot for Male')
qqline(resid(SLRObjects$male), col='red')
```

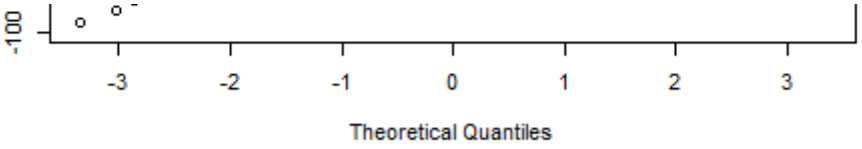
Hide

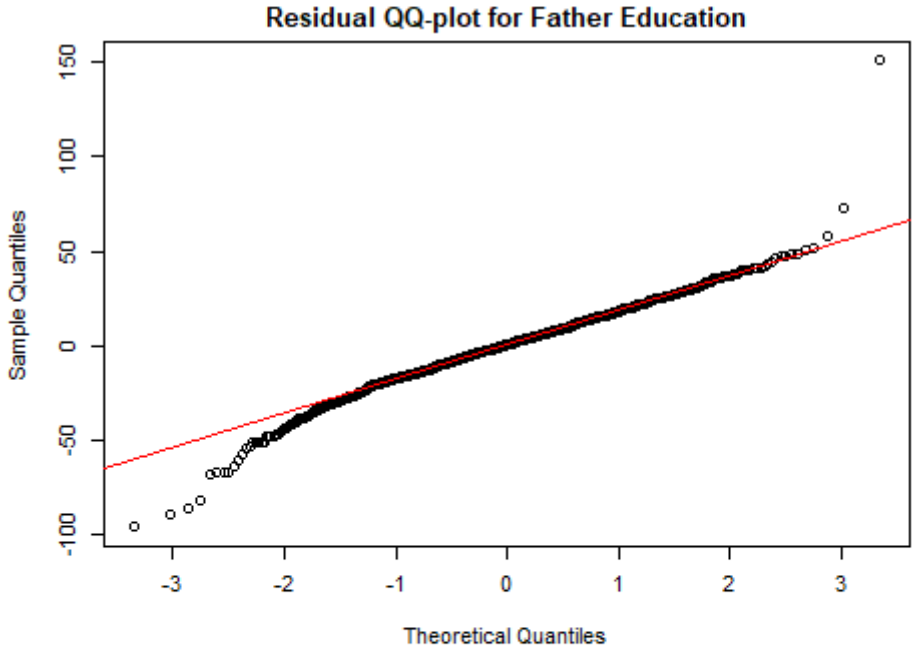
```
qqnorm(resid(SLRObjects$white), main = 'Residual QQ-plot for White')
qqline(resid(SLRObjects$white), col='red')
```

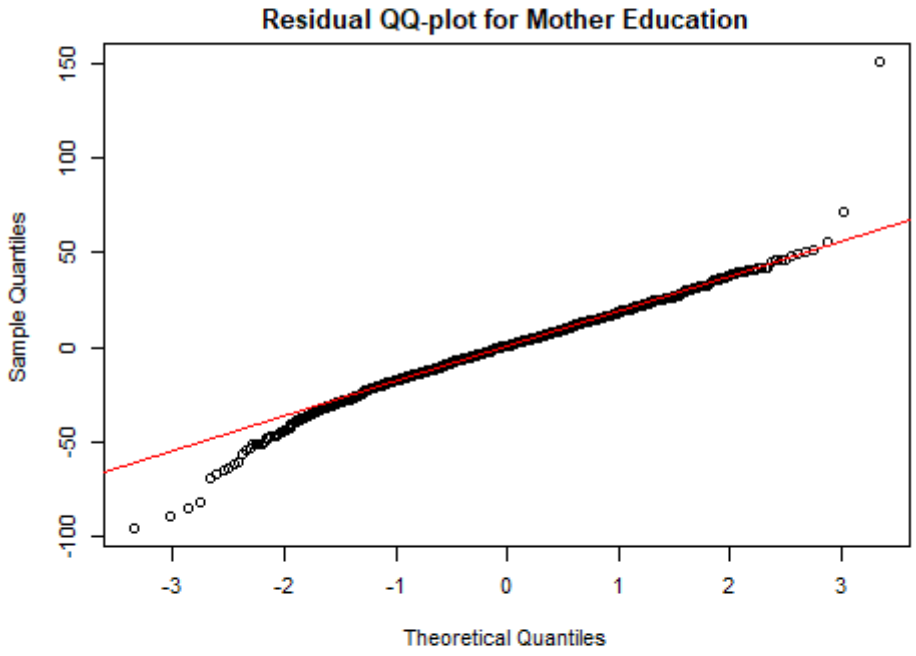
Hide

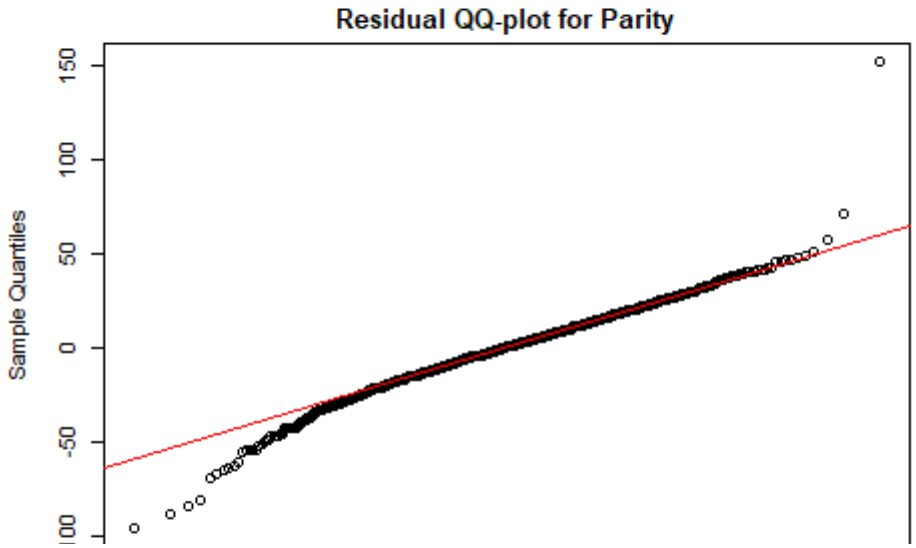
```
qqnorm(resid(SLRObjects$cigs), main = 'Residual QQ-plot for Cigs')
qqline(resid(SLRObjects$cigs), col='red')
```

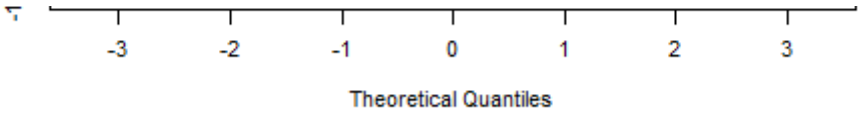


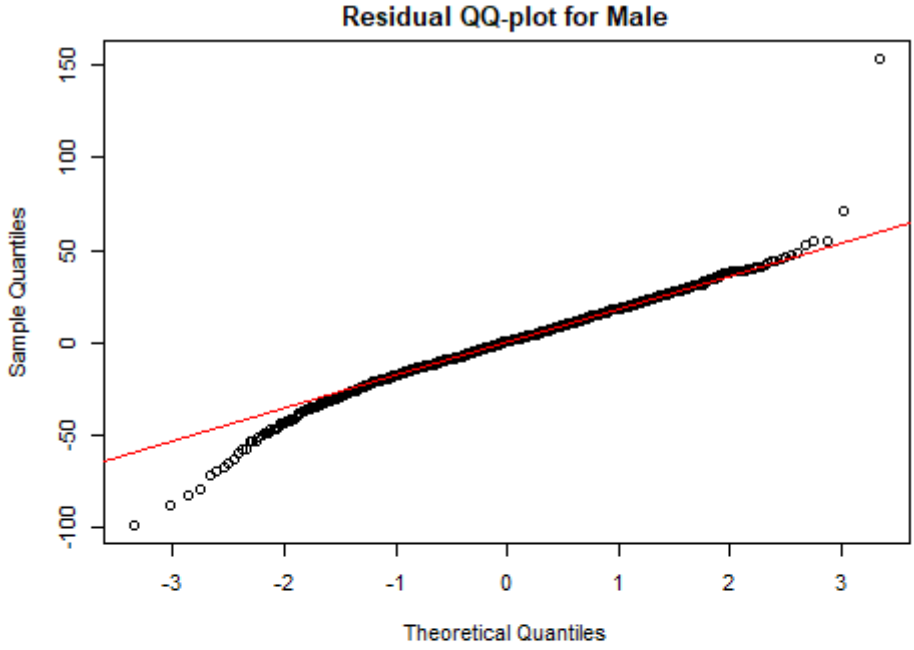

-100
-3
-2
-1
0
1
2
3
Theoretical Quantiles

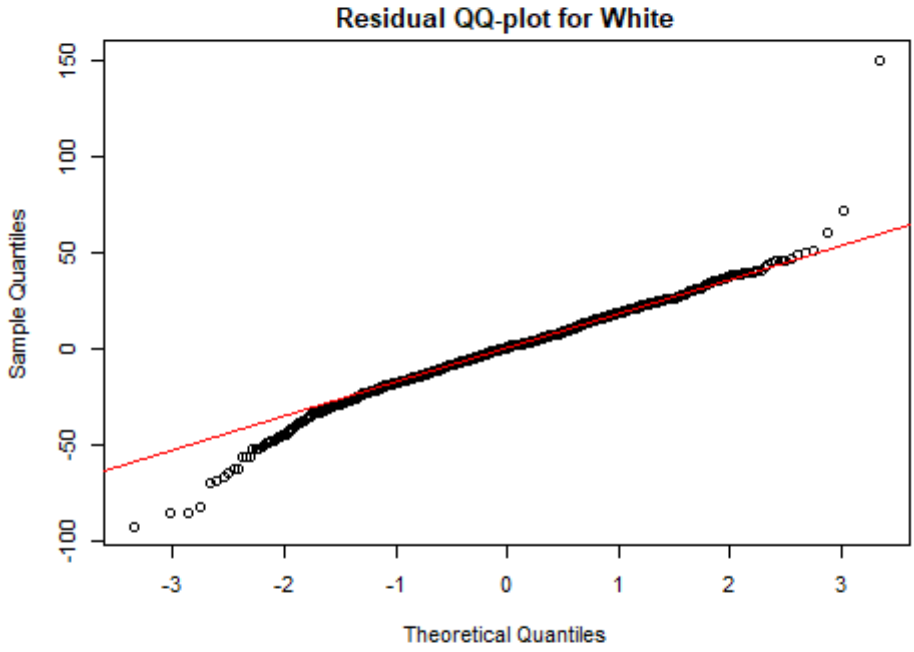

Residual QQ-plot for Father Education
Sample Quantiles
150
100
50
0
-50
-100
-3
-2
-1
0
1
2
3
Theoretical Quantiles

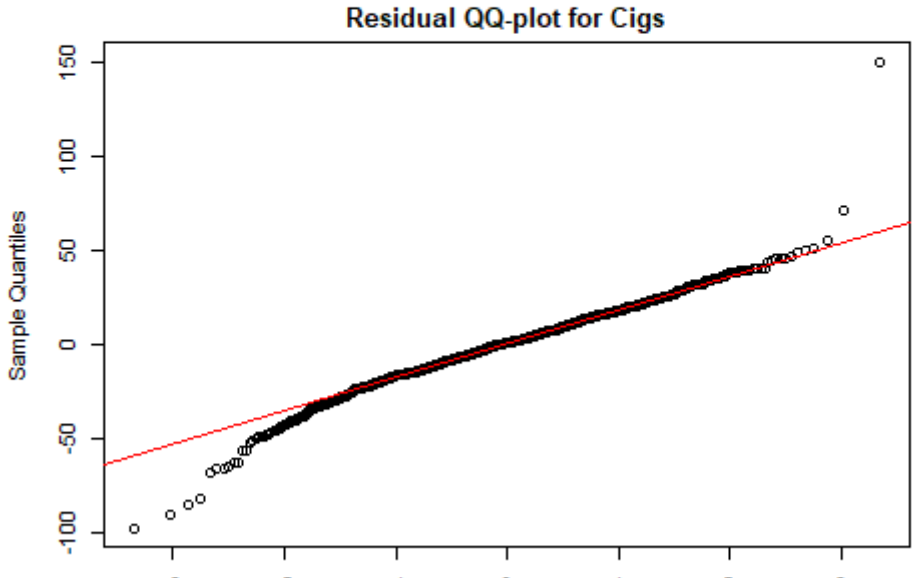

Residual QQ-plot for Mother Education
Sample Quantiles
150
100
50
0
-50
-100
-3
-2
-1
0
1
2
3
Theoretical Quantiles

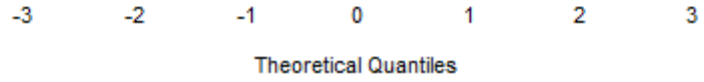

Residual QQ-plot for Parity
Sample Quantiles
150
100
50
0
-50
-100


-1
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Residual QQ-plot for Male
Sample Quantiles
150
100
50
0
-50
-100
-3
-2
-1
0
1
2
3
Theoretical Quantiles

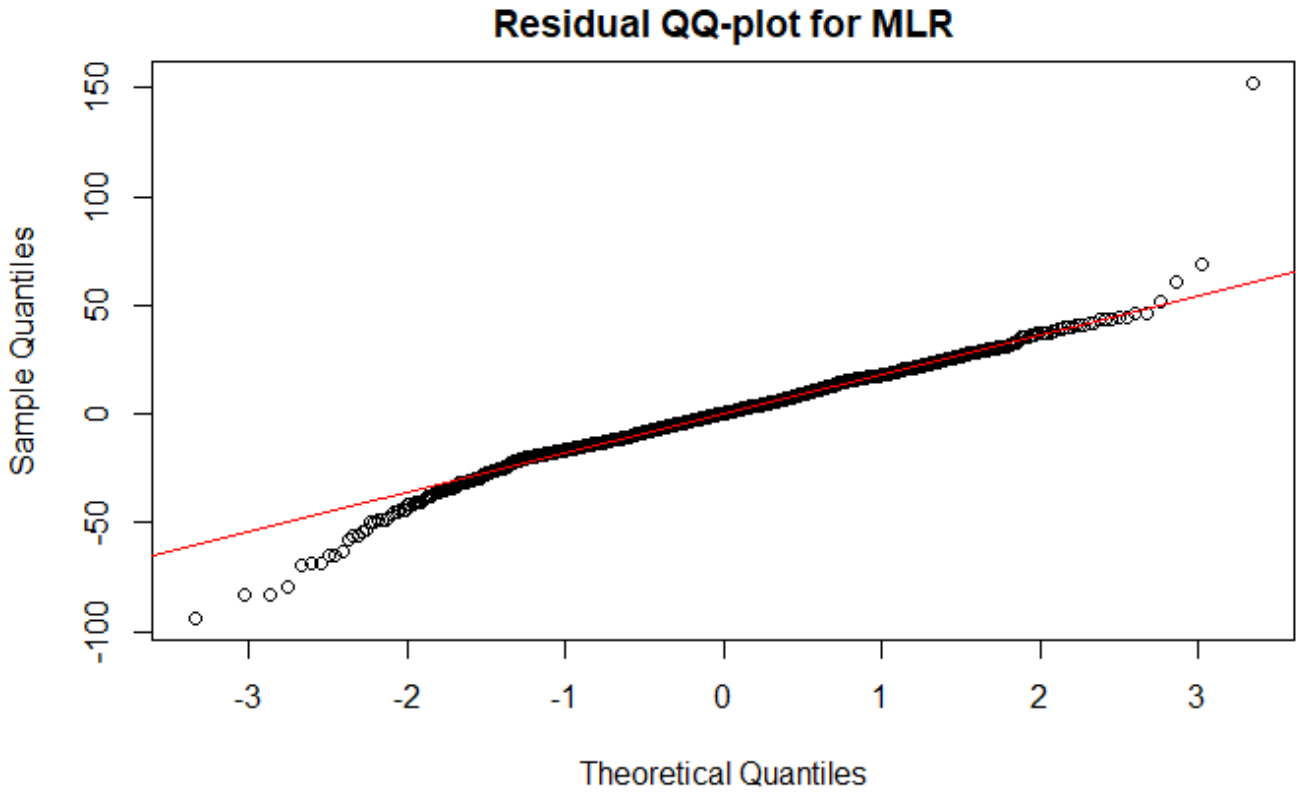

Residual QQ-plot for White
Sample Quantiles
150
100
50
0
-50
-100
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Residual QQ-plot for Cigs
Sample Quantiles
150
100
50
0
-50
-100



Hide

```
qqnorm(resid(linearRegression), main = 'Residual QQ-plot for MLR')
qqline(resid(linearRegression), col='red')
```



Explain Beta Coefficients for all the regressions

when the family income increases by 1 unit then, on average, the birth weight of the baby increases by 0.08965 grams, ceteris paribus

when the father's education increases by 1 unit then, on average, the birth weight of the baby increases by 0.6102 grams, ceteris paribus

when the mother's education increases by 1 unit then, on average, the birth weight of the baby increases by 0.3755 grams, ceteris paribus

when the number of children already born increases by 1 unit then, on average, the birth weight of the baby increases by 1.601 grams, ceteris paribus

when the child's sex is male, on average, the birth weight of the baby increases by 3.759 grams, ceteris paribus

when the child's race is white then, on average, the birth weight of the baby increases by 4.865 grams, ceteris paribus

when the number of cigs smoked by the mother while pregnant increases by 1 unit then, on average, the birth weight of the baby increases by -0.6203 grams, ceteris paribus

```
```