# HW3

*Chong Zhang*

*September 26, 2018*

## Load data

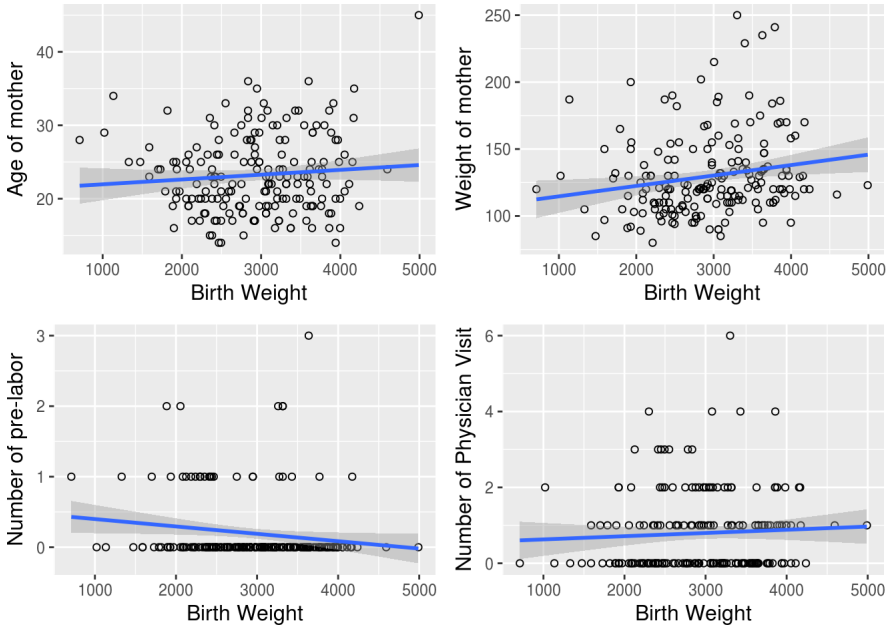
```
library(ggplot2)
library(gridExtra)
data_table = read.table('Birthweight.csv', header = T, sep = ',')
head(data_table)
```

```
##   BirthWt Age MotherWt  Race Smoke NumPreLabor Hypertension UterineIrr
## 1     709  28      120 Other   Yes           1           No        Yes
## 2    1021  29      130 White    No           0           No        Yes
## 3    1135  34      187 Black   Yes           0          Yes         No
## 4    1330  25      105 Other    No           1          Yes         No
## 5    1474  25       85 Other    No           0           No        Yes
## 6    1588  27      150 Other    No           0           No         No
##   NumPhysicianVt
## 1              0
## 2              2
## 3              0
## 4              0
## 5              0
## 6              0
```

## Question 1: Exploratory Data Analysis.

**a**. Using a scatterplot describe the relationship between BirthWt and the numeric independent variables; Age, MotherWt, NumPreLabor, NumPhysicianVt. Describe the general trend (direction and form). What are the values of the correlation coefficients? Please interpret.

```
p1 = ggplot(data = data_table, aes(x=BirthWt, y = Age))+geom_point(shape=1)+geom_smooth(method = 'lm')+xlab('Birt
h Weight')+ylab('Age of mother')+theme(axis.title = element_text(size=12))
p2 = ggplot(data = data_table, aes(x=BirthWt, y = MotherWt))+geom_point(shape=1)+geom_smooth(method = 'lm')+xlab(
'Birth Weight')+ylab('Weight of mother')+theme(axis.title = element_text(size=12))
p3 = ggplot(data = data_table, aes(x=BirthWt, y = NumPreLabor))+geom_point(shape=1)+geom_smooth(method = 'lm')+xl
ab('Birth Weight')+ylab('Number of pre-labor')+theme(axis.title = element_text(size=12))
p4 = ggplot(data = data_table, aes(x=BirthWt, y = NumPhysicianVt))+geom_point(shape=1)+geom_smooth(method = 'lm')
+xlab('Birth Weight')+ylab('Number of Physician Visit')+theme(axis.title = element_text(size=12))
grid.arrange(p1,p2,p3,p4)
```




```
round(cor(data_table[c(1,2,3,6,9)]),2)
```

```
##                BirthWt  Age MotherWt NumPreLabor NumPhysicianVt
## BirthWt           1.00 0.09     0.19       -0.15           0.06
## Age               0.09 1.00     0.18        0.07           0.22
## MotherWt          0.19 0.18     1.00       -0.14           0.14
## NumPreLabor      -0.15 0.07    -0.14        1.00          -0.04
## NumPhysicianVt    0.06 0.22     0.14       -0.04           1.00
```

```
Cor_co1 = round(cor(data_table$BirthWt, data_table$Age),2)
Cor_co2 = round(cor(data_table$BirthWt, data_table$MotherWt),2)
Cor_co3 = round(cor(data_table$BirthWt, data_table$NumPreLabor),2)
Cor_co4 = round(cor(data_table$BirthWt, data_table$NumPhysicianVt),2)
```

It seems that there is **no** linear relationship between the birth weight and the age of the monther. The distibution of birth weight is similar across all ages. The correlation coefficient of this pair of data is **`0.09`** . It means that there is no linear relationship.

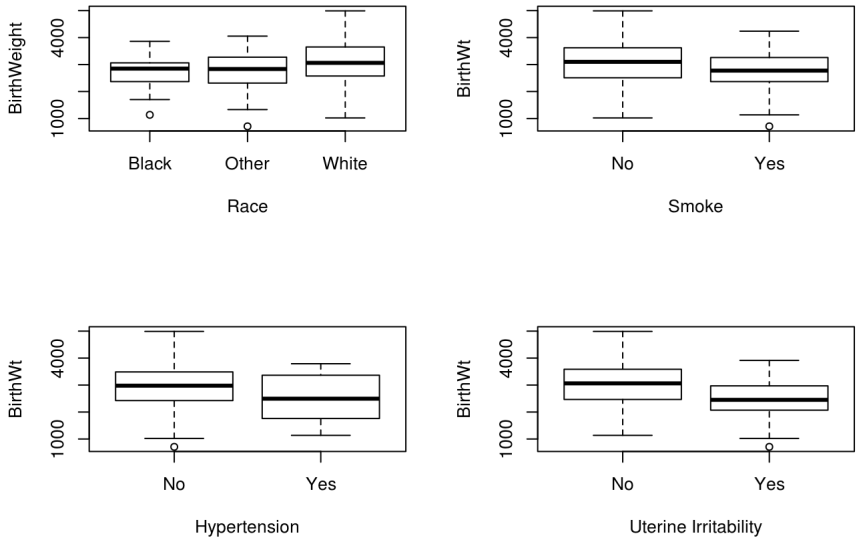
It seems that there is a **very very mild** linear relationship between the birth weight and the weight of the monther. It seems that the heavier the mother, the heavier of the baby. The correlation coefficient of this pair of data is **`0.19`** . It indicates a very very mild linear relationship.

It seems that there is **no** linear relationship between the birth weight and the weight of the monther. The correlation coefficient of this pair of data is **`-0.15`** . It suggests that there is no linear relationship.

It seems that there is **no** linear relationship between the birth weight and the weight of the monther. The correlation coefficient of this pair of data is **`0.06`** . It suggest that there is no linear relationship.

**b**. Describe the relationship between BirthWt and the categorical independent variables Race, Smoke, Hypertension, UterineIrr. Does BirthWt vary with the categorical variables? Use boxplot function in R as follows:

```
par(mfrow=c(2,2))
boxplot(data_table$BirthWt~as.factor(data_table$Race),xlab="Race", ylab='BirthWeight')
boxplot(data_table$BirthWt~as.factor(data_table$Smoke), xlab="Smoke", ylab="BirthWt")
boxplot(data_table$BirthWt~as.factor(data_table$Hypertension), xlab="Hypertension", ylab="BirthWt")
boxplot(data_table$BirthWt~as.factor(data_table$UterineIrr), xlab="Uterine Irritability", ylab="BirthWt")
```




For the variable **Race**, it seems that mothers with the race of **White** gave birth to babies with higher brith weight. For the variable **Smoke**, it seems that the mothers who **smoke** gave birth to babies with lower brith weight. For the variable **Hypertension**, it seems that the mothers with **Hypertension** gave birth to babies with lower brith weight. For the variable **Uterine Irritability**, it seems that the mothers with **Uterine Irritability** gave birth to babies with lower brith weight.

**c**. Based on this exploratory analysis, is it reasonable to assume a linear regression model? Would you suggest that BirthWt varies with all or only some of the independent variables? Would you recommend using the qualitative variables Race, Smoke, Hypertension, UterineIrr in the model? Why?

```
summary(aov(lm(data_table$BirthWt~data_table$Race)))
```

```
##                  Df   Sum Sq Mean Sq F value  Pr(>F)
## data_table$Race   2  5015725 2507863   4.913 0.00834 **
## Residuals       186 94953931  510505
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(lm(data_table$BirthWt~data_table$Smoke)))
```

```
##                   Df   Sum Sq Mean Sq F value  Pr(>F)
## data_table$Smoke   1  3625946 3625946   7.038 0.00867 **
## Residuals        187 96343710  515207
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(lm(data_table$BirthWt~data_table$Hypertension)))
```

```
##                           Df   Sum Sq Mean Sq F value Pr(>F)
## data_table$Hypertension   1  2130425 2130425   4.072  0.045 *
## Residuals               187 97839231  523204
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(lm(data_table$BirthWt~data_table$UterineIrr)))
```

```
##                         Df   Sum Sq Mean Sq F value   Pr(>F)
## data_table$UterineIrr   1  8059031 8059031    16.4 7.52e-05 ***
## Residuals             187 91910625  491501
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the exploratory analysis above, I think it is reasonable to assume a linear regression model. The reponse variable BirthWt only varies with limited number of independent variables, such as MotherWt. I would use the all the qualitative variables. Because ANOVA test suggests that the means of response variable are statistically different with respect to the qualitative variable.

## Question 2: Fitting the Linear Regression Model.

Fit a linear regression to evaluate the relationship between BirthWt and all the predictors.

```
model = lm(data_table$BirthWt~., data=data_table)
summary(model)
```

```
##
## Call:
## lm(formula = data_table$BirthWt ~ ., data = data_table)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1825.26  -435.21    55.91   473.46  1701.20
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2439.534    327.234   7.455 3.71e-12 ***
## Age               -3.570      9.620  -0.371 0.711012
## MotherWt           4.354      1.736   2.509 0.013007 *
## RaceOther        133.350    159.393   0.837 0.403925
## RaceWhite        488.428    149.985   3.257 0.001349 **
## SmokeYes        -352.045    106.476  -3.306 0.001142 **
## NumPreLabor      -48.402    101.972  -0.475 0.635607
## HypertensionYes -592.827    202.321  -2.930 0.003830 **
## UterineIrrYes   -516.081    138.885  -3.716 0.000271 ***
## NumPhysicianVt   -14.058     46.468  -0.303 0.762598
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 650.3 on 179 degrees of freedom
## Multiple R-squared:  0.2427, Adjusted R-squared:  0.2047
## F-statistic: 6.376 on 9 and 179 DF,  p-value: 7.891e-08
```

```
coef = round(model$coefficients,2)
```

### (i) What are the model parameters and what are their estimates?

There are **10** model parameters as shown above. The estimates of those parameters are also shown above.

### (ii) Write down the equation for the regression line;

$$BirthWt = 2439.53 - 3.57 * Age + 4.35 * MotherWt + 133.35 * RaceOther + 488.43 RaceWhite - 352.04 * SmokeYes - 48.40 * NumPreLabor - 5$$

### (iii) Interpret the estimated value of the parameters corresponding with Age, MotherWt and Race in the context of the problem.

For the parameter corresponding to **Age**, it means that with one year increase in the age of the mother, there will be **3.57 gram decrease** in the Birth weight of the baby when holding all other variables constant.

For the parameter corresponding to **MotherWt**, it means that with one pound increase in the weight of mother, there will be **4.35 gram increase** in the Birth weight of the baby when holding all other variables constant.

For the parameter of Race, the regression model used **Black** as its baseline, so for **RaceOther**, it means that when the monther's race is **Other**, there will be a **133.35 gram increase** in the Birth weight of the baby compared to the Birth weight of a baby whose mother is **Black** when holding all other variables contant.

Similarly, for **RaceWhite**, it means that when the monther's race is **White**, there will be a **488.43 gram increase** in the Birth weight of the baby compared to the Birth weight of a baby whose mother is **Black** when holding all other variables contant.

(iv) Find a 95% confidence interval for the parameters corresponding to all predictors and the intercept.

```
confint(object = model,level = 0.95)
```

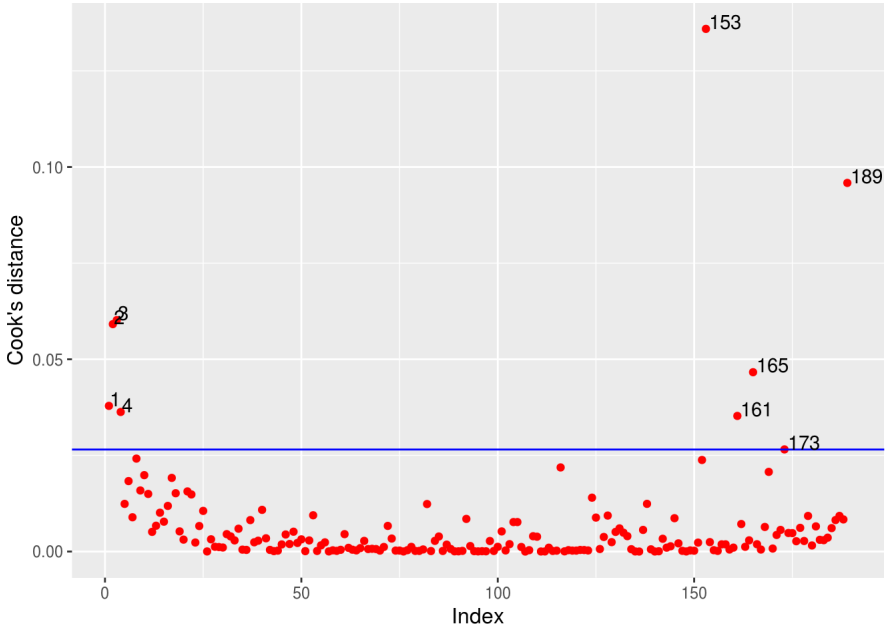
```
##                        2.5 %      97.5 %
## (Intercept)     1793.8016986 3085.267098
## Age              -22.5535895   15.413721
## MotherWt           0.9291722    7.778853
## RaceOther       -181.1807101  447.881573
## RaceWhite        192.4622407  784.392836
## SmokeYes        -562.1550309 -141.934036
## NumPreLabor     -249.6231426  152.819074
## HypertensionYes -992.0688887 -193.586000
## UterineIrrYes   -790.1441980 -242.017757
## NumPhysicianVt  -105.7536827   77.637574
```

The 95% confidence intervals for all predictors and intercept are shown above.

Question 3: Outliers. Based on your analysis in the first two question, are there any possible outliers in the data set? Would you consider taking them out? Provide relevant plots and reasoning behind this decision.

Based on the plots above, it seems that there are some outliers in the data. However besides visually inspect outliers, I also performed Cook's test.

```
cook = cooks.distance(model)
cook = as.data.frame(cbind(index=1:189, cooksd = cook))
p = ggplot(data = cook,aes(x=index, y = cooksd)) + geom_point(col='red')+xlab('Index')+ylab('Cook\'s distance')+t
heme(axis.title = element_text(size=12))
p = p + geom_hline(yintercept = 4*mean(cook$cooksd), col='blue')
p = p + geom_text(aes(label = ifelse(cooksd>4*mean(cooksd), index,'')),hjust=-0.1, vjust=0.01)
p
```



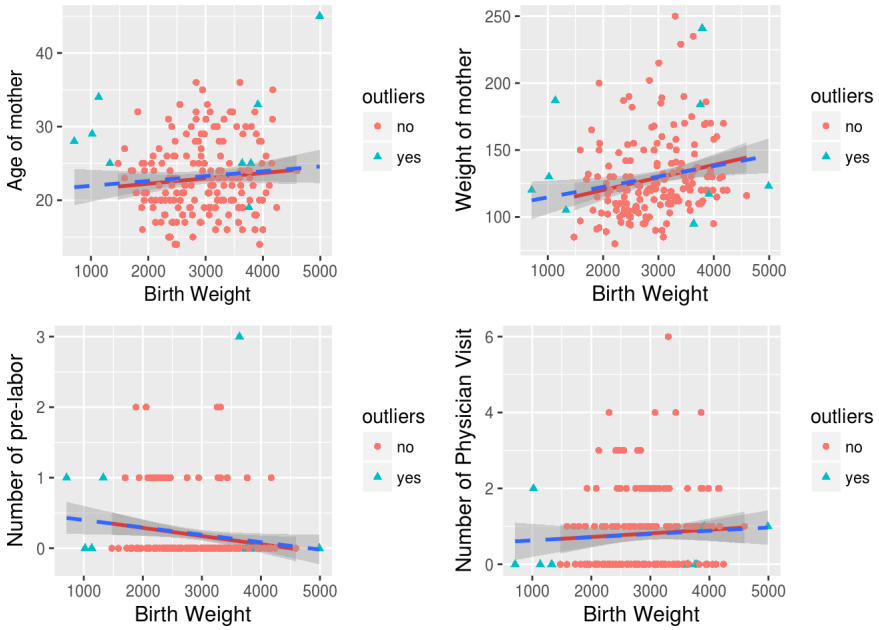

In general, those observations that have a cook's distance greater than 4 times the mean may be classified as influential. Thus data points **1,2,3,4,153,161,165,173,189** are considered as influential points. They could be outliers.

Let's take a look at those data points.

```
outliers = data_table[c(1,2,3,4,153,161,165,173,189),]
outliers
```

```
##     BirthWt Age MotherWt  Race Smoke NumPreLabor Hypertension UterineIrr
## 1       709  28      120 Other   Yes           1           No        Yes
## 2      1021  29      130 White    No           0           No        Yes
## 3      1135  34      187 Black   Yes           0          Yes         No
## 4      1330  25      105 Other    No           1          Yes         No
## 153    3637  25       95 White   Yes           3           No        Yes
## 161    3756  19      184 White   Yes           0          Yes         No
## 165    3790  25      241 Black    No           0          Yes         No
## 173    3912  33      117 White    No           0           No        Yes
## 189    4990  45      123 White    No           0           No         No
##     NumPhysicianVt
## 1                0
## 2                2
## 3                0
## 4                0
## 153              0
## 161              0
## 165              0
## 173              1
## 189              1
```

```
new_data = cbind(data_table, cook)

new_data$outliers = ifelse(new_data$index%in%c(1,2,3,4,153,161,165,173,189),'yes','no')


p1 = ggplot(data = new_data, aes(x=BirthWt, y = Age))+geom_smooth(data = subset(new_data,outliers=='no'), aes(x=B
irthWt,y=Age),method = 'lm',col='red')+xlab('Birth Weight')+ylab('Age of mother')+geom_point(aes(col = outliers,s
hape=outliers))+geom_smooth(method='lm',linetype=2)

p2 = ggplot(data = new_data, aes(x=BirthWt, y = MotherWt))+geom_smooth(data = subset(new_data,outliers=='no'), ae
s(x=BirthWt,y=MotherWt),method = 'lm',col='red')+xlab('Birth Weight')+ylab('Weight of mother')+theme(axis.title =
 element_text(size=12))+geom_point(aes(col = outliers,shape=outliers))+geom_smooth(method='lm',linetype=2)

p3 = ggplot(data = new_data, aes(x=BirthWt, y = NumPreLabor))+geom_smooth(data = subset(new_data,outliers=='no'),
 aes(x=BirthWt,y=NumPreLabor),method = 'lm',col='red')+xlab('Birth Weight')+ylab('Number of pre-labor')+theme(axi
s.title = element_text(size=12))+geom_point(aes(col = outliers,shape=outliers))+geom_smooth(method='lm',linetype=
2)

p4 = ggplot(data = new_data, aes(x=BirthWt, y = NumPhysicianVt))+geom_smooth(data = subset(new_data,outliers=='n
o'), aes(x=BirthWt,y=NumPhysicianVt),method = 'lm',col='red')+xlab('Birth Weight')+ylab('Number of Physician Visi
t')+theme(axis.title = element_text(size=12))+geom_point(aes(col = outliers,shape=outliers))+geom_smooth(method=
'lm',linetype=2)

grid.arrange(p1,p2,p3,p4)
```




As we can see, nearly all the potential outliers are on the edge of the distribution of the data points. I also plot the regression lines for all the none-outliers (red) and original data(blue dash line). It seems that after removing all the potential outliers, the regression line is still almost the same. Besides, all the variables included in the study are realtively easy to get and not likely to be wrong, thus all the data tend to be trustful. Even if there are some seemly extrem data points, it could still be the nature of the system. In conclusion, it is not necessary to remove all the potential outliers.

Question 4: Testing the significance of the linear relationship observed in the data. Using the model created in question 2, test whether there is a significant linear relationship, use alpha=0.05, between BirthWt and MotherWt; indicate the parameter corresponding to MotherWt with $\beta\_MotherWt$. From the output, answer the following for the test $\beta\_MotherWt$ H0: = 0

```
summary(model)
```

```
##
## Call:
## lm(formula = data_table$BirthWt ~ ., data = data_table)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1825.26  -435.21    55.91   473.46  1701.20
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2439.534    327.234   7.455 3.71e-12 ***
## Age               -3.570      9.620  -0.371 0.711012
## MotherWt           4.354      1.736   2.509 0.013007 *
## RaceOther        133.350    159.393   0.837 0.403925
## RaceWhite        488.428    149.985   3.257 0.001349 **
## SmokeYes        -352.045    106.476  -3.306 0.001142 **
## NumPreLabor      -48.402    101.972  -0.475 0.635607
## HypertensionYes -592.827    202.321  -2.930 0.003830 **
## UterineIrrYes   -516.081    138.885  -3.716 0.000271 ***
## NumPhysicianVt   -14.058     46.468  -0.303 0.762598
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 650.3 on 179 degrees of freedom
## Multiple R-squared:  0.2427, Adjusted R-squared:  0.2047
## F-statistic: 6.376 on 9 and 179 DF,  p-value: 7.891e-08
```

(i) What is the P-value of the test?

Based on the P value for MontherWt which is **0.013007**.

((ii) What does the actual value of the P-value tell you?

It is smaller than alpha = 0.5. Which means the null hypothesis $\beta\_MotherWt$ H0: = 0 is rejected. The alternative hypothesis is Ha: $\neq$ 0.

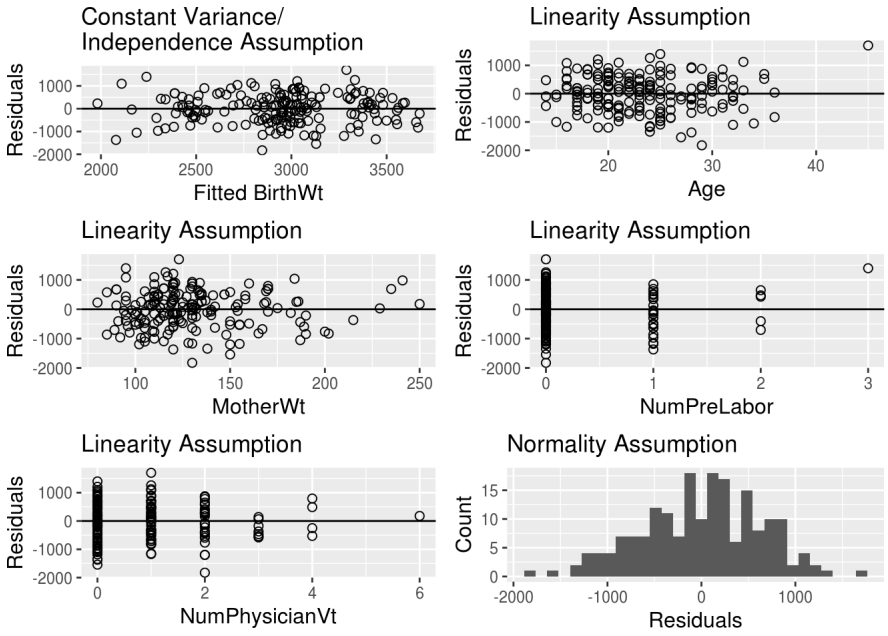
(iii) State your conclusion in the context of the problem.

Thus there is a significant linear relationship between BirthWt and MotherWt.

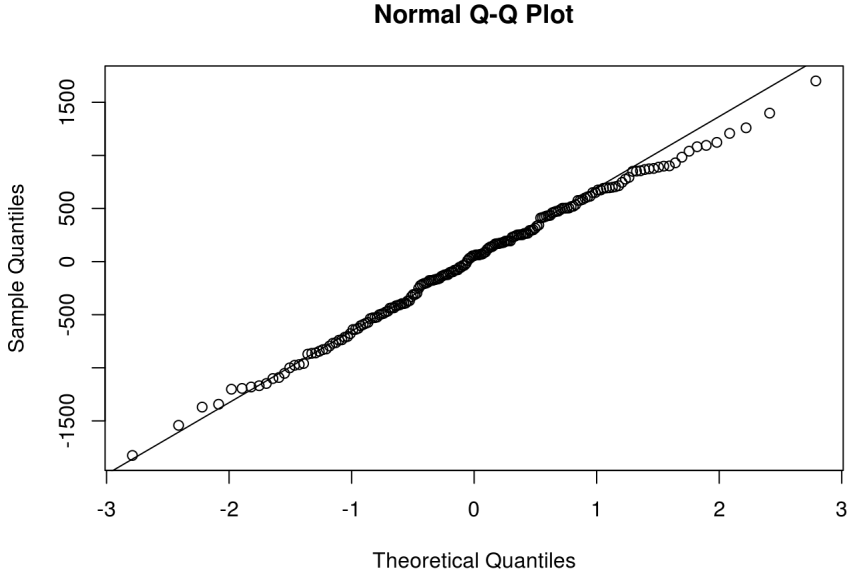
Question 5: Checking the Assumptions of the Model. Plot the relevant residual plots to check the model assumptions. Enumerate the assumptions and describe what graphical techniques you used. Interpret the displays with respect to the assumptions of the linear regression model. In other words, comment on whether there are any apparent departures from the assumptions of the linear regression model.

```
p1=ggplot(data= model, aes(x=model$fitted.values, y=model$residuals))+geom_point(shape=1, size=2)+xlab('Fitted Bi
rthWt')+ylab('Residuals')+theme(axis.title=element_text(size=12))+geom_hline(yintercept = 0)+ggtitle('Constant Va
riance/\nIndependence Assumption')

p2 =ggplot(data= data_table,aes(x=data_table$Age,y=model$residuals))+geom_point(shape=1,size=2)+xlab('Age')+ylab(
'Residuals')+theme(axis.title = element_text(size=12))+geom_hline(yintercept=0)+ggtitle('Linearity Assumption')

p3 =ggplot(data= data_table,aes(x=data_table$MotherWt,y=model$residuals))+geom_point(shape=1,size=2)+xlab('Mother
Wt')+ylab('Residuals')+theme(axis.title=element_text(size=12))+geom_hline(yintercept=0)+ggtitle('Linearity Assump
tion')

p4 =ggplot(data= data_table,aes(x=data_table$NumPreLabor,y=model$residuals))+geom_point(shape=1,size=2)+xlab('Num
PreLabor')+ylab('Residuals')+theme(axis.title = element_text(size=12))+geom_hline(yintercept = 0)+ggtitle('Linear
ity Assumption')

p5 =ggplot(data= data_table,aes(x=data_table$NumPhysicianVt,y=model$residuals))+geom_point(shape=1,size=2)+xlab(
'NumPhysicianVt')+ylab('Residuals')+theme(axis.title =element_text(size=12))+geom_hline(yintercept = 0)+ggtitle(
'Linearity Assumption')

p6 = ggplot(data= data_table,aes(model$residuals))+geom_histogram()+xlab('Residuals')+ylab('Count')+theme(axis.ti
tle =element_text(size=12))+ggtitle('Normality Assumption')

grid.arrange(p1,p2,p3,p4,p5,p6)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
qqnorm(model$residuals)
qqline(model$residuals)
```




**Constant Variance Assumption**: I plot the residuals against the fitted values. All the residuals are distributed around 0. It also seems that the variance is constant across all the fitted values. Thus the **Constant Variance Assumption holds**.

**Independence Assumption** It seems that there are 3 clusters of residuals which means that the error terms are correlated. Thus **Independence Assumption does not hold**.

**Linearity Assumption** I plot the residuals against each predictiing variable. All the plots show residuals distribut equally around the y=0 line. Thus it indicates that **Linearity Assumption holds**.

**Normality assumption** I use the QQ plot to study the normality assumption. Plot the residuals against the theoretical quantiles. The residuals approximately follow a straight line. Thus it suggests that the **Normality assumption holds**.