FINAL PROJECT

DATA 1204 - STATISTICAL AND PREDICTIVE MODELING I

ANALYZING FACTORS INFLUENCING NEWBORN BIRTHWEIGHTS: A STATISTICAL PERSPECTIVE

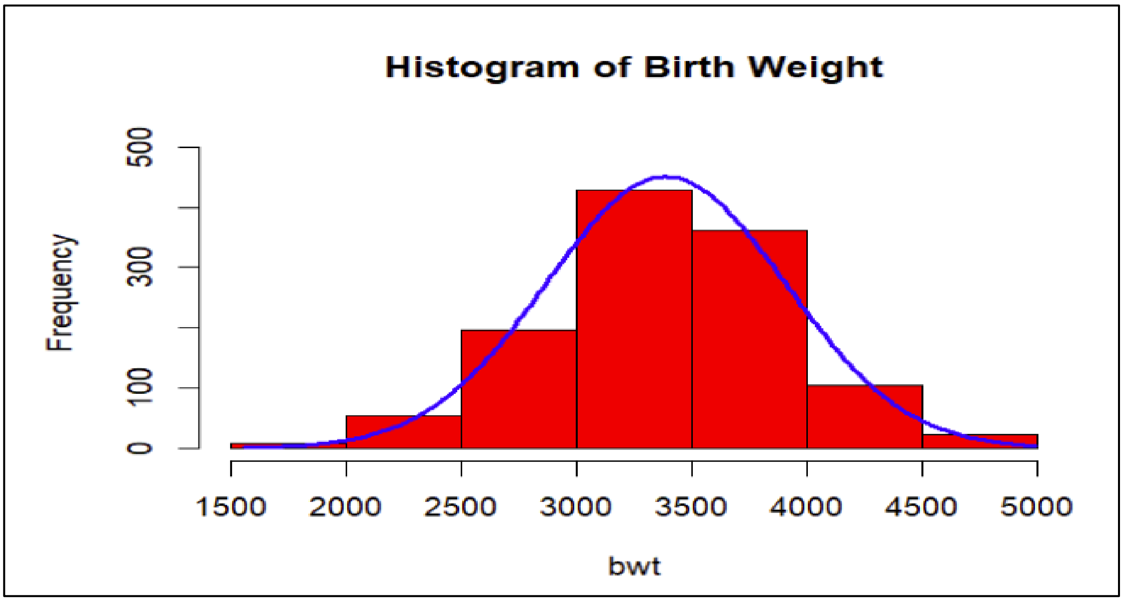
GROUP - 6
Professor: Emad Sheikh

R CODE LINES

Code to compute and state the basic statistics (i.e., Mean, SD, Min/Max):

#Exporting Library

```
> install.packages("lattice")
package ‘lattice’ successfully unpacked and MD5 sums checked
The downloaded binary packages are in
      C:\Users\naguk\AppData\Local\Temp\RtmpKaVKcu\downloaded_packages
> library(gmodels)
> library(lattice)
> library(psych)
Attaching package: ‘psych’
The following objects are masked from ‘package:ggplot2’:
> #View Dataset
> data <- Birthweights

> #Set calculations to 3 digits
> options(digits=3)
> describe(data$bwt)
   vars    n mean  sd median trimmed mad  min  max range  skew kurtosis
se
X1    1 1174 3387 520   3402    3393 462 1559 4990  3430 -0.13     0.43
15.2
> describe(data$gestation)
   vars    n mean sd median trimmed  mad min max range  skew kurtosis   se
X1    1 1174  279 16    280     280 11.9 148 353   205 -0.85     6.74 0.47
> describe(data$age)
   vars    n mean   sd median trimmed  mad min max range skew kurtosis
se
X1    1 1174 27.3 5.93     26    26.9 5.93  14  46    32 0.57    -0.26
0.17
> describe(data$height)
   vars    n mean   sd median trimmed  mad min max range  skew kurtosis
se
X1    1 1174  163 6.51    163     163 7.41 135 183    48 -0.13     0.08
0.19
> describe(data$weight)
   vars    n mean  sd median trimmed  mad  min max range skew kurtosis
se
X1    1 1174 58.3 9.4   56.7    57.4 8.01 39.5 113  73.9 1.21     2.71
0.27
> describe(data$smoke)
   vars    n mean   sd median trimmed mad min max range skew kurtosis   se
X1    1 1174 0.39 0.49      0    0.36   0   0   1     1 0.45     -1.8 0.01

> #Histogram of bwt
> x=data$bwt
> h<-hist(x, breaks=10, col="red", xlab="bwt",
+        main="Histogram of Birth Weight")
> xfit<-seq(min(x),max(x),length=40)
> yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
> yfit <- yfit*diff(h$mids[1:2])*length(x)
> lines(xfit, yfit, col="blue", lwd=2)
```



Conduct a T-test that the mean for "bwt" is equal to 3400:

```
> library(ggplot2)
> library(readr)
> data <- read_csv("C:\\Users\\ajayk\\Downloads\\Birthweights.csv")
> mean_bwt <- mean(data$bwt, na.rm = TRUE)
> sd_bwt <- sd(data$bwt, na.rm = TRUE)
> n <- nrow(data)
> cat("Mean birthweight:", mean_bwt, "\n")
Mean birthweight: 3387
> cat("Standard deviation of birthweight:", sd_bwt, "\n")
Standard deviation of birthweight: 520
> cat("Sample size:", n, "\n")
Sample size: 1174
>
> t_test_result <- t.test(data$bwt, mu = 3400, alternative = "two.sided")
>
> print(t_test_result)
```

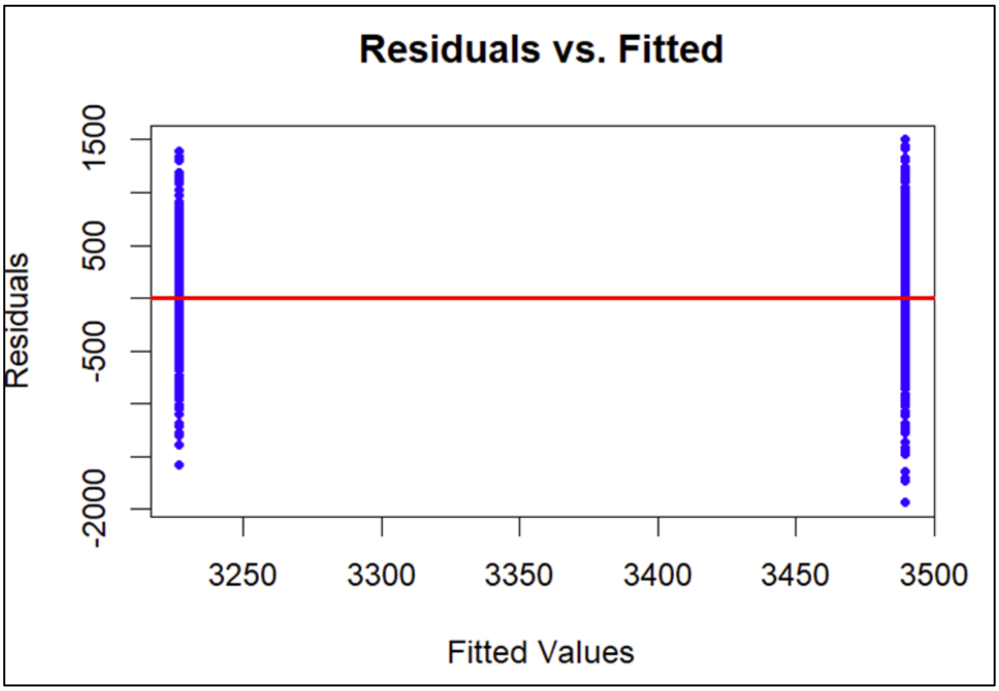
One Sample t-test

```
> data:  data$bwt
t = -0.9, df = 1173, p-value = 0.4
alternative hypothesis: true mean is not equal to 3400
95 percent confidence interval:
 3357 3417
sample estimates:
mean of x
     3387
```

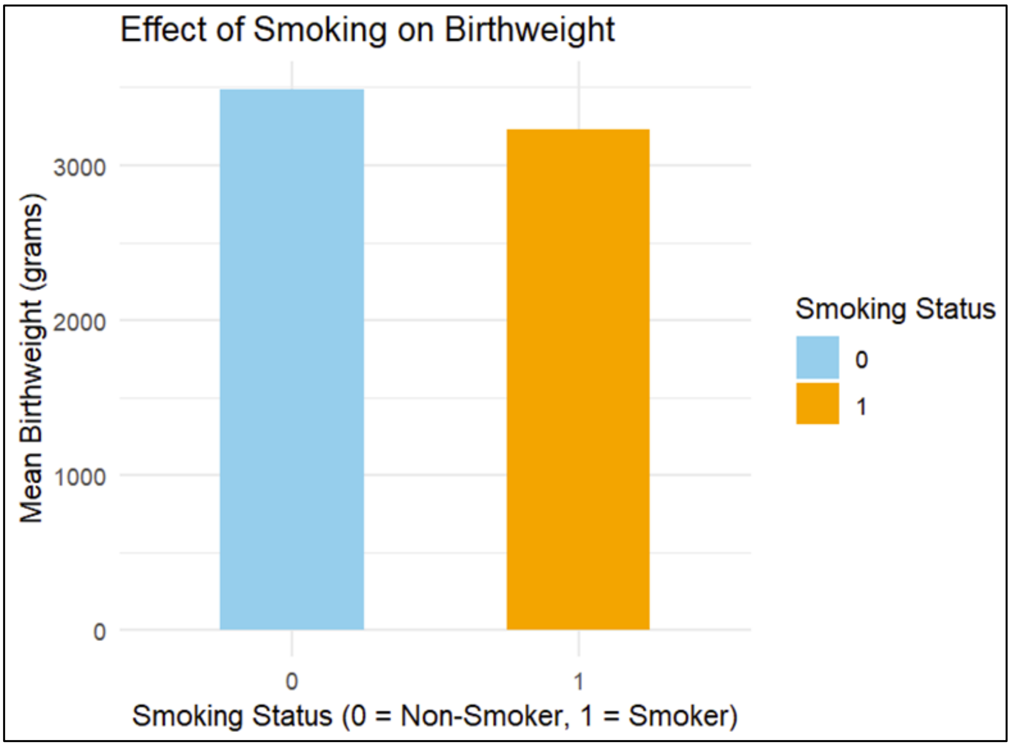
Code for Linear Regression

```
> # Load required libraries
> library(MASS)
> library(ggplot2)
>
> # Load the dataset
> Birthweights <- read.csv("C:/Users/antos/Downloads/Birthweights.csv")
> View(Birthweights)
>
> # Create the relationship model
> model <- lm(bwt ~ smoke, data = Birthweights)
> print(model)

Call:
lm(formula = bwt ~ smoke, data = Birthweights)

Coefficients:
(Intercept)        smoke
     3489.5       -262.7


> # Summary of the Linear Model
> model_summary <- summary(model)
> print(model_summary)

Call:
lm(formula = bwt ~ smoke, data = Birthweights)

Residuals:
     Min       1Q   Median       3Q      Max
-1930.19  -314.29    25.91   317.00  1500.11

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3489.49      18.84 185.222   <2e-16 ***
smoke        -262.69      30.13  -8.719   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 503.8 on 1172 degrees of freedom
Multiple R-squared:  0.06091,	Adjusted R-squared:  0.06011
F-statistic: 76.02 on 1 and 1172 DF,  p-value: < 2.2e-16

> # Extract regression coefficients
> cat("Regression Model: bwt =", round(model$coefficients[1], 2), "+",
+     round(model$coefficients[2], 2), "* smoke\n")

Regression Model: bwt = 3489.49 + -262.69 * smoke

> # Prediction Example
> newdata <- data.frame(smoke = 1)
> predicted_bwt <- predict(model, newdata)
> cat("Predicted Birthweight for a Smoker:", round(predicted_bwt, 2),
"grams\n")
Predicted Birthweight for a Smoker: 3226.8 grams
>
> # Residual vs. Fitted Plot
> plot(model$fitted.values, model$residuals,
+      main = "Residuals vs. Fitted",
+      xlab = "Fitted Values",
+      ylab = "Residuals",
+      pch = 20, col = "blue")
> abline(h = 0, col = "red", lwd = 2)
```

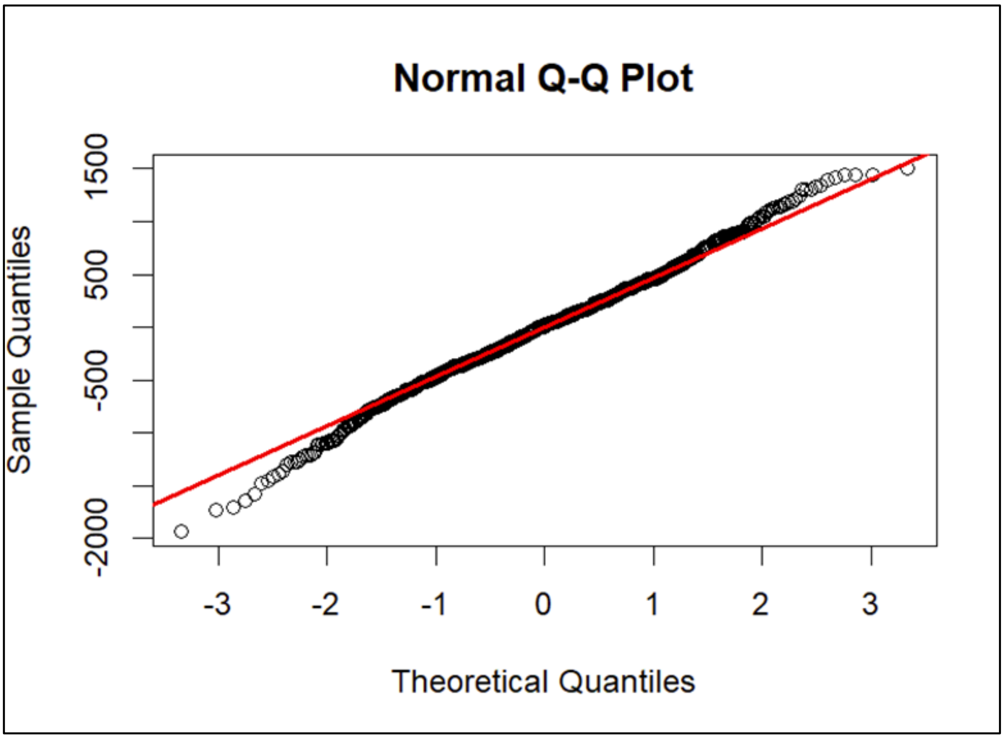
```
>
> # Aggregate mean birthweight by smoking status
> smoke_effect <- aggregate(bwt ~ smoke, data = Birthweights, FUN = mean)
>
```



```
> # Create bar chart of mean birthweight by smoking status
> ggplot(smoke_effect, aes(x = factor(smoke), y = bwt, fill =
factor(smoke))) +
+     geom_bar(stat = "identity", width = 0.5) +
+     labs(title = "Effect of Smoking on Birthweight",
+          x = "Smoking Status (0 = Non-Smoker, 1 = Smoker)",
+          y = "Mean Birthweight (grams)") +
+     scale_fill_manual(values = c("skyblue", "orange"), name = "Smoking
Status") +
+     theme_minimal() +
+     theme(legend.position = "none")
```



```
>
> # Q-Q Plot for Normality of Residuals
> qqnorm(model$residuals, main = "Normal Q-Q Plot")
> qqline(model$residuals, col = "red", lwd = 2)
>
```



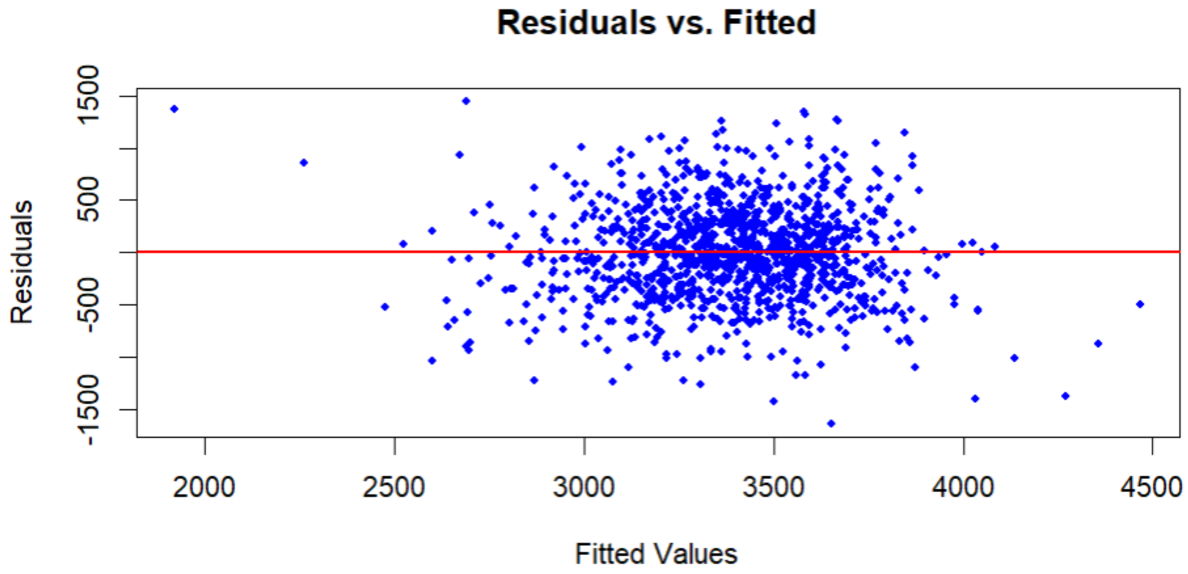
Code for Multiple Regression

```
> library(MASS)
Warning message:
package ‘MASS’ was built under R version 4.4.2
> Birthweights <- read.csv("C:/Users/Jani Begam/Downloads/Birthweights.csv
")
>   View(Birthweights)
> # Create the relationship model
> model <- lm(bwt ~ gestation + region + age + height + weight + smoke, da
ta = Birthweights)
> print(model)

Call:
lm(formula = bwt ~ gestation + region + age + height + weight + 
    smoke, data = Birthweights)

Coefficients:
    (Intercept)        gestation  regionnorthwest  
      -2182.074           12.378          -31.361  
regionsoutheast  regionsouthwest              age  
        -14.172          -30.795            1.960  
         height           weight            smoke  
         12.059            3.595         -235.434  

> # Fit the multiple linear regression model
> multi_model <- lm(bwt ~ gestation + region + age + height + weight + smo
ke, data = Birthweights)
> model_summary <- summary(multi_model)
> print(model_summary)
```

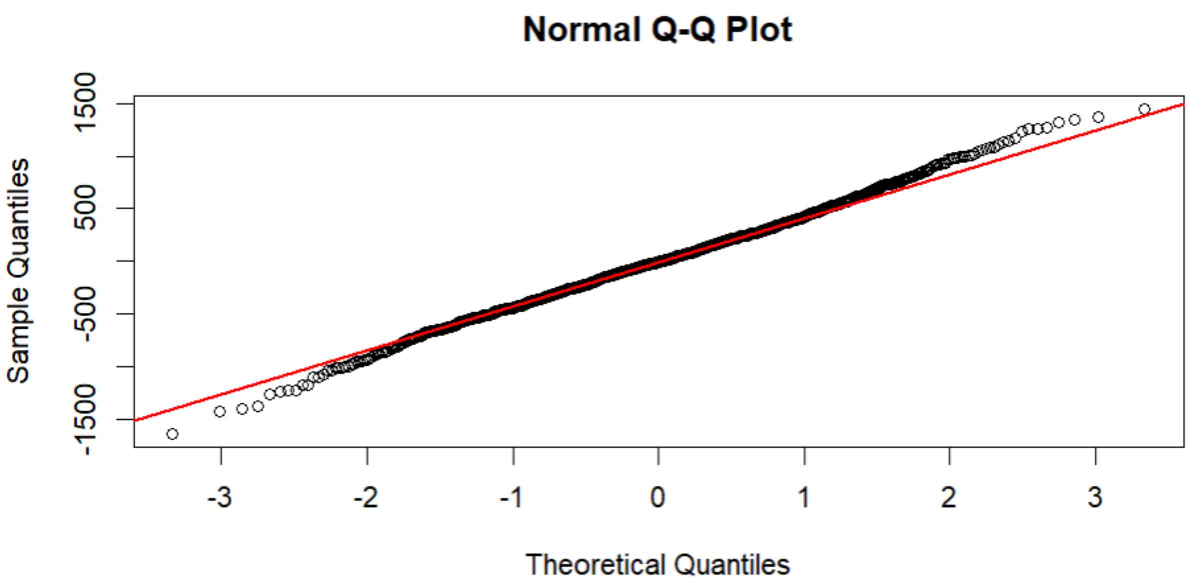
```
Call:
lm(formula = bwt ~ gestation + region + age + height + weight + 
    smoke, data = Birthweights)

Residuals:
     Min       1Q   Median       3Q 
-1640.40  -293.00    -3.12   270.92 
     Max 
 1450.25 

Coefficients:
                  Estimate Std. Error
(Intercept)     -2182.0740   408.1697
gestation          12.3779     0.8294
regionnorthwest   -31.3607    38.0497
regionsoutheast   -14.1721    36.8045
regionsouthwest   -30.7947    38.0228
age                 1.9596     2.2585
height             12.0588     2.2668
weight              3.5946     1.5849
smoke            -235.4339    27.1811
                t value Pr(>|t|)    
(Intercept)      -5.346 1.08e-07 ***
gestation        14.924  < 2e-16 ***
regionnorthwest  -0.824   0.4100    
regionsoutheast  -0.385   0.7003    
regionsouthwest  -0.810   0.4182    
age               0.868   0.3858    
height            5.320 1.24e-07 ***
weight            2.268   0.0235 *  
smoke            -8.662  < 2e-16 ***
---
Signif. codes:  
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
  0.1 ‘ ’ 1

Residual standard error: 450.7 on 1165 degrees of freedom
Multiple R-squared:  0.2529,	Adjusted R-squared:  0.2478 
F-statistic: 49.29 on 8 and 1165 DF,  p-value: < 2.2e-16
```

```
> # Backward Stepwise Regression
> backward_model <- stepAIC(multi_model, direction = "backward", trace = F
ALSE)
> summary(backward_model)

Call:
lm(formula = bwt ~ gestation + height + weight + smoke, data = Birthweight
s)

Residuals:
     Min       1Q   Median       3Q 
-1599.20  -293.69   -10.34   279.92 
     Max 
 1460.84 

Coefficients:
              Estimate Std. Error t value
(Intercept) -2157.3715   394.2329  -5.472
gestation      12.3966     0.8248  15.031
height         12.0368     2.2515   5.346
weight          3.7445     1.5581   2.403
smoke        -236.5044    27.0610  -8.740
            Pr(>|t|)    
(Intercept) 5.43e-08 ***
gestation    < 2e-16 ***
height      1.08e-07 ***
weight        0.0164 *  
smoke        < 2e-16 ***
---


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

Residual standard error: 450.2 on 1169 degrees of freedom
Multiple R-squared:  0.2518,	Adjusted R-squared:  0.2492 
F-statistic: 98.34 on 4 and 1169 DF,  p-value: < 2.2e-16


> # Forward Stepwise Regression
> forward_model <- stepAIC(multi_model, direction = "forward", trace = FAL
SE)
> summary(forward_model)

Call:
lm(formula = bwt ~ gestation + region + age + height + weight + 
    smoke, data = Birthweights)

Residuals:
     Min       1Q   Median       3Q 
-1640.40  -293.00    -3.12   270.92 
     Max 
 1450.25 

Coefficients:
                  Estimate Std. Error
(Intercept)     -2182.0740   408.1697
gestation          12.3779     0.8294
regionnorthwest   -31.3607    38.0497
regionsoutheast   -14.1721    36.8045
regionsouthwest   -30.7947    38.0228
age                 1.9596     2.2585
height             12.0588     2.2668
weight              3.5946     1.5849
smoke            -235.4339    27.1811
                t value Pr(>|t|)    
(Intercept)      -5.346 1.08e-07 ***
gestation        14.924  < 2e-16 ***
```

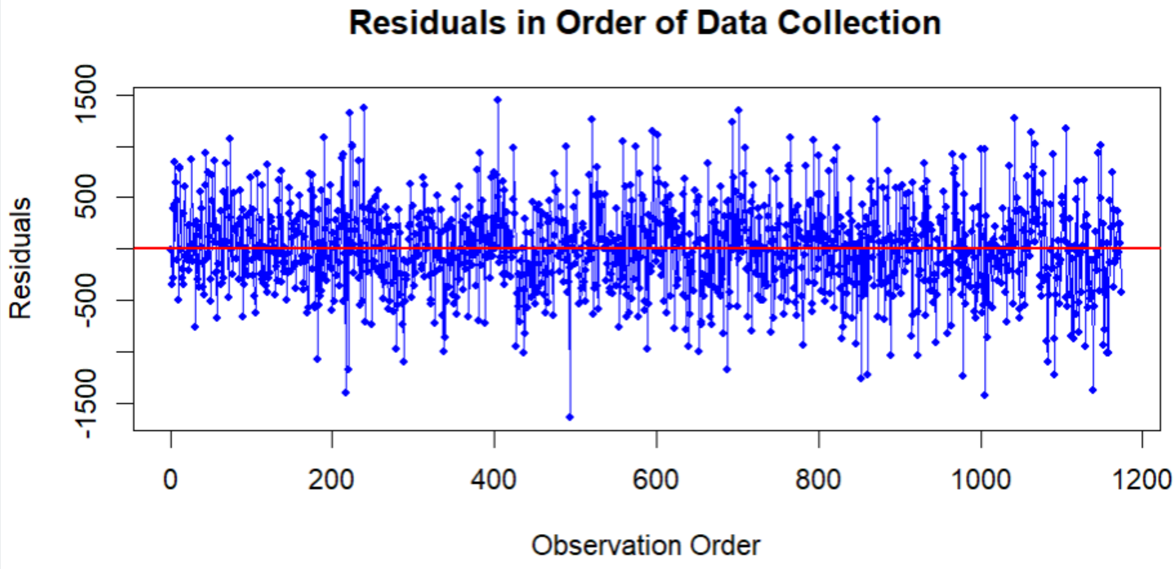
```
regionnorthwest  -0.824   0.4100    
regionsoutheast  -0.385   0.7003    
regionsouthwest  -0.810   0.4182    
age               0.868   0.3858    
height            5.320 1.24e-07 ***
weight            2.268   0.0235 *  
smoke            -8.662  < 2e-16 ***
---
Signif. codes:  
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
  0.1 ‘ ’ 1

Residual standard error: 450.7 on 1165 degrees of freedom
Multiple R-squared:  0.2529,	Adjusted R-squared:  0.2478 
F-statistic: 49.29 on 8 and 1165 DF,  p-value: < 2.2e-16

>  # Prediction to forecast new birthweight values based on new input data
> newdata <- data.frame(gestation = 280, region = 'southwest', age = 30, h
eight = 165, weight = 70, smoke = 1)
> predicted_bwt <- predict(model, newdata)
> print(predicted_bwt)
       1 
3317.615 

> # Residual vs. Fitted plot
> plot(multi_model$fitted.values, multi_model$residuals,
+      main = "Residuals vs. Fitted",
+      xlab = "Fitted Values",
+      ylab = "Residuals",
+      pch = 20, col = "blue")
> abline(h = 0, col = "red", lwd = 2)  # Add a horizontal line at 0
```



```
> # Q-Q Plot for Normality of Residuals
> qqnorm(multi_model$residuals, main = "Normal Q-Q Plot")
> qqline(multi_model$residuals, col = "red", lwd = 2)
```



```
> # Residuals in Order of Data Collection
> plot(multi_model$residuals, type = "o",
+      main = "Residuals in Order of Data Collection",
+      xlab = "Observation Order",
+      ylab = "Residuals",
+      col = "blue", pch = 20)
> abline(h = 0, col = "red", lwd = 2)
```


Residuals in Order of Data Collection
Residuals
Observation Order