LINEAR REGRESSION IN R

BY : Kikonyogo Steven

Definition

Linear regression is a data analysis technique for predicting and modelling the relationship between a response dependent quantitative variable and either one independent variable (Simple linear regression) or several explanatory variables (multiple linear regression).

The main assumptions of the model are;

(i) Linearity of two variables.

(ii) Normality of residuals.

(iii) Constant variability (homoscedasticity)

About the project

We are to use data from 619 new born babies.We are interested in predicting their birth weight in terms of gestation period, weight of mother, mother and sex of a baby.

(i) Importing data and loading packages

```
birthweight <- rio::import( here::here("Birth_baby_weight.csv"))

if(!require(pacman)) install.packages("pacman")
pacman::p_load( gt,car, dplyr, lessR, ggplot2, magrittr, janitor, plotrix, kableExtra )
```

Data manipulation and an overview of the data.

```
bweight <- birthweight %>%
  mutate(birthwt = bwt_ounce*0.02835,
         mheight = mheight_inches*0.0254,
         mweight = mweight_pounds*0.4536)

bweight <- bweight %>%
  select(Gender,birthwt,gestation_days,mother_age,
         mweight) %>%
  filter(gestation_days > 200 & gestation_days < 350)

bweight$Gender <- as.factor(bweight$Gender)

levels(bweight$Gender) <- c("male","female")


str(bweight)
```

Table 1: Sample data

Gender	birthwt	gestation_days	mother_age	mweight
male	3.40	284	27	45.36
male	3.20	282	33	61.24
male	3.63	279	28	52.16
male	3.06	282	23	56.70
male	3.86	286	25	42.18
male	3.91	244	33	80.74
female	3.74	245	23	63.50
female	3.40	289	25	56.70
male	4.05	299	30	61.69
male	4.08	282	32	56.25

```
## 'data.frame':    617 obs. of  5 variables:
##  $ Gender        : Factor w/ 2 levels "male","female": 1 1 1 1 1 1 2 2 1 1 ...
##  $ birthwt       : num  3.4 3.2 3.63 3.06 3.86 ...
##  $ gestation_days: int  284 282 279 282 286 244 245 289 299 282 ...
##  $ mother_age    : int  27 33 28 23 25 33 23 25 30 32 ...
##  $ mweight       : num  45.4 61.2 52.2 56.7 42.2 ...
```

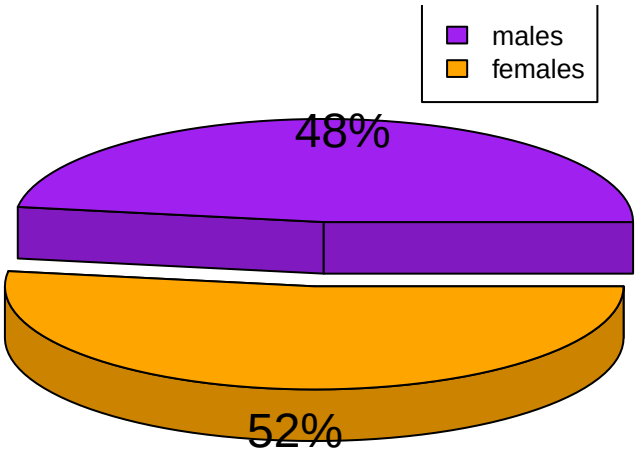
```
kable(head(bweight, 10),caption = "Sample data", digits = 2)
```

Descriptive statistics of the data

```
summary(bweight)
```

```
##     Gender       birthwt      gestation_days    mother_age       mweight
##  male  :294   Min.   :1.928   Min.   :223.0   Min.   :15.00   Min.   : 40.37
##  female:323   1st Qu.:3.090   1st Qu.:272.0   1st Qu.:23.00   1st Qu.: 52.16
##               Median :3.402   Median :279.0   Median :26.00   Median : 57.15
##               Mean   :3.408   Mean   :278.9   Mean   :26.97   Mean   : 58.72
##               3rd Qu.:3.714   3rd Qu.:288.0   3rd Qu.:30.00   3rd Qu.: 63.50
##               Max.   :4.990   Max.   :329.0   Max.   :44.00   Max.   :113.40
```

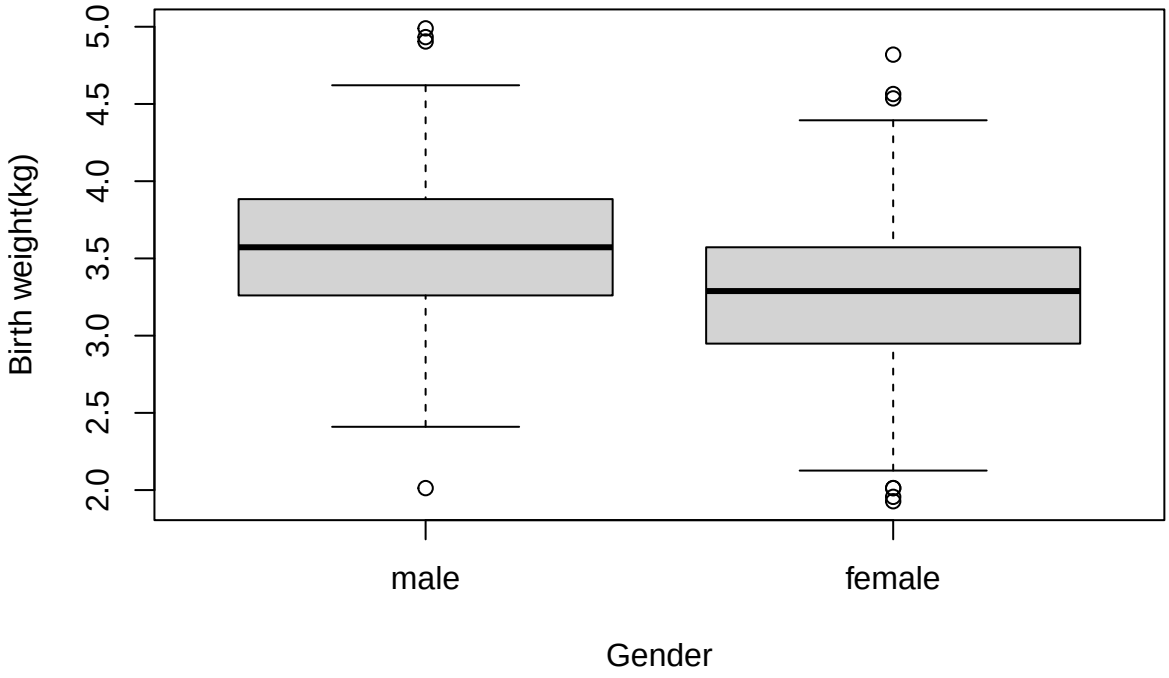
```
table1 <- table(bweight$Gender)


piepercent <-paste0(round(100 * table1/sum(table1)), "%")

plotrix::pie3D(table1,radius = 1.2,
               explode = 0.25,
               labels = piepercent,
               col = c("purple","orange"),
               main = "Births according to sex  ",
               col.main="blue")
legend("topright",
       c("males","females"),
       cex = 0.8,
       fill =c("purple","orange") )
```



Majority of babies were females (n=323, 52%) compared to males (n=294, 48%). The minimum birth weight was1.9kg, maximum was 5.0kg and mean was 3.4kg. The minimum mother's age was 15 years, maximum was 44 years and mean was 27 years. The minimum mother's weight was 40.4kg, maximum was 113.4kg and mean was 58.7kg.

Testing of hypotheses

```
# Testing for the diff in mean birth weight between male babies and female babies

boxplot(birthwt ~ Gender, data = bweight,
        main="BOX PLOT OF BIRTH WEIGHT BY GENDER",
             ylab = "Birth weight(kg)",
             col.main = "blue" )
```




```
shapiro.test(bweight$birthwt) # Test for normality
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bweight$birthwt
## W = 0.99647, p-value = 0.1894
```

```
car::leveneTest(bweight$birthwt ~ bweight$Gender) # Test for homogeneity of variance
```

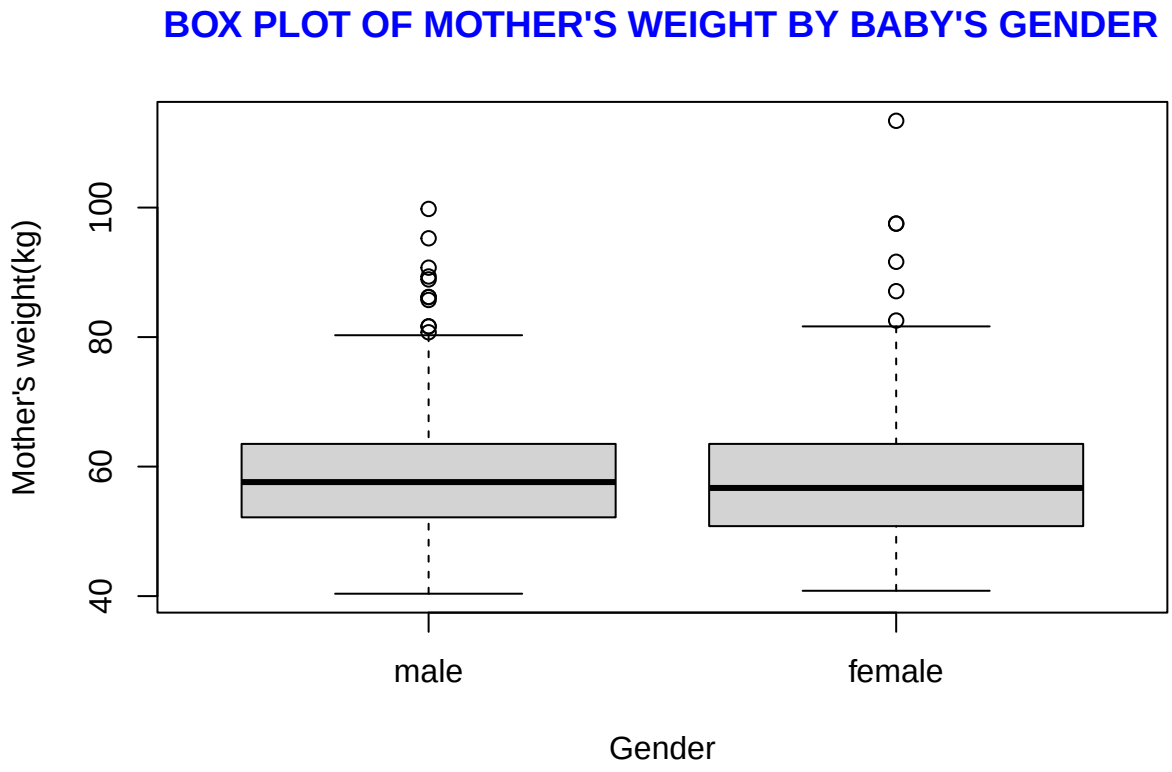
```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1   0.092 0.7618
##       615
```

```
t.test(data = bweight, birthwt ~ Gender,
       alt = "two.sided", var.eq = T)
```

```
##
##  Two Sample t-test
##
## data:  birthwt by Gender
## t = 8.0521, df = 615, p-value = 0.000000000000004226
## alternative hypothesis: true difference in means between group male and group female is not equal to
```

```
## 95 percent confidence interval:
##  0.227013 0.373463
## sample estimates:
##   mean in group male mean in group female
##             3.564964             3.264726
```

```
# Testing diff in mean mother's weight and baby's gender.

boxplot(mweight ~ Gender, data = bweight,
        main="BOX PLOT OF MOTHER'S WEIGHT BY BABY'S GENDER",
             ylab = "Mother's weight(kg)",
             col.main = "blue" )
```




```
car::leveneTest(bweight$mweight ~ bweight$Gender) # Test for homogeneity of variance
```

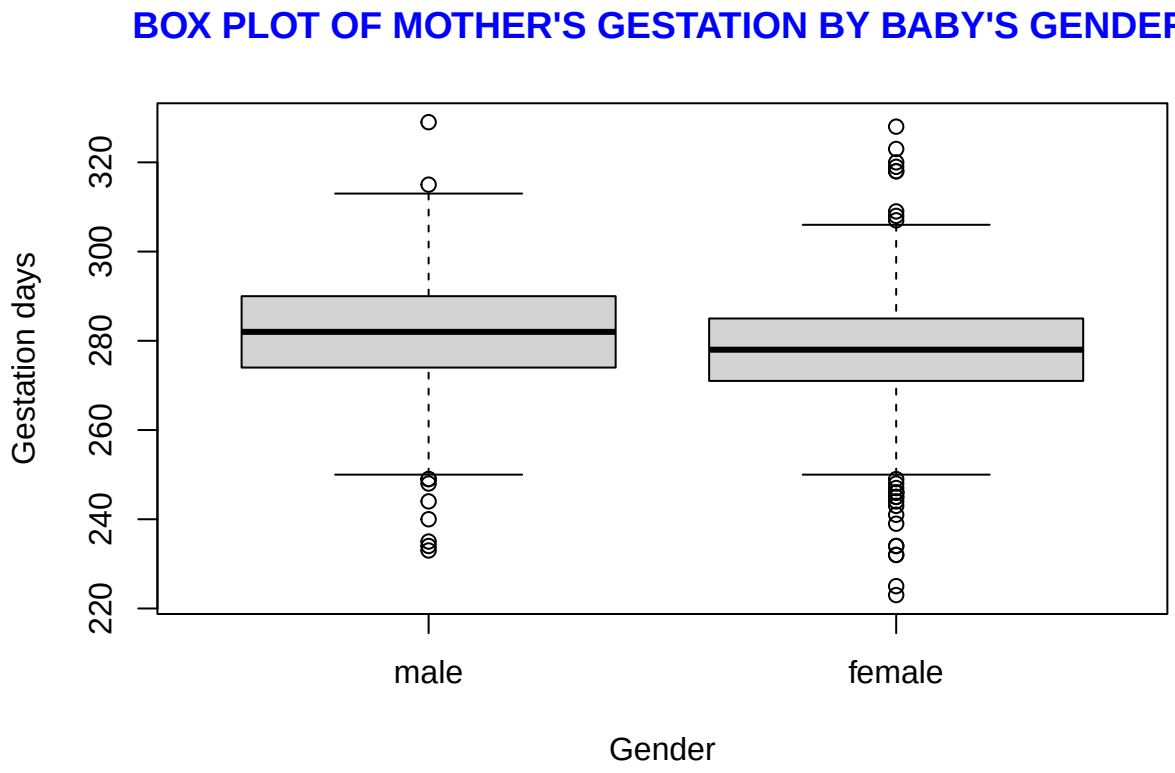
```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.0146 0.9039
##       615
```

```
t.test(data = bweight, mweight ~ Gender,
       alt = "two.sided", var.eq = T)
```

```
##
```

```
##  Two Sample t-test
##
## data:  mweight by Gender
## t = 0.96448, df = 615, p-value = 0.3352
## alternative hypothesis: true difference in means between group male and group female is not equal to
## 95 percent confidence interval:
##  -0.7906587  2.3168071
## sample estimates:
##   mean in group male mean in group female
##             59.11457             58.35150
```

```
# Testing diff between mother's gestation days and baby's gender.

boxplot(gestation_days ~ Gender, data = bweight,
        main="BOX PLOT OF MOTHER'S GESTATION BY BABY'S GENDER",
             ylab = "Gestation days",
             col.main = "blue" )
```




```
car::leveneTest(bweight$gestation_days ~ bweight$Gender) # Test for homogeneity of variance
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.4467 0.5042
##       615
```

```
t.test(data = bweight, gestation_days ~ Gender,
       alt = "two.sided", var.eq = T)
```

```
##
##  Two Sample t-test
##
## data:  gestation_days by Gender
## t = 2.7925, df = 615, p-value = 0.005392
## alternative hypothesis: true difference in means between group male and group female is not equal to
## 95 percent confidence interval:
##  0.9590464 5.5044443
## sample estimates:
##   mean in group male mean in group female
##             280.6156             277.3839
```

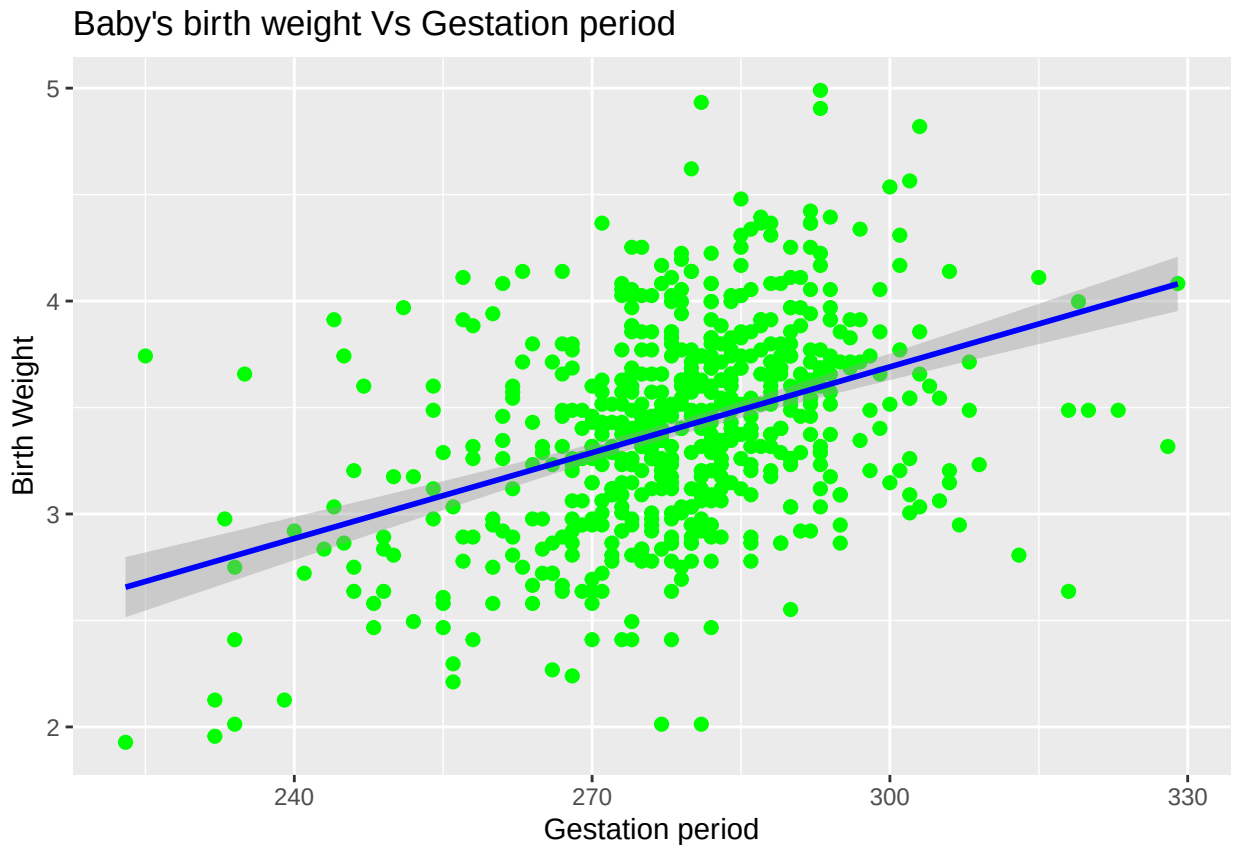
(i) The mean male birth weight = 3.6kg and for females = 3.3kg showing a statistical significant difference in mean birth weight (p<0.05).

(ii) The mean gestation period of mothers that birth male babies = 281days and females =277days showing a statistical significant difference in mean birth gestation period (p=0.005).

(iii) The mean weight of mothers that birth male babies = 59.1kg and females =58.3kg showing no statistical significant difference in mean birth gestation period (p=0.034).

fitting a simple linear regression model

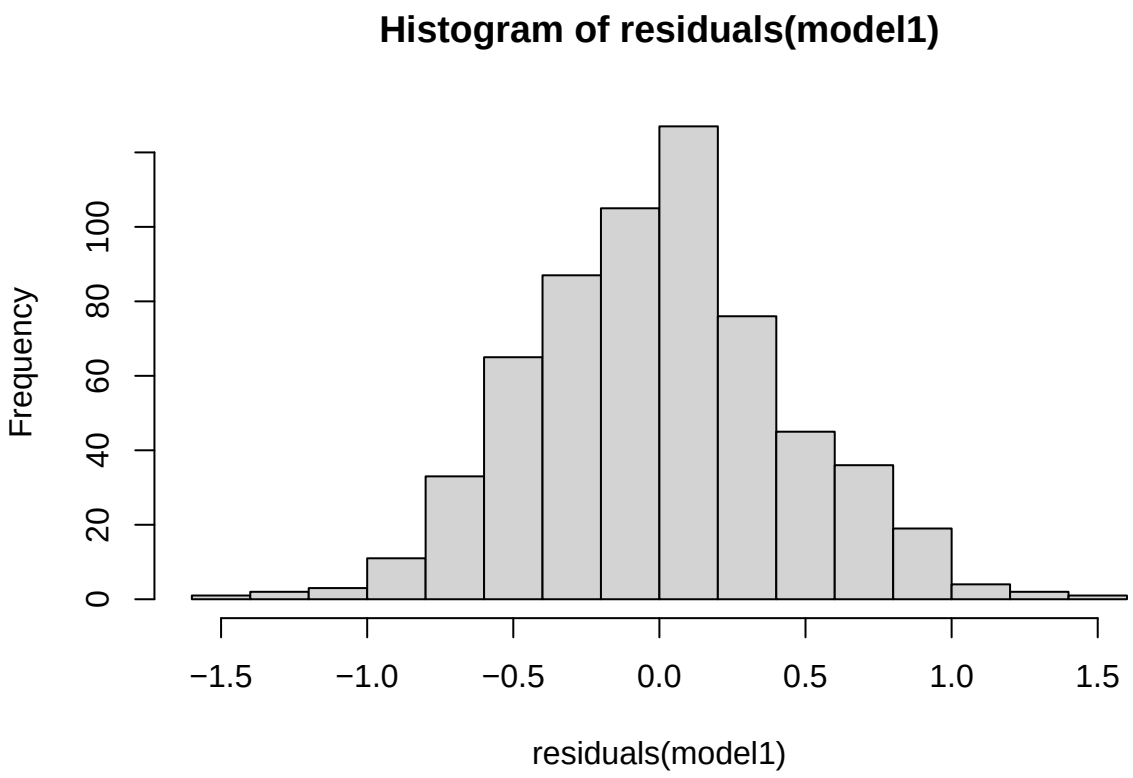
```
ggplot(data = bweight,
       mapping=aes(x=gestation_days,
                   y=birthwt))+
  geom_point(col='green',
             size=2)+
  geom_smooth(method='lm',
              col= 'blue')+
  labs(title="Baby's birth weight Vs Gestation period",
       x="Gestation period",
       y="Birth Weight")
```



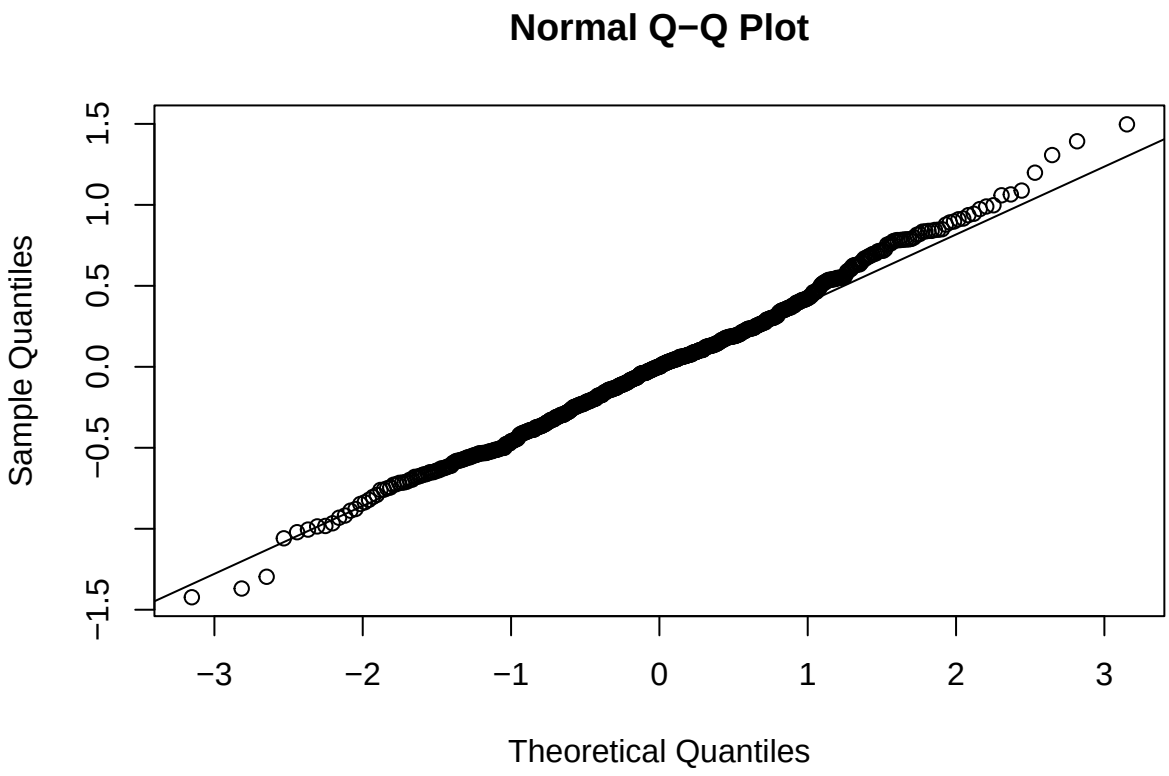

```
model1 <- lm(bweight$birthwt~bweight$gestation_days)

cor.test(bweight$birthwt,
         bweight$gestation_days,method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  bweight$birthwt and bweight$gestation_days
## t = 10.798, df = 615, p-value < 0.00000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3306937 0.4635334
## sample estimates:
##       cor
## 0.3992065
```

```
# Test for normality of residuals
hist(residuals(model1))
```




```
qqnorm(resid(model1))
qqline(resid(model1))
```

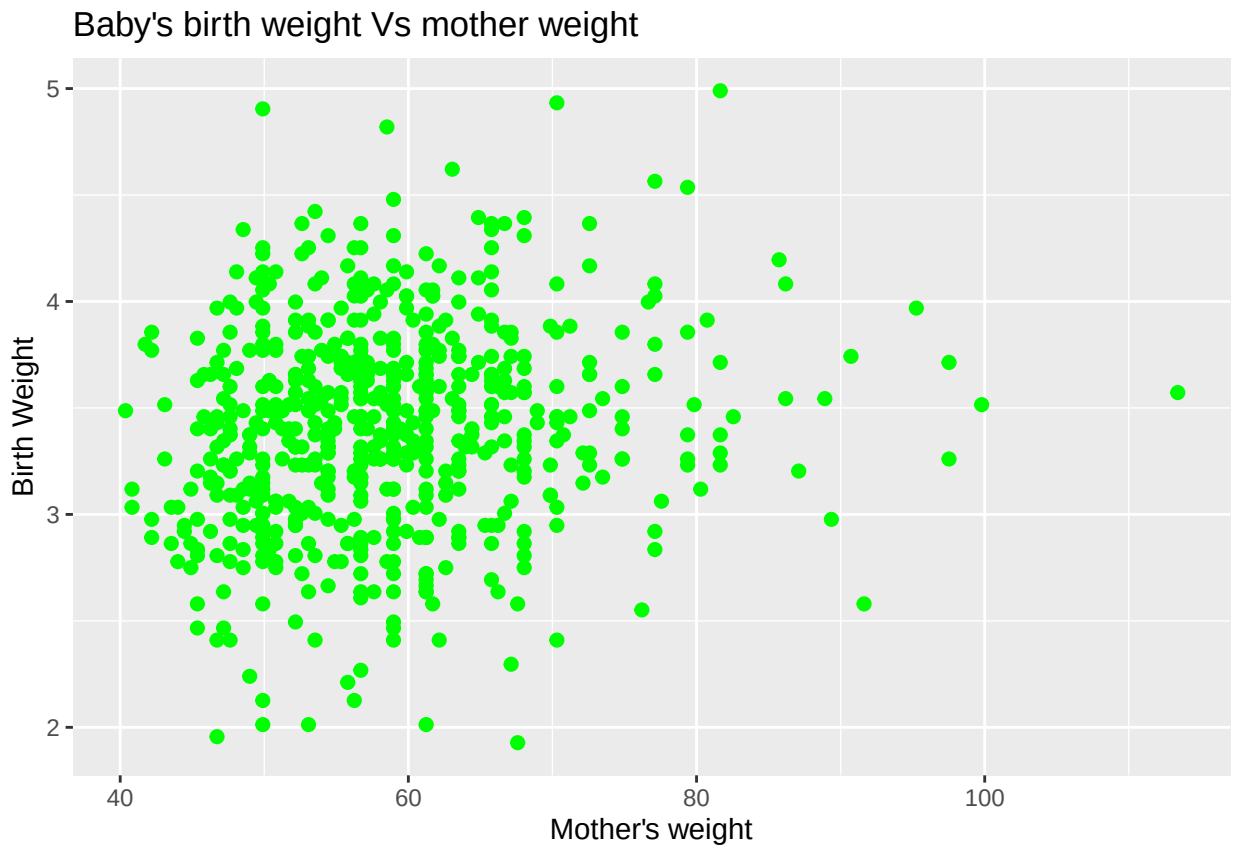



```
summary(model1)
```

```
##
## Call:
## lm(formula = bweight$birthwt ~ bweight$gestation_days)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.42284 -0.30375  0.00167  0.26160  1.49721
##
## Coefficients:
##                          Estimate Std. Error t value            Pr(>|t|)
## (Intercept)            -0.340466   0.347598  -0.979               0.328
## bweight$gestation_days  0.013438   0.001245  10.798 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4459 on 615 degrees of freedom
## Multiple R-squared:  0.1594, Adjusted R-squared:  0.158
## F-statistic: 116.6 on 1 and 615 DF,  p-value: < 0.00000000000000022
```

```
ggplot(data = bweight,
       mapping=aes(x=mweight,
                   y=birthwt))+
  geom_point(col='green',
```

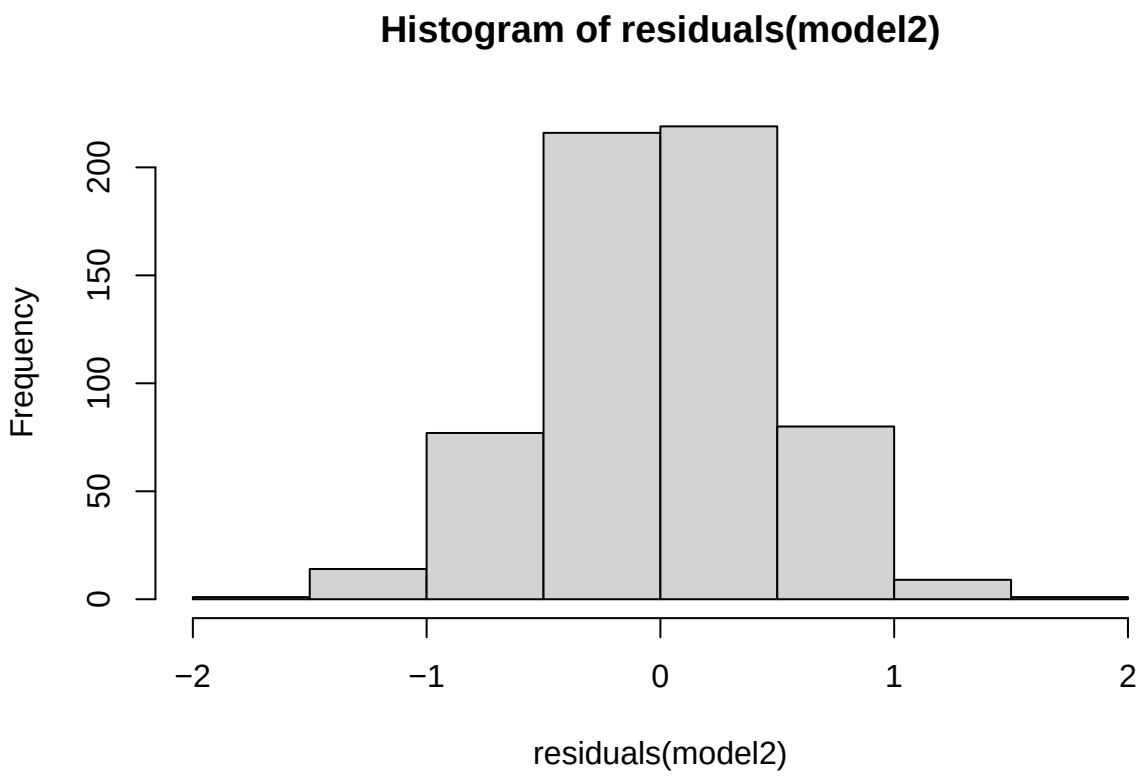
```
            size=2)+
  labs(title="Baby's birth weight Vs mother weight",
       x="Mother's weight",
       y="Birth Weight")
```



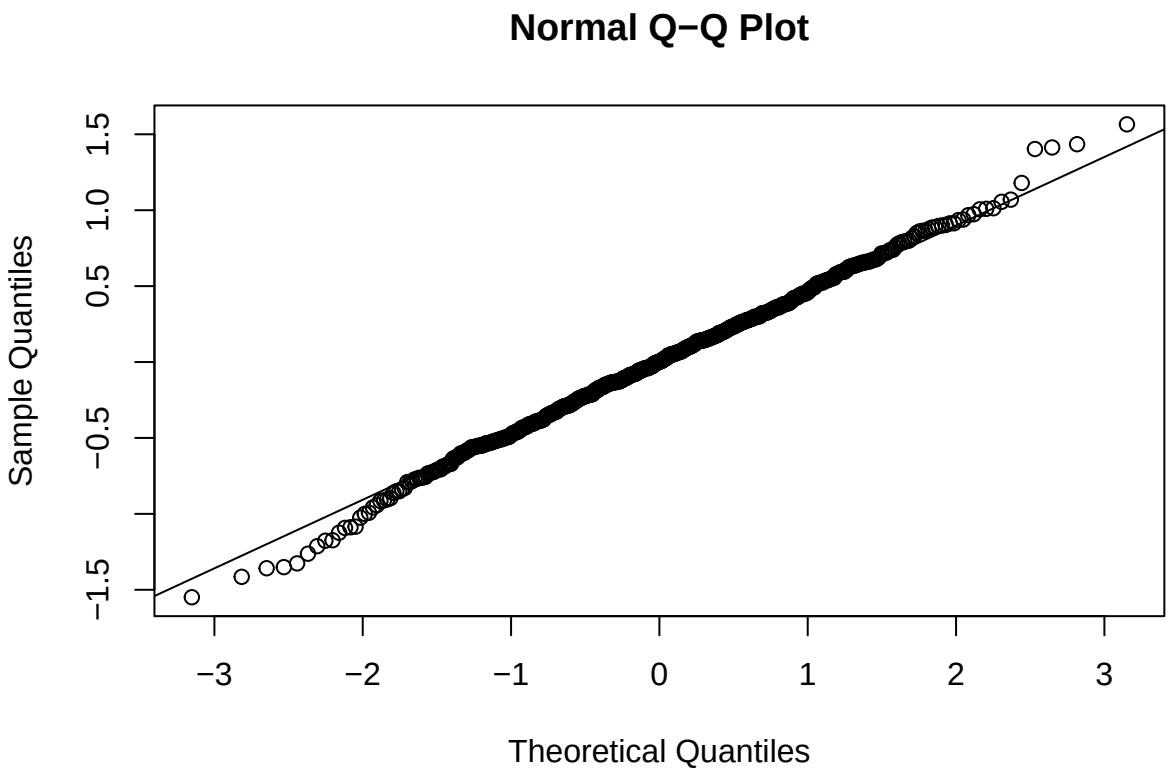

```
cor.test(bweight$birthwt,
         bweight$mweight,method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  bweight$birthwt and bweight$mweight
## t = 3.9535, df = 615, p-value = 0.00008596
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.07948783 0.23346518
## sample estimates:
##       cor
## 0.1574332
```

```
model2 <- lm(bweight$birthwt~bweight$mweight)
# Test for normality of residuals
hist(residuals(model2))
```




```
qqnorm(resid(model2))
qqline(resid(model2))
```




```
summary(model2)
```

```
##
## Call:
## lm(formula = bweight$birthwt ~ bweight$mweight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54914 -0.30862  0.00266  0.30049  1.56551
##
## Coefficients:
##                 Estimate Std. Error t value             Pr(>|t|)
## (Intercept)     2.950106   0.117370  25.135 < 0.0000000000000002 ***
## bweight$mweight 0.007795   0.001972   3.954             0.000086 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4803 on 615 degrees of freedom
## Multiple R-squared:  0.02479,    Adjusted R-squared:  0.0232
## F-statistic: 15.63 on 1 and 615 DF,  p-value: 0.00008596
```

```
cor.test(bweight$birthwt,
         bweight$mweight,method = "pearson")
```

```
##
```

```
##  Pearson's product-moment correlation
##
## data:  bweight$birthwt and bweight$mweight
## t = 3.9535, df = 615, p-value = 0.00008596
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.07948783 0.23346518
## sample estimates:
##       cor
## 0.1574332
```

Fitting a multiple linear regression model

```
model3 <- lm(birthwt ~ mweight + gestation_days,
             data = bweight)

model <- lm(birthwt ~ mweight + gestation_days + Gender,data = bweight)
summary(model)
```

```
##
## Call:
## lm(formula = birthwt ~ mweight + gestation_days + Gender, data = bweight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.51267 -0.29415 -0.01634  0.27587  1.28121
##
## Coefficients:
##                 Estimate Std. Error t value             Pr(>|t|)
## (Intercept)    -0.364342   0.348954  -1.044                0.297
## mweight         0.007320   0.001731   4.229    0.000027107816233 ***
## gestation_days  0.012460   0.001183  10.529 < 0.0000000000000002 ***
## Genderfemale   -0.254383   0.034205  -7.437    0.000000000000349 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4214 on 613 degrees of freedom
## Multiple R-squared:  0.2518, Adjusted R-squared:  0.2482
## F-statistic: 68.78 on 3 and 613 DF,  p-value: < 0.00000000000000022
```

(i) There is a moderate positive correlation between birth weight and gestation period(r=0.4).The simple model is;

birthweight = -0.34+0.0013gestation days.

(ii) There is a low positive correlation between birth weight and mother's weight(r = 0.15).The simple model is;

birthweight = 2.95+0.008motherweight.

(iii) The multiple regression model is;

birthweight = -0.36+0.007*motherweight+0.0125*gestationdays-0.254*gender.

[e.g: lets assume motherweight = 55kg, gestation period = 280 days].

(a) if the baby is a boy.The weight is;

*-0.36+0.007*55+0.0125*280-0.25*0 = 3.52kg.*

(b) if the baby is a girl.The weight is;

*-0.36+0.007*55+0.0125*280-0.25*1 = 3.5kg.*

This means that adjusting by mother's weight and gestation period, baby boys are slightly heavier than baby girls.However, the difference is statistically not significant.(R^2 = 25%).