CSC465 Assignment 1

Zeyi Pan

2/27/2020

Table of Contents

Q1

(a)

Using standard matrix multiplication, the least squares estimates of regression coefficients β is given by $\hat{\beta} = (X^T X)^{-1} X^T y$. This will be the unique solution if $X^T X$ is invertible.

(b)

We can get

$$X = \begin{bmatrix} 1 & & & \\ \vdots & X_{i1} & X_{i2} & X_{i3} \\ 1 & & & \end{bmatrix}_{n\times 4}$$

$$X^T X = \begin{bmatrix} n & n_1 & n_2 & n_3 \\ n_1 & n_1 & 0 & 0 \\ n_2 & 0 & n_2 & 0 \\ n_3 & 0 & 0 & n_3 \end{bmatrix}_{4\times 4}$$

Compute the determinant of $X^T X$, we get

$$detA = n_1^2 n_2 n_3 + n_1 n_2^2 n_3 + n_1 n_2 (n n_3 - n_3^2)$$

Notice that $n_1 + n_2 + n_3 = n$
i) if $n_k \neq n \ \ (k = 1, 2, 3)$, $detA \neq 0$. The matrix is invertible.
ii) if any $n_k = n \ \ (k = 1, 2, 3)$, $detA = 0$. The matrix is not invertible.

(c)

Deleting any of the four terms associated with coefficients will result in deleting m row and m column of $X^T X$. No matter which term is deleted, if $n_k \neq n \ \ (k = 1, 2, 3)$, $X^T X$ will always be a full rank matrix. Therefore, it will be invertible.

(d)

As we know, least square estimation is the orthogonal projection of vector y onto the column space of X. Therefore, if the column spaces of these four models are the same, the fitted values will be the same. Their column spaces are the same. Specifically, the two matrix as below have the same column space. (1) They are both subspaces of R^n. (2) $e_0 = e_1 + e_2 + e_3$ because only one of X_{i1}, X_{i2}, X_{i3} is 1 and then they sum up to 1. This means their basis vector can be written as linear combination of the other.

$$\begin{matrix} e_0 & e_1 & e_2 \end{matrix}$$
$$\begin{bmatrix} 1 & & \\ \vdots & X_{i1} & X_{i2} \\ 1 & & \end{bmatrix}_{n\times 3}$$
$$\begin{matrix} e_1 & e_2 & e_3 \end{matrix}$$
$$\begin{bmatrix} X_{i1} & X_{i2} & X_{i3} \end{bmatrix}_{n\times 3}$$

Q2

(a)

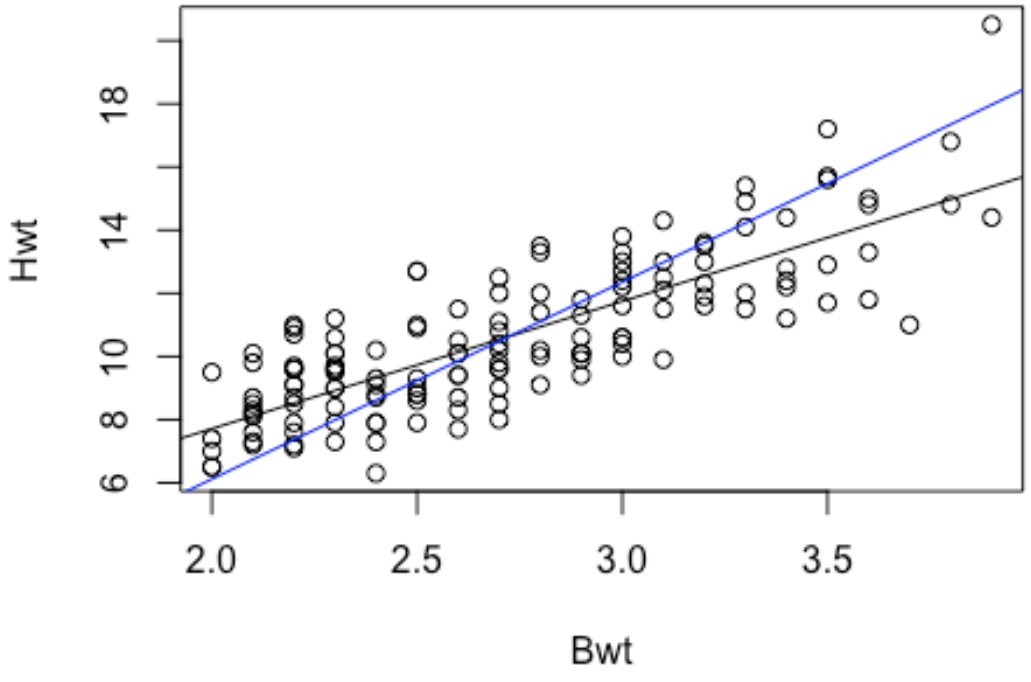
$$\beta_0' = -\frac{\beta_0}{\beta_1}, \beta_1' = \frac{1}{\beta_1}$$

(b)

The coefficients of these two models do not conform to the equivalence relationship given in (a). To estimate parameters of linear regression, we need to minimize the sum of squared errors. As the response and explanatory variables are switched (assume Hwt is always x axis), one model minimizes the vertical distance from points to a line and the other minimize the horizontal distance from points to a line. Obviously, the parameters will be different.

```
library(MASS)
Hwt = cats$Hwt
Bwt = cats$Bwt
# fit models
fit1 = lm(Hwt~Bwt)
fit2 =  lm(Bwt~Hwt)
#get coeficients
beta0 = summary(fit1)$coefficients[1,1]
beta1 = summary(fit1)$coefficients[2,1]
beta0_q = summary(fit2)$coefficient[1,1]
beta1_q = summary(fit2)$coefficients[2,1]
#calculate according to (a)
b0_q = -beta0/beta1
b1_q = 1/beta1
```

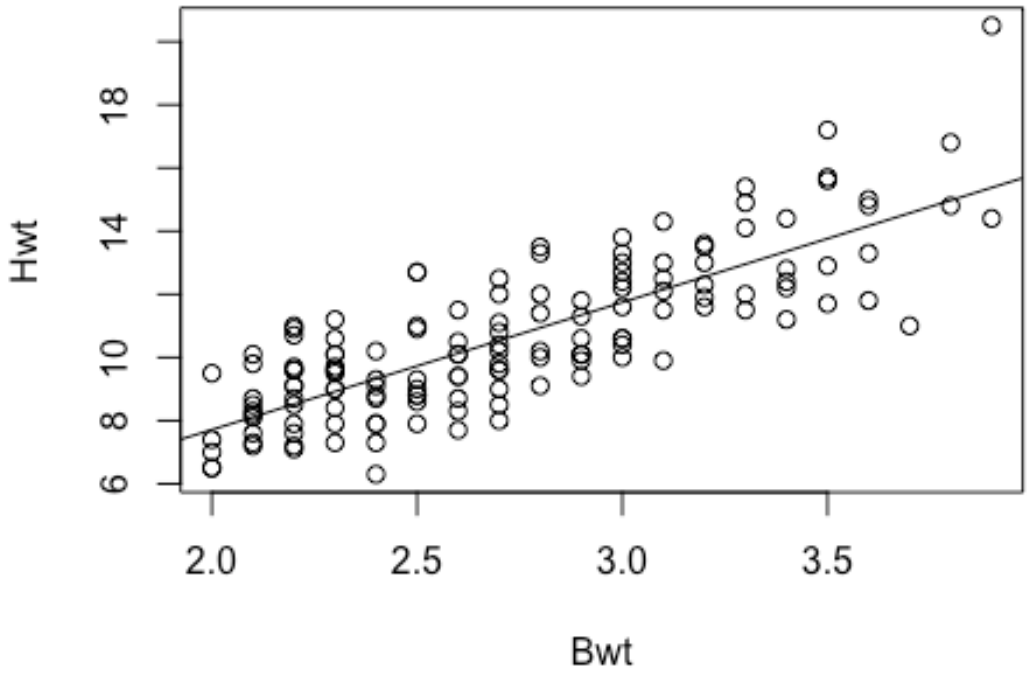
```r
#scatter plot
plot(Bwt, Hwt, xlab = "Bwt", ylab = "Hwt")
abline(fit1)
#calculate intercept and slope according to model2
i = -beta0_q/beta1_q
s = 1/beta1_q
#fit a line of model2
abline(a = i, b = s, col = "blue")
```



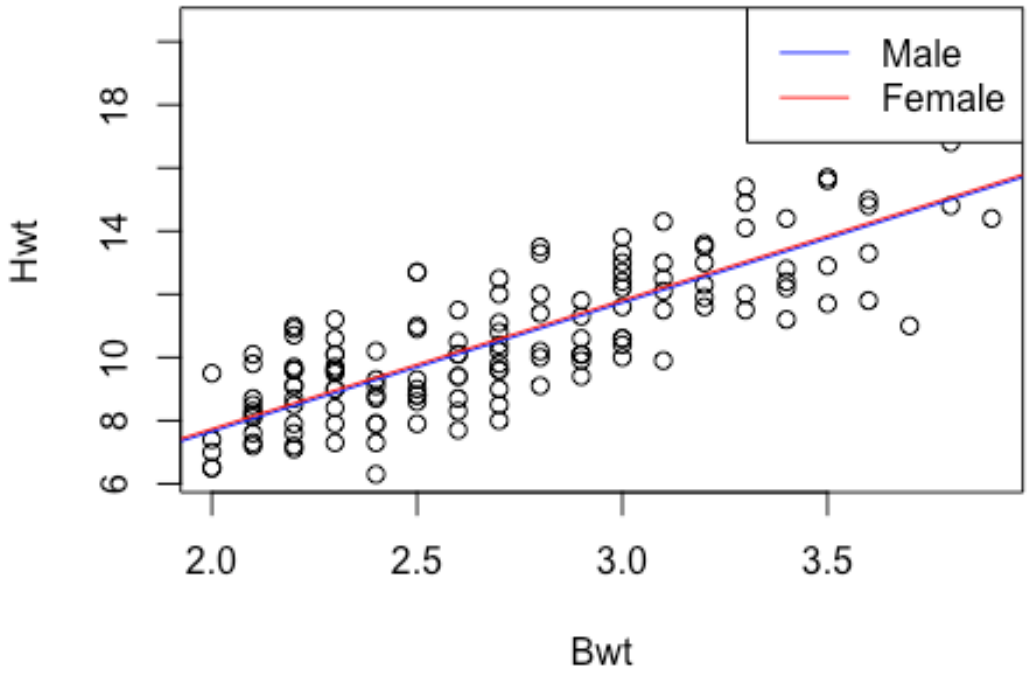
(c)

There is no statistical evidence that model 2 or model 3 improves model 1 at 0.95 significant level. The 1 - p is 0.7875 when comparing model 2 with model 1 and 0.1337 when comparing model 3 with model 1.

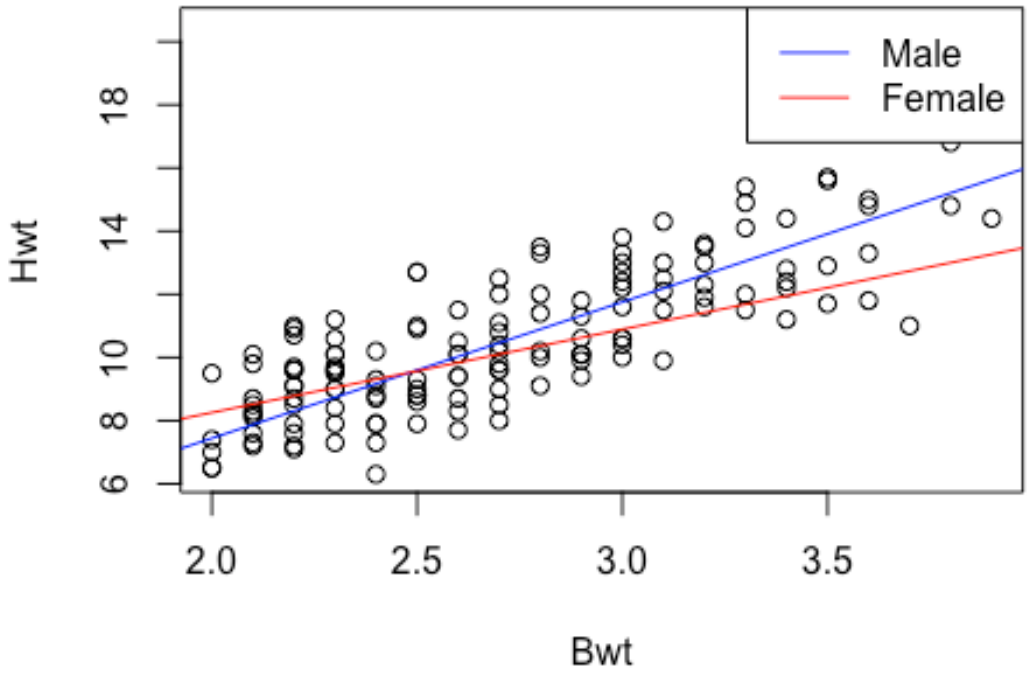
```r
#get varible sex
Sex = cats$Sex
#fit models
m1 = lm(Hwt~Bwt)
m2 = lm(Hwt~Bwt+Sex)
m3 = lm(Hwt~Bwt*Sex)
#get efficient of m2
m2_beta0 = summary(m2)$coefficients[1,1]
m2_beta1 = summary(m2)$coefficients[2,1]
m2_beta2 = summary(m2)$coefficients[3,1]
#get efficient of m3
m3_beta0 = summary(m3)$coefficients[1,1]
m3_beta1 = summary(m3)$coefficients[2,1]
m3_beta2 = summary(m3)$coefficients[3,1]
m3_beta3 = summary(m3)$coefficients[4,1]

#plot for model1
plot(Bwt, Hwt, xlab = "Bwt", ylab = "Hwt")
abline(m1)
```




```r
#plot for model2
plot(Bwt, Hwt, xlab = "Bwt", ylab = "Hwt")
##for male
abline(a = m2_beta0+m2_beta2, b = m2_beta1, col = "blue")
##for female
abline(a = m2_beta0, b = m2_beta1, col = "red")
#add legend
legend("topright", legend=c("Male", "Female"),
       col=c("blue", "red"), lty=c(1,1))
```




```r
#plot for model3
plot(Bwt, Hwt, xlab = "Bwt", ylab = "Hwt")
##for male
abline(a = m3_beta0+m3_beta2, b = m3_beta1+m3_beta3, col = "blue")
##for female
abline(a = m3_beta0, b = m3_beta1, col = "red")
#add legend
legend("topright", legend=c("Male", "Female"),
       col=c("blue", "red"), lty=c(1,1))
```




```r
#do F-test to compare model 2 to model 1
#get rss
m1_rss = sum(summary(m1)$residuals^2)
m2_rss = sum(summary(m2)$residuals^2)
m3_rss = sum(summary(m3)$residuals^2)

#get degree of freedom n - p - 1
df1 = summary(m1)$df[2]
df2 = summary(m2)$df[2]
df3 = summary(m3)$df[2]

#calculate f statistics
# model 1 and model 2
```

```r
F21 = (m1_rss-m2_rss)/(m2_rss/df2)
1 - pf(F21,1,df2)
```

```
## [1] 0.7875448
```

```r
#anova(m1, m2)
# model 1 and model 3
F31 = ((m1_rss-m3_rss)/2)/(m3_rss/df3)
1 - pf(F31,2,df3)
```

```
## [1] 0.1337349
```

```r
#anova(m1, m3)
```

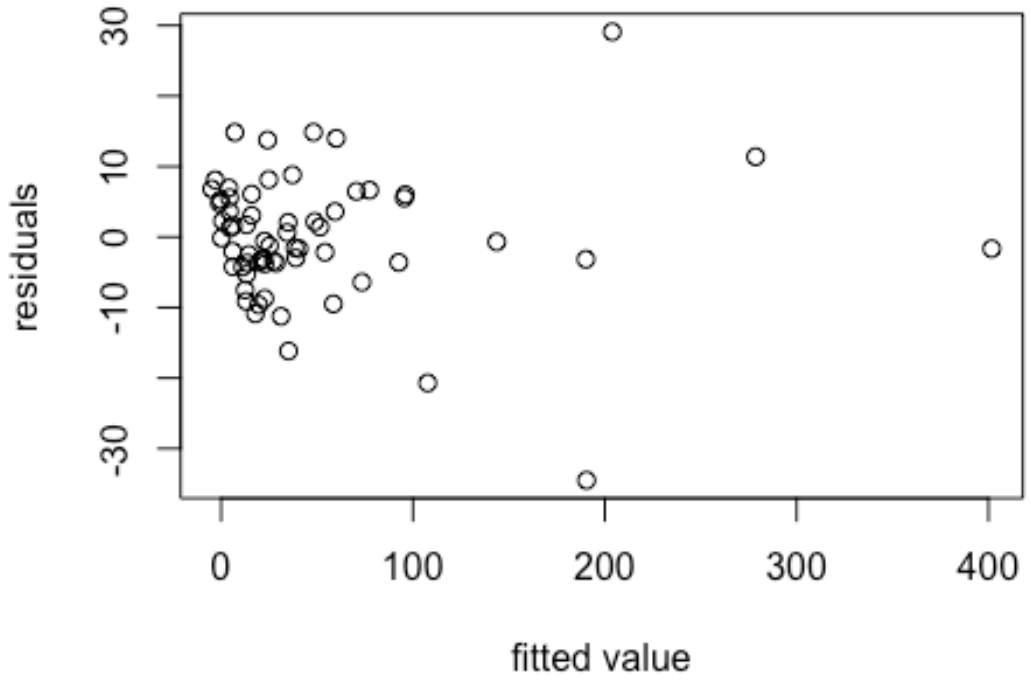
Q3

(a)

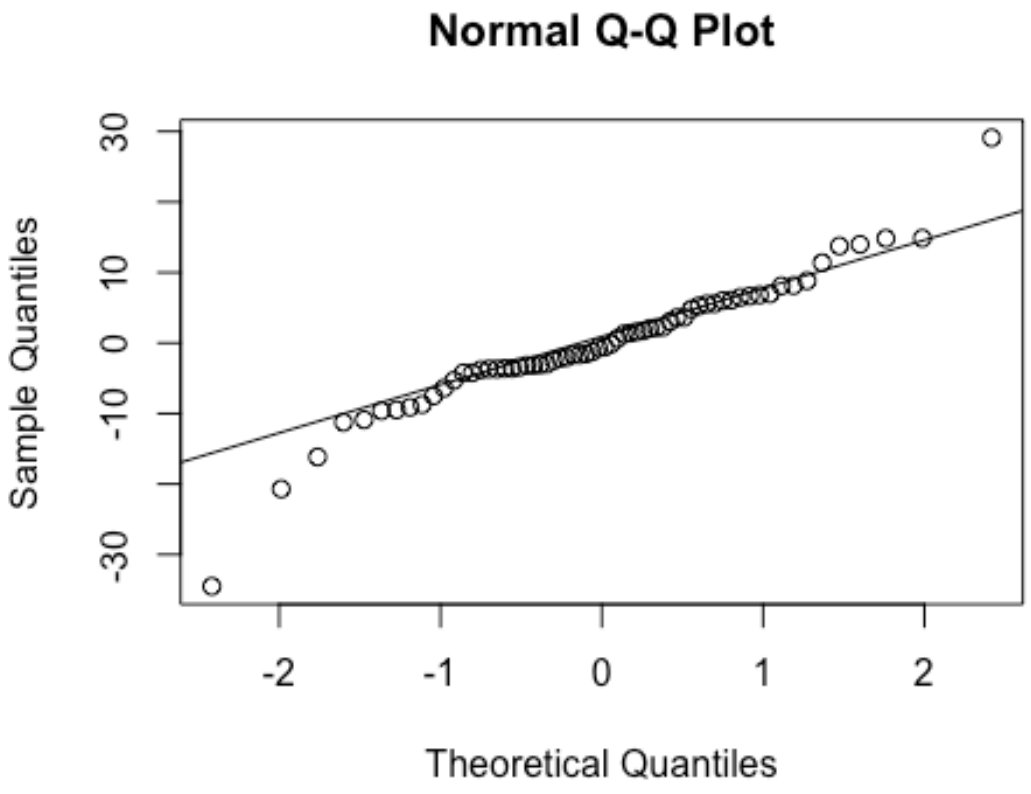
From residual plot, it is not randomly distributed. As Claims variable increases, the residual increases.

From normal qq plot, the relationship between the sample percentiles and theoretical percentiles is not linear. The condition that the error terms are normally distributed is not met.

```r
#get variables
Claims = Insurance$Claims
District = Insurance$District
Group = Insurance$Group
Age = Insurance$Age
Holders = Insurance$Holders
#fit a model
fit_q3 = lm(Claims ~ District + Group + Age + Holders)
#residual plot
plot(fit_q3$fitted.values, fit_q3$residuals, xlab = "fitted value",
ylab = "residuals")
```




```r
#normal quantile plot
qqnorm(fit_q3$residuals)
qqline(fit_q3$residuals)
```

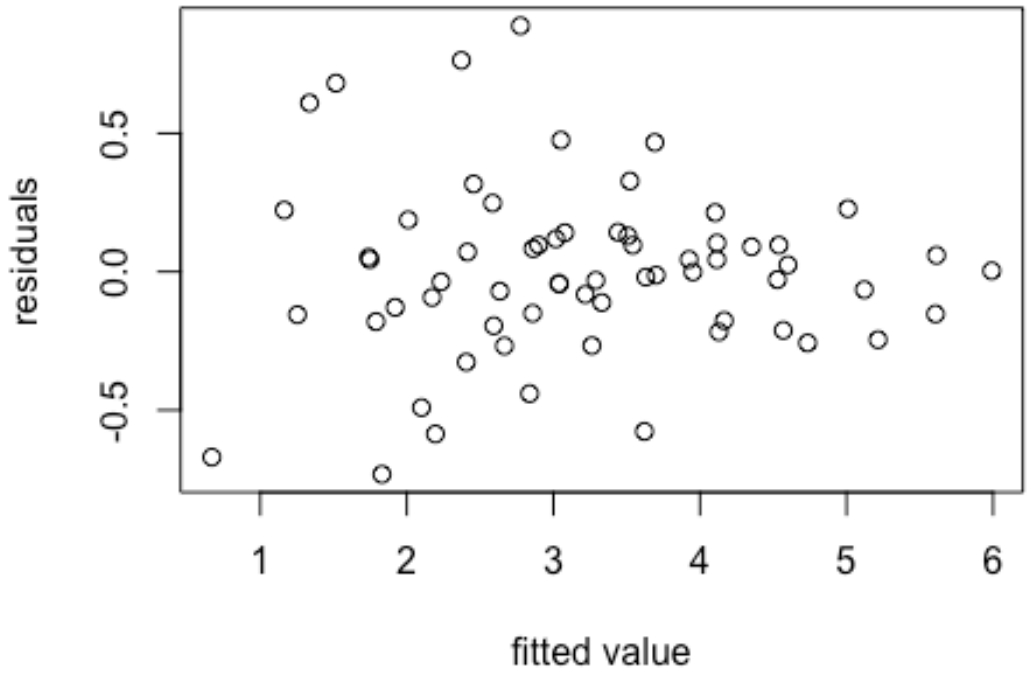


(b)

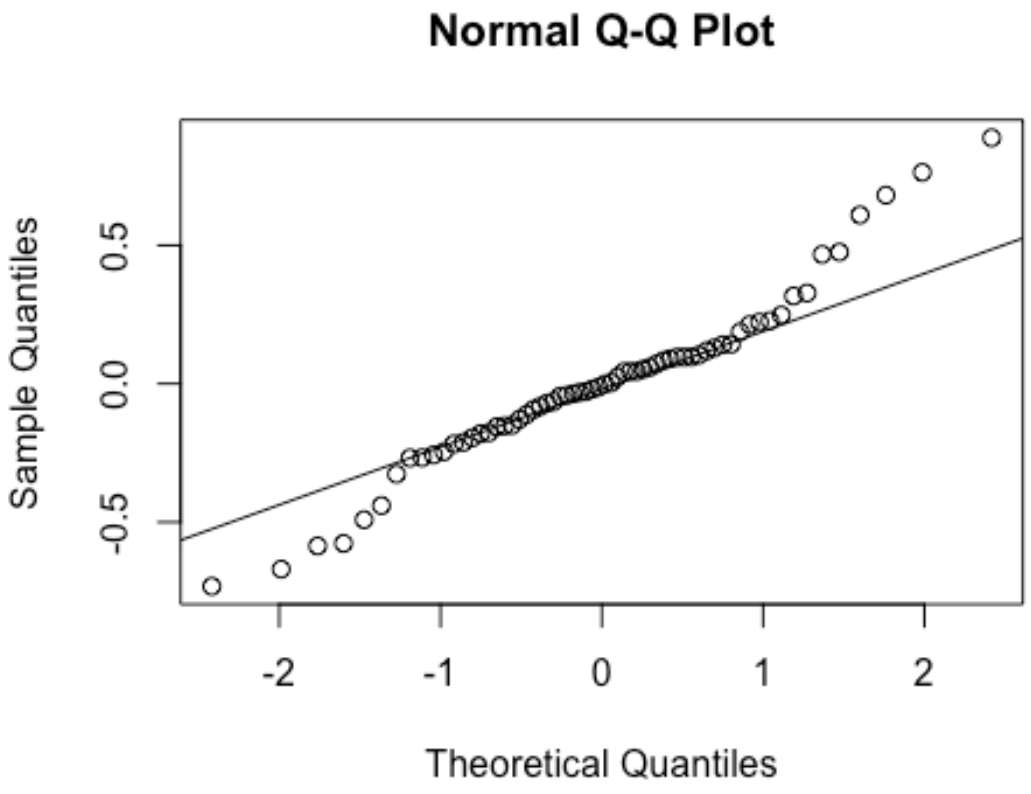
Because Claims varible has many zeros, and Log(0) is not defined, we need to add a constant a in order to avoid this situation.

When a = 10, it better normalized the residuals.

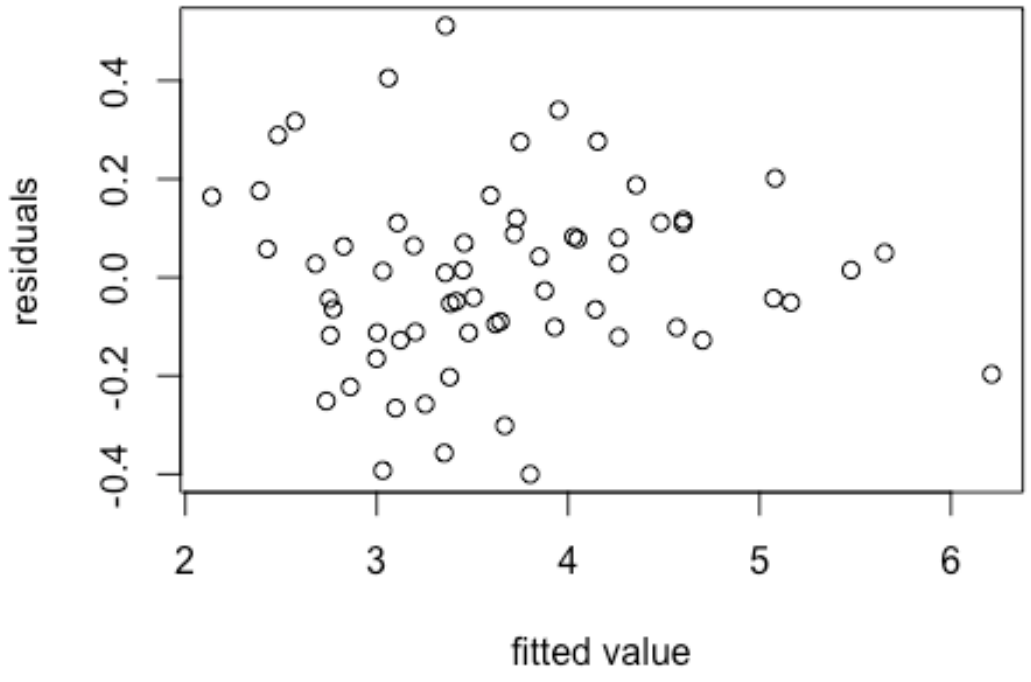
```
#a = 1
f_q31 = lm(log(Claims + 1) ~ District + Group + Age + Holders)
#residual plot
plot(f_q31$fitted.values, f_q31$residuals, xlab = "fitted value", ylab
= "residuals")
```



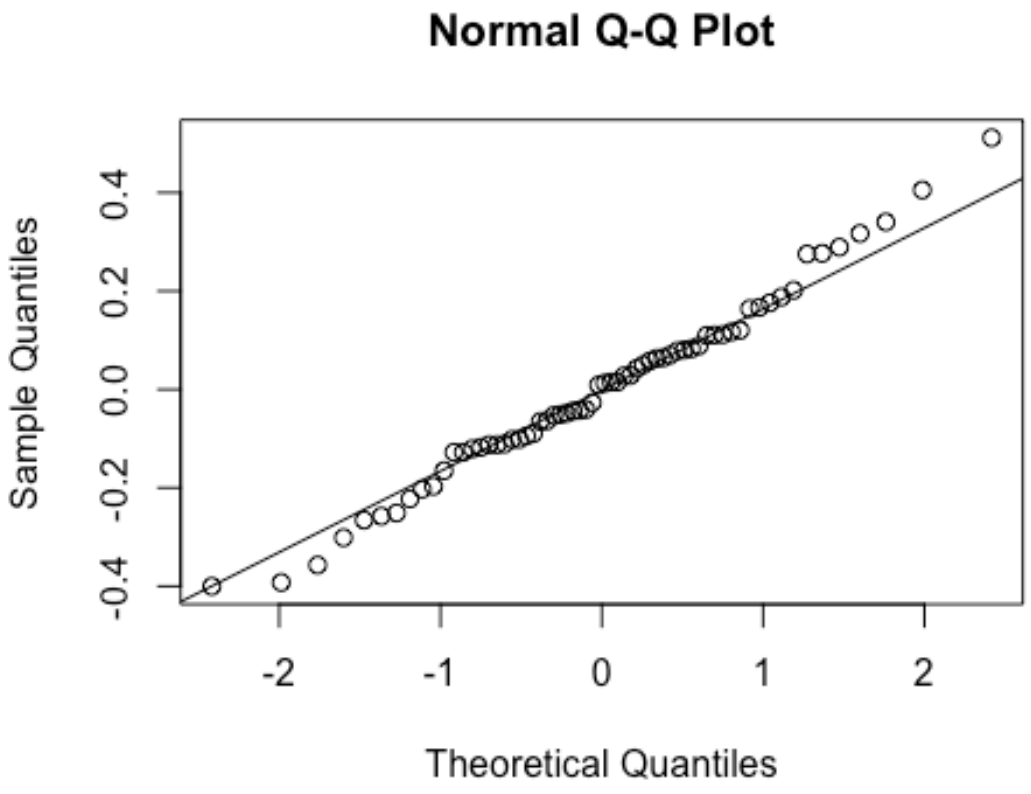

```r
#normal quantile plot
qqnorm(f_q31$residuals)
qqline(f_q31$residuals)
```




```r
#a = 10
f_q32 = lm(log(Claims + 10) ~ District + Group + Age + Holders)
#residual plot
plot(f_q32$fitted.values, f_q32$residuals, xlab = "fitted value", ylab
= "residuals")
```




```r
#normal quantile plot
qqnorm(f_q32$residuals)
qqline(f_q32$residuals)
```



(c)

There are $C_4^4 + C_4^3 + C_4^2 + C_4^1 + C_4^0 = 16$ models.

(d)

```
full = lm(log(Claims+10) ~ District + Group + Age + Holders)
full.formula = formula(terms(full))

upModel_d = function(m1){
  return(update(m1, ~ . -District))
}

upModel_g = function(m1){
  return(update(m1, ~ . -Group))
```

```
}

upModel_a = function(m1){
  return(update(m1, ~ . -Age))
}

upModel_h = function(m1){
  return(update(m1, ~ . -Holders))
}

applyUp = function(models){
  l = list()
  upmodels = list(upModel_a, upModel_g, upModel_d, upModel_h)
  for(f in upmodels){
    l = c(l, lapply(models, f))
  }
  return(l)
}
l4 = list(full.formula)
l3 = applyUp(l4)
l2 = applyUp(l3)
l1 = applyUp(l2)
l0 = applyUp(l1)
model_list = unique(c(l4, l3, l2, l1, l0))
model_list
```

```
## [[1]]
## log(Claims + 10) ~ District + Group + Age + Holders
##
## [[2]]
## log(Claims + 10) ~ District + Group + Holders
##
## [[3]]
## log(Claims + 10) ~ District + Age + Holders
##
## [[4]]
## log(Claims + 10) ~ Group + Age + Holders
##
```

```
## [[5]]
## log(Claims + 10) ~ District + Group + Age
##
## [[6]]
## log(Claims + 10) ~ District + Holders
##
## [[7]]
## log(Claims + 10) ~ Group + Holders
##
## [[8]]
## log(Claims + 10) ~ District + Group
##
## [[9]]
## log(Claims + 10) ~ Age + Holders
##
## [[10]]
## log(Claims + 10) ~ District + Age
##
## [[11]]
## log(Claims + 10) ~ Group + Age
##
## [[12]]
## log(Claims + 10) ~ Holders
##
## [[13]]
## log(Claims + 10) ~ District
##
## [[14]]
## log(Claims + 10) ~ Group
##
## [[15]]
## log(Claims + 10) ~ Age
##
## [[16]]
## log(Claims + 10) ~ 1
```

```
#get adj r square
r_sq = list()
```

```
for(m in model_list){
  r_sq = c(r_sq, summary(lm(m))$adj.r.squared)
}
#get the largest index and the corresponding model
model_list[which.max(r_sq)]

## [[1]]
## log(Claims + 10) ~ District + Group + Age + Holders
```

Q4

(a)

According to model 3 in Q2, we get

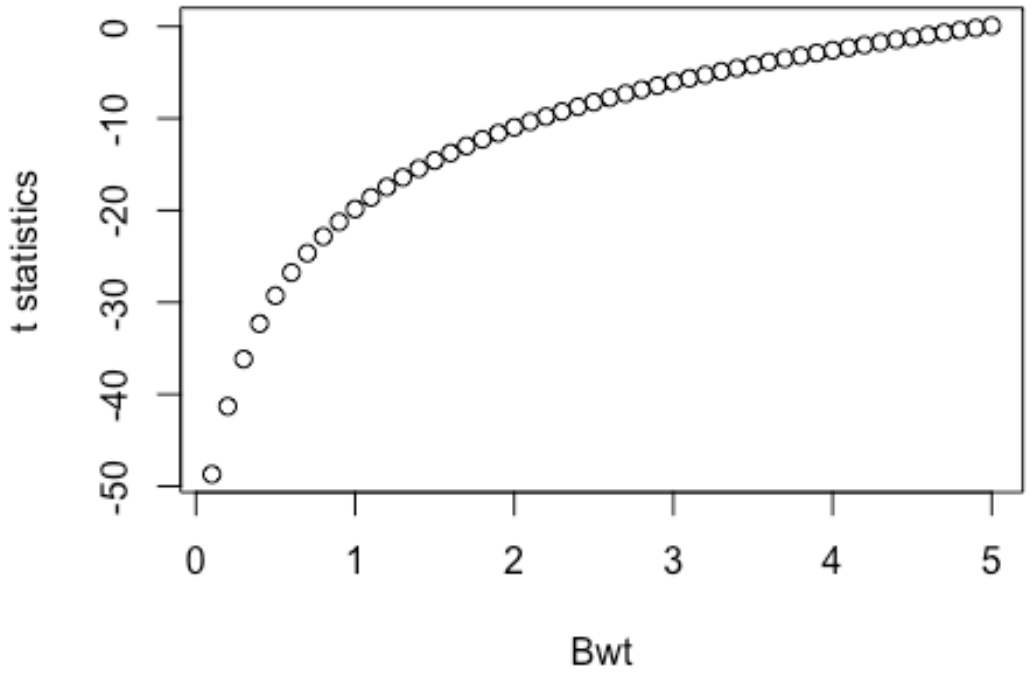
$$\mu_x^m = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x$$

$$\mu_x^f = \beta_0 + \beta_1 x$$

To test $H_0: \mu_x^m = \mu_x^f$, it is the same to do t-test on $\mu_x^m - \mu_x^f = \beta_2 + \beta_3 x$ From Q2, we can get estimated β_2 and β_3. Then the only variable is x and we can get a sequence of t as x ranges from 0 to 5.

From the plot, it appears that as Bwt is increasing, t statistics is increasing.

```
beta2_m3 = m3$coefficients[3]
beta3_m3 = m3$coefficients[4]
diff_mu = list()
t = list()
for(i in seq(0, 5, 0.1)){
  diff_mu = c(diff_mu, beta2_m3+beta3_m3*i)
  if(length(diff_mu)>1){
    ttest = t.test(as.numeric(diff_mu), mu=0, alternative =
"two.sided")
    t = c(t, ttest$statistic)
  }
}
plot(seq(0.1, 5, 0.1), t, xlab = "Bwt", ylab = "t statistics")
```



(b)

The P-value of F-test to compare model 1 and 3 is 0.8662651. Even at $\alpha = 0.1$, this value is much larger then α so that we cannot reject the null hypothesis and there might not be a significant improvement.

When $x = 3.5$, the P-value is 0.0001820016 which is much more smaller than 0.05. This means that we can reject the null hypothesis at 0.05 significant level.

This does not imply that model 1 is incorrect. In order to show that the mean values of heart weights are different, we cannot only depend on the test on a single x value. Instead, we need to take all the x value as whole and apply a t-test on it. We can see that as x value increases, the t value will increase. Therefore, there might be a point at which we cannot pass the hypothesis test.

```r
diff_mu = list()
for(i in seq(0, 3.5, 0.1)){
  diff_mu = c(diff_mu, beta2_m3+beta3_m3*i)
}
ttest = t.test(as.numeric(diff_mu), mu=0, alternative = "two.sided")
ttest$p.value
```

```
## [1] 0.0001820016
```

(c)

The issue in (b) is that when x value increases to a certain extend, we do not know if we can pass the hypothesis test with a corresponding large t statistics. We cannot draw the conclusion based on the test result at a single point or a part of data.

Using the hypothesis test method involving chi square will help to deal with the situation where t statistics increases as x value increases. When the t statistics increases and exceeds $(\chi^2_{p:\alpha})^{1/2}$, we can still reject the null hypothesis at α significant level.