Assignment #1: R code

Tristan De Alwis

2/27/2020

Loading Neccessary Libraries

```
suppressPackageStartupMessages(library(MASS))
```

Q2:

```
n = 144
```

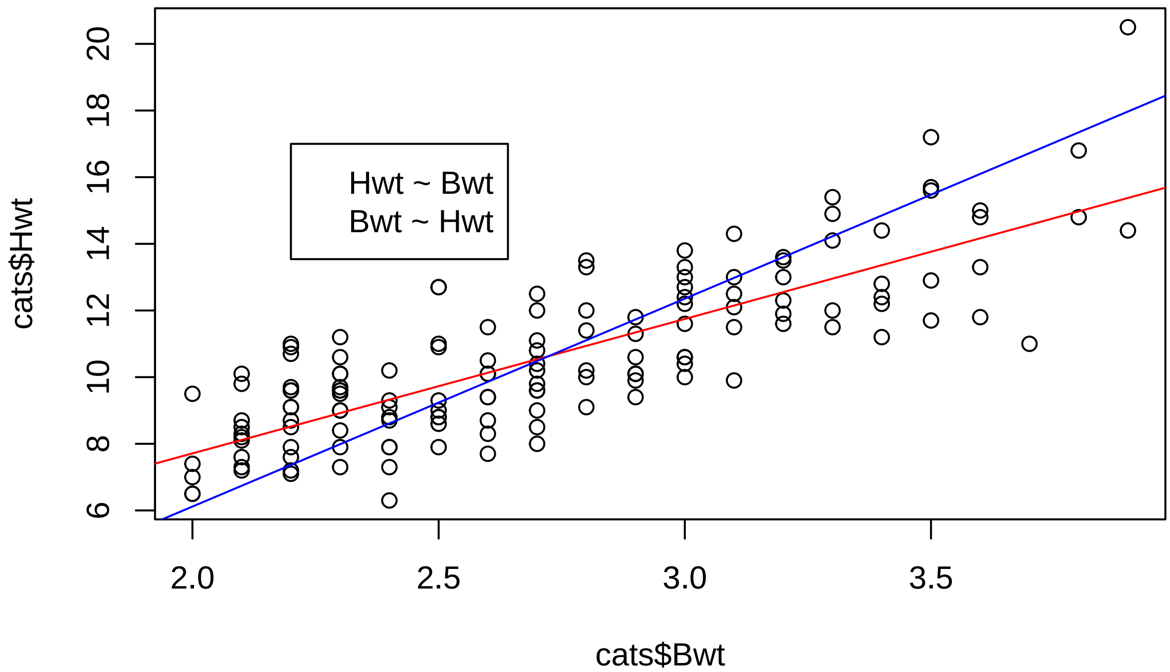
(b)

```
lm1 <- lm(cats$Hwt ~ cats$Bwt)
b0 <- summary(lm1)$coefficients[1, 1]
b1 <- summary(lm1)$coefficients[2, 1]
```

```
lm2 <- lm(cats$Bwt ~ cats$Hwt)
b0p <- summary(lm2)$coefficients[1, 1]
b1p <- summary(lm2)$coefficients[2, 1]
```

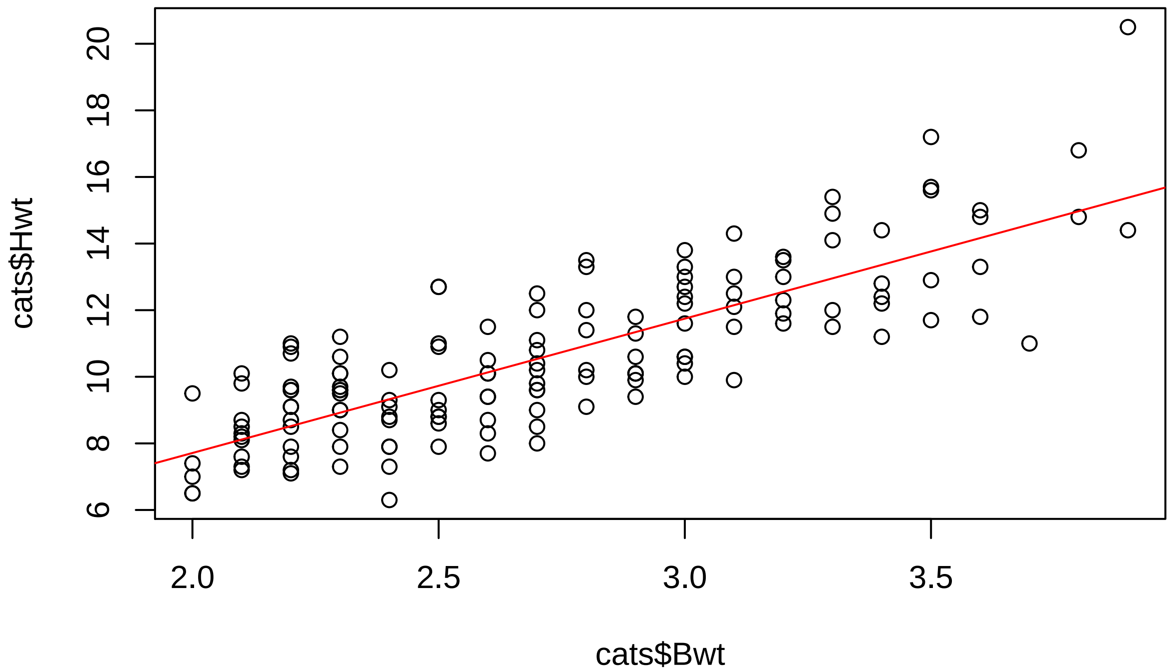
Because the coefficients are not equal to the inverse of the previous, they do not conform.

```
plot(cats$Bwt, cats$Hwt)
abline(b0, b1, col = "red")
abline(-b0p/b1p, 1/b1p, col = "blue")
legend(2.2, 17, legend = c("Hwt ~ Bwt", "Bwt ~ Hwt"), col = c("red", "blue"))
```



(c)

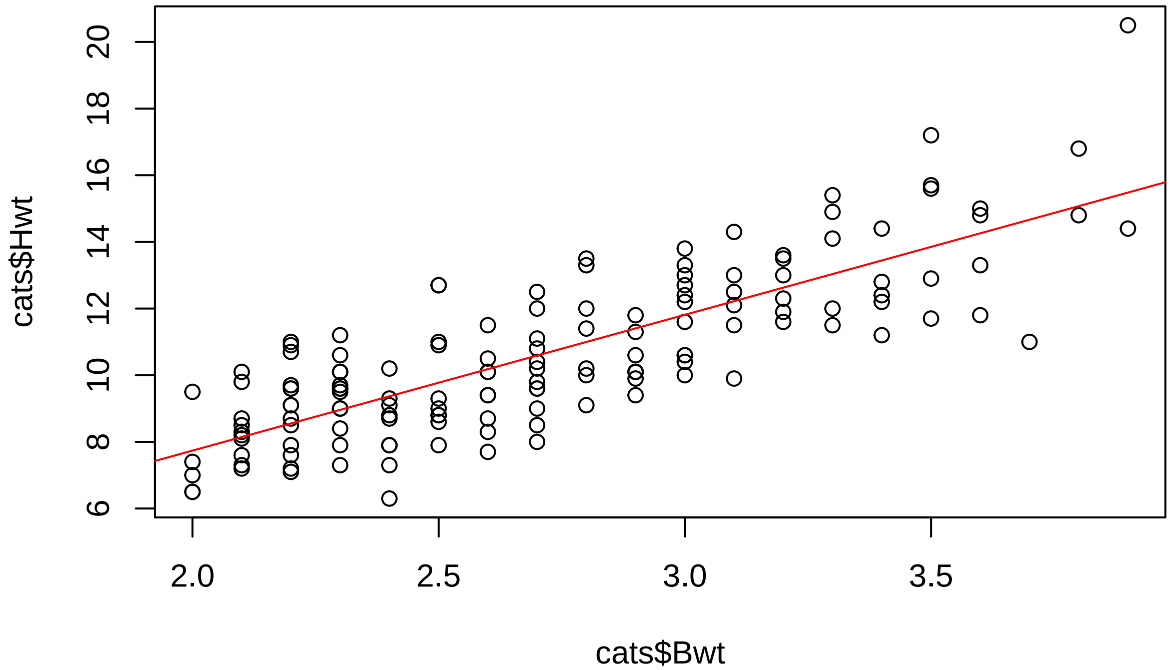
```
M1 <- lm(Hwt ~ Bwt, cats)
plot(cats$Bwt, cats$Hwt)
abline(M1, col = "red")
```




```
M2 <- lm(Hwt ~ Bwt + Sex, cats)
plot(cats$Bwt, cats$Hwt)
abline(M2, col = "red")
```

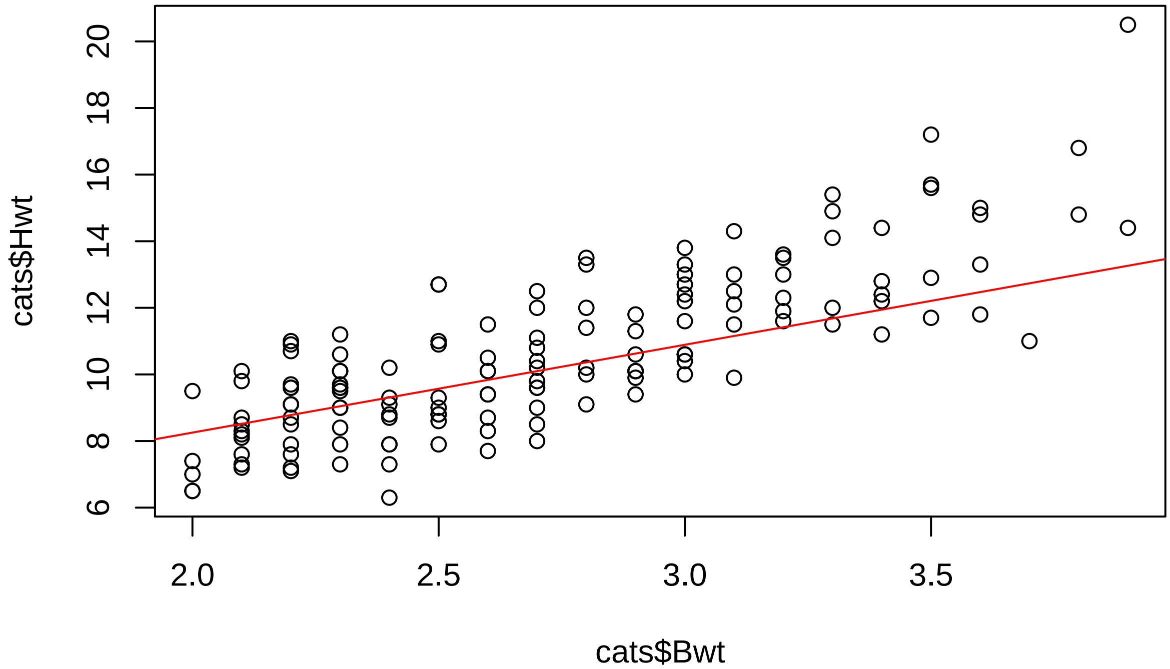
```
## Warning in abline(M2, col = "red"): only using the first two of 3
```

```
## regression coefficients
```




```r
M3 <- lm(Hwt ~ Bwt * Sex, cats)
plot(cats$Bwt, cats$Hwt)
abline(M3, col = "red")
```

```
## Warning in abline(M3, col = "red"): only using the first two of 4
## regression coefficients
```




```r
anova(M1, M2)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: Hwt ~ Bwt
## Model 2: Hwt ~ Bwt + Sex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    142 299.53
## 2    141 299.38  1    0.1548 0.0729 0.7875
```

With a p-value above 0.05 we fail to reject the null hypothesis and cannot conclude that Model 2 improves Model 1.

Q3:

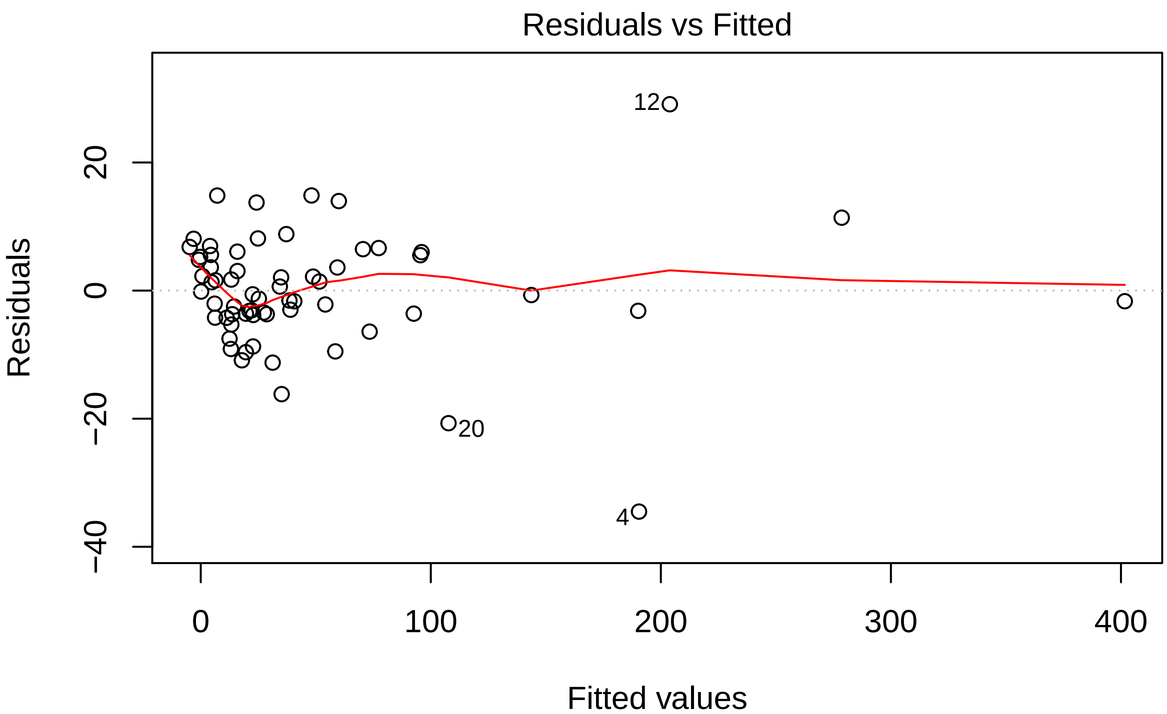
```
# Insurance
n = 64
```

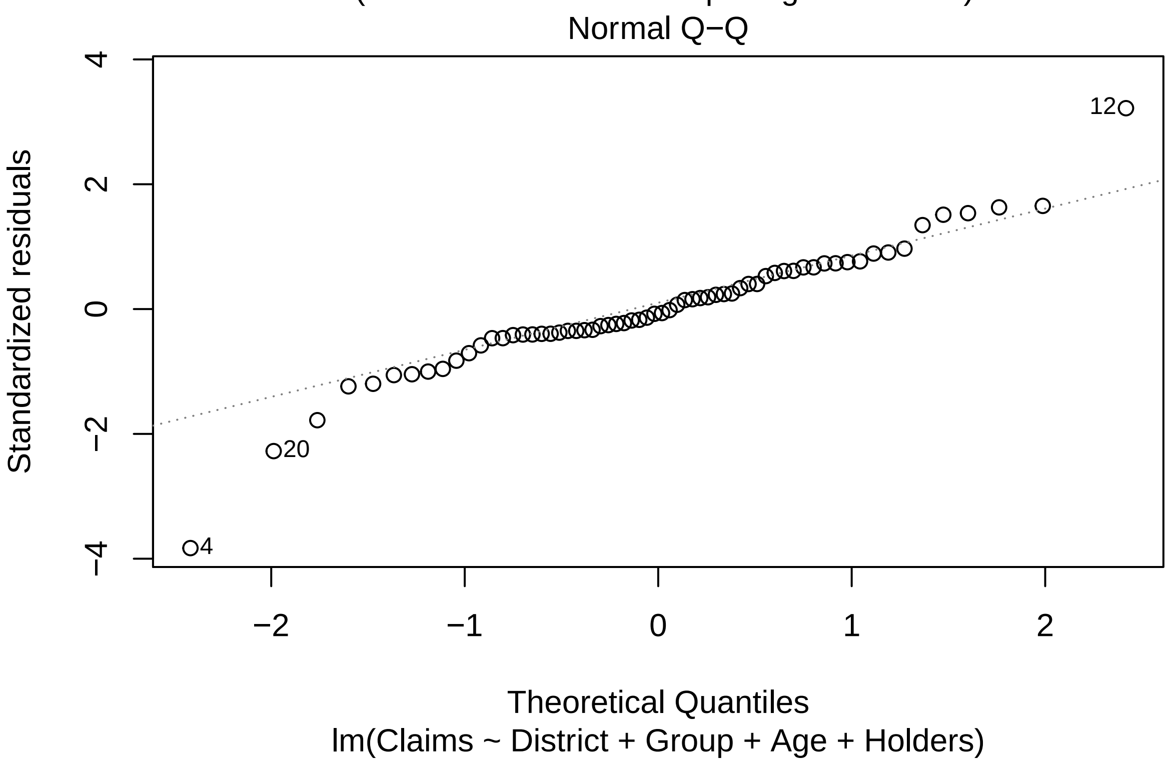
(a)

```
Claims.lm <- lm(Claims ~ District + Group + Age + Holders, Insurance)
Claims.lm
```

```
##
## Call:
## lm(formula = Claims ~ District + Group + Age + Holders, data = Insurance)
##
## Coefficients:
## (Intercept)    District2    District3    District4      Group.L
##     18.3162      -5.5286     -10.6234     -10.7989       5.0879
##     Group.Q      Group.C        Age.L        Age.Q        Age.C
##    -12.8630      -0.2430       8.6093       4.0472       3.8121
##     Holders
##      0.1032
```

```
plot(Claims.lm, which = c(1, 2))
```



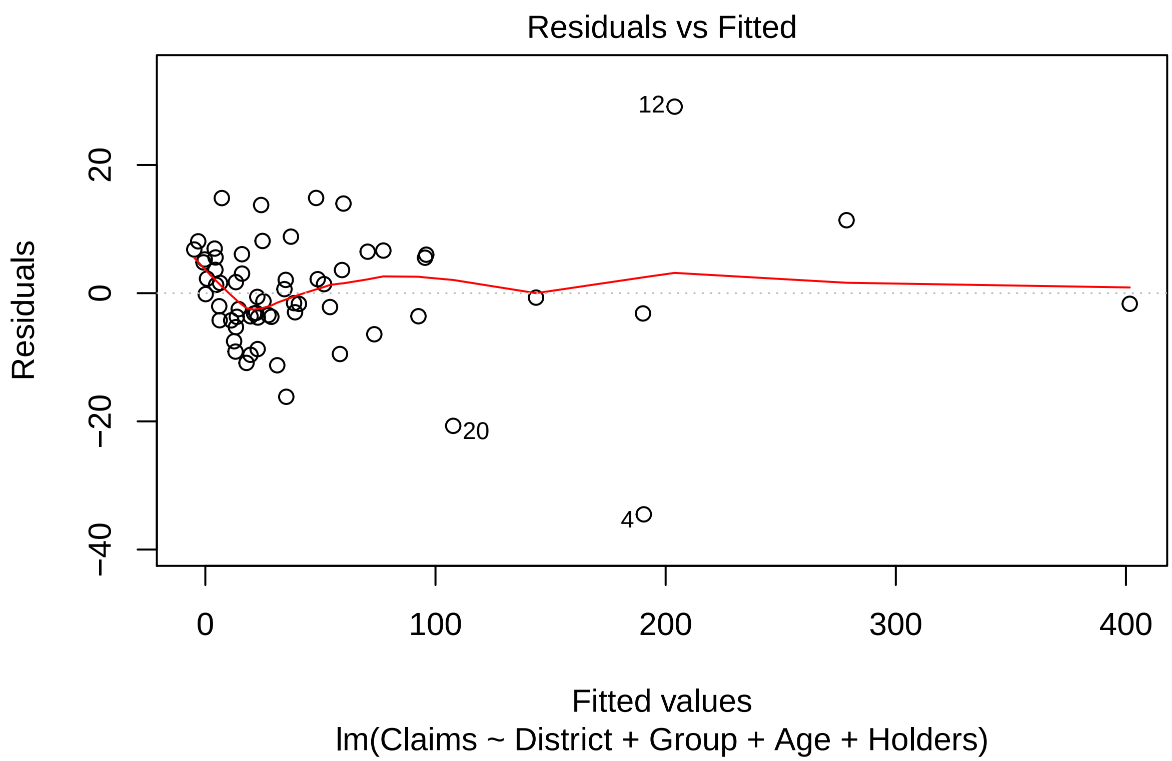


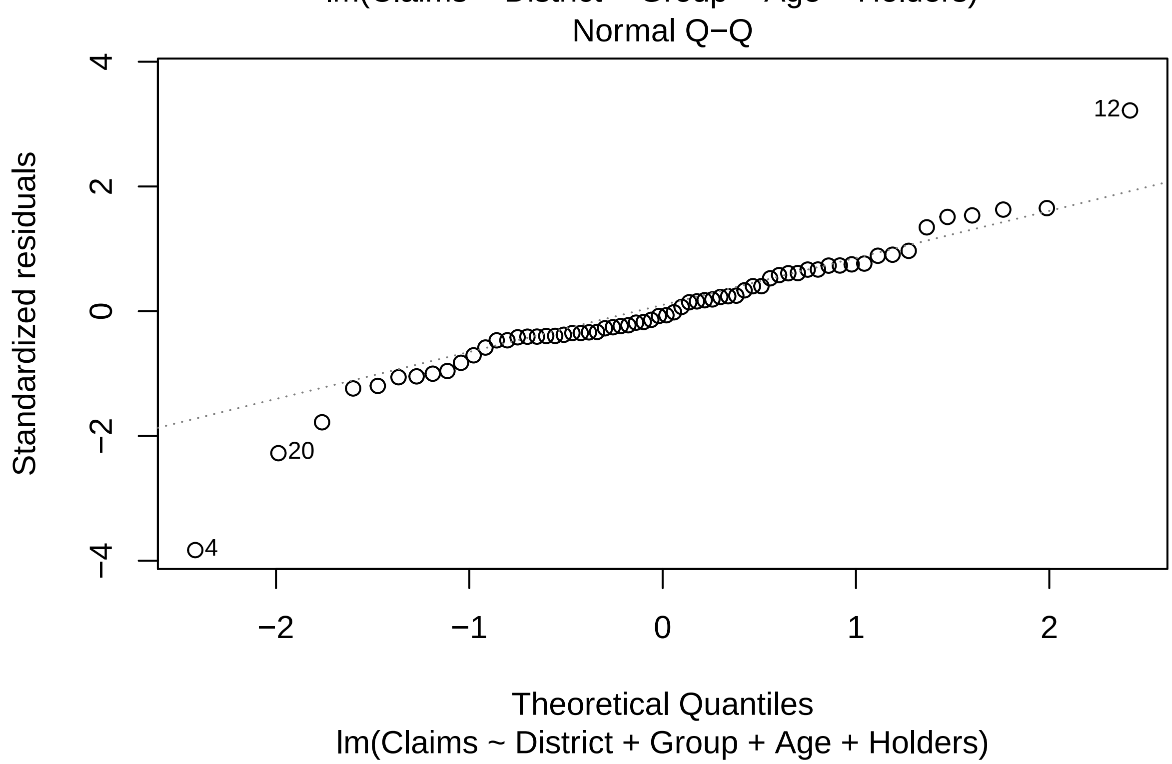
Since the data is closely fit to the line, the assumptions seem reasonable.

(b)

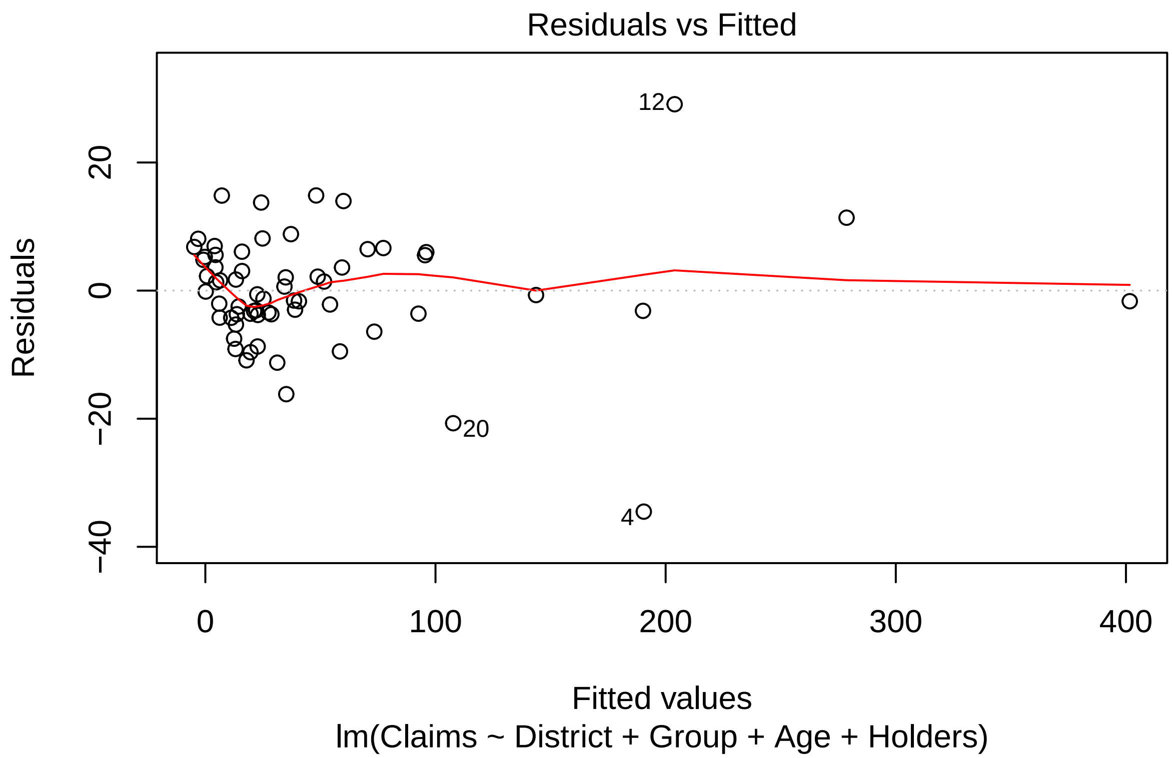
```
Claims <- log(Insurance$Claims)
Claims.lm <- lm(Claims ~ District + Group + Age + Holders, Insurance)
```

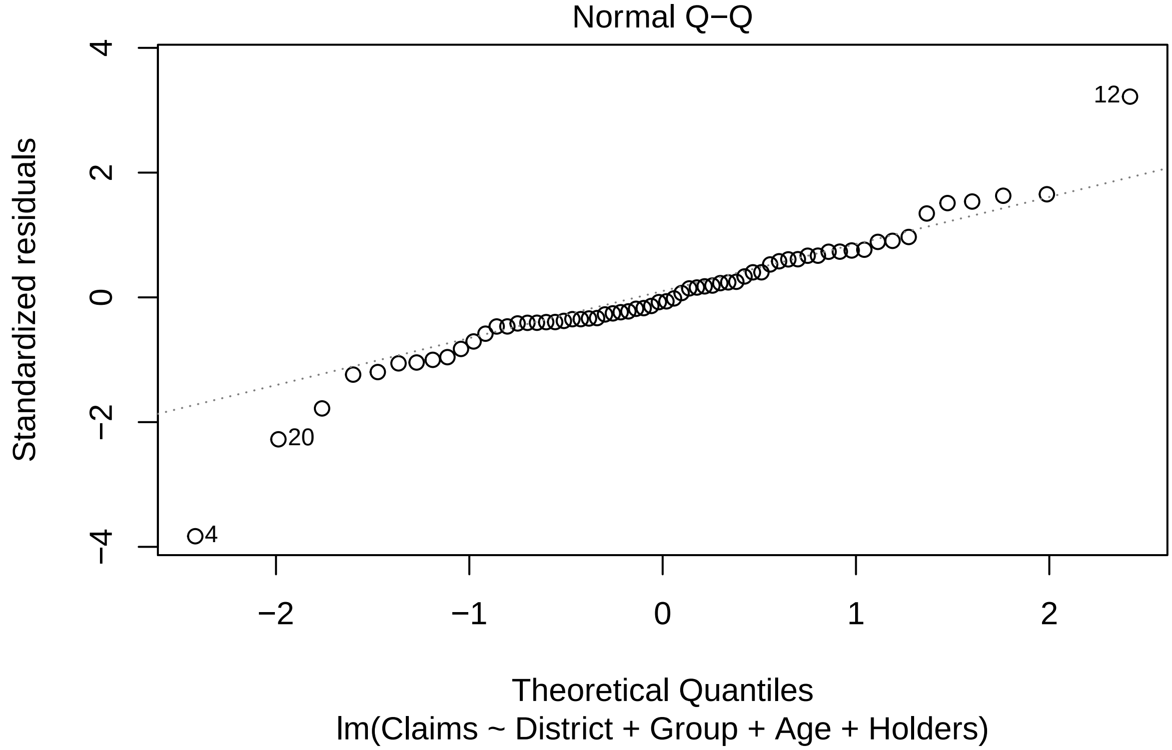
```r
plot(Claims.lm, which = c(1, 2))
```



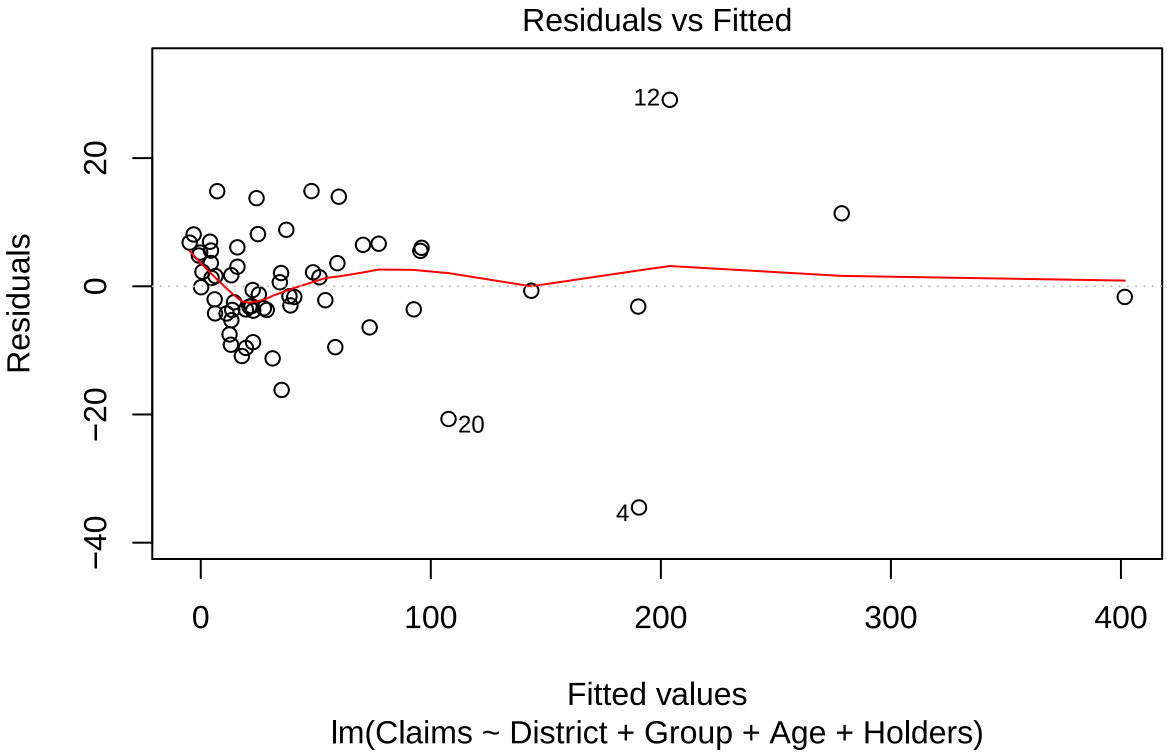


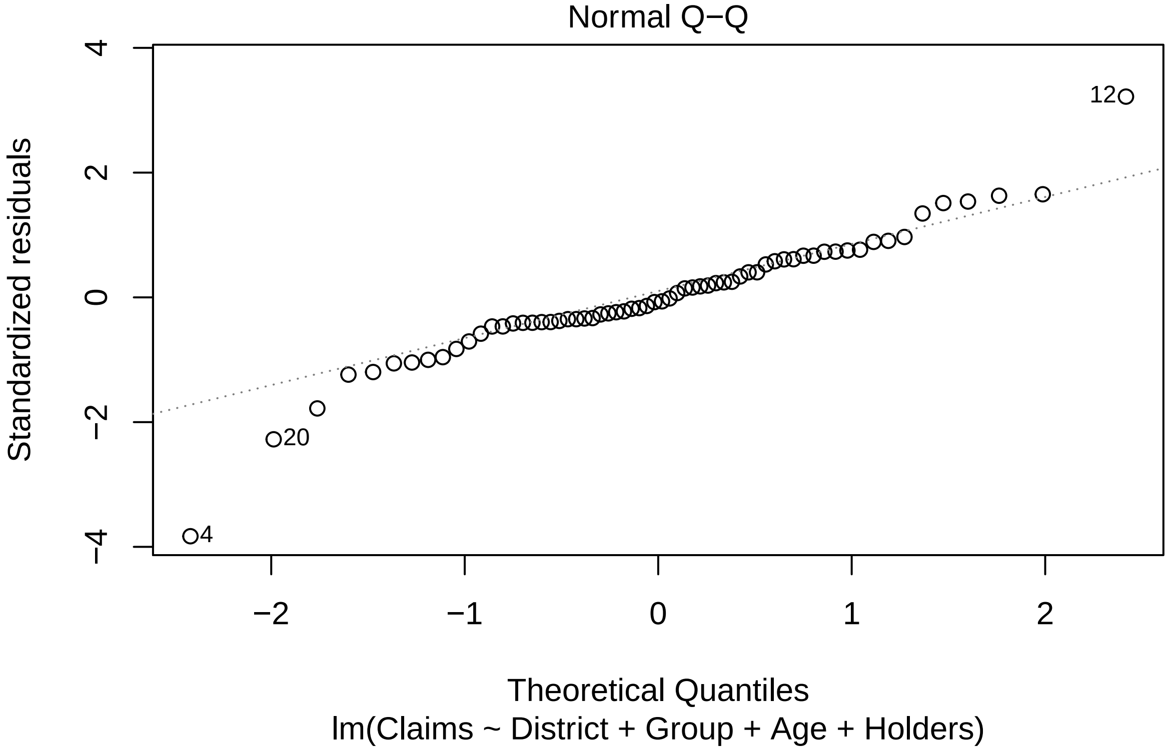

```r
Claims <- log(Insurance$Claims + 1)
Claims.lm <- lm(Claims ~ District + Group + Age + Holders, Insurance)
plot(Claims.lm, which = c(1, 2))
```






```
Claims <- log(Insurance$Claims + 10)
Claims.lm <- lm(Claims ~ District + Group + Age + Holders, Insurance)
plot(Claims.lm, which = c(1, 2))
```





We can't set a = 0 because log(0) does not exist nor would claims=0

(c)

$1 + 4 + 6 + 4 + 1 = 16$

(d)

```
themodel <- lm(log(Claims + 10) ~ ., data = Insurance)
full.formula <- formula(terms(themodel))

combs <- c(full.formula, update(full.formula, ~. - Group), update(full.formula,
    ~. - Holders), update(full.formula, ~. - District), update(full.formula,
    ~. - Age), update(full.formula, ~. - Group - Age), update(full.formula,
    ~. - Group - Holders), update(full.formula, ~. - Group - District), update(full.formula,
    ~. - Age - Holders), update(full.formula, ~. - Age - District), update(full.formula,
    ~. - Holders - District), update(full.formula, ~. - Group - Age - Holders),
    update(full.formula, ~. - Group - Age - District), update(full.formula,
        ~. - Group - Holders - District), update(full.formula, ~. - Age - Holders -
        District), update(full.formula, ~. - Group - Age - Holders - District))

Rsq <- list()
for (i in combs) {
    x <- summary(lm(i, data = Insurance))$adj.r.squared
    Rsq <- c(Rsq, x)

}
max_rsq <- max(sapply(Rsq, max))
max_rsq
```

```
## [1] 0.944299
```

```
index <- match(max_rsq, Rsq)
print(combs[index])
```

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

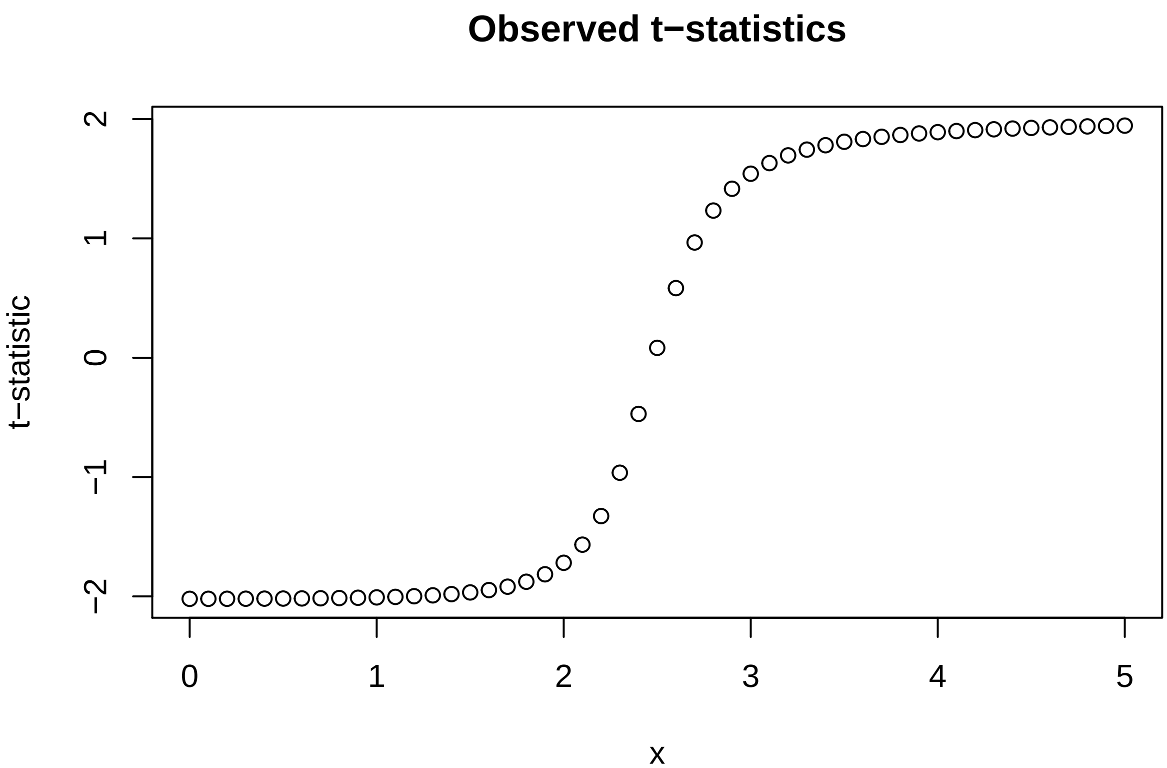
The model with all the variables has the largest R-squared adjusted value of 0.944299

Q4:

(a)

```
xMat <- matrix(nrow = nrow(cats), ncol = 4)
xMat[, 1] <- rep(1, nrow(xMat))
xMat[, 2] <- cats$Bwt
xMat[, 3] <- sapply(cats$Sex == "M", as.numeric)
xMat[, 4] <- cats$Bwt * xMat[, 3]

a <- function(x) {
    return(matrix(c(0, 0, 1, x), nrow = 4, ncol = 1))
}

sigN <- function(x) {
    return(sqrt(MSE * t(a(x)) %*% solve(t(xMat) %*% xMat) %*% a(x)))
}
MSE <- sum((M3$residuals)^2)/(nrow(cats) - 4)

tStat <- function(x) {
    (coef(M3)[[3]] + coef(M3)[[4]] * x)/sigN(x)
```

```r
}

x <- seq(0, 5, by = 0.1)

v <- c()
counter <- 1
for (i in x) {
    v[counter] <- tStat(i)
    counter <- counter + 1
}

plot(x, v, main = "Observed t-statistics", ylab = "t-statistic")
```



(b)

```r
tStat(3.5)
```

```
##          [,1]
## [1,] 1.809406
```

```r
2 * (1 - pt(tStat(3.5), 97 + 47 - 2))
```

```
##            [,1]
## [1,] 0.07250306
```

It is > 0.05 At alpha $= 0.1$ there is no significant improvement. This implies that sex matters, at least for some values.

(c)

```
boolList <- abs(v) > sqrt(qchisq(0.9, 4))

sum(boolList)/length(v)
```

```
## [1] 0
```

We cannot reject any of the hypotheses as their chi-squared critical value is more extreme than their t-statistic. The third model does not improve first model. This shows that the answer in part (b) is a problem, and that one should not conclude that Model 3 improves Model 1 based on individual values.