Problem 3

Part a:

```
baseball <- read.csv("Baseball-Salary-Data.csv", header = TRUE)
ls(baseball)
baseball$player <- NULL
par(mfrow = c(1, 2))
hist(baseball$salary, main = "Salary")
fit <- lm((salary)~., data = baseball)
head(baseball[, 14:17])
summary(fit)
```

```
Call:
lm(formula = (salary) ~ ., data = baseball)

Residuals:
    Min      1Q  Median      3Q     Max 
-1908.3  -463.0    10.9   340.7  3181.7 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           223.115    332.717   0.671 0.502970    
batting.average      3043.192   2712.536   1.122 0.262746    
on.base.percent     -3528.013   2376.084  -1.485 0.138581    
runs                    7.100      5.643   1.258 0.209259    
hits                   -2.698      3.312  -0.815 0.415788    
doubles                 1.368      8.611   0.159 0.873846    
triples               -17.922     21.647  -0.828 0.408339    
home.runs              19.483     12.583   1.548 0.122506    
rbi                    17.415      5.068   3.436 0.000668 ***
walks                   5.815      4.523   1.285 0.199548    
strike.outs            -9.586      2.151  -4.457 1.15e-05 ***
stolen.bases           13.044      4.714   2.767 0.005988 ** 
errors                 -9.553      7.500  -1.274 0.203693    
free.agent.eligible  1372.886    108.594  12.642  < 2e-16 ***
free.agent           -280.790    137.640  -2.040 0.042168 *  
arbitration.eligible  783.592    118.289   6.624 1.48e-10 ***
arbitration           352.114    241.829   1.456 0.146361    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 694.3 on 320 degrees of freedom
Multiple R-squared:  0.7014,	Adjusted R-squared:  0.6865 
F-statistic: 46.99 on 16 and 320 DF,  p-value: < 2.2e-16
```

Part b:

As we can see by the multiple R-squared value, the amount of the variation in salaries explained by model used is 70.14%

Part c:

The coefficient for hits in the model used is negative, which does not match my intuition about the relationship between hits and salary. I would generally expect the salary of a player who would be expected to deliver more hits and consequently be more valued offensively speaking to be compensated in a direct relationship with the aforementioned increase of value.

Part d:

From the summary output above, we can see that the p-value for the model utility test is below the .05 significance level, thus we reject the null hypothesis that none of the 16 predictors is related to salary. Consequently, we conclude that this model is useful.

Part f:

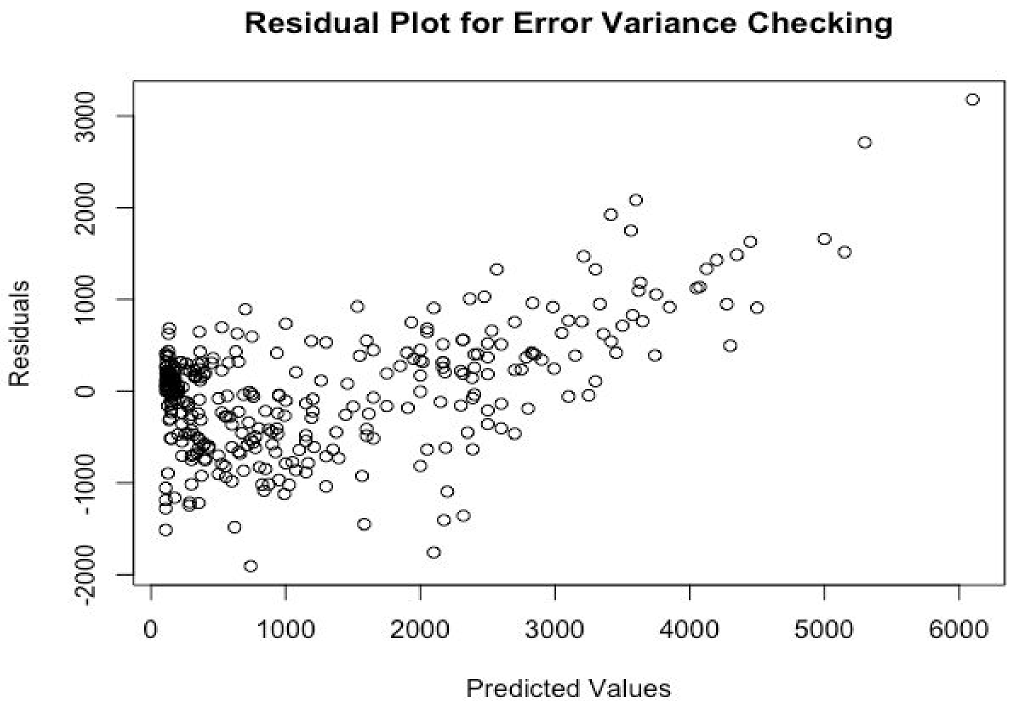
```
fit_part_f <- lm((salary) ~. - batting.average - on.base.percent -
     hits - doubles -
     triples, data = baseball)

summary(fit_part_f)
```

```
Call:
lm(formula = (salary) ~ . - batting.average - on.base.percent -
    hits - doubles - triples, data = baseball)

Residuals:
    Min      1Q  Median      3Q     Max
-1861.7  -467.9    43.3   330.5  3231.6

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -100.767     89.198  -1.130  0.25944
runs                    3.675      3.825   0.961  0.33732
home.runs              25.020      9.778   2.559  0.01096 *
rbi                    16.153      3.818   4.231 3.03e-05 ***
walks                   2.368      2.912   0.813  0.41672
strike.outs            -9.718      1.935  -5.023 8.43e-07 ***
stolen.bases           12.568      4.574   2.748  0.00633 **
errors                 -9.579      7.256  -1.320  0.18776
free.agent.eligible  1357.942    105.133  12.916  < 2e-16 ***
free.agent           -272.826    136.961  -1.992  0.04721 *
arbitration.eligible  776.478    116.676   6.655 1.20e-10 ***
arbitration           343.844    240.852   1.428  0.15436
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 692.7 on 325 degrees of freedom
Multiple R-squared:  0.6981,	Adjusted R-squared:  0.6879
F-statistic: 68.33 on 11 and 325 DF,  p-value: < 2.2e-16
```

From the summary statistics, we can see that the percentage of variation in salary explained by the 11 variables not name in part e is 69.81%

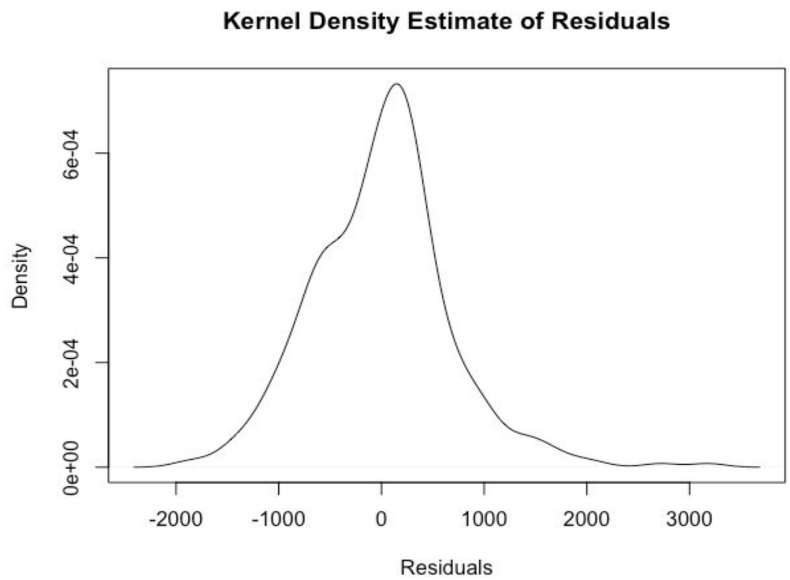
Part g:

```
## i. residuals plot
resids <- fit$residuals
preds <- fit$fitted.values
plot(baseball$salary, resid, xlab = "Predicted Values", ylab =
"Residuals", main = "Residual Plot for Error Variance Checking")
```



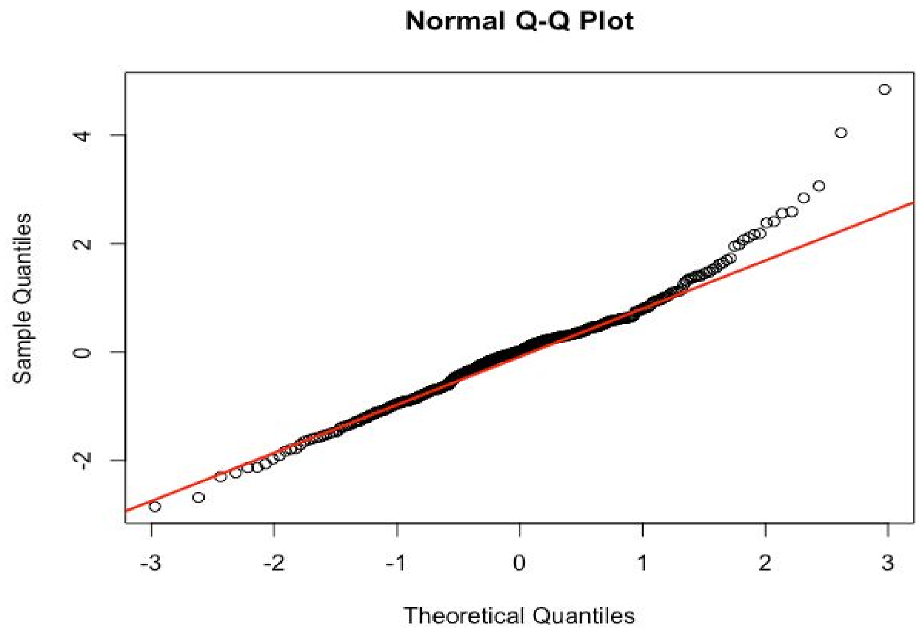
There is a definite positive trend in the data as our predicted values increase and some of the residuals are extremely large suggesting that the linear model we used might not be the best fit for the data in question.

```
## ii. kernel density estimate of the residuals
plot(density(resids), xlab = "Residuals", ylab = "Density", main =
"Kernel Density Estimate of Residuals")
```



The Kernel Density plot of the residuals shows a right-skew trend in the data which suggests that our dataset might not be strictly normal, which would validate one of our main assumptions in using linear regression.

```
## iii. Q-Q plot of the standardized residuals
stdresids = rstandard(fit)
qqnorm(stdresids, main = "Normal Q-Q Plot")
qqline(stdresids, col = "RED", lwd = 2)
```



The Q-Q plot for the standardized residuals shows a definite curvature suggesting that we're seeing values that we wouldn't expect from a normal sample.

<u>Problem 4:</u>
Part a:

```
library(leaps)

y = (baseball$salary)
X = baseball[,2:17]

out = leaps(X, y, method = 'r2', nbest = 1)

aic = 1:16
bic = 1:16
for(j in 1:16) {
  vec=(1:16)[out$which[j,] == TRUE]

  Data = baseball[, c(1, vec+1)]

  fit = lm((salary) ~., data = Data)
  aic[j] = AIC(fit)
  bic[j] = AIC(fit, k = log(nrow(baseball)))
```

```
}
cbind(aic,bic)

##choosing model based on aic
min_aic_degree <- which.min(aic)
Min_aic_degree
```

```
> ##choosing model based on aic
> min_aic_degree <- which.min(aic)
> min_aic_degree
[1] 9
```

```
variable_mask <- out$which[min_aic_degree, ]
varnames <- colnames(baseball)
varnames[2:length(varnames)][variable_mask]

leaps_X <-X[, variable_mask]
leaps_model_aic <- lm(y ~., data = leaps_X)
summary(leaps_model_aic)
```

```
Call:
lm(formula = y ~ ., data = leaps_X)

Residuals:
    Min      1Q  Median      3Q     Max 
-1921.2  -460.1    28.1   328.2  3226.4 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           -106.227     87.615  -1.212  0.22622    
home.runs               27.469      9.514   2.887  0.00415 ** 
rbi                     17.222      3.180   5.416 1.18e-07 ***
walks                    3.849      2.466   1.561  0.11957    
strike.outs            -10.300      1.892  -5.444 1.03e-07 ***
stolen.bases            15.124      3.668   4.124 4.73e-05 ***
free.agent.eligible   1376.629    104.445  13.180  < 2e-16 ***
free.agent            -299.979    135.802  -2.209  0.02787 *  
arbitration.eligible   765.440    115.616   6.621 1.47e-10 ***
arbitration            370.961    239.010   1.552  0.12161    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 693.1 on 327 degrees of freedom
Multiple R-squared:  0.696,	Adjusted R-squared:  0.6876 
F-statistic: 83.18 on 9 and 327 DF,  p-value: < 2.2e-16
```

```
leaps_X2 <- subset(leaps_X, select = - walks)
```

```
leaps_model_aic2 <- lm(y ~., data = leaps_X2)
summary(leaps_model_aic2)
```

```
Call:
lm(formula = y ~ ., data = leaps_X2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1926.2  -465.2    19.3   321.1  3346.0 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -103.078     87.783  -1.174  0.24115    
home.runs              26.180      9.499   2.756  0.00618 ** 
rbi                    19.027      2.968   6.410 5.06e-10 ***
strike.outs            -9.572      1.838  -5.209 3.36e-07 ***
stolen.bases           16.472      3.572   4.611 5.75e-06 ***
free.agent.eligible  1411.617    102.234  13.808  < 2e-16 ***
free.agent           -320.086    135.485  -2.363  0.01873 *  
arbitration.eligible  765.602    115.868   6.608 1.58e-10 ***
arbitration           366.636    239.517   1.531  0.12680    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 694.6 on 328 degrees of freedom
Multiple R-squared:  0.6937,	Adjusted R-squared:  0.6862 
F-statistic: 92.86 on 8 and 328 DF,  p-value: < 2.2e-16
```

```
leaps_X3 <- subset(leaps_X2, select = - arbitration)
leaps_model_aic3 <- lm(y ~., data = leaps_X3)
summary(leaps_model_aic3)
```

```
Call:
lm(formula = y ~ ., data = leaps_X3)

Residuals:
    Min      1Q  Median      3Q     Max 
-1928.2  -450.0    16.3   328.0  3335.6 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -109.277     87.869  -1.244   0.2145    
home.runs              25.447      9.506   2.677   0.0078 ** 
rbi                    19.335      2.967   6.516 2.71e-10 ***
strike.outs            -9.537      1.841  -5.180 3.88e-07 ***
stolen.bases           16.387      3.579   4.578 6.65e-06 ***
free.agent.eligible  1408.415    102.421  13.751  < 2e-16 ***
free.agent           -320.084    135.761  -2.358   0.0190 *  
arbitration.eligible  818.333    110.855   7.382 1.29e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 696 on 329 degrees of freedom
Multiple R-squared:  0.6915,	Adjusted R-squared:  0.685 
F-statistic: 105.4 on 7 and 329 DF,  p-value: < 2.2e-16
```

```
##choosing model based on bic
min_bic_degree <- which.min(bic)
Min_bic_degree
```

```
> min_bic_degree
[1] 6
```

```
bic_variable_mask <- out$which[min_bic_degree, ]
varnames <- colnames(baseball)
varnames[2:length(varnames)][variable_mask]

bic_leaps_X <- X[, variable_mask]
leaps_model_bic <- lm(y ~., data = leaps_X)
summary(leaps_model_bic)
```

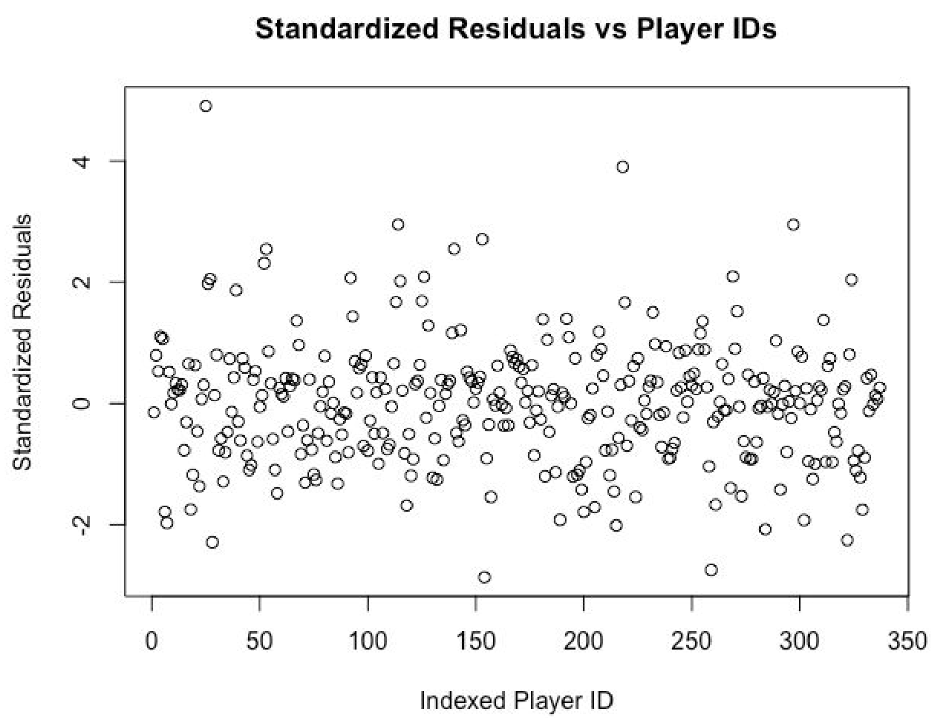
```
Call:
lm(formula = y ~ ., data = leaps_X)

Residuals:
    Min      1Q  Median      3Q     Max 
-1921.2  -460.1    28.1   328.2  3226.4 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -106.227     87.615  -1.212  0.22622    
home.runs             27.469      9.514   2.887  0.00415 ** 
rbi                   17.222      3.180   5.416 1.18e-07 ***
walks                  3.849      2.466   1.561  0.11957    
strike.outs          -10.300      1.892  -5.444 1.03e-07 ***
stolen.bases          15.124      3.668   4.124 4.73e-05 ***
free.agent.eligible 1376.629    104.445  13.180  < 2e-16 ***
free.agent          -299.979    135.802  -2.209  0.02787 *  
arbitration.eligible 765.440    115.616   6.621 1.47e-10 ***
arbitration          370.961    239.010   1.552  0.12161    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 693.1 on 327 degrees of freedom
Multiple R-squared:  0.696,	Adjusted R-squared:  0.6876 
F-statistic: 83.18 on 9 and 327 DF,  p-value: < 2.2e-16
```

The main reason that I decided to select my model based on AIC instead of BIC, because AIC is typically better suited for higher-dimensional, extremely complex scenarios (because BIC more seriously penalizes complexity), and as I briefly touched on before, I think this problem fits

that description in the sense that there are lots of variables to start with and I think there is missing, useful information. I also choose to only include variables that were statistically significant to reduce unnecessary model complexity and for general simplicity. Moreover, there was a relatively small difference between the R-Squared values of the AIC and BIC models.

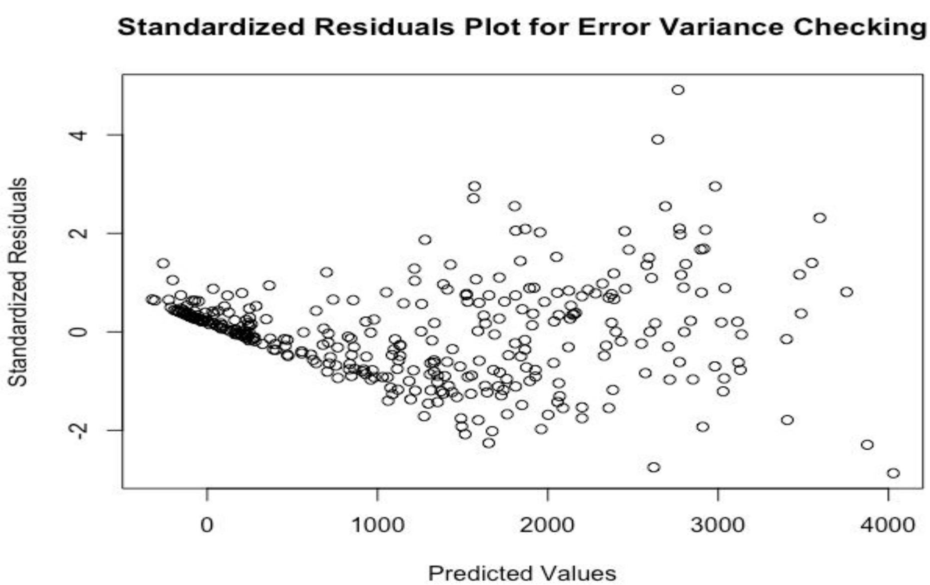
Part b:

```
resids <- leaps_model_aic3$residuals
preds <- leaps_model_aic3$fitted.values
stdresids = rstandard(leaps_model_aic3)

start = 1
index <- seq(start, 337, 1)

plot(index, stdresids, xlab = "Indexed Player ID", ylab =
"Standardized Residuals", main = "Standardized Residuals vs Player
IDs")
```



```
low_index <- which(stdresids < -3)
baseball[low_index, ]

## no residuals less than -3

high_index <- which(stdresids > 3)
baseball[high_index, ]
    salary batting.average on.base.percent runs hits doubles triples
25    6100           0.302           0.391  102  174      44       6
218   5300           0.316           0.397   78  153      35       3
    home.runs rbi walks strike.outs stolen.bases errors free.agent.eligible
25         18 100    90          67            2     15                   1
218        31 100    65         121            6      7                   1
    free.agent arbitration.eligible arbitration
25           1                    0           0
218          1                    0           0
## 2 residuals greater than 3
## 25 - Bobby Bonilla
## 218 - Danny Tartabull

## The obvious things that jump out about these players is that their
respective salaries are basically the league minimum and it appears
from their other offensive statistics that they did not play in very
many games. Therefore they are not “normal” players in the MLB, and
it makes sense that they are more or less outliers in this dataset.
```

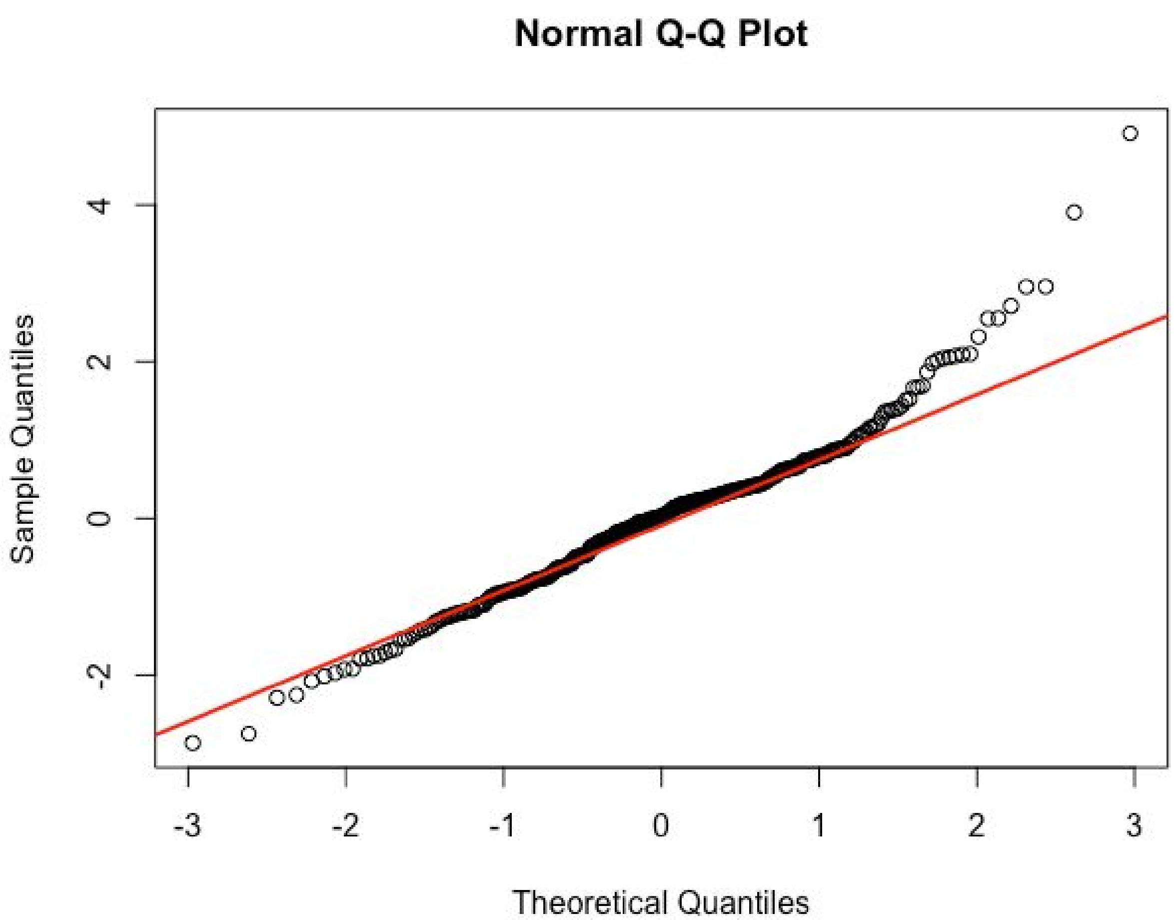
Part c:



This plot is relatively random despite the clear cluster on the left side of the graph, but that "boundary" of sorts makes sense given the structure of the salaries and other data. As such, the assumption of equal variances is obviously violated; however, compared to the residuals plot from Problem 3, I think it's safe to say that this model is a better fit for the data.

Part d:

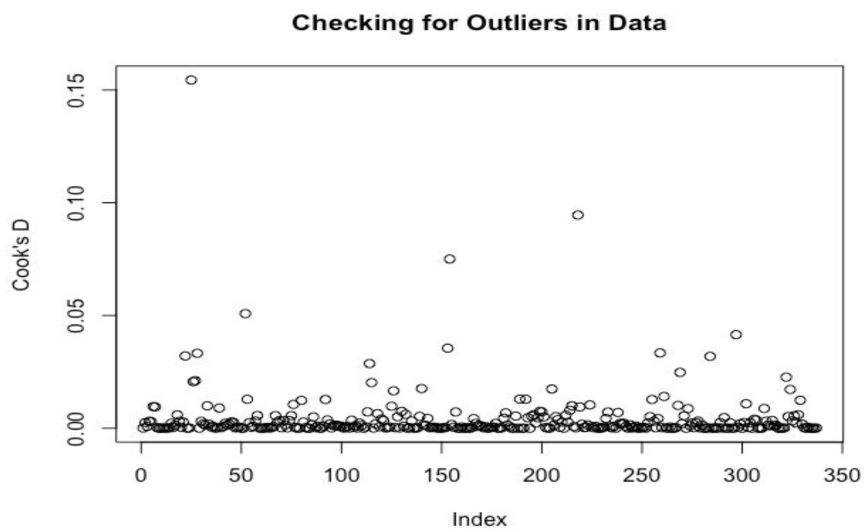
```
qqnorm(stdresids, main = "Normal Q-Q Plot")
qqline(stdresids, col = "RED", lwd = 2)
```



The Q-Q plot for the standardized residuals shows a definite curvature suggesting that we're seeing values that we wouldn't expect from a normal distributed sample - though this sort of makes sense when considering athlete salaries as the top performers typically make exceptionally more than their, relatively average counterparts - voiding our assumption of normality in the sample.

Part e:

```
cdists <- cooks.distance(leaps_model_aic3)
plot(cdists, ylab = "Cook's D", main = "Checking for Outliers in
Data")
```



The data points do not seem to be influential because all of them are far less than 1.