Baseball Salary Data Study

1.

```
# (1)
Data = read.csv('baseball.csv')
baseball=Data[,1:17] # This takes data from x1..x16.

fit = lm(salary ~.,data=baseball) # use the data frame baseball
# salary on the left of ~ has R use salary column of baseball as response
# . on right of ~ tells R to use all other columns of baseball as independent variables
```

```
> 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
```

2.

```
# (2)
# coefficient of independent variable "hits": -2.698
# according to this, getting 10 extra hits means a salary DECREASE of 10*2.7 = $27.
# This does not make sense. By my intuition, one would expect that more hits means better payday.
# There should be a positive coefficient for hits, not negative.

variation_percentage = summary(lm(salary ~.,data=baseball))$r.squared
# 70.14% of variation in salaries explained by linear model
```

```
> variation_percentage = summary(lm(salary ~.,data=baseball))$r.squared
> variation_percentage
[1] 0.7014386
```

3. Based on the summary of the fit in problem 1, we found that the p-value when testing the independent variables against the response (salary) is 2.2 * 10^-16. This is much less than the level of significance 0.05 and would be lower than most level of significances anyway. Even when we tested individual response variables, the largest p-value attained was 0.02, still less than 0.05. The linear model's variables are all at least somewhat related to the salary.

```
# (3)
baseball_none = Data[,c(1,13)] # testing free agent status vs salary had the highest p-value
# But this p-value of 0.02 was still < 0.05.
fit_none = lm(salary ~.,data=baseball_none) # use data from the eleven columns above
```

```
> baseball_none = Data[,c(1,13)]
> fit_none = lm(salary ~.,data=baseball_none) # use data from the eleven columns above
> summary(fit_none)

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

Residuals:
    Min      1Q  Median      3Q     Max 
-1455.6  -969.1  -528.8   821.2  4644.4 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1078.11     102.05  10.565   <2e-16 ***
errors         25.17      11.35   2.218   0.0272 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1233 on 335 degrees of freedom
Multiple R-squared:  0.01447,	Adjusted R-squared:  0.01153 
F-statistic: 4.919 on 1 and 335 DF,  p-value: 0.02723
```

4. From the summary of the fitted model below using the 11 chosen independent variables, the p value attained = 2.2e^-16 which is much less than the level of significance of 0.05. Based on this, we can agree with the null hypothesis and say that batting average, OBP, hits, doubles, and triples are not needed in the same model with the other 11 variables. This result is very surprising because you would expect the best players (the ones with the higher salaries) to have many more hits and a better batting average. We take these pieces away but still get a relatively strong model even with the other 11 variables like home runs, walks, strikeouts, etc.

```
# (4)
baseball_eleven = Data[,c(1,8:17)] # This takes data from x8..x17 and compares with salary in x1
fit_eleven = lm(salary ~.,data=baseball_eleven) # use data from the eleven columns above
```

```
> baseball_eleven = Data[,c(1,8:17)] # This takes data from x8..x17 and compares with salary in x1
> fit_eleven = lm(salary ~.,data=baseball_eleven)
> > variation_percentage_eleven = summary(lm(salary ~.,data=baseball_eleven))$r.squared
>  > variation_percentage_eleven
Ca [1] 0.6972869
lm(formula = salary ~ ., data = baseball_eleven)

Residuals:
    Min      1Q  Median      3Q     Max 
-1856.2  -463.6    42.9   349.0  3260.0 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           -90.563     88.553  -1.023   0.3072    
home.runs              24.786      9.774   2.536   0.0117 *  
rbi                    18.068      3.257   5.547 6.01e-08 ***
walks                   3.857      2.465   1.565   0.1185    
strike.outs            -9.837      1.931  -5.095 5.92e-07 ***
stolen.bases           15.196      3.666   4.145 4.33e-05 ***
errors                 -8.491      7.167  -1.185   0.2370    
free.agent.eligible  1367.017    104.696  13.057  < 2e-16 ***
free.agent           -280.462    136.714  -2.051   0.0410 *  
arbitration.eligible  782.842    116.474   6.721 8.06e-11 ***
arbitration           373.259    238.871   1.563   0.1191    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 692.6 on 326 degrees of freedom
Multiple R-squared:  0.6973,	Adjusted R-squared:  0.688 
F-statistic: 75.09 on 10 and 326 DF,  p-value: < 2.2e-16
```

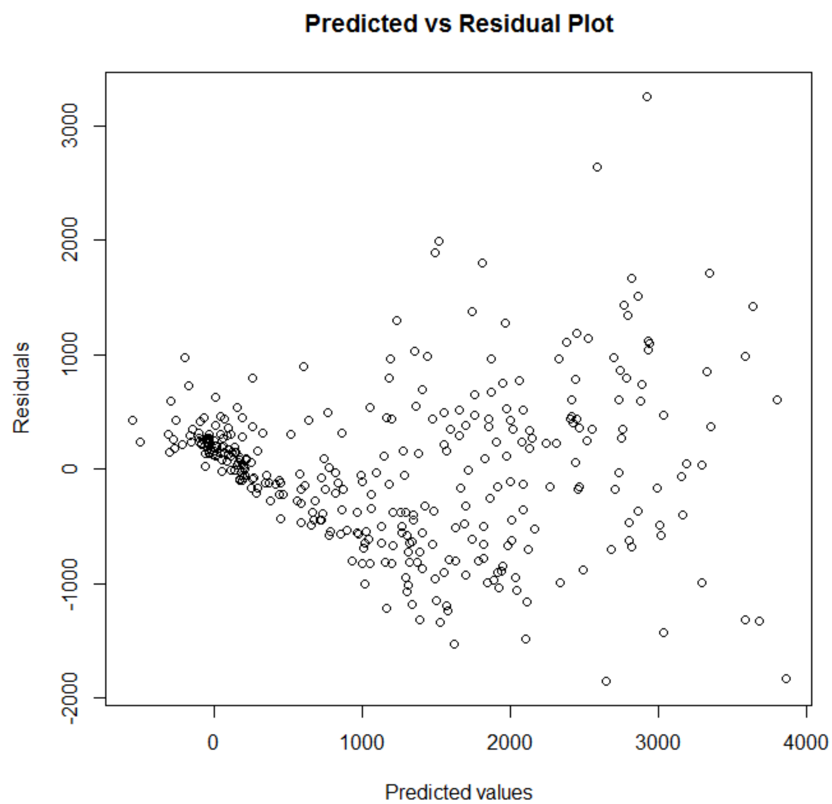
5.

```
# (5)
variation_percentage_eleven = summary(lm(salary ~.,data=baseball_eleven))$r.squared
# 69.73% of variation in salaries explained by linear model of given 11 columns
```

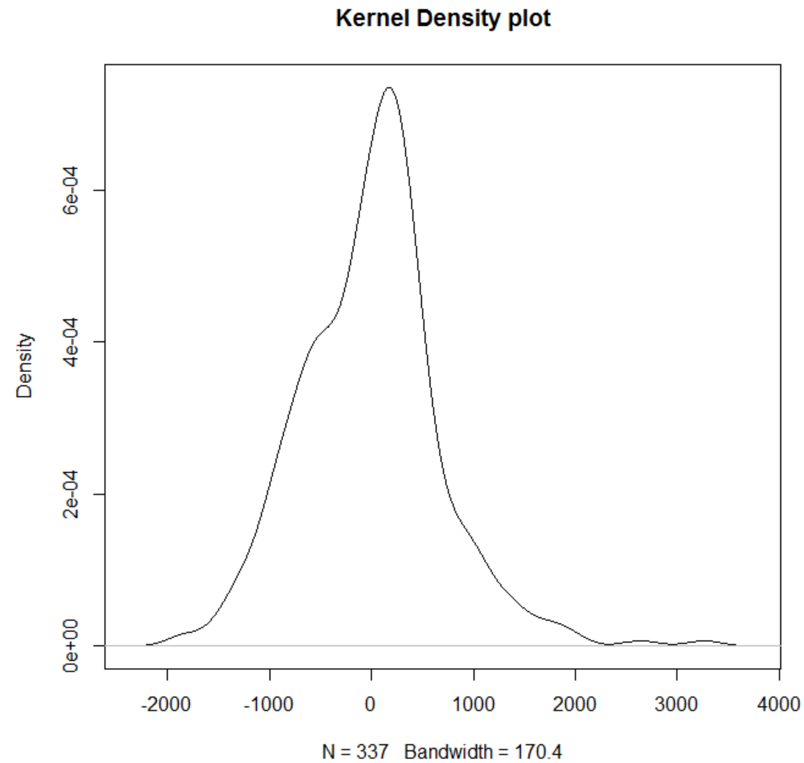
6.

```
# (6)
resid = fit_eleven$residuals
# put predicted values into a predict vector
predict = fit$fitted.value

plot(predict,resid, main="Predicted vs Residual Plot", xlab="Predicted values", ylab="Residuals")
# plot predicted values vs residuals

plot(density(fit_eleven$residuals), main = "Kernel Density plot") # kernel density estimate

out = qqnorm((resid-mean(resid))/sd(resid)) # normal probability plot
x = range(out$x)
lines(c(x[1],x[2]),c(x[1],x[2]))
```

- In the predicted vs residual plot, as the predicted values increase, there appears to be more large residual values. This would lead us to say that the model we have fits well for predicted values less than 1000 but after 1000 there's much more fluctuation in terms of how accurately our model fits the given dataset.



- The kernel density plot to the left shows that the data is distributed normally. You can see a rough normal curve on the plot with one slight dent between -1000 and 0 on the x-axis on the curve. Based on this plot, we can assume normality for the data and proceed with any calculations based on this assumption.



- The normal probability plot to the right shows a very linear relationship between the theoretical quantiles and the sample quantiles. The plotted values closely follow a 45 degree line here so just like we concluded earlier in the kernel density plot, we can safely say that the data in the baseball file are approximately normally distributed.