STATS 212
Homework3
March 7, 2025

Samuel Molero
samueljosemolero@tamu.edu
Section: 501

---

1. Q1?

Ans:

(a)

```
df <- read.csv("Baseball-Salary-Data.csv")
head(df)
m1 <- lm(salary ~ . - player, data = df)
summary(m1)
```

Running the above code output the following results:

```
                    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
```

(b) The R square of 0.7014, which translates to 70.14%

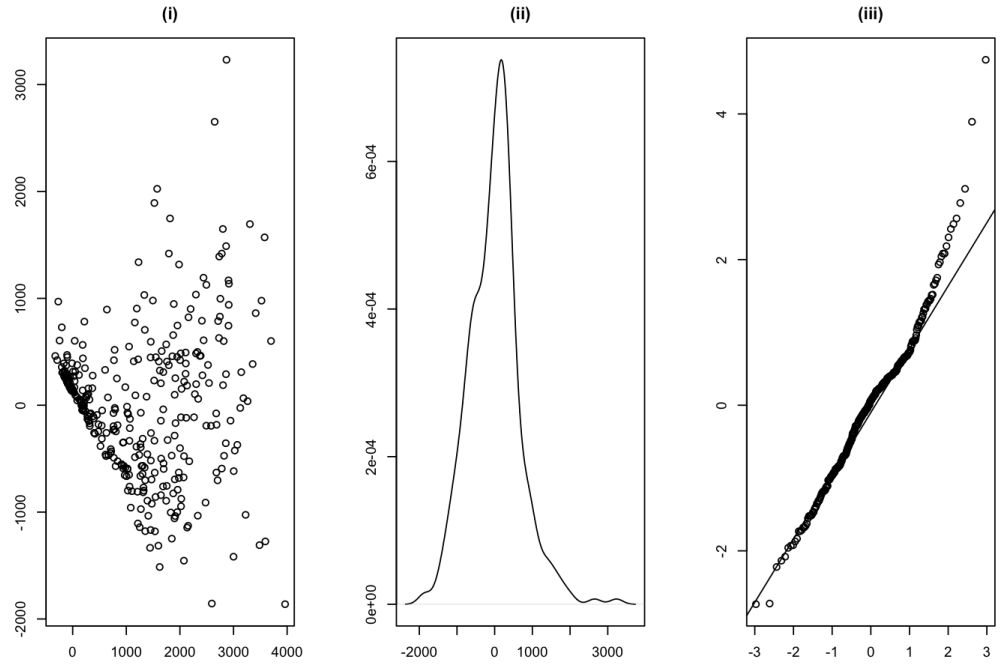
(c) The coefficient for the predictor is -2.698 which predicts a decrease in salaries holding all the other variables constant.This could be due to the variation in salary data is explained by other variables such as home run and stolen bases.

(d) Given that the P-value is 2.2e-16, which is smaller than 0.05 thus rejecting the null. This means that the model has good utility, and is a helpful indicative at predicting salary data.

(e) Give that the F statistic is 0.619 and that is bigger than the give alpha, the second model does not have a significant improvement at predicting the Salary compared to the first model. This result is surprising became you would assume that having less predictors would make the prediction of salary worse given that there is less information.

(f) This value will be the R-squared which is 0.6981 about 69.81%

(g)



(i) The is no clear pattern, but the spread of residuals increases in proportion to fitted values, indicating no constant variance. Additionally, non-linearity exits which demonstrate that a linear model may not be the best fit.
(ii) The distribution is highly skewed, indicating for no normality assumption.
(iii) The points mostly follow a straight line in the middle, but deviate at

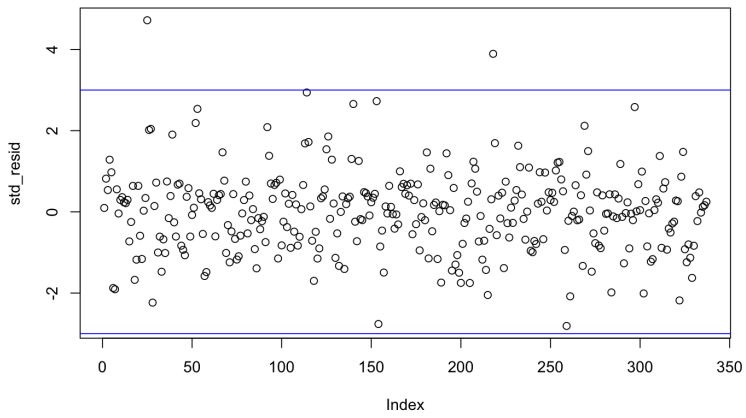
extremes demonstrating a non-normality at outliers.

(h)

```
source("AIC-Leaps.R")
df <- read.csv("Baseball-Salary-Data.csv")
df <- df[, -18]

library(leaps)
leaps_ic <- leaps.AIC(df[,2:17], df[, 1])
    # includes all predictors             includes the
        response variable
leaps_output <- leaps(df[, 2:17], y = df[, 1], nbest =
    1) # perform AIC\BIC calc

var_names <- colnames(df[, 2:17]) # get predictors
   name
best_model_index <- which.min(leaps_output$Cp) # Get
   selected variables
var_mask <- leaps_output$which[best_model_index, ]
model_vars <- var_names[var_mask] # extract the best
   predictor names
model_vars
```

By the code above,the best predictors are "home.runs", "rbi","walks", "strike.outs", "stolen.bases","free.agent.eligible","free.agent","arbitration.eligible", and "arbitration"

The rationale for selecting these predictors was based on the combination of player performance metrics and contractual factors that ultimately influence the salary. The leaps() function alongside Mallows CP helps identify the best subset, by comparing the models' goodness of fit and complexity. A lower CP will indicate that the model is both accurate and not complex, which is why the models with the lowest CP were selected as the best subset, to ensure that the salary variations are explained without unnecessary predictors, and thus show the strongest statistical relation with salary.

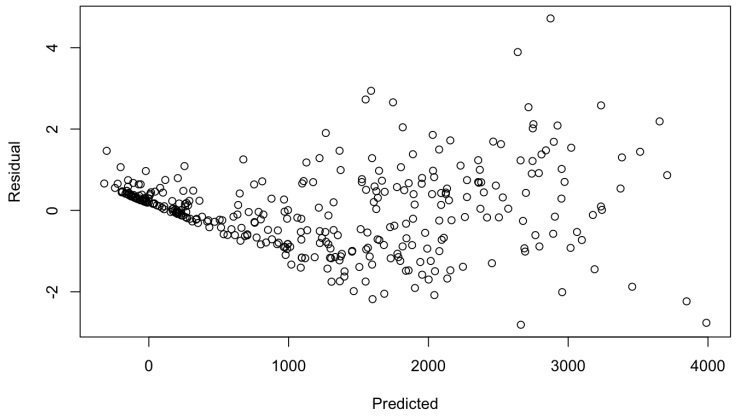


(i)

Player with standardized residual greater than the 3 are:

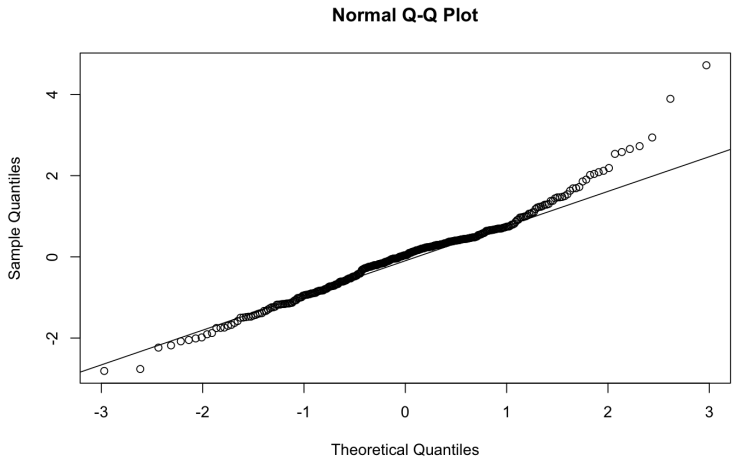
	salary	batting.average	on.base.percent	runs	hits	doubles	triples	home.runs	rbi	walks	strike.outs
25	6100	0.302	0.391	102	174	44	6	18	100	90	67
218	5300	0.316	0.397	78	153	35	3	31	100	65	121

Demonstrating who the difference in salaries does not correlate with the batting average, hits, and triples.



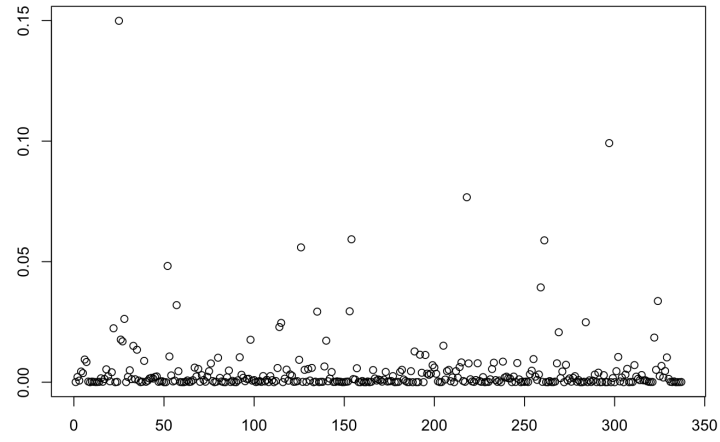
(j)

The plot exhibits a curved pattern, suggesting no linearity in the data.



(k)

Mostly linear but demonstrate non-normality by outliers.



(l)

Most points are near zero, demonstrating little influence, with a few being

above the zero range.
Running the following code to determine the outliers

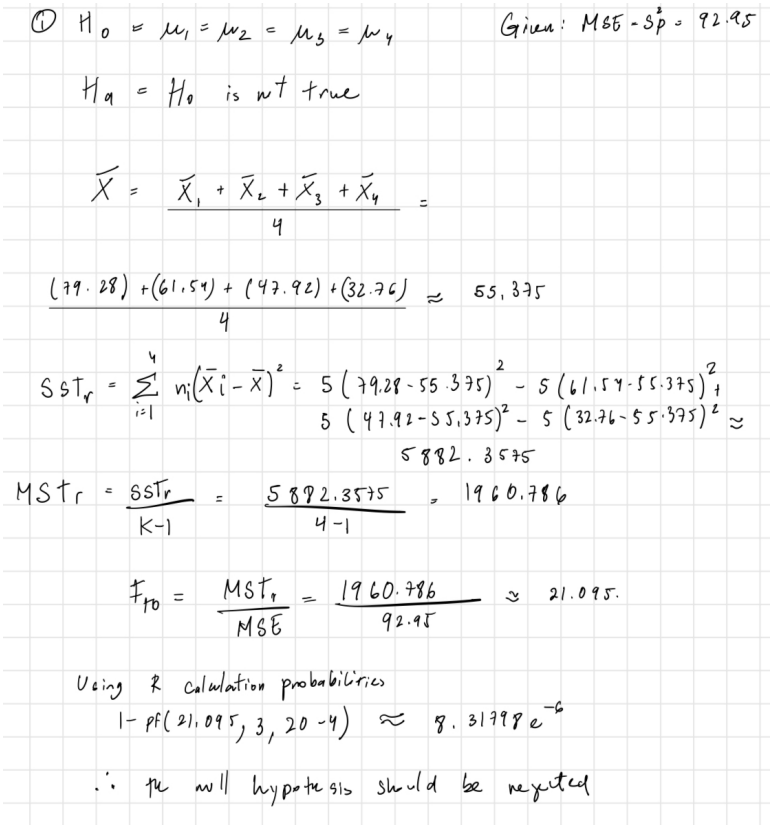
```
influential_points <- which(cd > (4 / length(cd)))
print(influential_points)
df[influential_points, ]

output (a few):
stolen.bases      errors
22            76      6
25             2     15
26            10      7
27            37      3
```

Which demonstrates that predictors such as stolen.bases and errors have a high impact on the fitted model

2. Q2?

Ans:

(a)



(b)

```
dta2 <- read.table("SleepRem.txt", header = TRUE, sep = "")
attach(dta2)
fit <- aov(values ~ as.factor(ind), data = dta2)
anova(fit)
```

Using the above code snippet we get the following result:

```
Response: values
           Df  Sum Sq  Mean Sq  F value  Pr(>F)
find        3  5881.7 1960.58  21.093 8.322e-06
Residuals 16 1487.1   92.95
```

Given that is confirmed that the p-value is $8.322e-06$ the null hypothesis should be rejected.

(c)

```
#To test variance
```

```
anova(aov(resid(aov(values ~ ind))**2 ~ ind))
```

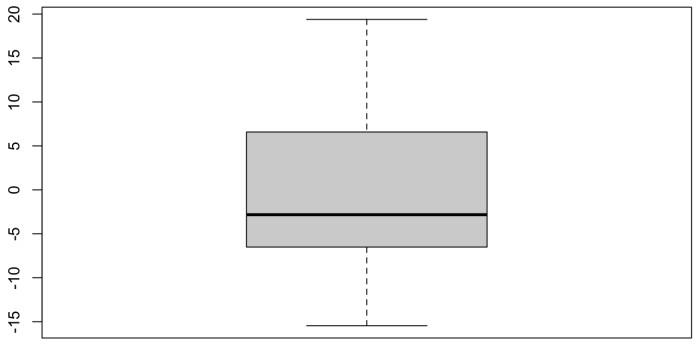
Given that the p-value is 0.621. This suggests that the assumption of equal variance is approximately valid.

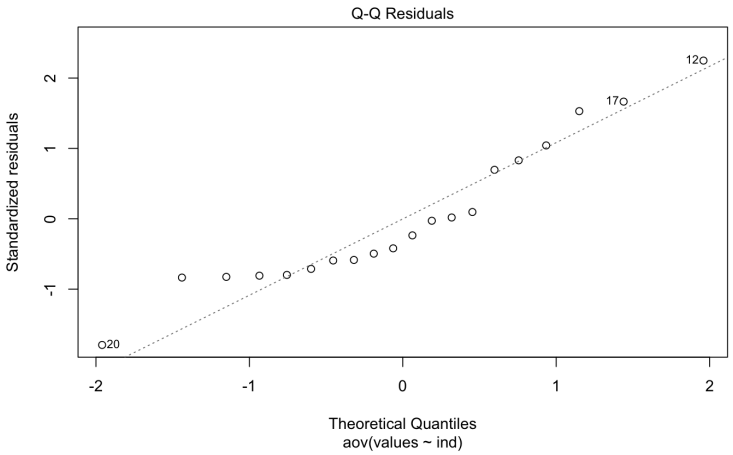
```
#to test stability
shapiro.test(resid(aov(values ~ ind)))
```

p-value = 0.1285 suggests that the normality assumption approximately holds.

```
fit = aov(values ~ find)
boxplot(resid(fit))
plot(fit, which=2)
```

The Code snippets creates the following graphs:





plots also suggest that the normality assumption is approximately satisfied, in agreement with the Shapiro-Wilk test p-value.