Modeling

Group 6

Predicting WAR

```
# Simpler implementation?
#plot + stat_smooth(mapping = aes(x = Year, y = prop), data = couldabeens_post, method = "lm", formula
```

```
#pitchers <- df_pit_rkes
#pitchers1 <- drop_na(pitchers)
#pitchers1_trn <- pitchers1 %>% sample_frac(0.7)
#pitchers1_tst <- pitchers1 %>% anti_join(pitchers1_trn)
```

```
#library(leaps)
#ss1 <- regsubsets(WAR~. - Rk - Player, data = pitchers1_trn, nvmax = 49, method = "forward")
```

```
# remove troublesome variables
wrangle_lm <- function(dataset){
  dataset[,-c(1,2,5,6,7,8,25,26)] %>% drop_na()
}
dataset <- wrangle_lm(df_pit_rkes)
# select significant variables
select_vars <- function(dataset){
  dataset[,c(1,3,4,12,13,14,19,26,30,32,37,38,40)] %>% drop_na()
}
pitchers <- select_vars(dataset)
```

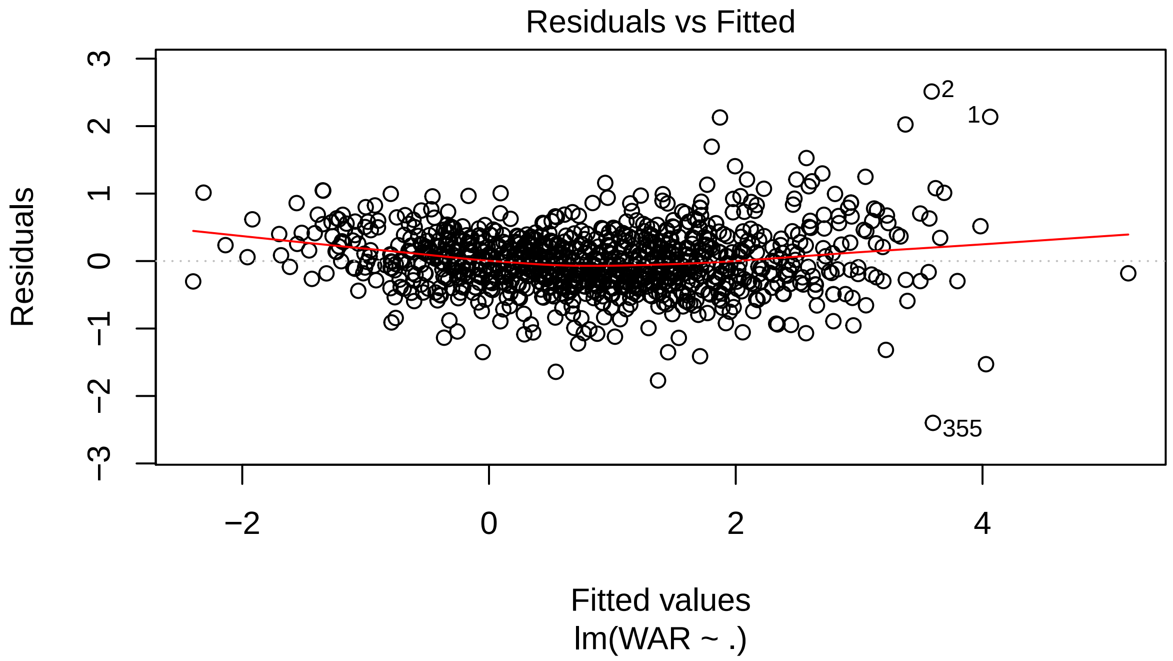
```
pitchers_trn <-  pitchers %>% sample_frac(0.7)
pitchers_tst <- pitchers %>% anti_join(pitchers_trn)
```

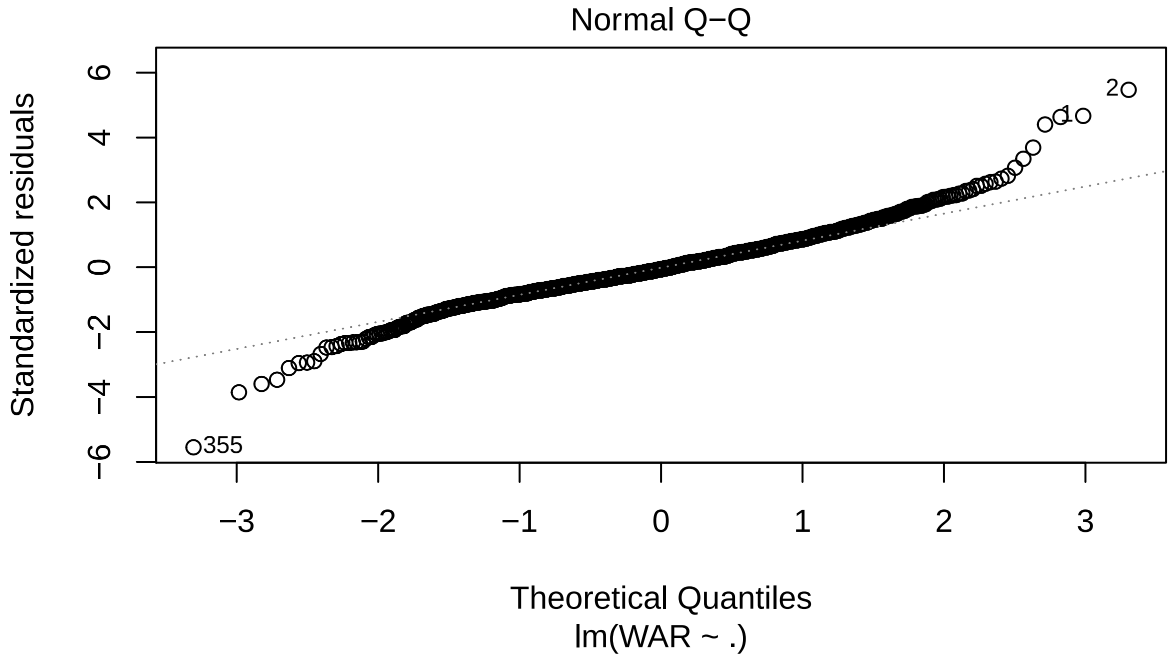
```
linear_model <- lm(WAR ~ ., data = pitchers)
summary(linear_model)
```

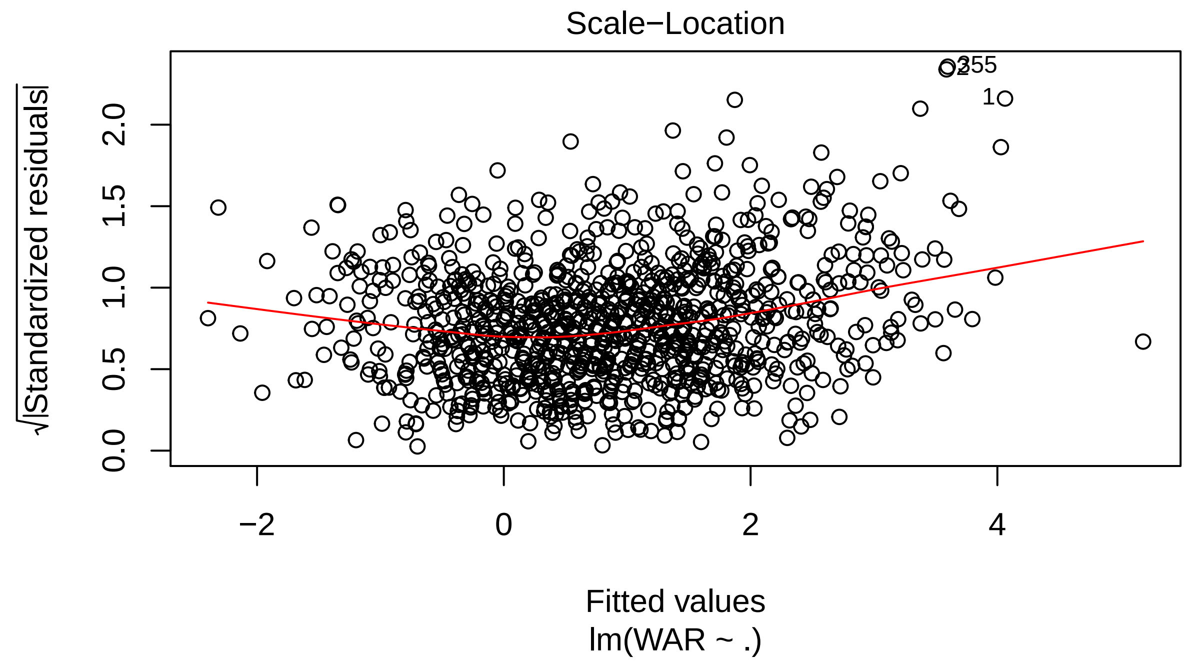
```
##
## Call:
## lm(formula = WAR ~ ., data = pitchers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.39814 -0.26531 -0.03036  0.25217  2.51213
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.4214065  0.3080594  -1.368  0.17163
## G            0.0088687  0.0015426   5.749 1.18e-08 ***
## GS           0.0847422  0.0056713  14.942  < 2e-16 ***
## H            0.0292305  0.0015373  19.014  < 2e-16 ***
## R           -0.1019976  0.0055552 -18.361  < 2e-16 ***
## ER           0.0321035  0.0061658   5.207 2.32e-07 ***
## `ERA+`       0.0047342  0.0007226   6.552 8.93e-11 ***
## IBB          0.0148943  0.0076841   1.938  0.05285 .
## GDP          0.0080564  0.0043426   1.855  0.06385 .
## CS           0.0278851  0.0084695   3.292  0.00103 **
## OBP          5.0432213  0.9027225   5.587 2.95e-08 ***
## SLG          6.7642786  0.6565909  10.302  < 2e-16 ***
## `OPS+`      -0.0440980  0.0024499 -18.000  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

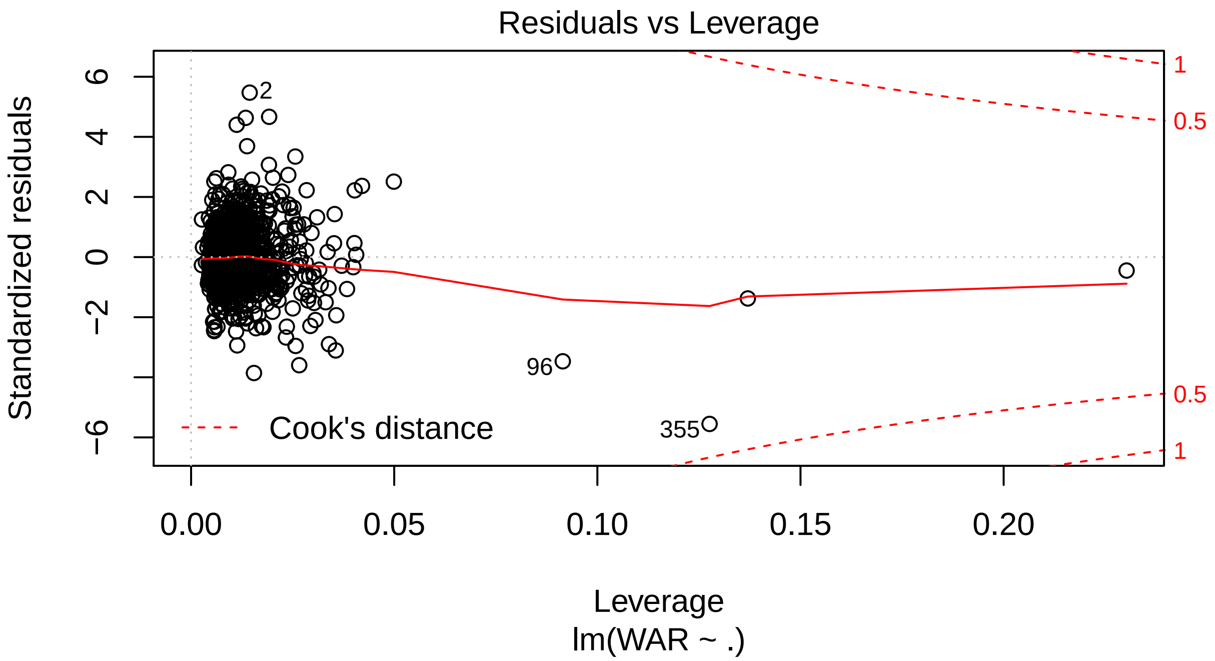
```
##
## Residual standard error: 0.4625 on 1042 degrees of freedom
## Multiple R-squared:  0.8521, Adjusted R-squared:  0.8504
## F-statistic: 500.2 on 12 and 1042 DF,  p-value: < 2.2e-16
```

```r
plot(linear_model)
```









```r
predictions <- predict(linear_model, pitchers_tst)
test_MSE <- mean(predictions - pitchers_tst$WAR)^2
test_MSE
```

```
## [1] 5.265982e-05
```

```r
#data.frame(model = 1:50, adjr2 = summary(ss1)$adjr2, rss = summary(ss1)$rss, cp = summary(ss1)$cp)%>%
```