Brian Chen
MATH 329
Summer 2024
6/12/2024

Hitters Complete Regression Analysis

Exploratory Analysis

“Hitters” is a dataset of Major League Baseball statistics about different players from the 1986 and 1987 seasons. This dataset contains observations for 322 players and has 20 variables, and our response variable is *Salary*, a continuous variable that represents each player’s 1987 annual salary on opening day in thousands of dollars.

The continuous variables *AtBat*, *Hits*, *HmRun*, *Runs*, *RBI*, *Walks*, *PutOuts*, *Assists*, and *Errors* measure the number of times at bat, hits, home runs, runs, runs batted, walks, putouts, assists, and errors, respectively, of each player in 1986. These variables, save *Errors*, measure some aspect of each player’s performance.

The continuous variables *CAtBat*, *CHits*, *CHmRun*, *CRuns*, *CRBI*, and *CWalks* measure the number of times at bat, hits, home runs, runs, runs batted, and walks, respectively, of each player during their careers.

The continuous variable *Years* is the amount of years each player has been playing in the Major League.

The categorical variables *League*, *Division*, and *NewLeague* hold information about a players league at the end of 1986, division at the end of 1986, and league at the beginning of 1987, respectively.

Data Cleaning

The categorical variables “League”, “Division”, and “NewLeague” are factors by default.

To remove any missing data, the dataset was copied without entries with missing data into the dataset *ds*, with

```
ds = na.omit(Hitters)
```

As a result, the data now has 263 observations as 59 observations had missing data. This adjusted data set, *ds*, is what we will use for the rest of our analysis.

Numeric and Visual Analysis

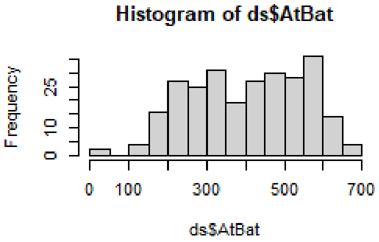
To check the correlations of each continuous variable relative to *Salary*,

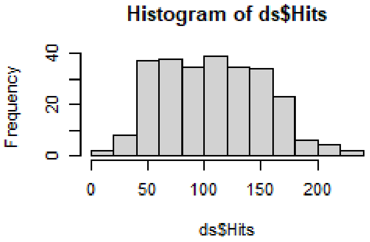
```
cor(ds[,c(-14,-15,-20)])[17,]
```

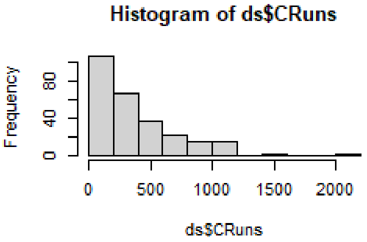
14, 15, 20 are the indexes of the categorical variables, and 17 is the index of “Salary” after indexes have been adjusted for removing indexes 14 and 15. The variables most correlated to *Salary* are *CRBI*(0.5670), *CRuns*(0.5647), *CHits*(0.5489), *CAtBat*(0.5261), *CHmRun*(0.5249).

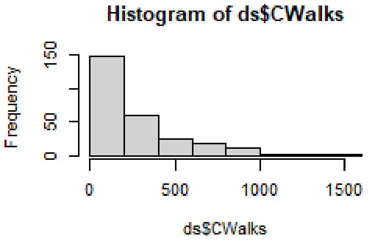
However, the variables that we will be analyzing, determined to be the most significant predictors of *Salary* in the Regression Analysis section, are *AtBat*, *Hits*, *Walks*, *CRuns*, *CWalks*, and *PutOuts*.

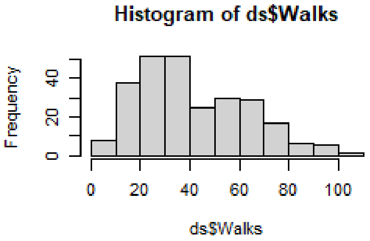
Using the “summary()” function, *AtBat* has a similar mean(403.6) and median(413.0), and the distribution is normal. *Hits* also has a similar mean(107.8) and median(103.0) with a normal distribution.

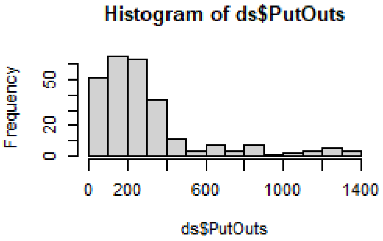






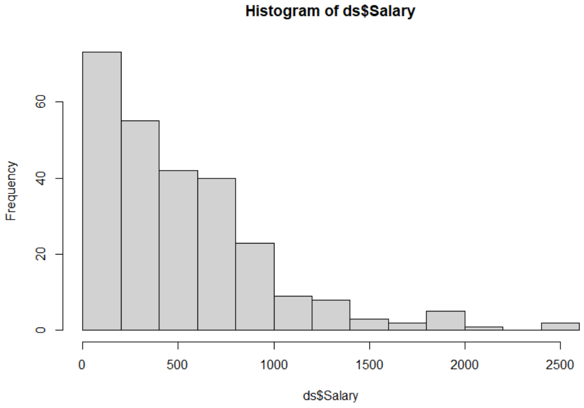




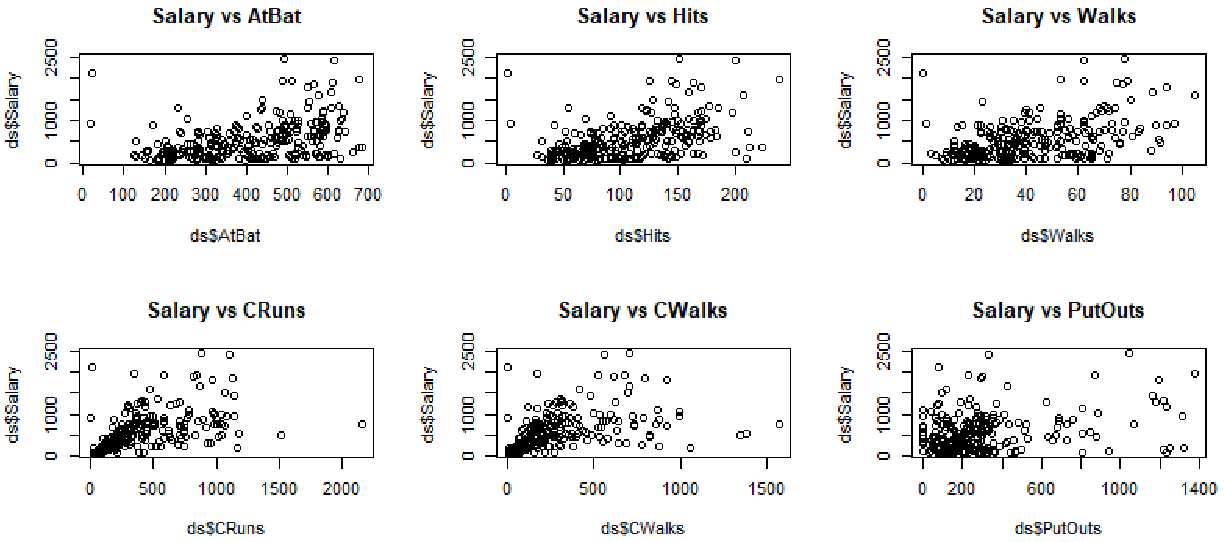


CRuns, *CWalks*, *Walks*, and *PutOuts* have distributions that are skewed right. Their means are noticeably larger than their medians and are 361.2 to 250.0(*CRuns*), 260.3 to 174.09(*CWalks*), 41.11 to 37.00(*Walks*), and 290.7 to 224.0(*PutOuts*), respectively.

Similar to most of our significant variables, our response variable, *Salary*, also has a distribution that is skewed right, and a noticeably higher mean(535.9) than median(425.0).

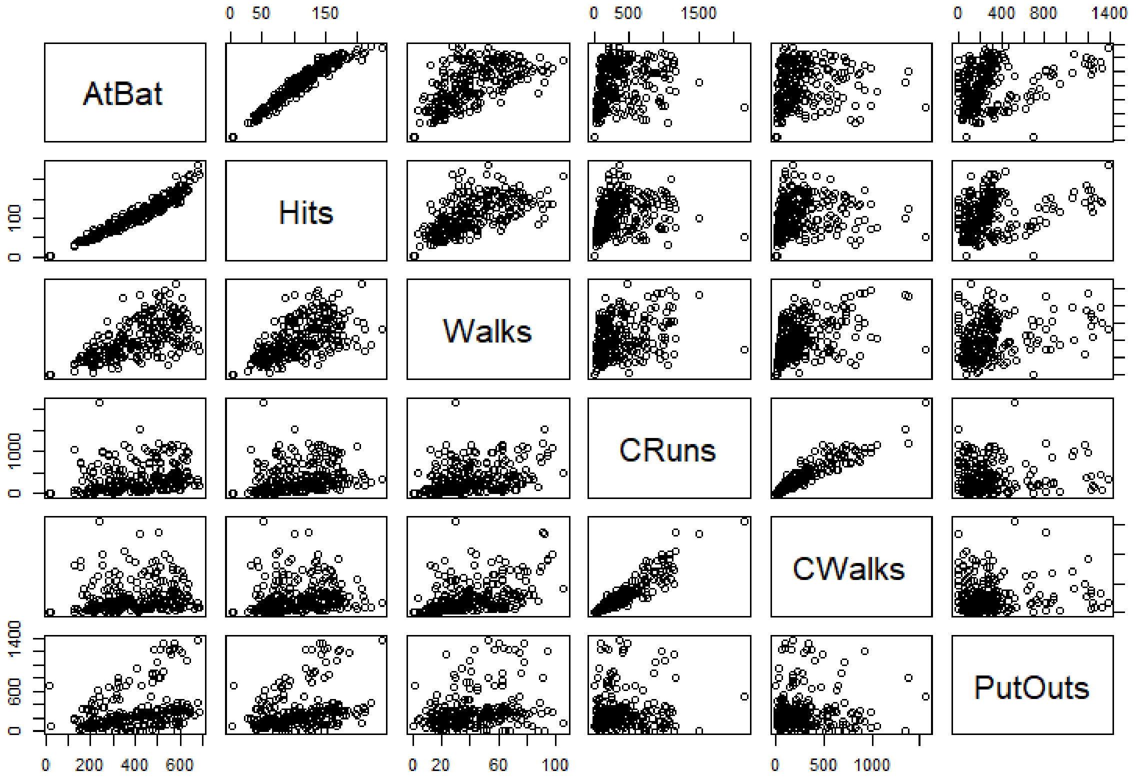


These significant variables are all positively correlated to *Salary*, especially in the lower ranges.



These variables measure different aspects of performance, which explain the positive relationship these variables have with Salary, as better performance is often associated with better compensation.

When examining these variables against each other, most combinations have some positive correlation. AtBat and Hits and CRuns and CWalks have a very strong positive correlation. Intuitively, more chances at bat(*AtBat*) means more chances for a player to gain a hit(*Hits*).



PutOuts, *CRuns*, and *CWalks* tend to cluster to the left and the lower ranges. *CRuns* and *CWalks* especially lack correlation with other variables, and this is likely because these variables represent statistics from players' entire careers and the other statistics are from the 1986 season, meaning that these variables have completely different time frames. In addition, there is a strong correlation between the *CRuns* and *CWalks*, which are career variables.

Regression Analysis

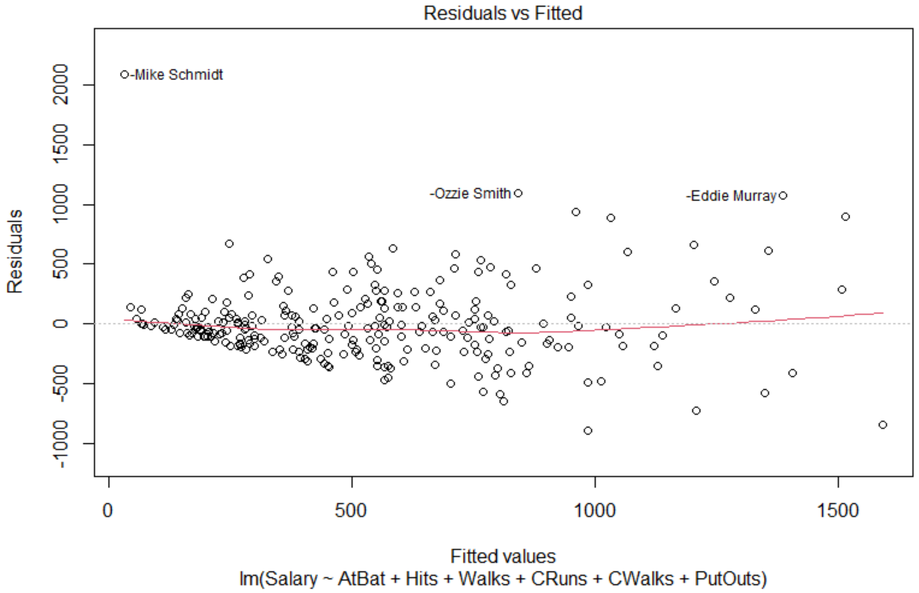
The predictors included in my linear regression model are *AtBat*, *Hits*, *Walks*, *CRuns*, *CWalks*, and *PutOuts*. These predictors were chosen by creating a linear model and using the "summary()" function on that model to examine which variables had acceptable p-values, which indicate if a variable was a significant predictor of the response variable.

```
lr=lm(Salary~AtBat+Hits+HmRun+Runs+RBI+Walks+CAtBat+CHits+CHmRun
+CRuns+CRBI+CWalks+PutOuts+Assists+Errors+Years, data=ds)
summary(lr)
```

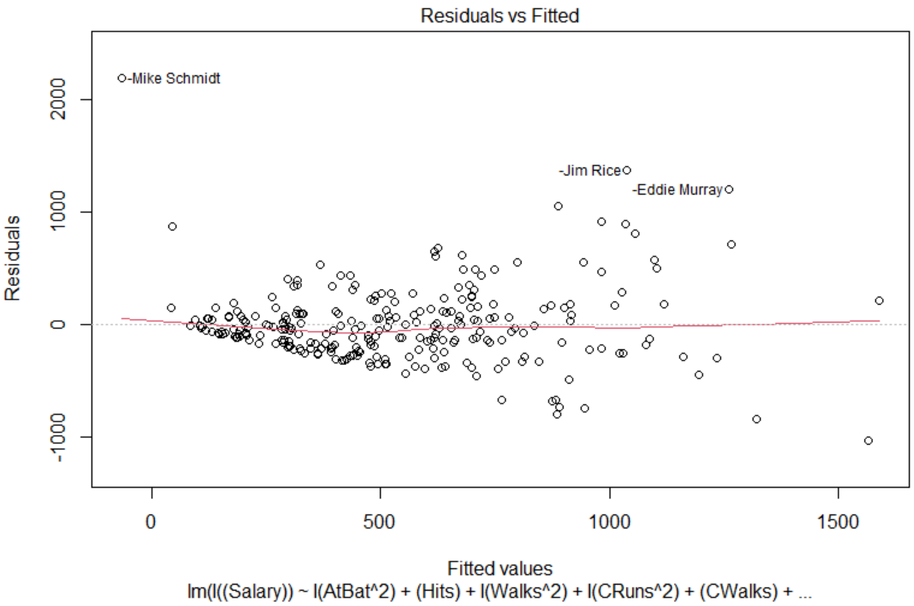
The new linear model includes those six most significant predictor variables,

```
lr1 = lm(Salary~AtBat+Hits+Walks+CRuns+CWalks+PutOuts, data=ds)
```

The first assumption for regression that we look at is **linearity**. By examining the **Residuals vs Fitted** plot, we can tell that the red line is not linear.



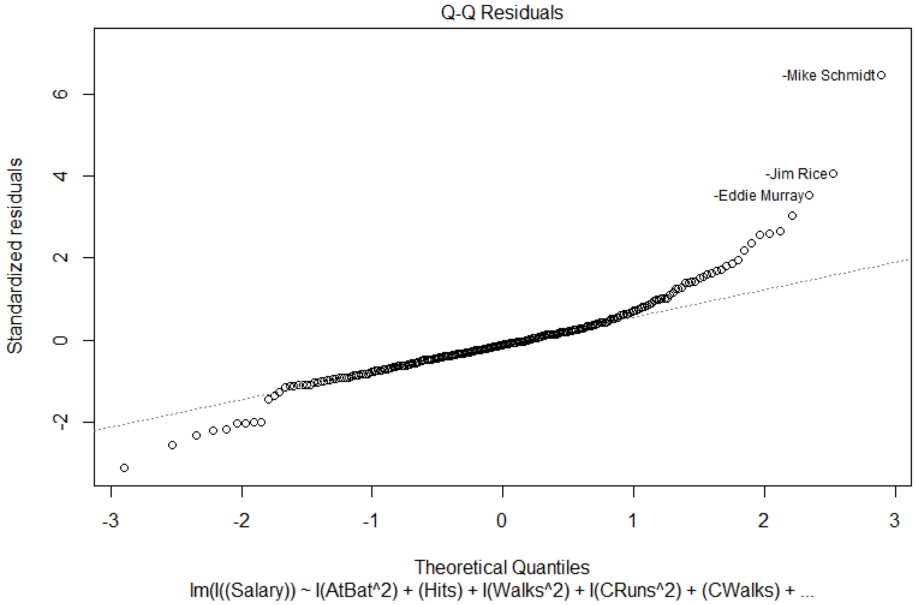
Violations of linearity are fixed by altering the predictor variables. The red line has a parabolic shape, indicating that squared terms may be useful. Squaring the variables *AtBat*, *Walks*, *CRuns*, and *PutOuts* creates a more linear red line on the **Residuals vs Fitted** plot.



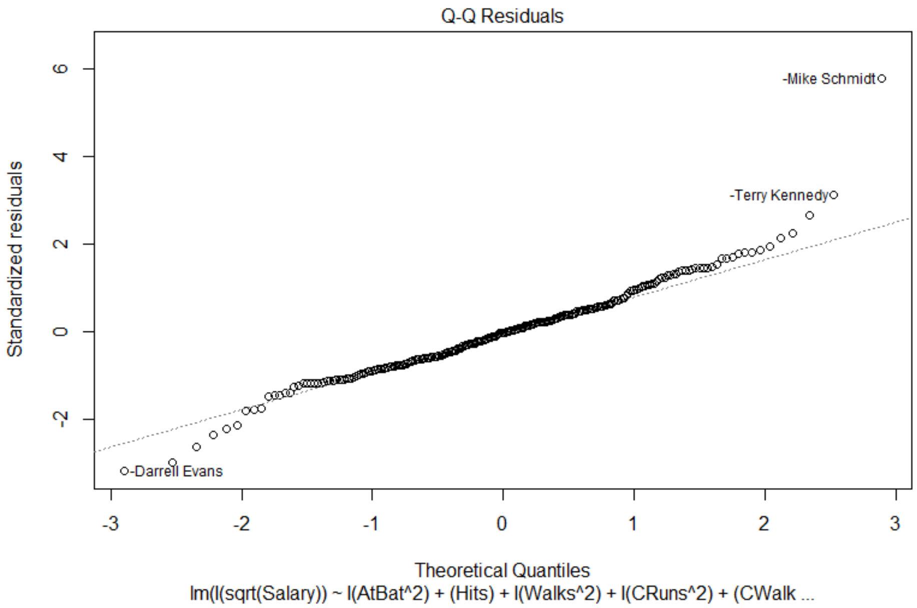
This results in the following linear model,

```
lr1=lm(I((Salary))~I(AtBat^2)+(Hits)+I(Walks^2)+I(CRuns^2)+(CWalks)+I(PutOuts^2), data=ds)
```

Normality can be examined in the **Q-Q Residuals** plot. Most of the data points are on the y=x line, indicating normality.



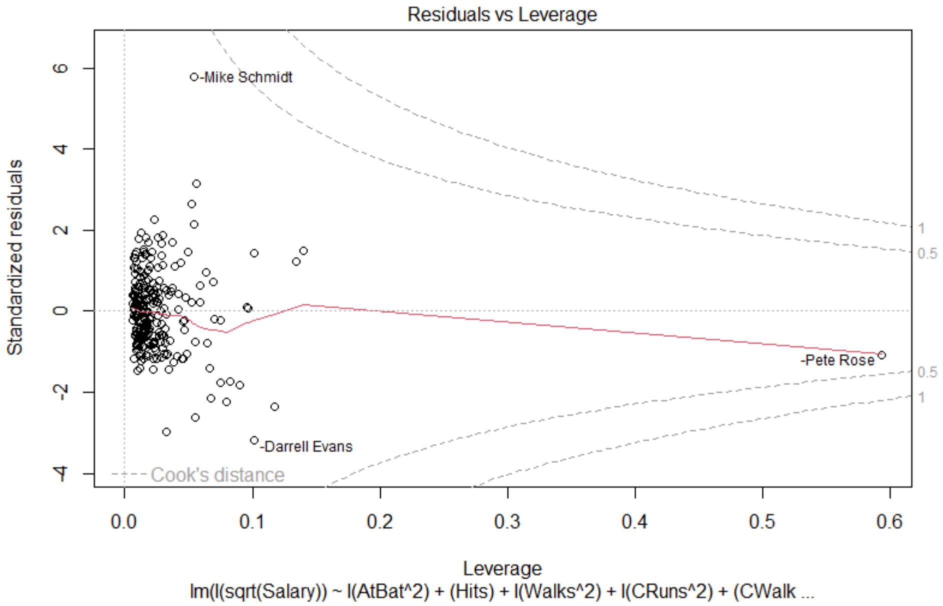
However, there are many points on either end of the plot that are not on this line. We can fix this by changing the response variable. By taking the square root of the response variable, more of the points are closer to the y=x line.



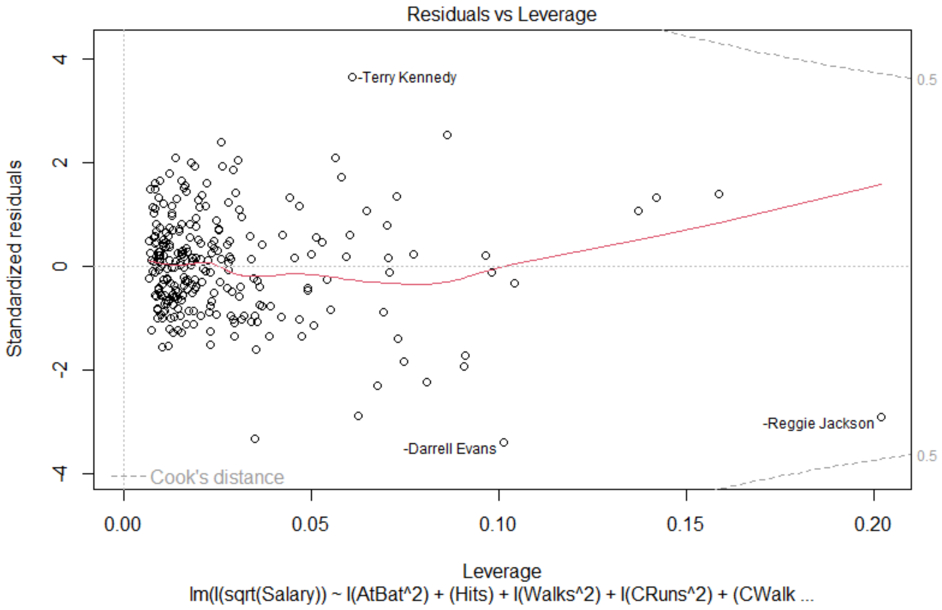
This results in the linear model,

```
lr1=lm(I(sqrt(Salary))~I(AtBat^2)+(Hits)+I(Walks^2)+I(CRuns^2)+(CWalks)+I(PutOuts^2), data=ds)
```

Outliers are found in the **Residuals vs Leverage** plot. High standardized residuals and leverages, more specifically residuals and leverages past the Cook's distance function, are indication of an outlier.



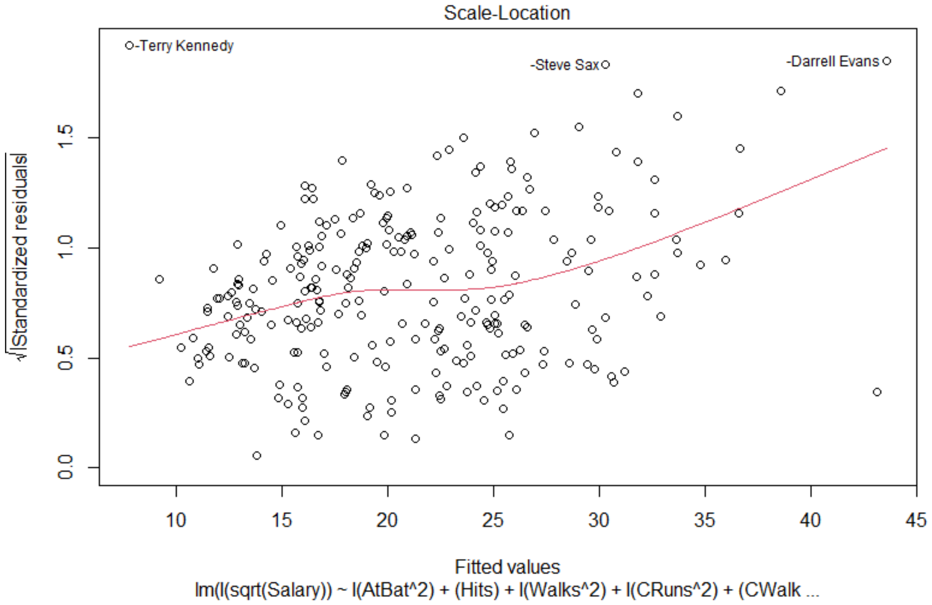
No points are on or past the Cook's distance function, but "Mike Schmidt" and "Pete Rose" exhibit an extremely high standardized residual and leverage, respectively.



Although there are points with high leverages and standardized residuals, they are of a lesser magnitude than the two previously removed outliers. I removed these points by creating a vector of the outlier names and remove these in the data I use for my linear model,

```
outliers = c('-Mike Schmidt', '-Pete Rose')
lr1=lm(I(sqrt(Salary))~I(AtBat^2)+(Hits)+I(Walks^2)+I(CRuns^2)+(
CWalks^2)+I(PutOuts^2), data=ds[!(row.names(ds) %in%
outliers),])
```

For **homoscedasticity**, we examine the Scale-Location plot. There is no discernable pattern in the spread of points, meaning that homoscedasticity is not violated.



Thus, my final model is

```
lr1=lm(I(sqrt(Salary))~I(AtBat^2)+(Hits)+I(Walks^2)+I(CRuns^2)+(
CWalks^2)+I(PutOuts^2), data=ds[!(row.names(ds) %in%
outliers),])
```

The final coefficients result in the equation,

$$\mathbf{Salary = 5.741 - 1.814e\text{-}5(AtBat)^2 + 1.248e\text{-}1(Hits) + 4.057e\text{-}5(Walks)^2 - 1.099e\text{-}6(CRuns)^2 + 1.908e\text{-}2(CWalks) + 3.300e\text{-}6(PutOuts)^2}$$

The adjusted R-squared for this model is 0.5058, and the multiple R-squared is 0.5172. This means the model explains around half of the variability in *Salary* and this model is a good fit for the data.

Conclusion

Creating a model for *Salary* variable in the "Hitters" dataset required creating a linear model using significant predictor variables and altering this linear model to fix any violated model assumptions and create a final model.

To find the significant variables, I created a linear model of all the numeric variables and used the "summary()" function on the linear model and chose the variables with low enough p-values. Interestingly, these variables did not align with the variables that had the most correlation to *Salary*.

In the linear model of these significant variables, linearity and normality were violated and were fixed with alterations to the function used in the linear model. In addition, there were outliers that were removed from the dataset.

The resulting function of my final model,
I(sqrt(Salary))~I(AtBat^2)+(Hits)+I(Walks^2)+I(CRuns^2)+(CWalks^2)+I(PutOuts^2),
had an adjusted R-squared value of 0.5058, which means that my final model is generally a good fit and can account for around half of the data.

However, although this model is a good fit, there is a lot of room for improvement in the adjusted R-squared value. Beyond errors in analysis that may have caused this, this can show that data sets are predictions and may not always accurately represent all the factors at play. For example, the popularity of a player could be a potential factor in their pay, which is something that is not in this dataset.