MTH-245 Project

Noah Johnson, Ian Martens, Caroline Vickery
"Walter's Ballpark Estimators'

1 Abstract

Background: Baseball has been called "America's Pastime" for generations, and this form of entertainment would not be possible without the players. Players build up statistical profiles based on their performance in games throughout the duration of their careers. These profiles can be helpful to team general managers when deciding how much to pay a player for their yearly salary. The purpose of this paper is to develop a model that best predicts a player's salary according to their performance statistics. **Methods:** We develop ten linear mixed models of varying levels of inclusion of variables and interaction terms. We begin with a first-order model with all terms in our dataset, then refine that model through stepwise AIC and interaction analyses. We prioritize the following variables: batting average, OBP, RBIs, hits, and home runs; however, our final model finds other variables to be more important predictors of salary. **Bottom Line:** After developing ten models, we conclude that Model 9 is our best model based on it having the highest RSE and Adjusted R^2 values.

2 Introduction

Baseball has provided countless hours of entertainment throughout America's history, contributing to culture and social relationships. Players dedicated to the game have become icons, racking up fame and fortune for providing an exciting and accessible form of entertainment. How should these players' contributions to the game and to culture be valued? Baseball players are employed by teams that pay them salaries, but how do the teams' general managers determine how much to pay their players? Players' performance can be evaluated on multiple criteria, including batting average, on-base percentage (OBP), RBIs, hits, and home runs.

In this paper, we aim to model how these and additional variables predict a baseball player's salary. We will explore whether or not better performing players have higher salaries, as well as if salary is a valid predictor of a player's on-field performance. We found our dataset, entitled "Pay for Play: Are Baseball Salaries Based on Performance?" from Journal of Statistics Education Data Archive. It enumerates Major League Baseball players' salaries from 1992 corresponding with their 1991 season performance statistics. The study was originally designed to assess the worth of performance statistics, as well as the financial benefit available to players if they change employers. We hypothesize that players' salaries will correlate highly with on-field performance, focusing on the following variables: batting average, OBP, RBIs, hits, and home runs. We anticipate that players with higher performance statistics in these categories will have higher salaries. Congruently, we expect players with high records of errors and strikeouts to have lower salaries.

The code below loads our data into `R` and formats it properly. Our data has 20 columns and 337 rows. We add column names and ensure the appropriate variables are read as a factors. The following variables are all continuous quantitative variables: salary, batting

average, OBP, runs, hits, doubles, triples, home runs, RBI, walks, strike outs, stolen bases, and errors. Our categorical variables are free agency eligibility, free agent standing, arbitration eligibility, and arbitration standing. Player name is also a categorical variable, but we exclude it from our models because it is more of a label than a variable.

```
baseball.datt <- read.csv("~/Downloads/baseball.dat.formatted2")
baseball.dat.okay <- baseball.datt %>%
  dplyr::select(-"X", -"free_agent_eligibility")
columns<-c("salary", "batting_avg", "OBP", "runs", "hits", "doubles",
  "triples", "home_runs", "RBI", "walks", "strike_outs",
  "stolen_bases", "errors", "free_agency_eligibility", "free_agent",
  "arbitration_eligibility", "arbitration", "name")
colnames(baseball.dat.okay)<-columns
dim(baseball.dat.okay)
```

```
## [1] 337  18
```

```
baseball.dat.okay$free_agent <- as.factor(baseball.dat.okay$free_agent)
baseball.dat.okay$free_agency_eligibility <-
  as.factor(baseball.dat.okay$free_agency_eligibility)
baseball.dat.okay$arbitration <- as.factor(baseball.dat.okay$arbitration)
baseball.dat.okay$arbitration_eligibility <-
  as.factor(baseball.dat.okay$arbitration_eligibility)
```

3 Exploratory Analysis

To familiarize ourselves with the data values and distribution, we calculated basic statistics and created plots, as shown below.

```
summarize(baseball.dat.okay, mean_sal = mean(salary), min_sal = min(salary),
          max_sal = max(salary), mean_BA = mean(batting_avg),
          min_BA = min(batting_avg), max_BA = max(batting_avg),
          mean_hits = mean(hits), min_hits = min(hits), max_hits = max(hits),
          mean_OBP = mean(OBP), min_OBP = min(OBP), max_OBP = max(OBP),
          mean_HR = mean(home_runs), min_HR = min(home_runs),
          max_HR = max(home_runs), mean_RBI = mean(RBI), min_RBI = min(RBI),
          max_RBI = max(RBI))

baseball.dat.okay %>% count(free_agency_eligibility)
baseball.dat.okay %>% count(free_agent)
baseball.dat.okay %>% count(arbitration)
baseball.dat.okay %>% count(arbitration_eligibility)
```

Variable	Mean	Minimum	Max
Salary	1248.528	109	6100
Batting Average	0.258	0.063	0.457
Hits	92.834	1	216
OBP	0.324	0.063	0.486
Home Runs	9.098	0	44
RBI	44.021	0	133

Table 1: Mean, minimum, and maximum values for variables of focus.

Variable	Yes	No	Percentage of Yeses
Free Agency Eligibility	134	203	39.763%
Free Agent	39	298	11.573%
Arbitration Eligibility	65	272	19.288%
Arbitration	10	327	2.97%

Table 2: Counts and percentages of yeses (considered a success) for categorical variables.

From Table 2, we see that the percentage of "successes"–defined as having received a 'yes' (or 1) in the appropriate categorical variable–are much smaller than the percentages of "failures." Furthermore, we see that, of those salaried baseball players eligible for free agency (134), a small number (39; 29.1%) actually have free-agent status. Of those salaried baseball players eligible for arbitration (65), only 10 (15.4%) have actually achieved arbitration status. Eligibility for free agency and arbitration are prerequisites for achieving free agent and arbitration status.

After making these Table 1, we decided to visually convey the distributions of all continuous quantitative variables (Figure 1). We divided the continuous quantitative variables between those reported as percentages and those reported as observed statistics. We included all quantitative variables in this analysis because we planned to evaluate models that include all quantitative variables in addition to our five primary quantitative variables.

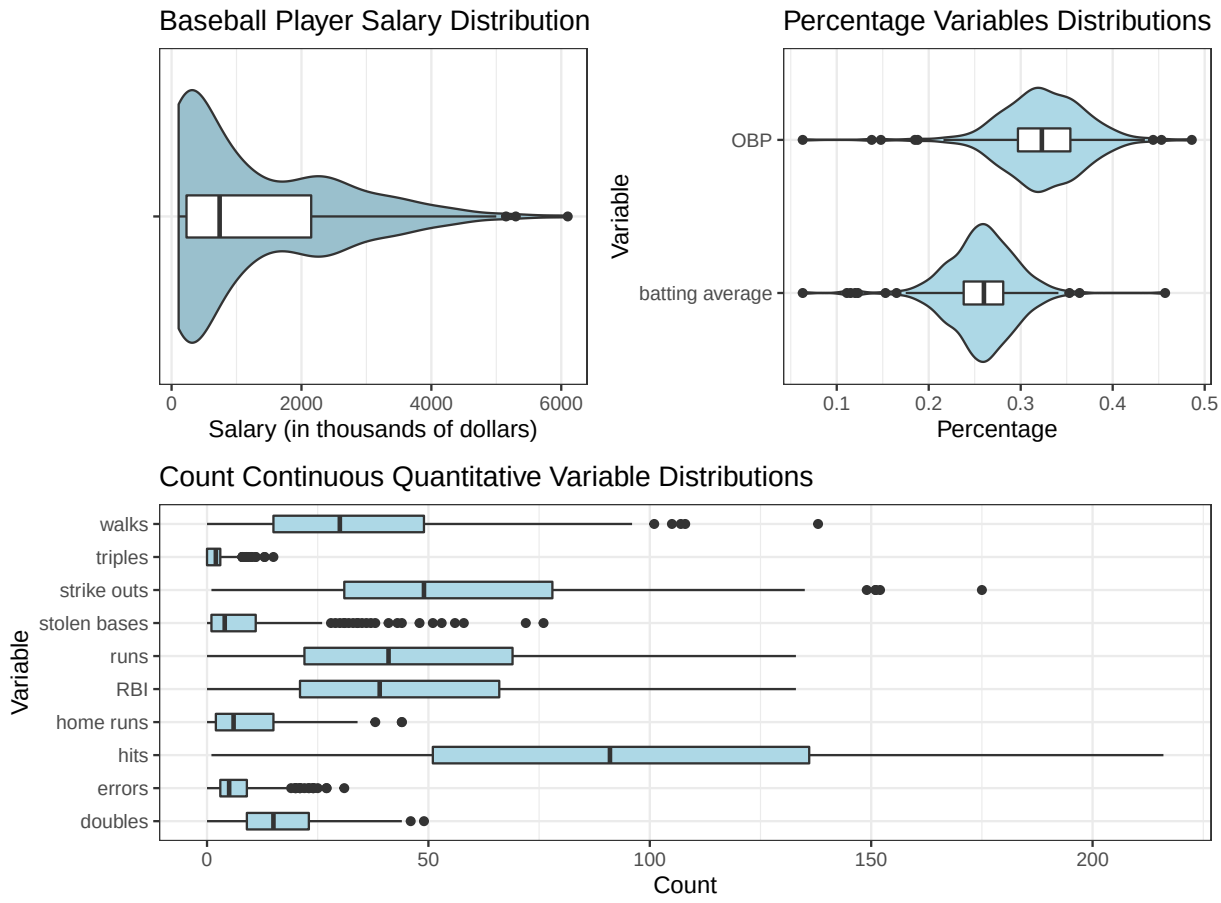

Figure 1: **Top Left:** A violin plot with an inset boxplot of baseball player salaries (in thousands of dollars) in 1992. **Top Right:** Violin plots with inset boxplots of variables conveyed as percentages. **Bottom:** A series of boxplots depicting the distributions of continuous quantitative variables.

From Figure 1, we see that the salary data has a severe right-skew, for which we will have to account later in the analysis. More players are paid lower salaries, which would be expected; the range of the top two quantiles (approximately 4852) is over four times larger than that of the lower half of the salaries (approximately 1139). Figure 4 further assesses the normality–rather, the lack thereof–of players' salaries. OBP and batting average are approximately Gaussian, and OBP has a higher mean value than batting average. From the bottom plot in Figure 1, we struggle to compare variables due to the inconsistency of data ranges. We, therefore, standardize the data to facilitate comparison between count continuous quantitative variables (Figure 2). The code below displays how we standardized these data. Describe the process: mean/sd. We also transformed our data into long format to facilitate plotting it in Figures 2 and 3.

```
numbers.dat.norm <- numbers.dat
numbers.dat.norm[, 1] <- numbers.dat[, 1]
numbers.dat.norm <- numbers.dat %>%
  mutate(runs = (numbers.dat[, 2] - mean(numbers.dat[, 2])) /
```

```
            sd(numbers.dat[, 2])) %>%
  mutate(hits = (numbers.dat[, 3] - mean(numbers.dat[, 3])) /
           sd(numbers.dat[, 3])) %>%
  mutate(doubles = (numbers.dat[, 4] - mean(numbers.dat[, 4])) /
           sd(numbers.dat[, 4])) %>%
  mutate(triples = (numbers.dat[, 5] - mean(numbers.dat[, 5])) /
           sd(numbers.dat[, 5])) %>%
  mutate(home_runs = (numbers.dat[, 6] - mean(numbers.dat[, 6])) /
           sd(numbers.dat[, 6])) %>%
  mutate(RBI = (numbers.dat[, 7] - mean(numbers.dat[, 7])) /
           sd(numbers.dat[, 7])) %>%
  mutate(walks = (numbers.dat[, 8] - mean(numbers.dat[, 8])) /
           sd(numbers.dat[, 8])) %>%
  mutate(strike_outs = (numbers.dat[, 9] - mean(numbers.dat[, 9])) /
           sd(numbers.dat[, 9])) %>%
  mutate(stolen_bases = (numbers.dat[, 10] - mean(numbers.dat[, 10])) /
           sd(numbers.dat[, 10])) %>%
  mutate(errors = (numbers.dat[, 11] - mean(numbers.dat[, 11])) /
           sd(numbers.dat[, 11]))

numbers.dat.norm.long <- numbers.dat.norm %>%
  rename("home runs" = "home_runs", "strike outs"="strike_outs",
         "stolen bases"="stolen_bases") %>%
  pivot_longer(cols = c("runs", "hits", "doubles", "triples",
                        "home runs", "RBI", "walks", "strike outs",
                        "stolen bases", "errors"), names_to = "number")
```

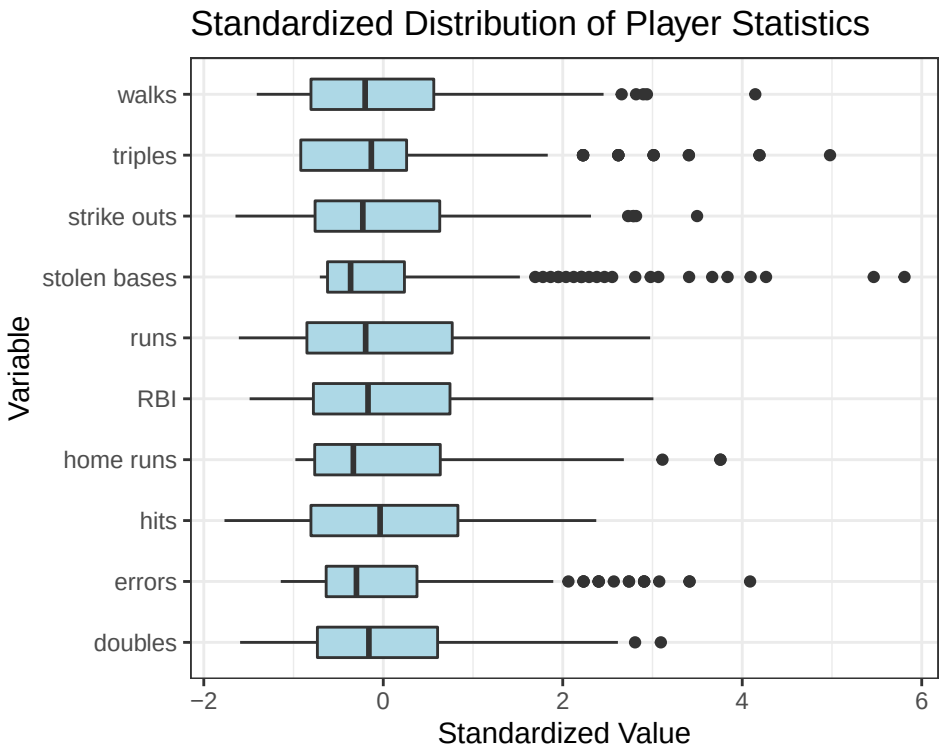

Figure 2: Standardized boxplot distributions of continuous quantitative variables originally communicated as count statistics.

These standardized statistics are now centered around zero, and from Figure 2, we see that many variables have a slight right skew. Among this group are the variables walks, triples, stolen bases, RBI, home runs, errors, and doubles. Strike outs, runs, and hits look approximately Gaussian according to our visual assessment, but we want to double check this assessment. Therefore, we create normal Q-Q plots of all standardized continuous quantitative variables (Figure 3). We also revisit here the distribution of salary by creating a separate Q-Q plot, further confirming our observation that salary is *not* normally distributed (Figure 4).

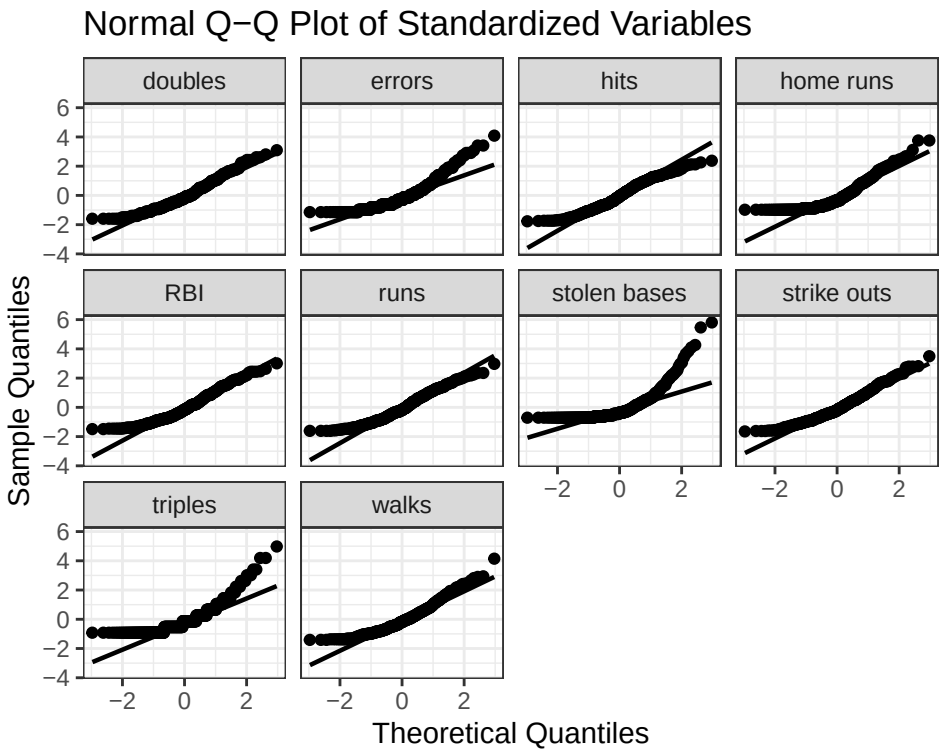

Figure 3: Normal Q-Q plots comparing theoretical to sample quantiles for standardized continuous quantitative variables.

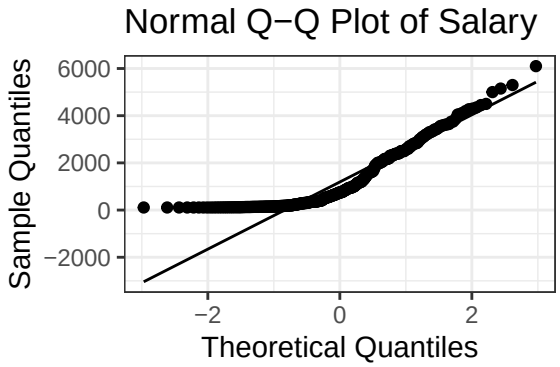

Figure 4: Normal Q-Q plot comparing theoretical to sample quantiles for baseball player salary.

We now assess the correlation by calculating Pearsons correlation coefficients between quantitative variables, using the normalized values for all but OBP, batting average, and salary. As seen in Figure 5, the only variables that are relatively correlated ($r > 0.60$) with salary are runs, hits, and RBI.

```
numbers.dat.cor <- numbers.dat.norm[, 2:11] %>%
  mutate(OBP = baseball.dat.okay$OBP,
         batting_avg = baseball.dat.okay$batting_avg,
         salary = baseball.dat.okay$salary) %>%
```

```
  rename("home runs"="home_runs",
         "strike outs"= "strike_outs",
         "stolen bases"="stolen_bases",
         "batting average" = "batting_avg")
```

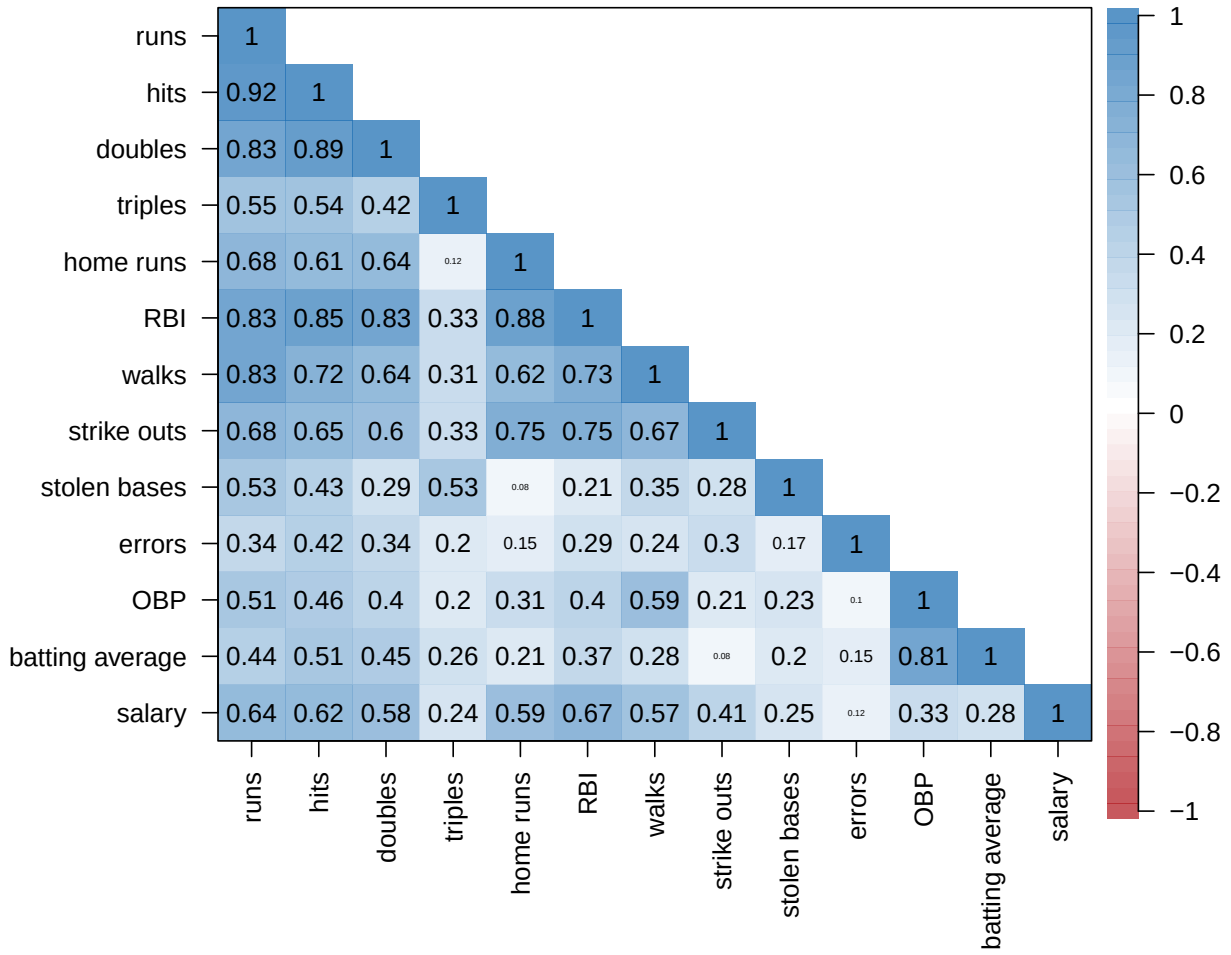

Figure 5: Pearsons correlation matrix of quantitative variables. Size of values and colors correspond to strength of correlation.

4 First-Order Model and Model Selection

4.1 First-Order Linear Model

We begin exploring the best model for this data with a first-order linear model with all predictor variables. Using the non-standardized data, the following equation gives the first-

order model:

$$\widehat{Y} = 223.115 + 3043.192I(\text{batting average}) - 3043.192I(\text{OBP}) + 7.1I(\text{runs})$$
$$-2.698I(\text{hits}) + 1.368I(\text{doubles}) - 17.922I(\text{triples}) + 19.483I(\text{home runs})$$
$$+17.415I(\text{RBI}) + 5.815I(\text{walks}) - 9.586I(\text{strike outs}) + 13.044I(\text{stolen bases})$$
$$-9.553I(\text{errors}) + 1372.886I(\text{free agency eligibility}) - 280.79I(\text{free agent})$$
$$+783.592I(\text{arbitration eligibility}) + 352.114I(\text{arbitration})$$

We will consider this Model 1.

```
baseball_all_model <- lm(salary ~ .-name,             #Model 1
                         data = baseball.dat.okay)
print(summary(baseball_all_model))
```

```
##
## Call:
## lm(formula = salary ~ . - name, data = baseball.dat.okay)
##
## 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_avg                3043.192   2712.536   1.122 0.262746
## OBP                       -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_agency_eligibility1   1372.886    108.594  12.642  < 2e-16 ***
## free_agent1                -280.790    137.640  -2.040 0.042168 *
## arbitration_eligibility1    783.592    118.289   6.624 1.48e-10 ***
## arbitration1                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
```

Using the standardized data, the following equation gives the first-order model, assigned as Model 2:

$$\widehat{Y} = 931.91 - 3043.19I(\text{batting average}) - 3528.01I(\text{OBP}) + 206.05I(\text{runs})$$
$$-140.03I(\text{hits}) + 14.3I(\text{doubles}) - 45.58I(\text{triples}) + 181I(\text{home runs})$$
$$+514.78I(\text{RBI}) + 144.45I(\text{walks}) - 324.27I(\text{strike outs}) + 152.15I(\text{stolen bases})$$
$$-56.63I(\text{errors}) + 1372.89I(\text{free agency eligibility}) - 280.79I(\text{free agent})$$
$$+783.592I(\text{arbitration eligibility}) + 352.11I(\text{arbitration})$$

```
numbers.dat.cor2<-numbers.dat.cor %>%
  mutate(free_agency_eligibility=baseball.dat.okay$free_agency_eligibility,
         free_agent=baseball.dat.okay$free_agent,
         arbitration_eligibility=baseball.dat.okay$arbitration_eligibility,
         arbitration=baseball.dat.okay$arbitration)

all_model <- lm(salary ~ ., data = numbers.dat.cor2)  #Model 2
print(summary(all_model))
```

```
##
## Call:
## lm(formula = salary ~ ., data = numbers.dat.cor2)
##
## 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)                 931.91     361.36   2.579 0.010357 *
## runs                        206.05     163.78   1.258 0.209259
## hits                       -140.03     171.86  -0.815 0.415788
## doubles                      14.30      90.00   0.159 0.873846
## triples                     -45.58      55.06  -0.828 0.408339
## `home runs`                 181.00     116.89   1.548 0.122506
## RBI                         514.78     149.81   3.436 0.000668 ***
## walks                       144.45     112.37   1.285 0.199548
## `strike outs`              -324.27      72.76  -4.457 1.15e-05 ***
## `stolen bases`              152.15      54.99   2.767 0.005988 **
## errors                      -56.63      44.46  -1.274 0.203693
## OBP                       -3528.01    2376.08  -1.485 0.138581
## `batting average`          3043.19    2712.54   1.122 0.262746
## free_agency_eligibility1   1372.89     108.59  12.642  < 2e-16 ***
## free_agent1                -280.79     137.64  -2.040 0.042168 *
## arbitration_eligibility1    783.59     118.29   6.624 1.48e-10 ***
```

```
## arbitration1               352.11     241.83   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
```

Table 3 gives the summary statistics of these first-order models, Model 1 and Model 2. We know that standardizing the data will not change R^2, adjusted-R^2, and RSE values, so Models 1 and 2 have the same values for these summary statistics. Their coefficients, however, do differ for some predictor variables. The first-order models start with a fairly high R^2 value (0.7014), indicating that a large portion of variability within the data can be explained with the first-order models. However, the RSE (694.3) is very large, which provides room for improvement. RSE Both models indicate significance for the following predictor variables: RBI, strike outs, stolen bases, free agency eligibility, free agent, and arbitration eligibility.

R-squared	Adjusted R-squared	RSE
0.7014	0.6865	694.3

Table 3: Summary statistics for first-order analysis Models 1 and 2.

While our summary statistics (Table 3) are relatively high, we must assess dependence of residual values on the response variable, salary, as well as the distribution of residuals. The first objective, we accomplish using the Breusch-Pagan test; the latter, we assess visually and with the Shapiro -Wilk test.

Breusch-Pagan Test: The Breusch-Pagan test assesses the hetero/homoskedasticity of residual values. Therefore, our hypotheses are as follows:

$$H_0 : \text{homoscedasticity of residuals}$$

$$H_a : \text{heteroscedasticity of residuals}$$

We reject the null hypothesis in favor of the alternative hypothesis, suggesting that there is sufficient evidence that the residuals are heteroscedastic ($bp = 74.434; p\text{-value} < 0.0001$).

```
(model1bp<-bptest(baseball_all_model))
```

```
##
##  studentized Breusch-Pagan test
##
## data:  baseball_all_model
## BP = 74.434, df = 16, p-value = 1.648e-09
```

Distribution of residuals: We now aim to assess whether the residuals are approximately Gaussian distributed or not. Therefore, our hypotheses are as follows:

$$H_0 : \text{the residuals are Gaussian distributed in the population}$$

$$H_a : \text{the residuals are } \textit{not} \text{ Gaussian distributed in the population}$$

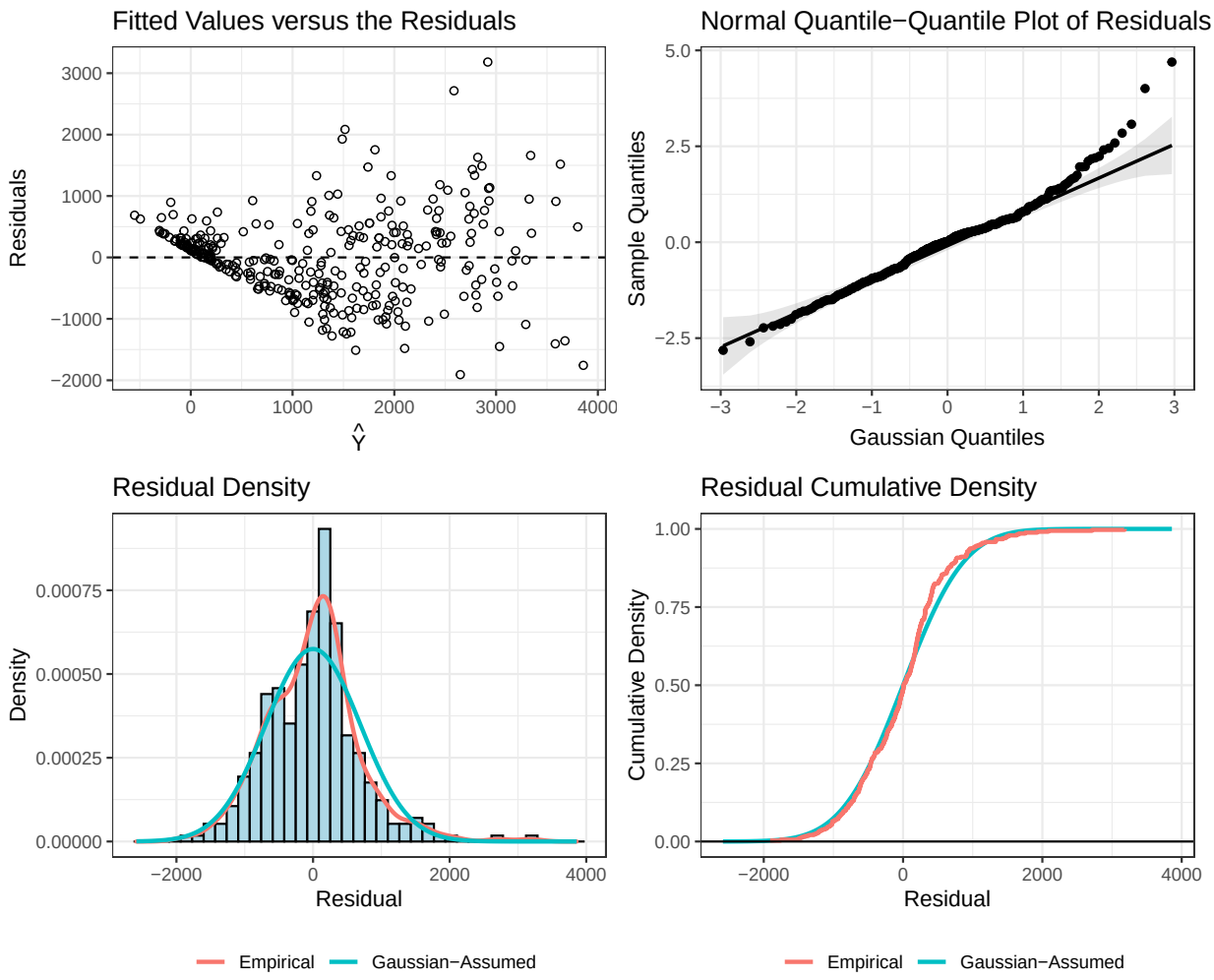

Figure 6: Side-by-side plots of residual values for Model 1.

From Figure 6, we see a megaphone shape to our residual plot, which indicates a lack of constant variance. We will revisit this later. More pertinently, we notice a slight deviation form normality in our q-q plot (top right of Figure 6). We will back up our visual assessment with the Shapiro-Wilk test so we can confidently accept or reject the null hypothesis.

```
(shapwk.model1<-shapiro.test(baseball_all_model$residuals))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  baseball_all_model$residuals
## W = 0.9727, p-value = 5.417e-06
```

From the code above, we see that there is sufficient evidence ($W = 0.9727, p\text{-value} < 0.0001$) to reject the null hypothesis in favor of the alternative, thereby suggesting that the residuals are not normally distributed in Model 1.

Multicollinearity: We calculate the Variance Inflation Factors (VIFs) of Model 1 to assess multicollinearity. We will consider VIFs between 1 and 5 to be *moderately correlated*, and VIFs greater than 5 to be *highly correlated*. From Table 4, we see that half of our predictor variables in Model 1 demonstrate high multicollinearity. While low multicollinearity would be ideal, we will focus on correcting other unmet assumptions in our models and see if multicollinearity will improve.

```
model1vif<-vif(baseball_all_model)
```

Variable	VIF	Correlation
Batting average	8.021	High
OBP	8.742	High
Runs	18.697	High
Hits	20.587	High
Doubles	5.646	High
Triples	2.113	Moderate
Home runs	9.524	High
RBI	15.645	High
Walks	8.802	High
Strike outs	3.690	Moderate
Stolen bases	2.108	Moderate
Errors	1.378	Moderate
Free agency eligibility	1.975	Moderate
Free agent	1.355	Moderate
Arbitration eligibility	1.523	Moderate
Arbitration	1.177	Moderate

Table 4: VIFs for Model 1 and their interpreted strength of correlation.

4.2 Tukey Analysis

We perform a Tukey Analysis to understand how to best transform our data in order to adjust our heavily skewed response variable, Y = salary. The Tukey Analysis reports $\lambda = 0.125$, which we interpret to suggest a transformation of $Y^{0.125}$. We transform and reevaluate our original first-order linear model (Model 1) in the next section.

```
baseball.dat.okay.tuk<-baseball.dat.okay
baseball.dat.okay.tuk$salary<-
  transformTukey(baseball.dat.okay$salary, , plotit=FALSE)
```

```
##
##     lambda     W Shapiro.p.value
## 406  0.125 0.936        7.26e-11
##
## if (lambda >  0){TRANS = x ^ lambda}
## if (lambda == 0){TRANS = log(x)}
## if (lambda <  0){TRANS = -1 * x ^ lambda}
```

4.3 Transformed First-Order Linear Model

```r
all_model_tuktransformed<-lm(salary ~ .-name, data = baseball.dat.okay.tuk)
print(summary(all_model_tuktransformed))     #Model 3    # C-Vick model
```

```
##
## Call:
## lm(formula = salary ~ . - name, data = baseball.dat.okay.tuk)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65055 -0.08063 -0.01601  0.09514  0.44060
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.9676277  0.0729115  26.987  < 2e-16 ***
## batting_avg                0.3478127  0.5944239   0.585 0.558876
## OBP                       -0.6839151  0.5206939  -1.313 0.189966
## runs                       0.0005472  0.0012367   0.442 0.658473
## hits                       0.0011532  0.0007257   1.589 0.113015
## doubles                   -0.0006213  0.0018870  -0.329 0.742180
## triples                   -0.0049796  0.0047437  -1.050 0.294630
## home_runs                  0.0018940  0.0027573   0.687 0.492643
## RBI                        0.0030074  0.0011107   2.708 0.007137 **
## walks                      0.0015011  0.0009912   1.514 0.130919
## strike_outs               -0.0017090  0.0004713  -3.626 0.000335 ***
## stolen_bases               0.0016356  0.0010331   1.583 0.114354
## errors                    -0.0023453  0.0016436  -1.427 0.154590
## free_agency_eligibility1   0.4493141  0.0237973  18.881  < 2e-16 ***
## free_agent1               -0.0784802  0.0301623  -2.602 0.009701 **
## arbitration_eligibility1   0.3594640  0.0259218  13.867  < 2e-16 ***
## arbitration1              -0.0067890  0.0529943  -0.128 0.898144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1521 on 320 degrees of freedom
## Multiple R-squared:  0.8021,Adjusted R-squared:  0.7922
## F-statistic: 81.07 on 16 and 320 DF,  p-value: < 2.2e-16
```

The following equation gives the the following equation gives the Tukey-transformed first-order model, assigned as Model 3:

$$\widehat{Y} = 1.968 - 0.348I(\text{batting average}) - 0.684I(\text{OBP}) + 0.001I(\text{runs}) + 0.001I(\text{hits})$$
$$-0.001I(\text{doubles}) - 0.005I(\text{triples}) + 0.002I(\text{home runs}) + 0.003I(\text{RBI}) + 0.002I(\text{walks})$$
$$-0.002I(\text{strike outs}) + 0.002I(\text{stolen bases}) - 0.002I(\text{errors}) + 0.449I(\text{free agency eligibility})$$
$$-0.078I(\text{free agent}) + 0.359I(\text{arbitration eligibility}) - 0.007I(\text{arbitration})$$

Table 5 gives the summary statistics of the Tukey-transformed first-order model, Model 3. Model three has a higher R^2 (0.8021) than Models 1 and 2 (0.7014). The RSE (0.1521) is

much smaller than the RSE for Models 1 and 2 (694.3), indicating an improvement from the first two models to Model 3. Model three indicates that the following variables are significant: RBI, strike outs, free agency eligibility, free agent, and arbitration eligibility. Models 1 and 2 indicate additional significance for stolen bases.

R-squared	Adjusted R-squared	RSE
0.8021	0.7922	0.1521

Table 5: Summary statistics for first-order analysis Model 3.

For Model 3, we repeat the Breusch-Pagan and Shapiro-Wilk tests.

Breusch-Pagan Test: Our hypotheses are as follows:

$$H_0 : \text{homoscedasticity of residuals}$$

$$H_a : \text{heteroscedasticity of residuals}$$

Under a 5% significance level, we fail to reject the null hypothesis ($bp = 26.245; p\text{-value} = 0.507$). We, therefore, have sufficient evidence that the residuals are homoscedastic.

```
(model3bp<-bptest(all_model_tuktransformed))
```

```
##
##  studentized Breusch-Pagan test
##
## data:  all_model_tuktransformed
## BP = 26.245, df = 16, p-value = 0.05067
```

Distribution of Residuals: Our hypotheses are as follows:

$$H_0 : \text{the residuals are Gaussian distributed in the population}$$

$$H_a : \text{the residuals are } \textit{not} \text{ Gaussian distributed in the population}$$

```
(shapwk.model3<-shapiro.test(all_model_tuktransformed$residuals))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  all_model_tuktransformed$residuals
## W = 0.96644, p-value = 5.112e-07
```

From the code above, we see that there is sufficient evidence ($W = 0.9664, p\text{-value} < 0.0001$) to reject the null hypothesis in favor of the alternative, thereby suggesting that the residuals are not Gaussian distributed in Model 3. The normal Q-Q plot in Figure 7 supports this conclusion.

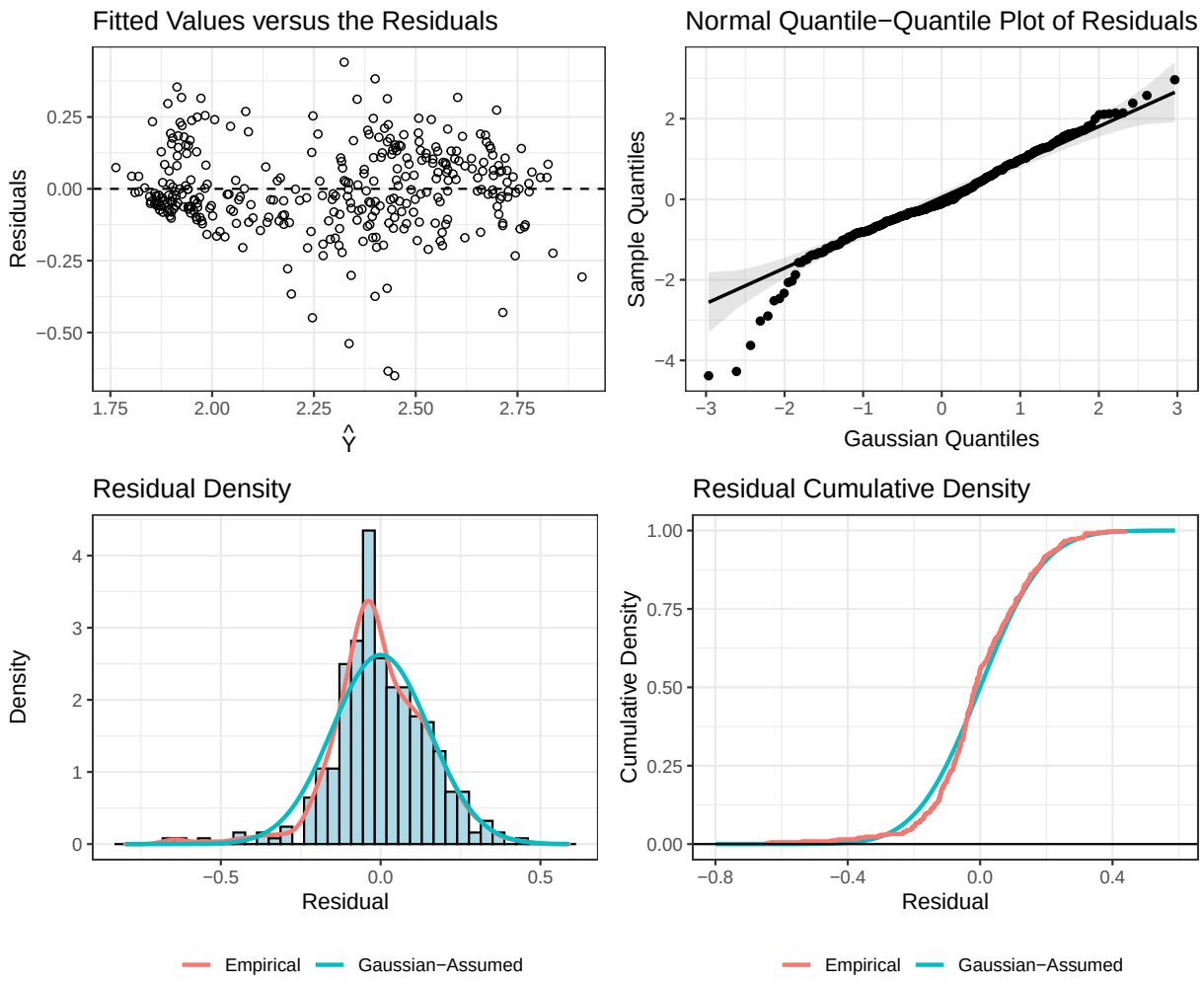

Figure 7: Side-by-side plots of residual values for Model 3.

Multicollinearity: We again calculate the Variance Inflation Factors (VIFs) of Model 3 to assess multicollinearity. We will consider VIFs between 1 and 5 to be *moderately correlated*, and VIFs greater than 5 to be *highly correlated*. We get the same VIFs as were found for Model 1, listed in Table 4. We see that half of our predictor variables in Model 1 demonstrate high multicollinearity. WHY IS THIS??

```
model3vif<-vif(all_model_tuktransformed)
```

Summary of diagnostics: Model 3 (the Tukey-transformed model) has the highest R^2(0.8021) and lowest RSE (0.1521) of Models 1-4 (the last of which will be explored below). The homoscedasticity of residuals assumption is met, but we cannot conclude that the residuals are Gaussian distributed. Figure 7, however, illustrates that the majority of residual Q-Q points follow a linear relationship, with only the lowest points deviating from normality. **We, therefore, decide to remove the bottom 10% of salary data points and focus on finding a model for the top 90% of salaries in the distribution. This will be explored in the next subsection.**

4.4 Tukey Transformed Model for top 90% of Salaries

The following code selects the top 90% of our original data. We then Tukey-transform this data and visualize it (Figure 8). The Tukey Analysis suggests a transformation of $Y^{0.175}$. The summary statistics for this model are given in Table 6. This Tukey-transformed model of the top 90% of salary data will be designated as Model 4.

```
top90.dat.okay<-baseball.dat.okay %>%
  filter(baseball.dat.okay$salary > quantile(baseball.dat.okay$salary, 0.1))
top90.dat.okay$salary<-
  transformTukey(top90.dat.okay$salary, , plotit=FALSE)
```

```
##
##     lambda      W Shapiro.p.value
## 408  0.175 0.9484       9.082e-09
##
## if (lambda >  0){TRANS = x ^ lambda}
## if (lambda == 0){TRANS = log(x)}
## if (lambda <  0){TRANS = -1 * x ^ lambda}
```

```
top90_all_tuk<-lm(salary ~ .-name, data = top90.dat.okay) #Model 4
print(summary(top90_all_tuk))
```

```
##
## Call:
## lm(formula = salary ~ . - name, data = top90.dat.okay)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77148 -0.18241 -0.01516  0.19598  0.80504
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.5711141  0.1585768  16.214  < 2e-16 ***
## batting_avg                 1.1187727  1.5591780   0.718  0.47363
## OBP                        -1.4561193  1.3678690  -1.065  0.28800
## runs                        0.0030866  0.0025262   1.222  0.22279
## hits                        0.0003782  0.0015343   0.247  0.80546
## doubles                    -0.0005872  0.0035361  -0.166  0.86823
## triples                    -0.0083178  0.0088207  -0.943  0.34649
## home_runs                   0.0026067  0.0052214   0.499  0.61801
## RBI                         0.0054910  0.0020674   2.656  0.00836 **
## walks                       0.0023428  0.0022174   1.057  0.29162
## strike_outs                -0.0022867  0.0009182  -2.490  0.01333 *
## stolen_bases                0.0039409  0.0019143   2.059  0.04044 *
## errors                     -0.0055813  0.0030588  -1.825  0.06911 .
## free_agency_eligibility1    0.9082903  0.0454055  20.004  < 2e-16 ***
## free_agent1                -0.1516931  0.0560000  -2.709  0.00716 **
## arbitration_eligibility1    0.7138653  0.0491550  14.523  < 2e-16 ***
```

```
## arbitration1              0.0770864  0.1027680   0.750  0.45382
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2786 on 283 degrees of freedom
## Multiple R-squared:  0.7992,Adjusted R-squared:  0.7878
## F-statistic: 70.39 on 16 and 283 DF,  p-value: < 2.2e-16
```

R-squared	Adjusted R-squared	RSE
0.7992	0.7878	0.2786

Table 6: Summary statistics for Model 4, which uses the top 90% of salary data.

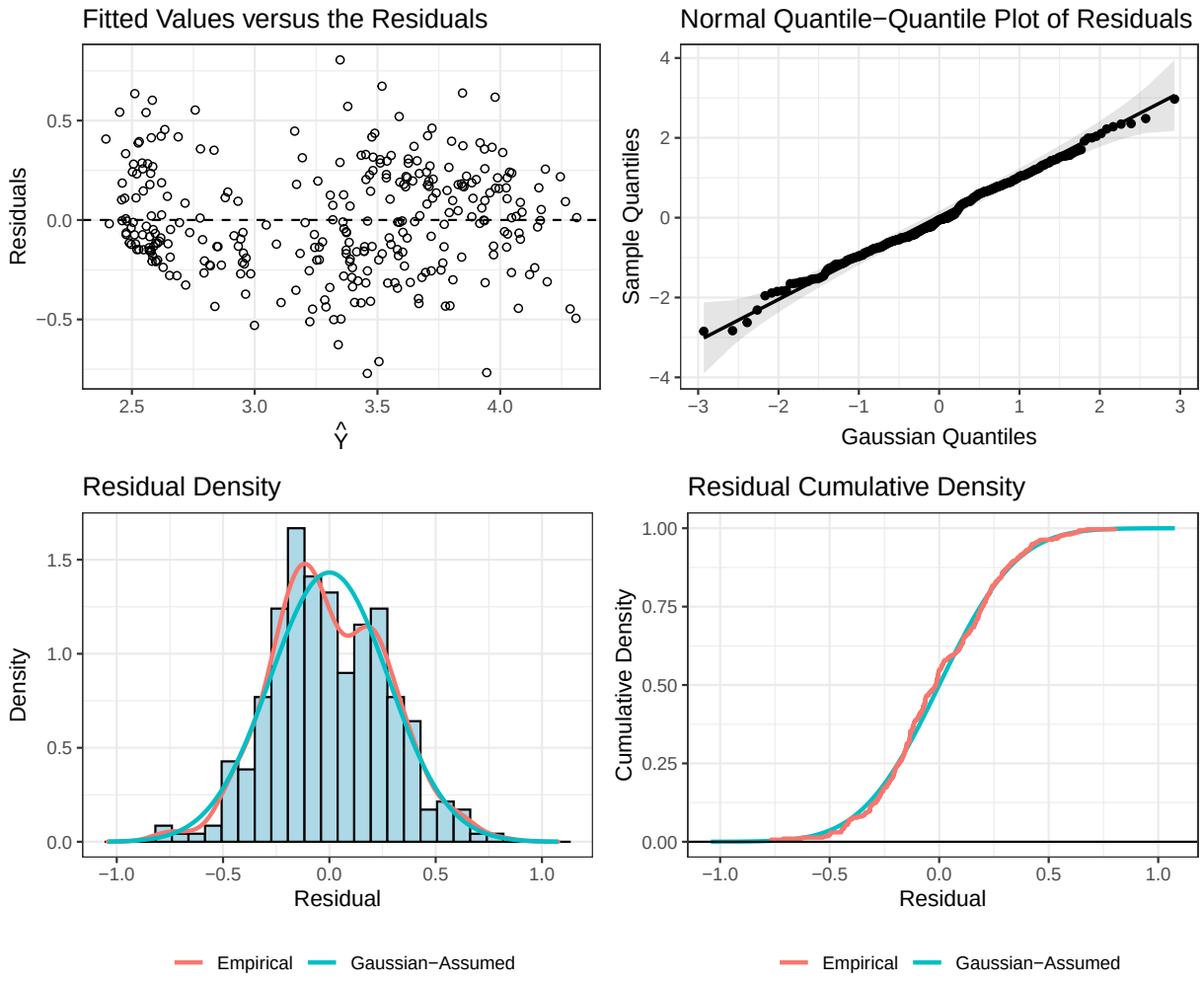

Figure 8: Side-by-side plots of residual values for Model 4.

Visually, we see that the homoscedasticity and normality assumptions for the residuals for Model 4 are much better than those of Models 1-3. We repeat the Breusch-Pagan and

Shapiro-Wilk tests to affirm our visual assessments.

Breusch-Pagan Test: Our hypotheses are as follows:

$$H_0 : \text{homoscedasticity of residuals}$$

$$H_a : \text{heteroscedasticity of residuals}$$

Under a 5% significance level, we fail to reject the null hypothesis ($bp = 22.307; p\text{-value} = 0.1335$). We, therefore, have sufficient evidence that the residuals are homoscedastic.

```
(model4bp<-bptest(top90_all_tuk))
```

```
##
##  studentized Breusch-Pagan test
##
## data:  top90_all_tuk
## BP = 22.307, df = 16, p-value = 0.1335
```

Distribution of Residuals: Our hypotheses are as follows:

$$H_0 : \text{the residuals are Gaussian distributed in the population}$$

$$H_a : \text{the residuals are } \textit{not} \text{ Gaussian distributed in the population}$$

From the code above, we fail to reject the null hypothesis ($W = 0.9664, p\text{-value} = 0.3441$), thereby suggesting that the residuals are Gaussian distributed in Model 4. The normal Q-Q plot in Figure 8 supports this conclusion.

```
(shapwk.model4<-shapiro.test(top90_all_tuk$residuals))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  top90_all_tuk$residuals
## W = 0.99443, p-value = 0.3441
```

Multicollinearity: We again calculate the Variance Inflation Factors (VIFs) of Model 4 to assess multicollinearity. We will consider VIFs between 1 and 5 to be *moderately correlated*, and VIFs greater than 5 to be *highly correlated*. From Table 7, we see that the same variables are considered highly correlated for Model 4 as were highly correlated for Models 1-3.

```
model4vif<-vif(top90_all_tuk)
```

Variable	VIF	Correlation
Batting average	12.913	High
OBP	14.515	High
Runs	18.730	High
Hits	20.813	High
Doubles	4.823	High
Triples	1.996	Moderate
Home runs	9.237	High
RBI	13.182	High
Walks	11.209	High
Strike outs	3.354	Moderate
Stolen bases	2.021	Moderate
Errors	1.279	Moderate
Free agency eligibility	1.960	Moderate
Free agent	1.341	Moderate
Arbitration eligibility	1.550	Moderate
Arbitration	1.188	Moderate

Table 7: VIFs for Model 4 and their interpreted strength of correlation.

Summary of diagnostics: For Model 4, which predicts the top 90% of salary data, the homoscedasticity and Gaussian distribution assumptions are met for the residuals. These two attributes are desired. However, we see high multicollinearity (VIF$>$ 5) for half of our predictor variables. In the next section, we run some interaction analyses to try to improve our model, as well as employ model selection techniques.

5 Interaction Analysis and Model Selection

5.1 Initial Variables of Interest

Based on our Introduction, we claim that we want to focus on the following variables: batting average, OBP, RBI, hits, and home runs. We run this model below, with both no transformation on salary and a $Y^{0.175}$ Tukey transformation. We see in table 8 that this manual selection of variables results in by far the worst summary statistics of any model in this analysis. We, therefore, abandon these models, bypassing the assessment of assumptions to try more statistically rigorous variable selection techniques.

```
var.of.int.nottuk<-lm(salary ~ batting_avg + OBP + RBI + hits + home_runs,
         data=top90.dat.okay.nottuk)  #no Tukey transformation
summary(var.of.int.nottuk)
```

```
##
## Call:
## lm(formula = salary ~ batting_avg + OBP + RBI + hits + home_runs,
##     data = top90.dat.okay.nottuk)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2740.8  -572.5   -32.3   526.8  3356.2
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -284.474    439.747  -0.647  0.51820
## batting_avg -4826.580   2859.103  -1.688  0.09244 .
## OBP          4391.174   2219.252   1.979  0.04879 *
## RBI            10.494      6.593   1.592  0.11252
## hits            7.689      2.523   3.048  0.00252 **
## home_runs      21.201     14.162   1.497  0.13544
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 953.1 on 294 degrees of freedom
## Multiple R-squared:  0.4242,Adjusted R-squared:  0.4144
## F-statistic: 43.31 on 5 and 294 DF,  p-value: < 2.2e-16
```

```
var.of.int.tuk<-lm(salary ~ batting_avg + OBP + RBI + hits + home_runs,
         data=top90.dat.okay)  #Y^0.175 Tukey transformation
summary(var.of.int.tuk)
```

```
##
## Call:
## lm(formula = salary ~ batting_avg + OBP + RBI + hits + home_runs,
##     data = top90.dat.okay)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.17509 -0.33077  0.05705  0.35940  1.17114
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.599814   0.217690  11.943  < 2e-16 ***
## batting_avg -3.388957   1.415359  -2.394   0.0173 *
## OBP          2.522704   1.098610   2.296   0.0224 *
## RBI          0.002717   0.003264   0.832   0.4058
## hits         0.005595   0.001249   4.480 1.07e-05 ***
## home_runs    0.007679   0.007011   1.095   0.2742
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4718 on 294 degrees of freedom
## Multiple R-squared:  0.4016,Adjusted R-squared:  0.3914
## F-statistic: 39.46 on 5 and 294 DF,  p-value: < 2.2e-16
```

Model	R-squared	Adj R-squared	RSE
No Tranformation	0.4242	0.4144	953.1
Tukey Transformation	0.4016	0.3914	0.4718

Table 8: A comparison of summary statistics for linear models that only include our specified variables of interest.

5.2 Model Selection Techniques

For our model selection techniques, we perform a stepwise AIC for the Tukey transformed model of the top 90% of salaries, Model 4. The outputs are summarized in Table 9.

```
library(MASS)
#AIC, BIC, and Mallows for Model 4
AIC(top90_all_tuk)

## [1] 103.0338

BIC(top90_all_tuk)

## [1] 169.7019

ols_mallows_cp(top90_all_tuk, top90_all_tuk)

## [1] 17

#backward selection
backward_90_tuk<-stepAIC(top90_all_tuk,
                         direction="backward", trace = FALSE)
summary(backward_90_tuk)

##
## Call:
## lm(formula = salary ~ runs + RBI + strike_outs + stolen_bases +
##     errors + free_agency_eligibility + free_agent + arbitration_eligibility,
##     data = top90.dat.okay)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81375 -0.17261 -0.03649  0.20588  0.78157
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.4005599  0.0412063  58.257  < 2e-16 ***
## runs                        0.0037601  0.0013200   2.849  0.00470 **
## RBI                         0.0065753  0.0012173   5.402 1.37e-07 ***
## strike_outs                -0.0019175  0.0007279  -2.634  0.00889 **
## stolen_bases                0.0031966  0.0017882   1.788  0.07488 .
## errors                     -0.0052247  0.0028944  -1.805  0.07210 .
## free_agency_eligibility1    0.9158626  0.0428458  21.376  < 2e-16 ***
## free_agent1                -0.1537070  0.0549663  -2.796  0.00551 **
## arbitration_eligibility1    0.7259503  0.0462356  15.701  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2766 on 291 degrees of freedom
## Multiple R-squared:  0.7964,Adjusted R-squared:  0.7908
## F-statistic: 142.3 on 8 and 291 DF,  p-value: < 2.2e-16
```

```
AIC(backward_90_tuk)
```

```
## [1] 91.15056
```

```
BIC(backward_90_tuk)
```

```
## [1] 128.1884
```

```
ols_mallows_cp(backward_90_tuk, top90_all_tuk)
```

```
## [1] 4.910252
```

```
#forward selection
intercept.model<-lm(salary ~1, data=top90.dat.okay)
forward_90_tuk<-stepAIC(intercept.model,
        direction="forward",
        scope=list(upper=top90_all_tuk),
        trace=FALSE)
summary(forward_90_tuk)
```

```
##
## Call:
## lm(formula = salary ~ free_agency_eligibility + arbitration_eligibility +
##     runs + RBI + free_agent + strike_outs + errors + stolen_bases,
##     data = top90.dat.okay)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81375 -0.17261 -0.03649  0.20588  0.78157
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.4005599  0.0412063  58.257  < 2e-16 ***
## free_agency_eligibility1   0.9158626  0.0428458  21.376  < 2e-16 ***
## arbitration_eligibility1   0.7259503  0.0462356  15.701  < 2e-16 ***
## runs                       0.0037601  0.0013200   2.849  0.00470 **
## RBI                        0.0065753  0.0012173   5.402 1.37e-07 ***
## free_agent1               -0.1537070  0.0549663  -2.796  0.00551 **
## strike_outs               -0.0019175  0.0007279  -2.634  0.00889 **
## errors                    -0.0052247  0.0028944  -1.805  0.07210 .
## stolen_bases               0.0031966  0.0017882   1.788  0.07488 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.2766 on 291 degrees of freedom
## Multiple R-squared:  0.7964,Adjusted R-squared:  0.7908
## F-statistic: 142.3 on 8 and 291 DF,  p-value: < 2.2e-16
```

```
AIC(forward_90_tuk)
```

```
## [1] 91.15056
```

```
BIC(forward_90_tuk)
```

```
## [1] 128.1884
```

```
ols_mallows_cp(forward_90_tuk, top90_all_tuk)
```

```
## [1] 4.910252
```

```
#stepwise selection
both_90_tuk<-stepAIC(intercept.model,
                     direction="both",
                     scope=list(upper=top90_all_tuk),
                     trace=FALSE)
summary(both_90_tuk)
```

```
##
## Call:
## lm(formula = salary ~ free_agency_eligibility + arbitration_eligibility +
##     runs + RBI + free_agent + strike_outs + errors + stolen_bases,
##     data = top90.dat.okay)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81375 -0.17261 -0.03649  0.20588  0.78157
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.4005599  0.0412063  58.257  < 2e-16 ***
## free_agency_eligibility1   0.9158626  0.0428458  21.376  < 2e-16 ***
## arbitration_eligibility1   0.7259503  0.0462356  15.701  < 2e-16 ***
## runs                       0.0037601  0.0013200   2.849  0.00470 **
## RBI                        0.0065753  0.0012173   5.402 1.37e-07 ***
## free_agent1               -0.1537070  0.0549663  -2.796  0.00551 **
## strike_outs               -0.0019175  0.0007279  -2.634  0.00889 **
## errors                    -0.0052247  0.0028944  -1.805  0.07210 .
## stolen_bases               0.0031966  0.0017882   1.788  0.07488 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2766 on 291 degrees of freedom
## Multiple R-squared:  0.7964,Adjusted R-squared:  0.7908
```

```
## F-statistic: 142.3 on 8 and 291 DF,  p-value: < 2.2e-16
AIC(both_90_tuk)
## [1] 91.15056
BIC(both_90_tuk)
## [1] 128.1884
ols_mallows_cp(both_90_tuk, top90_all_tuk)
## [1] 4.910252
```

Model	R-squared	Adj R-squared	RSE	AIC	BIC	Mallows
Model 4	0.7992	0.7878	0.2786	103.0338	169.7019	17.0000
Backward	0.7964	0.7908	0.2766	91.1506	128.1884	4.9103
Forward	0.7964	0.7908	0.2766	91.1506	128.1884	4.9103
Stepwise (Model 5)	0.7964	0.7908	0.2766	91.1506	128.1884	4.9103

Table 9: A comparison of summary statistics for Model 4 and variable seletions in the backward, forward, and stepwise directions.

From these selection techniques, we see that all directions result in the same statistics. However, the summary statistics for the directionally selected models are all worse than those for Model 4. Selection techniques deem the following variables significant (under a 5% significance level): free agency eligibility, arbitration eligibility, runs, RBI, free agent, and strike outs. Errors and stolen bases are significant under a 10% significance level. The section below provides a summary of the model the stepwise model achieved through AIC.

5.3 Stepwise Model and Diagnostics

Designated as Model 5, the model below includes the variables selected by all directions of AIC, as all directions resulted in the same summary statistics (Table 9). We then assess the assumptions below again using the Breusch-Pagan and Shapiro-Wilk tests.

```
stepwise_top90 <- lm(salary ~ runs + RBI + strike_outs + stolen_bases
                     + errors + free_agency_eligibility + free_agent
                     + arbitration_eligibility,
                     data = top90.dat.okay) #Model 5
```

The following equation expresses the results of Model 5:

$$\widehat{Y} = 2.4006 + 0.0038I(\text{runs}) + 0.0066I(\text{RBI})$$
$$-0.0019I(\text{strike outs}) + 0.0032I(\text{stolen bases})$$
$$- - 0.0052I(\text{errors}) + 0.9159I(\text{free agency eligibility})$$
$$-0.1537I(\text{free agent}) + 0.7260I(\text{arbitration eligibility})$$

Breusch-Pagan Test: Our hypotheses are as follows:

$$H_0 : \text{homoscedasticity of residuals}$$

$$H_a : \text{heteroscedasticity of residuals}$$

Under a 5% significance level, we fail to reject the null hypothesis ($bp = 14.989; p\text{-value} = 0.5936$). We, therefore, have sufficient evidence that the residuals are homoscedastic.

```
bptest(stepwise_top90)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  stepwise_top90
## BP = 14.989, df = 8, p-value = 0.05936
```

Distribution of Residuals: Our hypotheses are as follows:

$$H_0 : \text{the residuals are Gaussian distributed in the population}$$

$$H_a : \text{the residuals are } \textit{not} \text{ Gaussian distributed in the population}$$

From the code above, we fail to reject the null hypothesis ($W = 0.9953, p\text{-value} = 0.5058$), thereby suggesting that the residuals are Gaussian distributed in Model 4. The normal Q-Q plot in Figure 9 supports this conclusion.

```
shapiro.test(stepwise_top90$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  stepwise_top90$residuals
## W = 0.99534, p-value = 0.5058
```

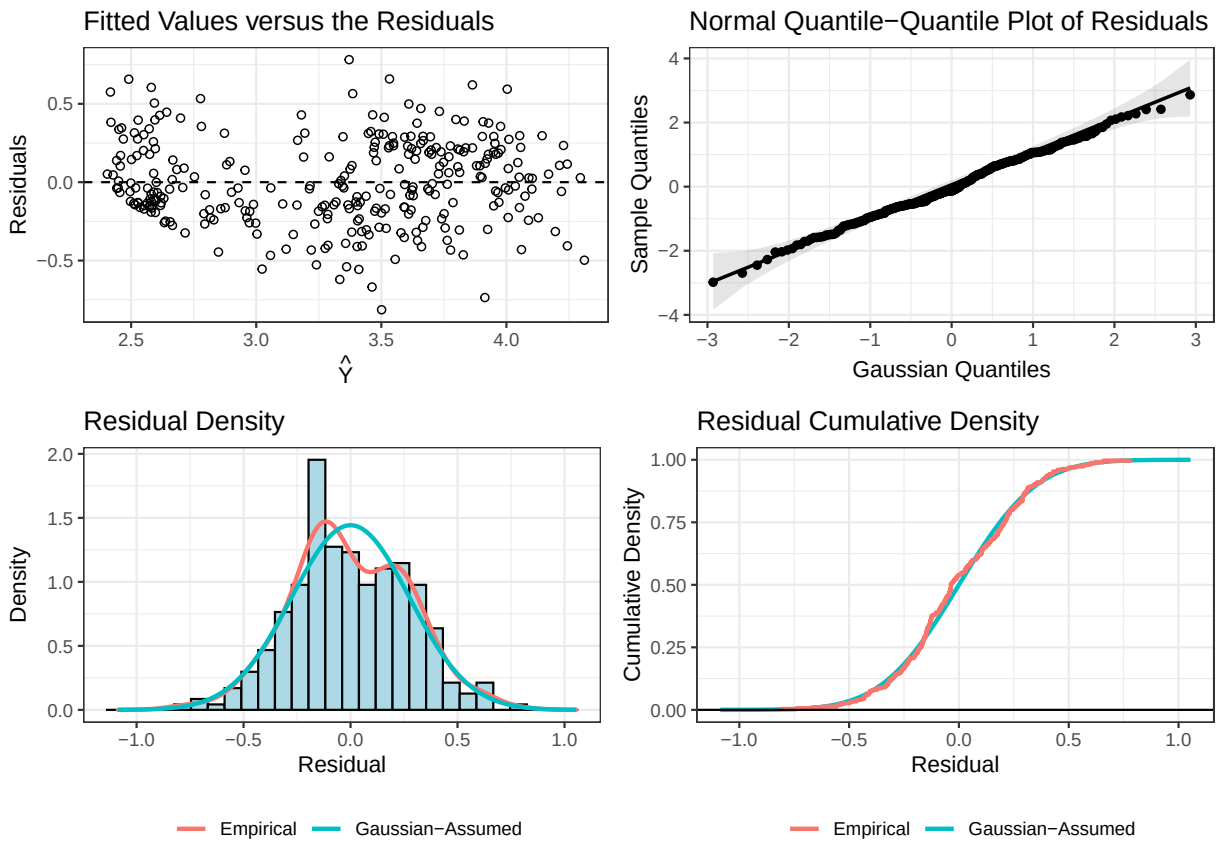

Figure 9: Side-by-side plots of residual values for Model 5.

Multicollinearity: We again calculate the Variance Inflation Factors (VIFs) of Model 5 to assess multicollinearity. We will consider VIFs between 1 and 5 to be *moderately correlated*, and VIFs greater than 5 to be *highly correlated*. From Table 10, we see that the multicollinearity in Model 5 is much more moderate than found in previous models. In the next section, we will run some interaction analyses to see if we can improve multicollinearity even more.

```
vif(stepwise_top90)
```

Variable	VIF	Correlation
Runs	5.187	High
RBI	4.635	Moderate
Strike outs	2.138	Moderate
Stolen bases	1.789	Moderate
Errors	1.162	Moderate
Free agency eligibility	1.771	Moderate
Free agent	1.310	Moderate
Arbitration eligibility	1.390	Moderate

Table 10: VIFs for Model 5 and their interpreted strength of correlation.

5.4 Full Interaction Model

We will now run a full interaction model with the variables selected by the stepwise AIC. We recall that this data is transformed according to the Tukey Analysis so that $Y^{0.175}$. Summary statistics are given in Table 11, and we again suppress the code used to find these table values for the sake of space. We will denote this model as Model 6.

```
top90.fullint<-lm(salary~runs*RBI +        #Model 6
                    runs*strike_outs +
                    runs*stolen_bases+
                    runs*errors +
                    runs*free_agency_eligibility +
                    runs*free_agent +
                    runs*arbitration_eligibility +
                    RBI*strike_outs +
                    RBI*stolen_bases +
                    RBI*errors +
                    RBI*free_agency_eligibility+
                    RBI*free_agent+
                    RBI*arbitration_eligibility +
                    strike_outs*stolen_bases +
                    strike_outs*errors +
                    strike_outs*free_agency_eligibility +
                    strike_outs*free_agent +
                    strike_outs*arbitration_eligibility +
                    stolen_bases*errors +
                    stolen_bases*free_agency_eligibility+
                    stolen_bases*free_agent +
                    stolen_bases*arbitration_eligibility+
                    errors*free_agency_eligibility +
                    errors*free_agent +
                    errors*arbitration_eligibility +
                    free_agency_eligibility*free_agent +
                    free_agency_eligibility*arbitration_eligibility+
                    free_agent*arbitration_eligibility,
                  data = top90.dat.okay)
top90.fullint$coefficients
```

```
##                                       (Intercept)
##                                      2.509874e+00
##                                              runs
##                                      9.103699e-03
##                                               RBI
##                                      2.657328e-03
##                                       strike_outs
##                                     -2.618567e-03
##                                       stolen_bases
##                                      4.612324e-03
```

```
##                                          errors
##                                   -9.363464e-03
##                         free_agency_eligibility1
##                                    4.834780e-01
##                                     free_agent1
##                                   -2.276768e-01
##                         arbitration_eligibility1
##                                    4.020240e-01
##                                        runs:RBI
##                                    8.040597e-06
##                                 runs:strike_outs
##                                   -9.889548e-05
##                                runs:stolen_bases
##                                   -6.179609e-05
##                                     runs:errors
##                                    1.269013e-04
##                       runs:free_agency_eligibility1
##                                   -1.260218e-04
##                                  runs:free_agent1
##                                    7.088413e-04
##                       runs:arbitration_eligibility1
##                                   -2.142209e-03
##                                   RBI:strike_outs
##                                    5.107975e-05
##                                  RBI:stolen_bases
##                                   -9.743425e-05
##                                     RBI:errors
##                                   -3.110480e-04
##                        RBI:free_agency_eligibility1
##                                    1.891379e-03
##                                   RBI:free_agent1
##                                    8.891635e-03
##                        RBI:arbitration_eligibility1
##                                    4.299447e-03
##                             strike_outs:stolen_bases
##                                    4.965054e-05
##                                  strike_outs:errors
##                                    1.828264e-04
##                  strike_outs:free_agency_eligibility1
##                                    5.023734e-03
##                                strike_outs:free_agent1
##                                   -6.219360e-03
##                  strike_outs:arbitration_eligibility1
##                                    2.163836e-03
##                                stolen_bases:errors
##                                    3.570302e-06
```

```
##             stolen_bases:free_agency_eligibility1
##                                      8.375614e-03
##                          stolen_bases:free_agent1
##                                     -1.240475e-03
##             stolen_bases:arbitration_eligibility1
##                                      1.090696e-02
##                   errors:free_agency_eligibility1
##                                      1.156874e-03
##                                errors:free_agent1
##                                      2.479915e-03
##                   errors:arbitration_eligibility1
##                                      4.904456e-03
##              free_agency_eligibility1:free_agent1
##                                                NA
## free_agency_eligibility1:arbitration_eligibility1
##                                                NA
##              free_agent1:arbitration_eligibility1
##                                                NA
```

R-squared	Adj R-squared	RSE	BP (p-value)	W (p-value)
0.8366	0.8164	0.2592	36.353 (0.3152)	0.99511 (0.4616)

Table 11: Summary statistics for Model 6. "BP" stands for Breusch-Pagan test, and "W" stands for Shapiro-Wilk test.

From Model 6, we get the highest R^2 (0.8366) and adjusted R^2 (0.8164) out of all models so far. Our RSE (0.2592) is low, and the assumptions of homoscedasticity and normality of residuals are met under a 5% significance level. We will next try to refine our interaction analysis by selecting specific variables. WHY DO WE SELECT THE VARIABLES THAT WE DO?

5.5 Select Interaction Analysis

Based on our knowledge of baseball and of this dataset, we run the following interaction model analyses. We recall that this form of our data includes the top 90% of salaries in the dataset and is transformed according to the Tukey Transformation $Y^{0.175}$. Table 12 gives the summary statistics for each interaction analysis below the R code. We suppress the code for model summaries, the Breusch-Pagan test, and the Shapiro-Wilk test for the sake of space.

```
#Interaction 1: RBI and free agency eligibility
top90.fs.int1 <- lm(salary ~ runs + RBI + strike_outs +
                                  stolen_bases + errors +
                                  free_agency_eligibility + free_agent +
                                  arbitration_eligibility +
                                  RBI:free_agency_eligibility,
                                data = top90.dat.okay)
top90.fs.int1$coefficients
```

```
##                   (Intercept)                          runs
##                   2.477500669                   0.003694237
##                           RBI                   strike_outs
##                   0.003977909                  -0.001663790
##                  stolen_bases                        errors
##                   0.003876666                  -0.005139489
##     free_agency_eligibility1                   free_agent1
##                   0.681889590                  -0.116422599
##     arbitration_eligibility1 RBI:free_agency_eligibility1
##                   0.767247526                   0.004885332
```

```
#Interaction 2: RBI and free agency eligibility, RBI and
#arbitration eligibility
top90.fs.int2 <- lm(salary ~ runs + RBI + strike_outs + stolen_bases + errors
                    + free_agency_eligibility + free_agent +
                      arbitration_eligibility + RBI:free_agency_eligibility
                     +  RBI:arbitration_eligibility, data = top90.dat.okay)
top90.fs.int2$coefficients
```

```
##                   (Intercept)                          runs
##                   2.549727151                   0.003925794
##                           RBI                   strike_outs
##                   0.001019991                  -0.001180929
##                  stolen_bases                        errors
##                   0.003160085                  -0.004404620
##     free_agency_eligibility1                   free_agent1
##                   0.597141297                  -0.116803309
##     arbitration_eligibility1 RBI:free_agency_eligibility1
##                   0.517896064                   0.007291344
## RBI:arbitration_eligibility1
##                   0.005501578
```

```
#Interaction 3: RBI and free agency eligibility, RBI and
#arbitration eligibility, runs and free agency eligibility
top90.fs.int3 <- lm(salary ~ runs + RBI + strike_outs + stolen_bases +
                      errors + free_agency_eligibility + free_agent +
                      arbitration_eligibility + RBI:free_agency_eligibility
                    + RBI:arbitration_eligibility +
                      runs:free_agency_eligibility, data = top90.dat.okay)
top90.fs.int3$coefficients
```

```
##                    (Intercept)                           runs
##                    2.554853750                    0.003289587
##                            RBI                    strike_outs
##                    0.001503899                   -0.001163808
##                   stolen_bases                         errors
##                    0.003140912                   -0.004200040
##      free_agency_eligibility1                    free_agent1
```

```
##                   0.579485331                   -0.113180485
##      arbitration_eligibility1  RBI:free_agency_eligibility1
##                   0.519594638                    0.006387647
##  RBI:arbitration_eligibility1 runs:free_agency_eligibility1
##                   0.005505690                    0.001200175
```

```
#Interaction 4: RBI and free agency eligibility, RBI and
#arbitration eligibility, runs and free agency eligibility, stolen bases
#and strike outs
top90.fs.int4 <- lm(salary ~ runs + RBI + strike_outs + stolen_bases +
                      errors + free_agency_eligibility + free_agent +
                      arbitration_eligibility + RBI:free_agency_eligibility
                      + RBI:arbitration_eligibility +
                      runs:free_agency_eligibility +
                      stolen_bases:strike_outs, data = top90.dat.okay)
top90.fs.int4$coefficients
```

```
##                   (Intercept)                          runs
##                  2.5036649510                 0.0030183145
##                           RBI                   strike_outs
##                  0.0012169328                -0.0000219699
##                  stolen_bases                        errors
##                  0.0123009977                -0.0034994826
##      free_agency_eligibility1                   free_agent1
##                  0.5659296503                -0.1141071095
##      arbitration_eligibility1  RBI:free_agency_eligibility1
##                  0.5101033901                 0.0070901050
##  RBI:arbitration_eligibility1 runs:free_agency_eligibility1
##                  0.0056038277                 0.0006864569
##      strike_outs:stolen_bases
##                 -0.0001181244
```

Model	R-squared	Adj R-squared	RSE	BP (p-value)	W (p-value)
Interaction 1	0.8076	0.8016	0.2694	16.110 (0.0646)	0.9934 (0.210)
Interaction 2	0.8144	0.8080	0.2650	19.732 (0.0329)	0.9946 (0.375)
Interaction 3	0.8147	0.8076	0.2653	19.840 (0.0476)	0.9947 (0.393)
Interaction 4	0.8199	0.8124	0.2620	22.307 (0.0342)	0.9941 (0.292)

Table 12: A comparison of summary statistics for Interaction Analyses. "BP" stands for Breusch-Pagan test, and "W" stands for Shapiro-Wilk test.

From Table 12, we can see that Interaction 4 produces the highest R^2 (0.8199) and adjusted R^2 (0.8124) values seen in any model so far; it also produces a low RSE (0.2620) and meets the normality assumption for the residuals through the Shapiro-Wilk test ($W = 0.9941; p\text{-value} = 0.292$). However, the Breusch-Pagan test presents sufficient evidence to reject the null hypothesis (H_0 : homoscedasticity of residuals) in favor of the alternative hypothesis (H_a : heteroscedasticity of residuals) under a 5% significance level ($bp = 22.307; p\text{-value} = 0.0342$). From visualizing the distribution of residuals in Figure 10, we see a slight diagonal pattern

in the lower portion of residuals but nothing that is extremely concerning. We will try a different transformation of salary to see if we can ameliorate this concern of heteroscedasticity but do not see this assumption being drastically violated.

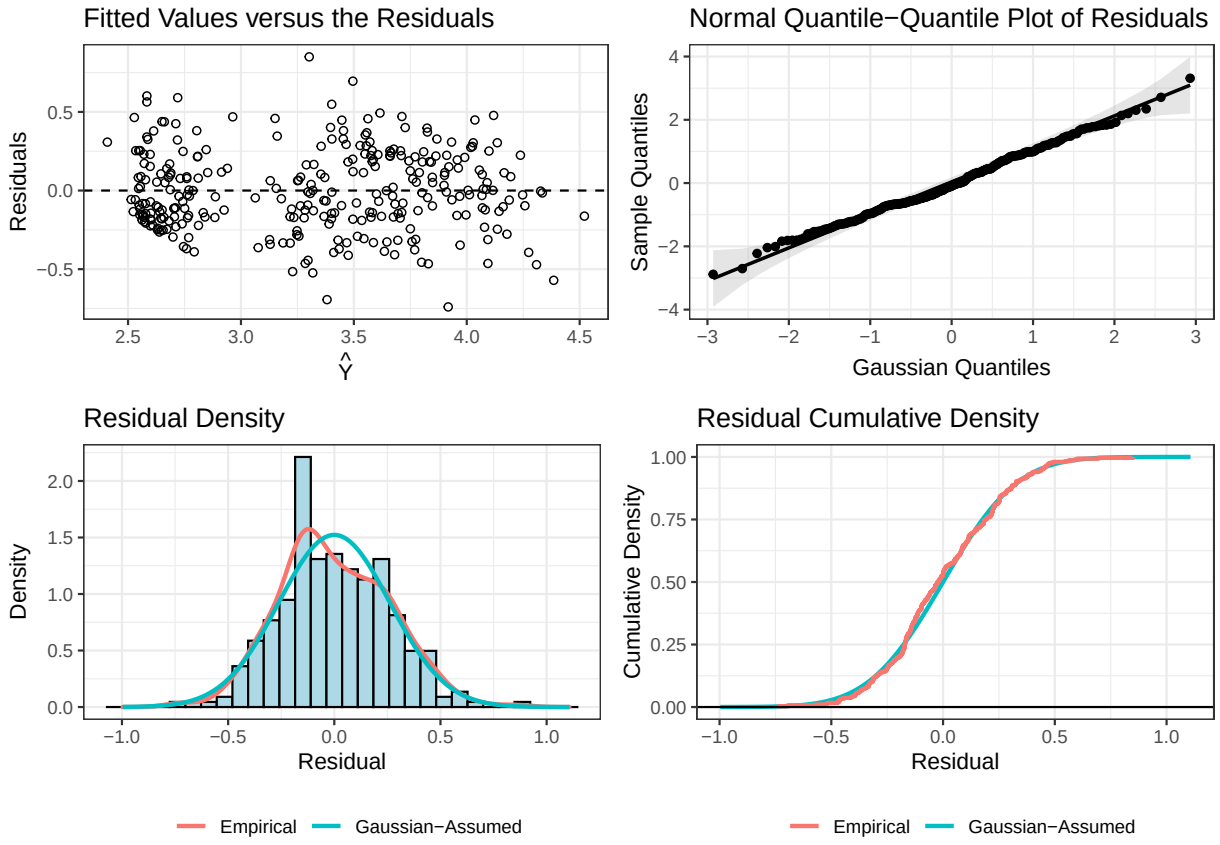

Figure 10: Side-by-side plots of residual values for Interaction Analysis 4.

5.6 Reevaluation of Tukey Transformation

If the Tukey transformation suggested for the top 90% of salaries does not completely ameliorate the heteroscedasticity of residuals, what if we try the Tukey transformation suggested for the complete salary data but still used the top 90% to satisfy the normality assumption? The code below reselects the top 90% of salary data but leaves it untransformed so that we can specify that we want a $Y^{0.125}$ transformation as opposed to the previous $Y^{0.175}$ transformation. We will call this interaction model "Model 7." Table 13 gives the summary statistics for Model 7.

```
top90.dat.okay.nottuk<-baseball.dat.okay %>%
  filter(baseball.dat.okay$salary > quantile(baseball.dat.okay$salary, 0.1))

top90.fs.int4.125 <- lm((salary)^0.125 ~ runs + RBI + strike_outs
                        + stolen_bases + errors + free_agency_eligibility
                        + free_agent + arbitration_eligibility
```

```
                        + RBI:free_agency_eligibility
                        + RBI:arbitration_eligibility
                        + runs:free_agency_eligibility
                        + stolen_bases:strike_outs,
                        data = top90.dat.okay.nottuk)  #Model 7
summary(top90.fs.int4.125)
```

```
##
## Call:
## lm(formula = (salary)^0.125 ~ runs + RBI + strike_outs + stolen_bases +
##     errors + free_agency_eligibility + free_agent + arbitration_eligibility +
##     RBI:free_agency_eligibility + RBI:arbitration_eligibility +
##     runs:free_agency_eligibility + stolen_bases:strike_outs,
##     data = top90.dat.okay.nottuk)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36714 -0.08679 -0.01233  0.09586  0.42085
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1.927e+00  2.644e-02  72.893  < 2e-16 ***
## runs                           1.597e-03  8.178e-04   1.952 0.051881 .
## RBI                            6.560e-04  8.952e-04   0.733 0.464245
## strike_outs                   -1.087e-05  4.114e-04  -0.026 0.978950
## stolen_bases                   6.191e-03  1.830e-03   3.383 0.000817 ***
## errors                        -1.771e-03  1.411e-03  -1.256 0.210275
## free_agency_eligibility1       3.030e-01  3.990e-02   7.595 4.37e-13 ***
## free_agent1                   -5.814e-02  2.695e-02  -2.157 0.031809 *
## arbitration_eligibility1       2.756e-01  4.467e-02   6.170 2.31e-09 ***
## RBI:free_agency_eligibility1   3.443e-03  1.013e-03   3.398 0.000775 ***
## RBI:arbitration_eligibility1   2.639e-03  8.468e-04   3.116 0.002018 **
## runs:free_agency_eligibility1  2.608e-04  9.715e-04   0.268 0.788508
## strike_outs:stolen_bases      -6.003e-05  2.076e-05  -2.892 0.004125 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1329 on 287 degrees of freedom
## Multiple R-squared:  0.8208,Adjusted R-squared:  0.8133
## F-statistic: 109.5 on 12 and 287 DF,  p-value: < 2.2e-16
```

```
bptest(top90.fs.int4.125)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  top90.fs.int4.125
```

```
## BP = 20.667, df = 12, p-value = 0.05548
```

```
shapiro.test(top90.fs.int4.125$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  top90.fs.int4.125$residuals
## W = 0.99406, p-value = 0.2897
```

R-squared	Adj R-squared	RSE	BP (p-value)	W (p-value)
0.8208	0.8133	0.1329	20.667 (0.05548)	0.9941 (0.2897)

Table 13: A comparison of summary statistics for Model 7, an interaction analysis. "BP" stands for Breusch-Pagan test, and "W" stands for Shapiro-Wilk test.

Model 7 presents high R^2 (0.8208) and adjusted R^2 (0.8133) values; it also produces the lowest RSE (0.1329). Model 7 also meets the normality assumption for the residuals through the Shapiro-Wilk test ($W = 0.9941; p\text{-value} = 0.2897$) and the homoscedasticity assumption for residuals through the Breusch-Pagan test under a 5% significance level ($bp = 20.667; p\text{-value} = 0.05548$). In Figure 11, we do not see significant differences in the residual distribution from Model 6 compared to the residual distribution of Interaction Analysis 4 (Figure 10). However, we choose to accept the results of the Breusch-Pagan test, not being overly concerned with the visualization of the residuals for Model 7.

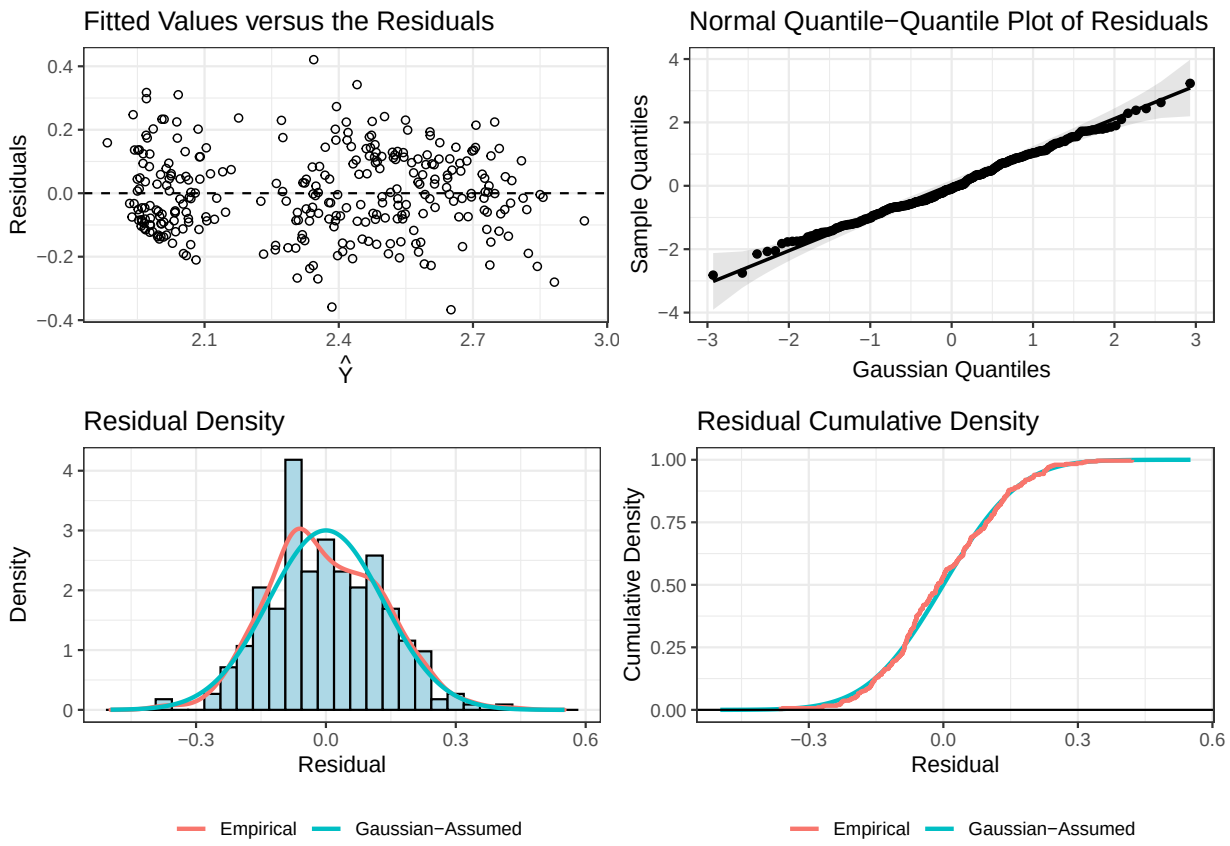

Figure 11: Side-by-side plots of residual values for Interaction Analysis 4.

Considering the success of the $Y^{0.125}$ transformation with only select interactions, what results from a full interaction analysis under a $Y^{0.125}$ transformation? The code below catalogs this result, and Table 14 summarizes the model statistics. We again suppress the code for model summaries, the Breusch-Pagan test, and the Shapiro-Wilk test for the sake of space. This full interaction model under a $Y^{0.125}$ transformation is denoted as Model 8.

```
top90.fullint.125<-lm((salary)^0.125~runs*RBI +      #Model 8
                      runs*strike_outs +
                      runs*stolen_bases+
                      runs*errors +
                      runs*free_agency_eligibility +
                      runs*free_agent +
                      runs*arbitration_eligibility +
                      RBI*strike_outs +
                      RBI*stolen_bases +
                      RBI*errors +
                      RBI*free_agency_eligibility+
                      RBI*free_agent+
                      RBI*arbitration_eligibility +
```

```
                    strike_outs*stolen_bases +
                    strike_outs*errors +
                    strike_outs*free_agency_eligibility +
                    strike_outs*free_agent +
                    strike_outs*arbitration_eligibility +
                    stolen_bases*errors +
                    stolen_bases*free_agency_eligibility+
                    stolen_bases*free_agent +
                    stolen_bases*arbitration_eligibility+
                    errors*free_agency_eligibility +
                    errors*free_agent +
                    errors*arbitration_eligibility +
                    free_agency_eligibility*free_agent +
                    free_agency_eligibility*arbitration_eligibility+
                    free_agent*arbitration_eligibility,
                    data = top90.dat.okay.nottuk)
top90.fullint.125$coefficients

##                                       (Intercept)
##                                      1.930453e+00
##                                              runs
##                                      4.711565e-03
##                                               RBI
##                                      1.571730e-03
##                                       strike_outs
##                                     -1.418478e-03
##                                       stolen_bases
##                                      2.402760e-03
##                                             errors
##                                     -5.140148e-03
##                           free_agency_eligibility1
##                                      2.583975e-01
##                                       free_agent1
##                                     -1.151368e-01
##                           arbitration_eligibility1
##                                      2.190526e-01
##                                          runs:RBI
##                                      2.270320e-06
##                                 runs:strike_outs
##                                     -4.852023e-05
##                               runs:stolen_bases
##                                     -3.316061e-05
##                                  runs:errors
##                                      6.410072e-05
##                         runs:free_agency_eligibility1
##                                     -1.574092e-04
```

```
##                                    runs:free_agent1
##                                        3.559780e-04
##                      runs:arbitration_eligibility1
##                                       -1.172217e-03
##                                    RBI:strike_outs
##                                        2.479588e-05
##                                   RBI:stolen_bases
##                                       -4.754056e-05
##                                      RBI:errors
##                                       -1.517990e-04
##                       RBI:free_agency_eligibility1
##                                        7.668680e-04
##                                    RBI:free_agent1
##                                        4.368939e-03
##                       RBI:arbitration_eligibility1
##                                        1.953511e-03
##                           strike_outs:stolen_bases
##                                        2.412172e-05
##                                 strike_outs:errors
##                                        9.076645e-05
##                 strike_outs:free_agency_eligibility1
##                                        2.621441e-03
##                              strike_outs:free_agent1
##                                       -3.071026e-03
##                 strike_outs:arbitration_eligibility1
##                                        1.217175e-03
##                                stolen_bases:errors
##                                        1.120244e-05
##               stolen_bases:free_agency_eligibility1
##                                        4.125317e-03
##                             stolen_bases:free_agent1
##                                       -5.473757e-04
##               stolen_bases:arbitration_eligibility1
##                                        5.375836e-03
##                     errors:free_agency_eligibility1
##                                        6.535586e-04
##                                  errors:free_agent1
##                                        1.462836e-03
##                     errors:arbitration_eligibility1
##                                        2.604448e-03
##                 free_agency_eligibility1:free_agent1
##                                                  NA
## free_agency_eligibility1:arbitration_eligibility1
##                                                  NA
##                 free_agent1:arbitration_eligibility1
##                                                  NA
```

R-squared	Adj R-squared	RSE	BP (p-value)	W (p-value)
0.8371	0.8169	0.1316	35.136 (0.3672)	0.99519 (0.4768)

Table 14: A comparison of summary statistics for Model 8, an interaction analysis. "BP" stands for Breusch-Pagan test, and "W" stands for Shapiro-Wilk test.

Model 8 presents the highest R^2 (0.8371) and adjusted R^2 (0.8169) values as well as the lowest RSE (0.1316) of all models explored so far. The homoscedasticity of residuals assumption is met ($bp = 35.136; p\text{-value} = 0.3672$); the Gaussian distribution of residuals assumption is also met ($W = 0.99519; p\text{-value} = 0.4768$).

Model Selection Techniques on Model 8:
While Model 8 gives us the best summary statistics so far, a full interaction model is not very realistic. We, therefore, perform a stepwise AIC to select the appropriate variables. Table 15 compares the summary statistics of Model 8 to those of the stepwise selected model, which we denote as Model 9.

```
#for Model 8
AIC(top90.fullint.125)

## [1] -331.3174

BIC(top90.fullint.125)

## [1] -201.685

ols_mallows_cp(top90.fullint.125, top90.fullint.125)

## [1] 34

intercept.model.125<-lm((salary)^0.125 ~1, data=top90.dat.okay.nottuk)
both_90_125<-stepAIC(intercept.model.125,       #Model 9
                     direction="both",
                     scope=list(upper=top90.fullint.125),
                     trace=FALSE)
both_90_125$coefficients

##                             (Intercept)              free_agency_eligibility1
##                            1.939091e+00                          3.301444e-01
##               arbitration_eligibility1                                  runs
##                            2.359826e-01                          1.435184e-03
##                                     RBI                           free_agent1
##                            1.354127e-03                         -1.971364e-01
##                            stolen_bases                           strike_outs
##                            2.952221e-03                         -5.114295e-04
##          free_agency_eligibility1:RBI          arbitration_eligibility1:RBI
##                            2.890228e-03                          2.607862e-03
##                        RBI:free_agent1                      RBI:stolen_bases
##                            2.679072e-03                         -7.342981e-05
## arbitration_eligibility1:stolen_bases free_agency_eligibility1:stolen_bases
##                            5.163288e-03                          3.562292e-03
```

```
AIC(both_90_125)
## [1] -354.0121
BIC(both_90_125)
## [1] -298.4554
ols_mallows_cp(both_90_125, top90.fullint.125)
## [1] 9.795209
```

Model	R-squared	Adj R-squared	RSE	AIC	BIC	Mallows
Model 8	0.8371	0.8169	0.1316	-331.3174	-201.685	34
Model 9	0.8274	0.8196	0.1307	-354.0121	-298.4554	9.795209

Table 15: A comparison of summary statistics for Model 8 (the full interaction analysis) and Moel 9 (the stepwise selected model).

We see in Table 15 that Model 9 gives us lower R^2 and adjusted R^2 values but also a lower RSE value, the latter of which is desirable. In the stepwise model, the following variables are indicated as significant under a 5% significance level:

free agency eligibility
arbitration eligibility
runs
free agent
free agency eligibility:RBI
arbitration eligibility:RBI
RBI:free agent
RBI:stolen bases
arbitration eligibility:stolen bases
free agency eligibility:stolen bases

A quick assessment of assumptions in the code below reveals that Model 9 meets the Gaussian distribution of residuals assumption ($W = 0.9952; p\text{-value} = 0.4731$) but cannot meet the homoscedasticity of residuals assumption $bp = 22.942; p\text{-value} = 0.04237$) under a 5% significance level.

```
bptest(both_90_125)     #homoscedasticity assumption
##
##  studentized Breusch-Pagan test
##
## data:  both_90_125
## BP = 22.942, df = 13, p-value = 0.04237
shapiro.test(both_90_125$residuals)    #Gaussian assumption
##
##  Shapiro-Wilk normality test
##
```

```
## data:  both_90_125$residuals
## W = 0.99517, p-value = 0.4731
```

With model selection techniques, interaction analysis, and reevaluated transformation in mind, we want to add back in some variables that are important to us as baseball fans. In the next section, we will experiment with another model based on factors that we would like to include in our analysis.

5.7 Variable Selection Based on Baseball Knowledge

We want to intentionally select certain variables based on what we know about baseball. I NEED A BOY TO EXPOUND ON THIS LOL. INSERT REASONING FOR CHOOSING THE VARIABLES INCLUDED IN THE MODEL BELOW.
We will still work with the top 90% of salary data so that the normality of residuals assumption can be met and maintain the $Y^{0.125}$ transformation so that the homoscedasticity of residuals assumption can be met. Table 16 gives the summary statistics for what we call "Model 10."

```
top90.varsel.int.125 <- lm((salary)^0.125 ~ OBP + hits + RBI + walks +
                       strike_outs + stolen_bases + errors +
                       free_agency_eligibility + free_agent +
                       arbitration_eligibility + RBI:free_agency_eligibility
                      +stolen_bases:strike_outs + RBI:arbitration_eligibility,
                      data = top90.dat.okay.nottuk)     #Model 10
top90.varsel.int.125$coefficients
```

```
##                   (Intercept)                           OBP
##                  1.991913e+00                 -2.512505e-01
##                          hits                           RBI
##                  8.298732e-04                  1.964019e-04
##                         walks                   strike_outs
##                  1.176774e-03                 -9.321068e-05
##                  stolen_bases                        errors
##                  6.646237e-03                 -2.270312e-03
##     free_agency_eligibility1                   free_agent1
##                  2.897928e-01                 -6.006053e-02
##     arbitration_eligibility1 RBI:free_agency_eligibility1
##                  2.647209e-01                  3.890102e-03
##     strike_outs:stolen_bases RBI:arbitration_eligibility1
##                 -6.192109e-05                  2.779019e-03
```

R-squared	Adjusted R-squared	RSE
0.8228	0.8147	0.1324

Table 16: Summary statistics for Model 10.

Model 10 produces high R^2 (0.8228) and adjusted R^2 (0.8147) values, as well as a low RSE (0.1324). The following variables are found to be significant under a 5% significance level:

hits
walks
stolen bases
free agency eligibility
free agent
arbitration eligibility
RBI:free agency eligibility
strike outs:stolen bases
RBI:arbitration eligibility

We will assess the required assumptions with the Breusch-Pagan and Shapiro-Wilk Tests.

Breusch-Pagan Test: Our hypotheses are as follows:

$$H_0 : \text{homoscedasticity of residuals}$$

$$H_a : \text{heteroscedasticity of residuals}$$

Under a 5% significance level, we fail to reject the null hypothesis ($bp = 21.503; p\text{-value} = 0.06355$). We, therefore, have sufficient evidence that the residuals are homoscedastic.

```
bptest(top90.varsel.int.125)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  top90.varsel.int.125
## BP = 21.503, df = 13, p-value = 0.06355
```

Distribution of Residuals: Our hypotheses are as follows:

$$H_0 : \text{the residuals are Gaussian distributed in the population}$$

$$H_a : \text{the residuals are } \textit{not} \text{ Gaussian distributed in the population}$$

From the code above, we fail to reject the null hypothesis ($W = 0.99366, p\text{-value} = 0.2406$), thereby suggesting that the residuals are Gaussian distributed in Model 10. The normal Q-Q plot in Figure 12 supports this conclusion.

```
shapiro.test(top90.varsel.int.125$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  top90.varsel.int.125$residuals
## W = 0.99366, p-value = 0.2406
```

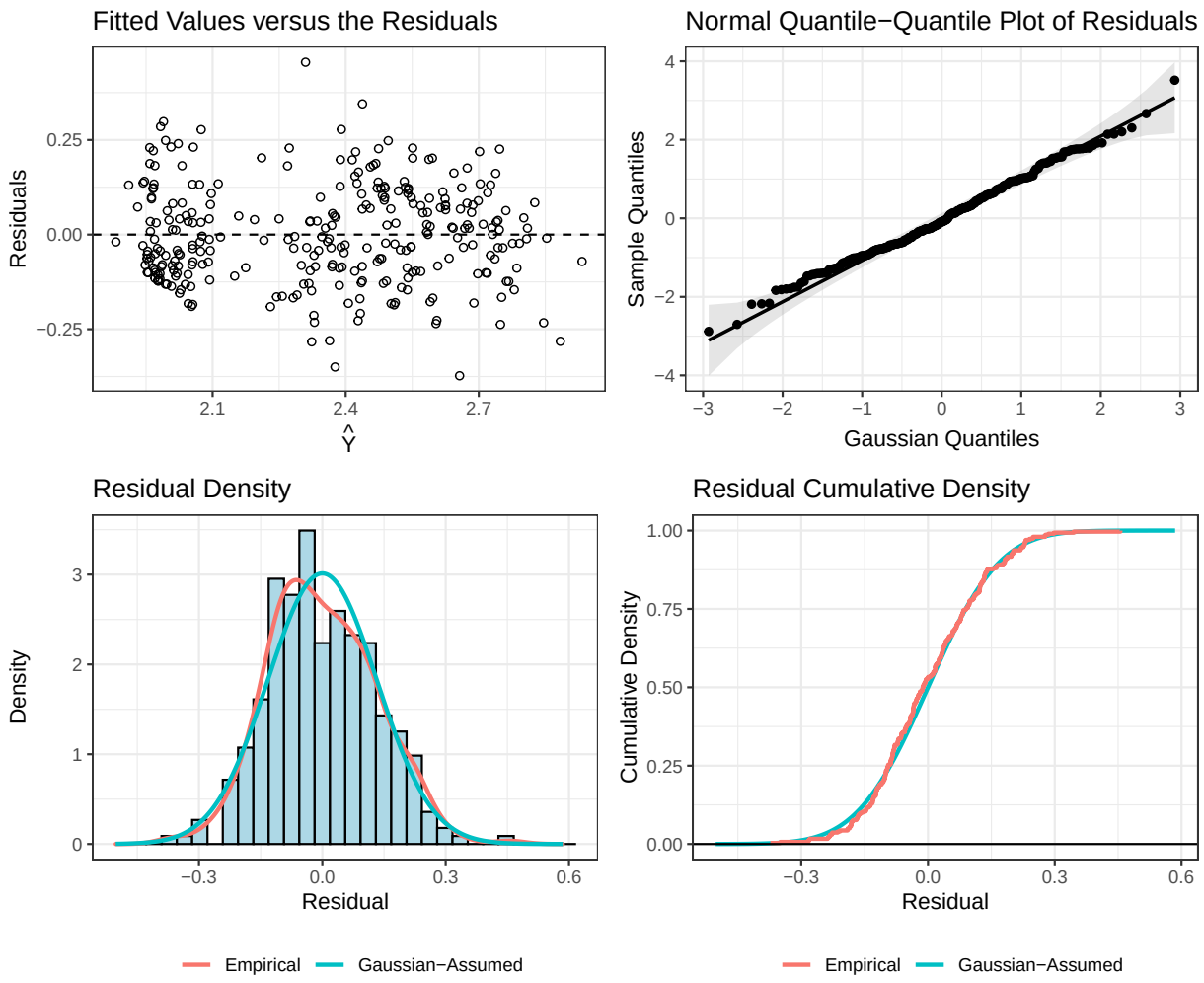

Figure 12: Side-by-side plots of residual values for Model 10.

Multicollinearity: We again calculate the Variance Inflation Factors (VIFs) of Model 10 to assess multicollinearity. We will consider VIFs between 1 and 5 to be *moderately correlated*, and VIFs greater than 5 to be *highly correlated*. From Table 17, we see that the multicollinearity in Model 10 is higher than the multicollinearity found in Model 5 (Table 10), which included no interaction terms and only AIC-selected variables. We will keep this distinction in mind when we select our final model.

```
vif(top90.varsel.int.125)
```

Variable	VIF	Correlation
OBP	2.055	Moderate
Hits	4.951	Moderate
RBI	10.208	High
Walks	3.403	Moderate
Strike outs	3.416	Moderate
Stolen bases	7.407	High
Errors	1.261	Moderate
Free agency eligibility	6.066	High
Free agent	1.350	Moderate
Arbitration eligibility	5.734	High
RBI:Free agency eligibility	9.359	High
Strike outs:Stolen bases	7.851	High
RBI:Arbitration eligibility	7.687	High

Table 17: VIFs for Model 10 and their interpreted strength of correlation.

5.8 Model with Maximized Complexity

The model below, designated as Model 11, includes 126 terms and is summarized in Table 18. Model 11 is then refined using stepwise AIC, creating Model 12. This variable selection technique results in a lower R^2. The adjusted R^2 for the stepwise model, however, is higher, and the RSE is lower for the stepwise model as well. The for loop below the model WHAT DOES THIS FOR LOOP DO????

```
top90.nottuk.noname<-top90.dat.okay.nottuk %>% dplyr::select(-"name")
model11 <- lm((salary)^0.175 ~ .^2, data = top90.nottuk.noname) #Model 11
bptest(model11)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model11
## BP = 102.91, df = 126, p-value = 0.9346
```

```
shapiro.test(model11$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model11$residuals
## W = 0.98627, p-value = 0.005882
```

```
model12 <- stepAIC(model11,  #Model 12
           direction = 'both', trace = FALSE)
model12$coefficients
```

```
##                          (Intercept)                           batting_avg
##                         2.584682e+00                         -1.802140e+00
##                                  OBP                                  runs
##                         9.324018e-01                         2.065569e-02
##                                 hits                               doubles
```

```
##                          6.360622e-03                          -2.422951e-02
##                               triples                              home_runs
##                          1.089049e-01                          -2.748209e-02
##                                   RBI                                 walks
##                          7.751012e-03                          -6.230664e-03
##                           strike_outs                           stolen_bases
##                         -4.367440e-03                          -7.212380e-02
##                                errors                free_agency_eligibility1
##                         -1.847120e-02                           1.700651e+00
##                           free_agent1                arbitration_eligibility1
##                         -1.031704e+00                           1.230494e+00
##                           arbitration1                       batting_avg:runs
##                         -4.193919e+00                           2.614474e-01
##                     batting_avg:hits                    batting_avg:doubles
##                         -1.472582e-01                           5.669141e-01
##                   batting_avg:triples                   batting_avg:home_runs
##                          1.519166e+00                          -8.529440e-01
##               batting_avg:free_agent1  batting_avg:arbitration_eligibility1
##                          1.608869e+01                          -2.734462e+00
##              batting_avg:arbitration1                               OBP:runs
##                          1.655379e+01                          -2.970880e-01
##                              OBP:hits                             OBP:doubles
##                          9.932483e-02                          -3.245367e-01
##                           OBP:triples                          OBP:home_runs
##                         -1.566400e+00                           7.408198e-01
##                              OBP:walks                       OBP:stolen_bases
##                          7.177428e-02                           2.576209e-01
##           OBP:free_agency_eligibility1                           OBP:free_agent1
##                         -4.097976e+00                          -9.101471e+00
##                             runs:hits                            runs:doubles
##                         -1.695923e-04                           8.917243e-04
##                        runs:home_runs                       runs:strike_outs
##                         -5.574157e-04                           8.790642e-05
##                           runs:errors          runs:free_agency_eligibility1
##                          7.713502e-04                           1.368108e-02
##                     runs:arbitration1                               hits:RBI
##                         -2.121693e-02                           1.602831e-04
##                      hits:free_agent1                        doubles:triples
##                         -1.378012e-02                          -6.069727e-03
##                            doubles:RBI      doubles:free_agency_eligibility1
##                         -8.019060e-04                          -1.477074e-02
##                            triples:RBI                          triples:walks
##                          1.186353e-03                           3.840420e-03
##                   triples:strike_outs      triples:free_agency_eligibility1
##                         -9.397071e-04                          -4.987597e-02
##                          home_runs:RBI                 home_runs:strike_outs
```

```
##                           6.185612e-04                           -2.631516e-04
##                 home_runs:stolen_bases    home_runs:free_agency_eligibility1
##                           8.862954e-04                            1.347525e-02
##                  home_runs:free_agent1    home_runs:arbitration_eligibility1
##                           3.575902e-02                            2.963812e-02
##                              RBI:walks                       RBI:stolen_bases
##                          -2.230275e-04                           -5.281110e-04
##                            RBI:errors                        RBI:free_agent1
##                          -2.612601e-04                            7.698784e-03
##                       RBI:arbitration1                      walks:stolen_bases
##                           1.810008e-02                           -3.092435e-04
##                           walks:errors                       walks:free_agent1
##                          -7.788602e-04                            1.666538e-02
##               strike_outs:stolen_bases                      strike_outs:errors
##                           1.839744e-04                            2.326586e-04
##                strike_outs:free_agent1  strike_outs:arbitration_eligibility1
##                          -6.601361e-03                           -3.932316e-03
## stolen_bases:free_agency_eligibility1 stolen_bases:arbitration_eligibility1
##                           1.453429e-02                            2.069284e-02
##                     errors:free_agent1
##                           2.372896e-02
```

```
bptest(model12)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model12
## BP = 65.696, df = 70, p-value = 0.6235
```

```
shapiro.test(model12$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model12$residuals
## W = 0.99341, p-value = 0.2133
```

```
folds_lin_top90_fs_int <- createFolds(top90.nottuk.noname$salary, k = 10)
preds_lin_top90_fs_int <- rep(NA, nrow(top90.nottuk.noname))
for (i in 1:10) {
  training = top90.nottuk.noname[-folds_lin_top90_fs_int[[i]],]
  testing = top90.nottuk.noname[folds_lin_top90_fs_int[[i]],]
  model = lm((salary)^0.175 ~ batting_avg + OBP + runs + hits + doubles +
               triples + home_runs + RBI + walks + strike_outs +
               stolen_bases + errors + free_agency_eligibility +
               free_agent + arbitration_eligibility + arbitration +
               batting_avg:runs + batting_avg:hits + batting_avg:doubles +
               batting_avg:triples + batting_avg:home_runs +
```

```
              batting_avg:free_agent + batting_avg:arbitration_eligibility +
              batting_avg:arbitration +
              OBP:runs + OBP:hits + OBP:doubles + OBP:triples +
              OBP:home_runs + OBP:walks + OBP:stolen_bases +
              OBP:free_agency_eligibility + OBP:free_agent +
              runs:hits + runs:doubles + runs:home_runs +
              runs:strike_outs + runs:errors + runs:free_agency_eligibility +
              runs:arbitration +
              hits:RBI + hits:free_agent + doubles:triples + doubles:RBI +
              doubles:free_agency_eligibility + triples:RBI + triples:walks +
              triples:strike_outs + triples:free_agency_eligibility +
              home_runs:RBI + home_runs:strike_outs + home_runs:stolen_bases +
              home_runs:free_agency_eligibility + home_runs:free_agent +
              home_runs:arbitration_eligibility + RBI:walks +
              RBI:stolen_bases + RBI:errors + RBI:free_agent +
              RBI:arbitration + walks:stolen_bases + walks:errors +
              walks:free_agent + strike_outs:stolen_bases +
              strike_outs:errors + strike_outs:free_agent +
              stolen_bases:free_agency_eligibility +
              stolen_bases:arbitration_eligibility + errors:free_agent,
            data = training)
  preds_lin_top90_fs_int[folds_lin_top90_fs_int[[i]]] = predict(model, testing)
}
rsquared_lin_top90_fs_int = 1 - sum(((top90.nottuk.noname$salary)^0.175 -
                                       preds_lin_top90_fs_int)^2 ) /
  (sum( ((top90.nottuk.noname$salary)^0.175 -
           mean((top90.nottuk.noname$salary)^0.175))^2))
rsquared_lin_top90_fs_int    #R-squared
```

```
## [1] 0.7933099
```

```
mae_lin_top90_fs_int = (1/nrow(top90.nottuk.noname))*
  sum(abs((top90.nottuk.noname$salary)^0.175 - preds_lin_top90_fs_int))
mae_lin_top90_fs_int     #Mean Absolute Error
```

```
## [1] 0.2169605
```

```
rmse_lin_top90_fs_int = sqrt(sum(((top90.nottuk.noname$salary)^0.175 -
                                    preds_lin_top90_fs_int)^2) /
                               (nrow(top90.nottuk.noname) - 70))
rmse_lin_top90_fs_int    #Root Mean Square Error
```

```
## [1] 0.3134924
```

Model	R-squared	Adjusted R-squared	RSE
Model 11	0.9054	0.8365	0.2245
Model 12	0.8969	0.8654	0.2218

Table 18: Summary statistics for Models 11 and 12.

After evaluating the Breusch-Pagan and Shapiro-Wilk tests, we fail to reject the null hypotheses in both these test (Breusch-Pagan: $bp = 65.696; p\text{-value} = 0.6235$) (Shapiro-Wilk: $w = 0.9934; p\text{-value} = 0.2133$) for Model 12. For Model 11, however, we fail to reject the null hypothesis for the Breusch-Pagan test ($bp = 102.91; p\text{-value} = 0.9346$), but do not have sufficient evidence to maintain the Gaussian assumption of the null for Shapiro-Wilk ($w = 0.98627; p\text{-value} = 0.005882$).
While these models give the best summary statistics of all models, these models are also by far the most complex. We show these models not to suggest their superiority in predictive ability, but rather to demonstrate how complex a model for this dataset could become. Model complexity will factor into our selection of our final model, and we assert that Models 11 and 12 far exceed our desired level of capacity.

6 Final Model and Conclusions

To discuss our models and reach a conclusion on our best model(s), we will refer to the models as described in the list below.

- Model 1: First-order linear model
  - All quantitative and categorical variables
  - Original scale for $Y = salary$
- Model 2: Standardized first-order linear model
  - All quantitative and categorical variables, but quantitative variables are standardized around 0
- Model 3: Tukey Transformed ($Y^{0.125}$) Model of all salary data
- Model 4: Tukey Transformed ($Y^{0.175}$) Model of top 90% of salary data
- Model 5: Stepwise selected model of top 90% ($Y^{0.175}$) Tukey-transformed salary data
  - The same variables were selected by forward, backward, and stepwise directions.
- Model 6: Full interaction model of top 90% of Tukey-transformed ($Y^{0.175}$) salary data
- Model 7: Interaction 4 Model of top 90% of Tukey-transformed ($Y^{0.125}$) salary data
  - Interactions:
    - RBI and free agency eligibility
    - RBI and arbitration eligibility
    - runs and free agency eligibility
    - stolen bases and strike outs
- Model 8: Full interaction model of top 90% of Tukey-transformed $Y^{0.125}$ salary data
- Model 9: Stepwise selection of Model 8
- Model 10: Model with select variables and interactions based on baseball knowledge
  - Tukey-transformed: $Y^{0.125}$

  - Top 90% of salary data
- Model 11: Model with maximized complexity
  - 126 predictive variables
  - Includes top 90% of Tukey-transformed $Y^{0.125}$ salary data
- Model 12: Stepwise selection of Model 11
  - 70 predictive variables

Table 19 below gives the summary statistics and assumptions for each model.

Model	R-squared	Adj R-squared	RSE	Homoscedasticity	Gaussian	M.E.	Interactions
1 and 2	0.7014	0.6865	694.3	Not Satisfied	Not Satisfied	16	0
3	0.8021	0.7922	0.1521	Satisfied	Not Satisfied	16	0
4	0.7992	0.7878	0.2786	Satisfied	Satisfied	16	0
5	0.7964	0.7908	0.2766	Satisfied	Satisfied	8	0
6	0.8366	0.8164	0.2592	Satisfied	Satisfied	0	28
7	0.8208	0.8133	0.1329	Satisfied	Satisfied	8	4
8	0.8371	0.8169	0.1316	Satisfied	Satisfied	0	28
9	0.8274	0.8196	0.1307	Not Satisfied	Satisfied	7	6
10	0.8228	0.8147	0.1324	Satisfied	Satisfied	10	3
11	0.9054	0.8365	0.2245	Satisfied	Not Satisfied	16	110
12	0.8969	0.8654	0.2218	Satisfied	Satisfied	16	54

Table 19: Summary table for all models included in this analysis. Columns "M.E." and "Interactions" denote the number of main effects and interaction terms found within the models. Models 1 and 2 are combined in the table because they differ only in that Model 2 is standardized and Model 1 is not.

6.1 Best Fitted Model

From Table 19, we see that different models have optimal statistics. Model 8 has the highest R^2 (0.8371), while Model 9 has the highest Adjusted R^2 (0.8196) and the lowest RSE (0.1307). We see, however, that Model 9 does not technically meet the assumption that the residuals are homoscedastic ($bp = 22.942; p\text{-value} = 0.04237$). The p-value for the Breusch-Pagan test is not far from 5%, so we below visualize the distribution of the residuals and see no extremely concerning pattern (Figure 13). We, therefore, choose to trust our visual assessment of homoscedasticity and assert that **Model 9 is our best fitted model** because it has the highest Adjusted R^2 and RSE of all models included in this analysis.

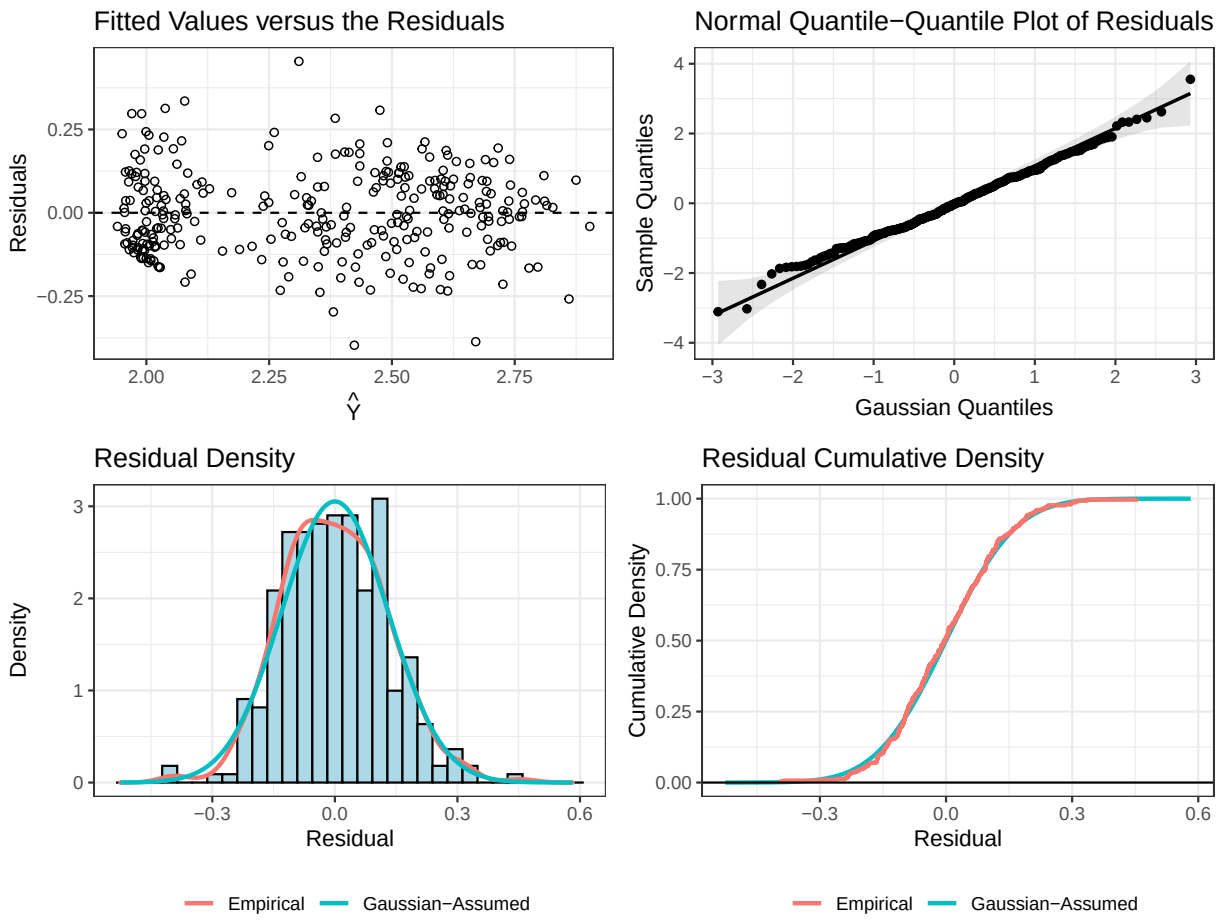

Figure 13: Side-by-side plots of residual values for Model 9.

The following gives the equation for Model 9:

$$\widehat{Y} = 1.939 + 0.3301I(\text{free agency eligibility}) + 0.2360I(\text{arbitration eligibility}) + 0.0014I(\text{runs})$$
$$+0.0014I(\text{RBI}) - 0.1971I(\text{free agent}) + 0.0030I(\text{stolen bases}) - 0.0005I(\text{strike outs})$$
$$+0.0029I(\text{free agency eligibility * RBI}) + 0.0026I(\text{arbitration eligibility * RBI})$$
$$+0.0027I(\text{RBI * free agent}) - 0.00007I(\text{RBI * stolen bases})$$
$$+0.0052I(\text{arbitration eligibility * stolen bases}) + 0.0037I(\text{free agency eligibility * stolen bases})$$

6.2 Best Model Diagnostics

Figure 13 above begins our summary of diagnostics for Model 9. We below repeat the output for the Breusch-Pagan and Shapiro-Wilk tests. A quick assessment of assumptions reveals that Model 9 meets the Gaussian distribution of residuals assumption ($W = 0.9952; p\text{-value} = 0.4731$) but does not meet the homoscedasticity of residuals assumption $bp = 22.942; p\text{-value} = 0.04237$) under a 5% significance level according to the Breusch-Pagan

test. We assert that because the p-value is so close to 5% and visually we do not see a drastic violation of homoscedasticity, this assumption can be reasonably claimed.

```
bptest(both_90_125)    #homoscedasticity assumption
```

```
##
##  studentized Breusch-Pagan test
##
## data:  both_90_125
## BP = 22.942, df = 13, p-value = 0.04237
```

```
shapiro.test(both_90_125$residuals)   #Gaussian assumption
```

```
##
##  Shapiro-Wilk normality test
##
## data:  both_90_125$residuals
## W = 0.99517, p-value = 0.4731
```

Multicollinearity: We again calculate the Variance Inflation Factors (VIFs) of Model 10 to assess multicollinearity. We will consider VIFs between 1 and 5 to be *moderately correlated*, and VIFs greater than 5 to be *highly correlated*. We include below the code to find the confidence intervals for each variable, which are also included in Table 20.

```
vif(both_90_125)
confint(both_90_125)
```

Variable	VIF	Correlation	Lower Bound (2.5%)	Upper Bound (97.5%)
Free agency eligibility	7.200	High	0.2498	0.4105
Arbitration eligibility	5.787	High	0.1483	0.3237
Runs	5.344	High	0.0002	0.0027
RBI	9.793	High	-0.0003	0.0030
Free agent	5.457	High	-0.3014	-.0928
Stolen bases	8.121	High	-0.0006	0.0006
Strike outs	2.284	Moderate	-0.0012	0.0002
Free agency eligibility:RBI	10.407	High	0.0015	0.0043
Arbitration eligibility:RBI	8.207	High	0.0009	0.0043
RBI:Free agent	4.988	Moderate	0.0008	0.0045
RBI:Stolen bases	7.728	High	-0.0001	-0.00002
Arbitration eligibility:Stolen bases	3.265	Moderate	0.0014	0.0089
Free agency eligibility:Stolen bases	3.284	Moderate	0.0006	0.0065

Table 20: VIFs for Model 9 and their interpreted strength of correlation, as well as the 95% confidence interval for each variable.

Multicollinearity in Model 9 is high for nine variables and moderate for four variables. While this multicollinearity is relatively high compared to, for example, Model 5 (Table 10), we expect our multicollinearity to be high due to the inclusion of interaction terms in this model. Additionally, we are not concerned about this high multicollinearity because our goal is prediction rather than to estimate parameters.

Linearity: Figure 14 depicts the correlation of main effect terms in Model 9 and the

Tukey-transformed ($Y^{0.125}$) response variable, salary. The code below formats the correlation coefficients for the model. Given

$$H_0 : \rho = 0 (\text{not linearly related}) \text{ versus } H_a : \rho \neq 0 (\text{linearly related})$$

or each pair of quantitative variables, where ρ denotes population correlation. In every case except stolen bases and free agency eligibility ($r = 0$), we reject the null hypothesis in favor of the alternative.

```
cor.dat.okay<-top90.nottuk.noname %>%
  dplyr::select(free_agency_eligibility, arbitration_eligibility, runs,
                RBI, free_agent, stolen_bases, strike_outs)
cor.dat.okay$free_agency_eligibility<-
  as.numeric(cor.dat.okay$free_agency_eligibility)
cor.dat.okay$arbitration_eligibility<-
  as.numeric(cor.dat.okay$arbitration_eligibility)
cor.dat.okay$runs<-as.numeric(cor.dat.okay$runs)
cor.dat.okay$RBI<-as.numeric(cor.dat.okay$RBI)
cor.dat.okay$free_agent<-as.numeric(cor.dat.okay$free_agent)
cor.dat.okay$stolen_bases<-as.numeric(cor.dat.okay$stolen_bases)
cor.dat.okay$strike_outs<-as.numeric(cor.dat.okay$strike_outs)
cor.dat.okay<-cor.dat.okay %>%
        rename("free agent eligibility"="free_agency_eligibility",
               "arbitration eligibility" = "arbitration_eligibility",
               "free agent"="free_agent",
               "stolen bases"="stolen_bases",
               "strike outs"="strike_outs")
```

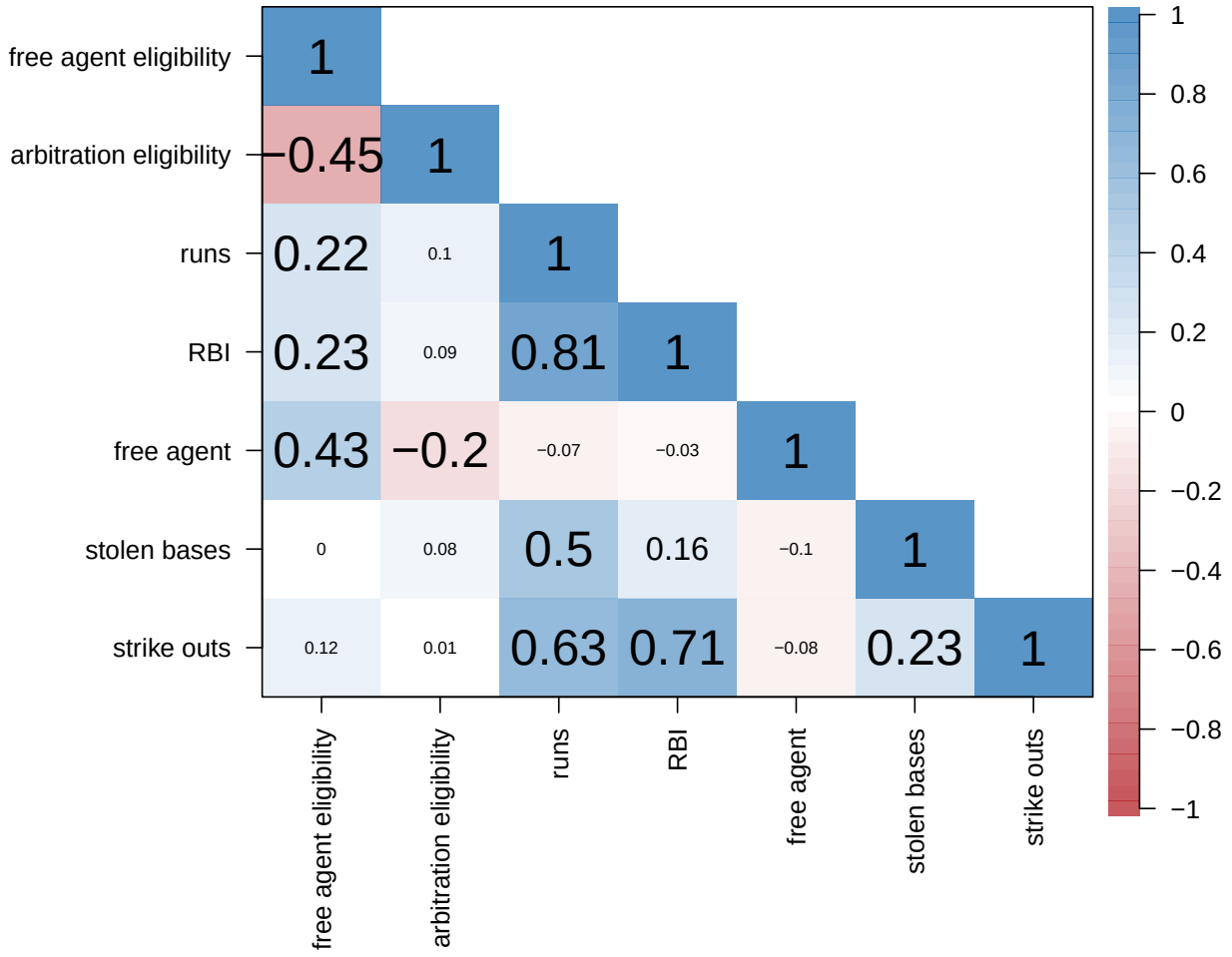

Figure 14: Pearson's correlation coefficients for Model 9 with Tukey transformed ($Y^{0.125}$) response variable.

Summary of Model Diagnostics: While the homoscedasticity of residuals assumption is borderline violated with the Breusch-Pagan test, we see no clear violation in the plot and therefore conclude that this assumption is satisfied. The distribution of the residuals is approximate Gaussian. Multicollinearity is high, which is to be expected and unconcerning for our analysis. The main effect variables are linearly related, with the exception of stolen bases and free agency eligibility.

6.3 Influence Analysis

To assess the influence of potential outliers and leverage points, we employ the following code to identify leverage points and outliers. We then calculate Cook's distance to assess if any points have a strong influence on our best model, Model 9.

```r
data.influence<- top90.dat.okay.nottuk %>%
  mutate(h.values=hatvalues(both_90_125),
         stdres=rstandard(both_90_125),
         studres=rstudent(both_90_125),
         cooks = cooks.distance(both_90_125),
         salary125 = (salary)^0.125) %>%
  dplyr::select(salary125, h.values, stdres, studres, cooks)

#Identification of leverage points
n<-nrow(data.influence)
cutoff.high<-4/n
cutoff.very.high<-6/n
data.influence %>% filter(h.values>cutoff.high) %>% count() #high leverage
```

```
##     n
## 1 292
```

```r
data.influence %>% filter(h.values>cutoff.very.high) %>%
  count() #very high leverage
```

```
##     n
## 1 237
```

```r
#Identification of outliers
data.influence %>% filter(abs(stdres)>2 | abs(studres)>2) %>%
  count() #moderate outlier
```

```
##    n
## 1 11
```

```r
data.influence %>% filter(abs(stdres)>3 | abs(studres)>3) %>%
  count()#strong outlier
```

```
##   n
## 1 3
```

```r
#Influence Analysis using Cook's Distance
data.influence %>% filter(cooks>.5) %>% count() #moderately influential
```

```
##   n
## 1 0
```

```r
data.influence %>% filter(cooks>1) %>% count() #strongly influential
```

```
##   n
## 1 0
```

We see from our leverage analysis that 292 data points have high leverage, and 237 of those have very high leverage. 11 points are moderate outliers, and 3 points are strong outliers. With such high returns on our leverage analysis and the presence of strong outliers, it is somewhat surprising that no points are found to be moderately or highly influential on our final model. However, this outcome is desirable, and we are satisfied with the results of our influence analysis.

6.4 Cross Validation

INSERT EXPLANATION OF CODE AND CROSS VALIDATION

```
folds_lin_orig <- createFolds(top90.nottuk.noname$salary, k = 10)
preds_lin_orig <- rep(NA, nrow(top90.nottuk.noname))
for (i in 1:10) {
  training = top90.nottuk.noname[-folds_lin_orig[[i]],]
  testing = top90.nottuk.noname[folds_lin_orig[[i]],]
  model = lm((salary)^0.125 ~ ., data = training)
  preds_lin_orig[folds_lin_orig[[i]]] = predict(model, testing)
}
rsquared_lin_orig = 1 - sum(((top90.nottuk.noname$salary)^0.125 -
    preds_lin_orig)^2 ) / (sum( ((top90.nottuk.noname$salary)^0.125 -
    mean((top90.nottuk.noname$salary)^0.125))^2))
rsquared_lin_orig
```

```
## [1] 0.7804013
```

```
mae_lin_orig = (1/nrow(top90.nottuk.noname))*
  sum(abs((top90.nottuk.noname$salary)^0.125 - preds_lin_orig))
mae_lin_orig
```

```
## [1] 0.1170543
```

```
rmse_lin_orig = sqrt(sum(((top90.nottuk.noname$salary)^0.125 -
    preds_lin_orig)^2) / (nrow(top90.nottuk.noname) - 13))
rmse_lin_orig
```

```
## [1] 0.1471363
```

6.5 Model Output Summary

Here, we summarize the outputs of our best model, Model 9. A more coehesive presentation of the statistics are presented in Table 21.

```
summary(both_90_125)$adj.r.squared
```

```
## [1] 0.8195547
```

```
AIC(both_90_125)
```

```
## [1] -354.0121
```

```
BIC(both_90_125)
```

```
## [1] -298.4554
```

```
ols_mallows_cp(both_90_125,top90.fullint.125)
```

```
## [1] 9.795209
```

Adj R-squared	AIC	BIC	Mallows	No. of Variables	No. of Parameters
0.8195547	-354.0121	-298.4554	9.7952	13	14

Table 21: The output summary of our best model, Model 6.

The following table (Table 22) presents the summary statistics of the coefficients. Significance levels are based on a 5% signicance level. 95% confidence intervals can be found in Table **??**.

Variable	Estimate	Standard Error	t-value	$Pr(>\|t\|)$	Significance
Intercept	1.939	2.485e-02	78.040	< 2e-16	Significant
Free agency eligibility	0.3301	4.082e-02	8.089	1.73e-14	Significant
Arbitration eligibility	0.2360	4.456e-02	5.296	2.37e-07	Significant
Runs	1.435e-03	6.330e-04	2.267	0.024	Significant
RBI	1.354e-03	8.358e-04	1.620	0.106	Not Significant
Free agent	-0.1971	5.299e-02	-3.720	2.39e-4	Significant
Stolen bases	2.952e-03	1.800e-03	1.640	0.102	Not Significant
Strike outs	-5.114e-04	3.554e-04	-1.439	0.151	Not Significant
Free agency eligibility:RBI	2.890e-03	7.277e-04	3.972	9.04e-05	Significant
Arbitration eligibility:RBI	2.608e-03	8.657e-04	3.012	2.83e-3	Significant
RBI:Free agent	2.679e-03	9.441e-04	2.838	4.87e-3	Significant
RBI:Stolen bases	-7.343e-05	2.804e-05	-2.619	9.29e-3	Significant
Arbitration eligibility:Stolen bases	5.163e-03	1.899e-03	2.719	6.94e-3	Significant
Free agency eligibility:Stolen bases	3.562e-03	1.513e-03	2.354	0.019	Significant

Table 22: VIFs for Model 9 and their interpreted strength of correlation, as well as the 95% confidence interval for each variable.

6.6 Interpretation of Significant Regression Parameters

6.7 Interaction Plots

6.8 Predictions

7 Conclusion

8 Bibliography