621 MoneyBall

Brandon Chung, Jiaxin Zheng, Andreina Arias, and Stephanie Chiang

2025-09-25

Introduction

In this homework assignment we will explore, analyze and model a data set containing 2276 professional baseball team records from the years 1871 to 2006. Our objective is to build a multiple linear regression model on the given training data to predict the number of wins for each team in the test data.

Data Exploration

Data Summary

The moneyball training data set contains 16 variables, excluding the index, and 2,276 observations. Each observational unit represents a single team's statistics for that year's performance. There are 15 predictor variables which are counts of various actions in baseball such as base hits, home runs, strikeouts, stolen bases, caught stealing, hits allows and more. The table in the introduction above provides a list of all variable definitions.

As seen below in our numerical summary the data contains NA values in certain variables (TEAM_BATTING_SO, TEAM_BASERUN_SB, TEAM_BASERUN_CS, TEAM_BATTING_HBP, TEAM_PITCHING_SO, and TEAM_FIELDING_DP). These NA values will be addressed in the data preparation. Notably TEAM_BATTING_HBP contains a large amount of NAs at a count of 2085. There is also certain variables with max and min values that deviate significantly from the interquartile ranges such as TEAM_PITCHING_H and TEAM_PITCHING_SO.

```r
glimpse(training)
```

```
## Rows: 2,276
## Columns: 16
## $ TARGET_WINS      <int> 39, 70, 86, 70, 82, 75, 80, 85, 86, 76, 78, 68, 72, 7~
## $ TEAM_BATTING_H   <int> 1445, 1339, 1377, 1387, 1297, 1279, 1244, 1273, 1391,~
## $ TEAM_BATTING_2B  <int> 194, 219, 232, 209, 186, 200, 179, 171, 197, 213, 179~
## $ TEAM_BATTING_3B  <int> 39, 22, 35, 38, 27, 36, 54, 37, 40, 18, 27, 31, 41, 2~
## $ TEAM_BATTING_HR  <int> 13, 190, 137, 96, 102, 92, 122, 115, 114, 96, 82, 95,~
## $ TEAM_BATTING_BB  <int> 143, 685, 602, 451, 472, 443, 525, 456, 447, 441, 374~
## $ TEAM_BATTING_SO  <int> 842, 1075, 917, 922, 920, 973, 1062, 1027, 922, 827, ~
## $ TEAM_BASERUN_SB  <int> NA, 37, 46, 43, 49, 107, 80, 40, 69, 72, 60, 119, 221~
## $ TEAM_BASERUN_CS  <int> NA, 28, 27, 30, 39, 59, 54, 36, 27, 34, 39, 79, 109, ~
## $ TEAM_BATTING_HBP <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ TEAM_PITCHING_H  <int> 9364, 1347, 1377, 1396, 1297, 1279, 1244, 1281, 1391,~
## $ TEAM_PITCHING_HR <int> 84, 191, 137, 97, 102, 92, 122, 116, 114, 96, 86, 95,~
## $ TEAM_PITCHING_BB <int> 927, 689, 602, 454, 472, 443, 525, 459, 447, 441, 391~
## $ TEAM_PITCHING_SO <int> 5456, 1082, 917, 928, 920, 973, 1062, 1033, 922, 827,~
## $ TEAM_FIELDING_E  <int> 1011, 193, 175, 164, 138, 123, 136, 112, 127, 131, 11~
## $ TEAM_FIELDING_DP <int> NA, 155, 153, 156, 168, 149, 186, 136, 169, 159, 141,~
```

```
colSums(is.na(training))
```

```
##      TARGET_WINS   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B
##                0                0                0                0
##  TEAM_BATTING_HR  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB
##                0                0              102              131
##  TEAM_BASERUN_CS TEAM_BATTING_HBP  TEAM_PITCHING_H TEAM_PITCHING_HR
##              772             2085                0                0
## TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E TEAM_FIELDING_DP
##                0              102                0              286
```

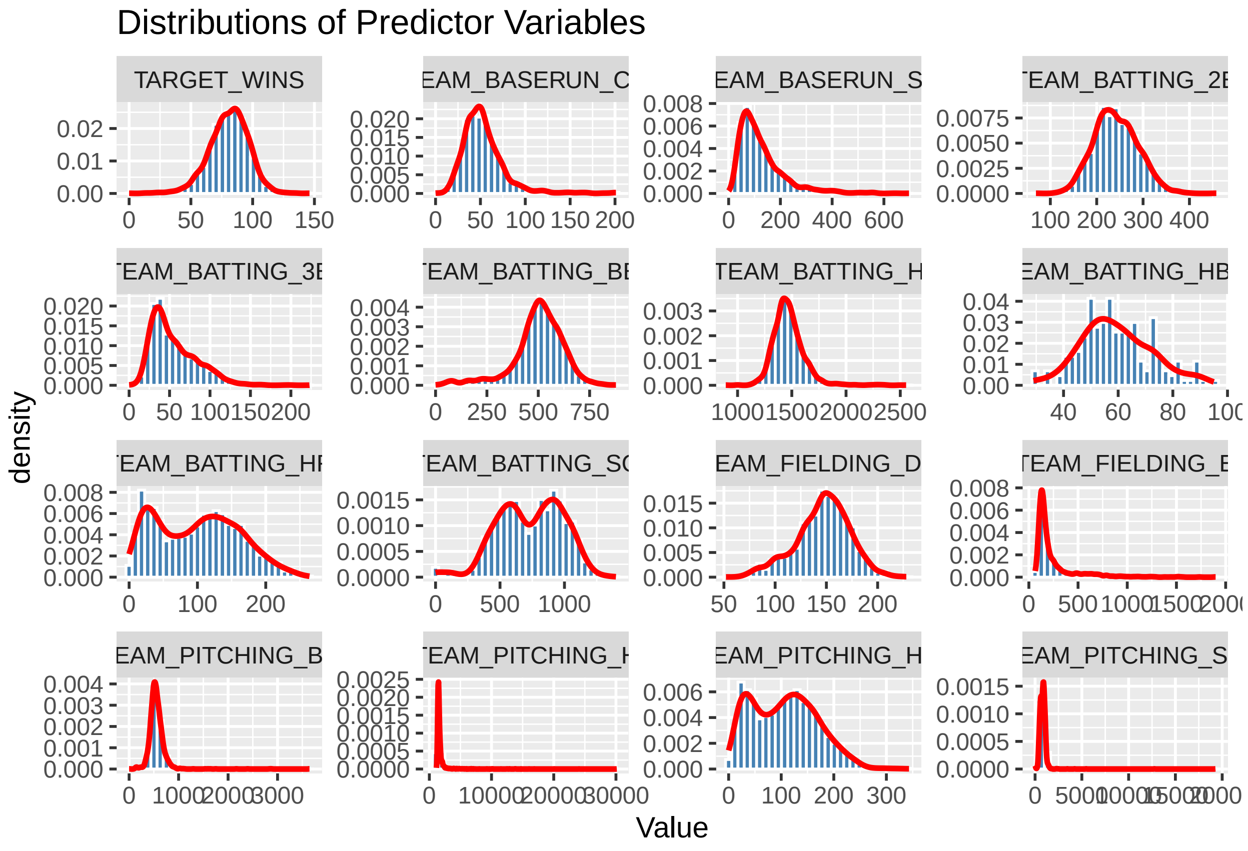
```
summary(training)
```

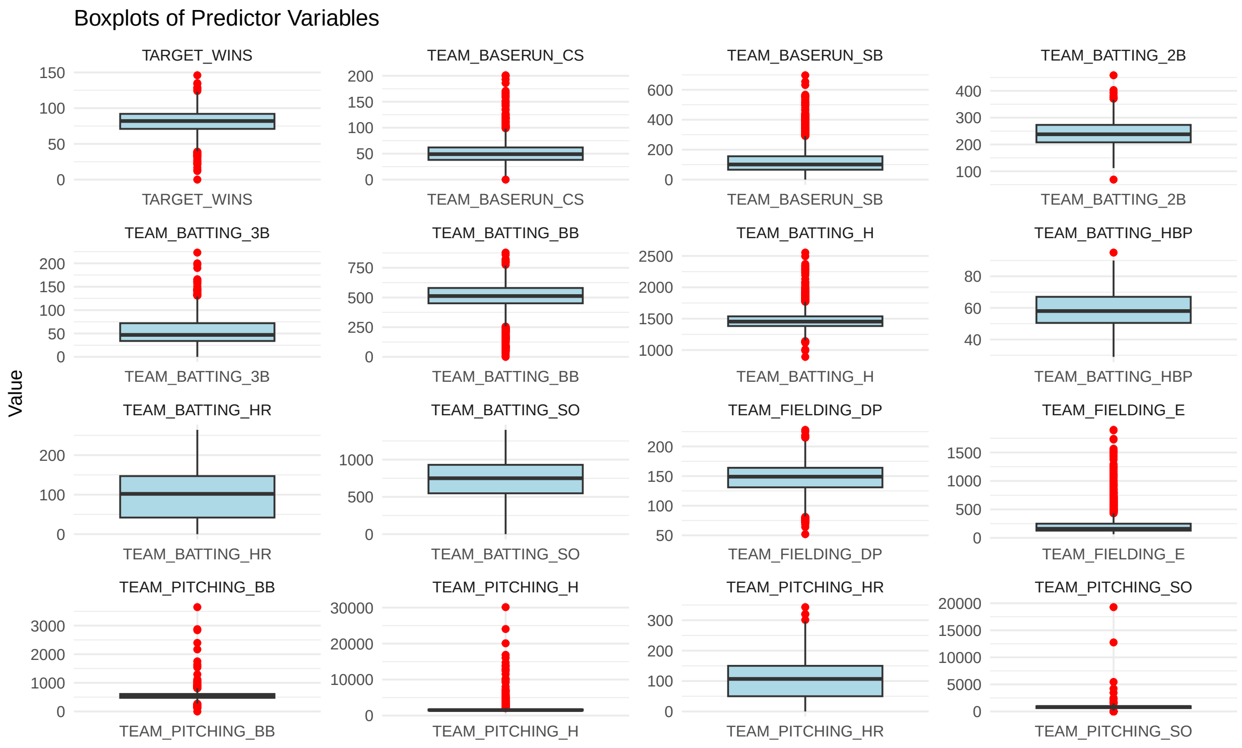
```
##   TARGET_WINS     TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
##  Min.   :  0.00   Min.   : 891   Min.   : 69.0   Min.   :  0.00
##  1st Qu.: 71.00   1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00
##  Median : 82.00   Median :1454   Median :238.0   Median : 47.00
##  Mean   : 80.79   Mean   :1469   Mean   :241.2   Mean   : 55.25
##  3rd Qu.: 92.00   3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 72.00
##  Max.   :146.00   Max.   :2554   Max.   :458.0   Max.   :223.00
##
##  TEAM_BATTING_HR  TEAM_BATTING_BB TEAM_BATTING_SO  TEAM_BASERUN_SB
##  Min.   :  0.00   Min.   :  0.0   Min.   :   0.0   Min.   :  0.0
##  1st Qu.: 42.00   1st Qu.:451.0   1st Qu.: 548.0   1st Qu.: 66.0
##  Median :102.00   Median :512.0   Median : 750.0   Median :101.0
##  Mean   : 99.61   Mean   :501.6   Mean   : 735.6   Mean   :124.8
##  3rd Qu.:147.00   3rd Qu.:580.0   3rd Qu.: 930.0   3rd Qu.:156.0
##  Max.   :264.00   Max.   :878.0   Max.   :1399.0   Max.   :697.0
##                                   NA’s   :102      NA’s   :131
##  TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR
##  Min.   :  0.0   Min.   :29.00    Min.   : 1137   Min.   :  0.0
##  1st Qu.: 38.0   1st Qu.:50.50    1st Qu.: 1419   1st Qu.: 50.0
##  Median : 49.0   Median :58.00    Median : 1518   Median :107.0
##  Mean   : 52.8   Mean   :59.36    Mean   : 1779   Mean   :105.7
##  3rd Qu.: 62.0   3rd Qu.:67.00    3rd Qu.: 1682   3rd Qu.:150.0
##  Max.   :201.0   Max.   :95.00    Max.   :30132   Max.   :343.0
##  NA’s   :772     NA’s   :2085
##  TEAM_PITCHING_BB TEAM_PITCHING_SO  TEAM_FIELDING_E  TEAM_FIELDING_DP
##  Min.   :   0.0   Min.   :     0.0   Min.   :  65.0   Min.   : 52.0
##  1st Qu.: 476.0   1st Qu.:   615.0   1st Qu.: 127.0   1st Qu.:131.0
##  Median : 536.5   Median :   813.5   Median : 159.0   Median :149.0
##  Mean   : 553.0   Mean   :   817.7   Mean   : 246.5   Mean   :146.4
##  3rd Qu.: 611.0   3rd Qu.:   968.0   3rd Qu.: 249.2   3rd Qu.:164.0
##  Max.   :3645.0   Max.   :19278.0   Max.   :1898.0   Max.   :228.0
##                   NA’s   :102                          NA’s   :286
```

```r
head(training)
```

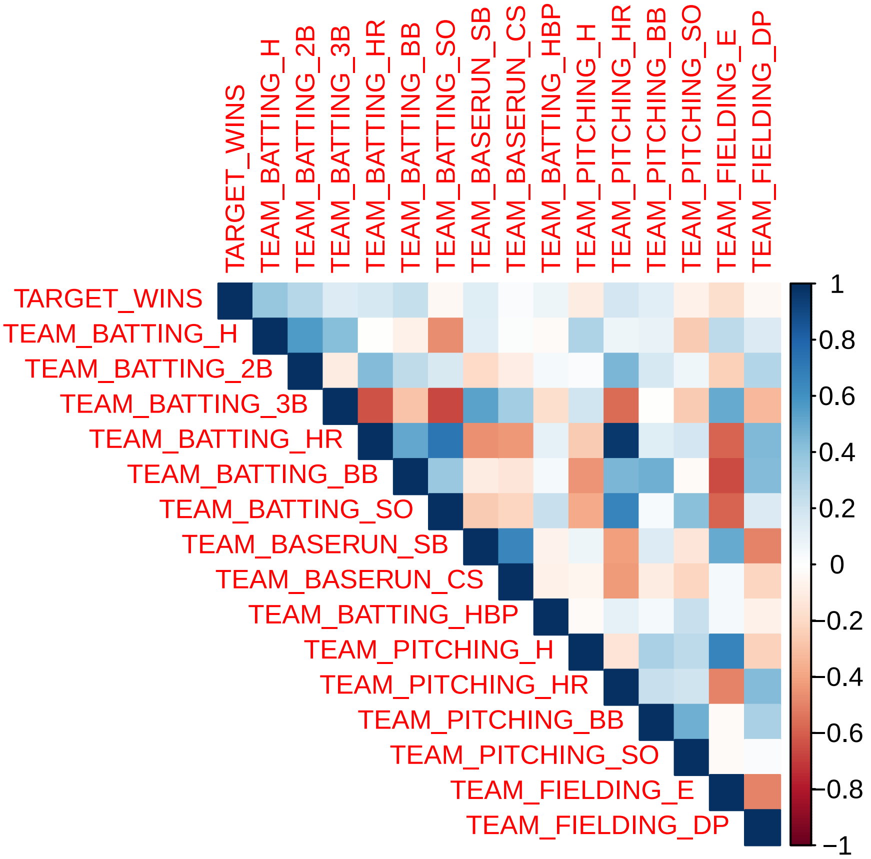
```
##   TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B TEAM_BATTING_HR
## 1          39           1445             194              39              13
## 2          70           1339             219              22             190
## 3          86           1377             232              35             137
## 4          70           1387             209              38              96
## 5          82           1297             186              27             102
## 6          75           1279             200              36              92
##   TEAM_BATTING_BB TEAM_BATTING_SO TEAM_BASERUN_SB TEAM_BASERUN_CS
## 1             143             842              NA              NA
## 2             685            1075              37              28
## 3             602             917              46              27
## 4             451             922              43              30
## 5             472             920              49              39
## 6             443             973             107              59
##   TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB
## 1               NA            9364               84              927
## 2               NA            1347              191              689
## 3               NA            1377              137              602
## 4               NA            1396               97              454
## 5               NA            1297              102              472
## 6               NA            1279               92              443
##   TEAM_PITCHING_SO TEAM_FIELDING_E TEAM_FIELDING_DP
## 1             5456            1011               NA
## 2             1082             193              155
## 3              917             175              153
## 4              928             164              156
## 5              920             138              168
## 6              973             123              149
```

Data Visualizations





The histogram and box plots above provide a better understanding of the distribution of our predictor variables. Most variables have a relatively normal distribution where others show strong left and right side skewing. The box plots also clue us into possible data entry errors as may be the case for TEAM_PITCHING_SO.



The correlation heatmap helps us to see the relationship of variables against the target variable and other predictors. Correlations are mostly what was expected based on the theoretical effect given in the introduction. An example of this can be seen with TEAM_BASERUN_CS where the correlation is slightly positive (0.02240407) when the theoretical effect is to have a negative impact on wins.

Data Preparation

The batter being hit by a pitch was removal as the influence is a factor outside of the batter's controls and it's not a repeatable skill.

```
Training_prep<-training|>
  select(-TEAM_BATTING_HBP)

str(Training_prep)
```

```
## 'data.frame':    2276 obs. of  15 variables:
##  $ TARGET_WINS     : int  39 70 86 70 82 75 80 85 86 76 ...
##  $ TEAM_BATTING_H  : int  1445 1339 1377 1387 1297 1279 1244 1273 1391 1271 ...
##  $ TEAM_BATTING_2B : int  194 219 232 209 186 200 179 171 197 213 ...
##  $ TEAM_BATTING_3B : int  39 22 35 38 27 36 54 37 40 18 ...
##  $ TEAM_BATTING_HR : int  13 190 137 96 102 92 122 115 114 96 ...
##  $ TEAM_BATTING_BB : int  143 685 602 451 472 443 525 456 447 441 ...
##  $ TEAM_BATTING_SO : int  842 1075 917 922 920 973 1062 1027 922 827 ...
##  $ TEAM_BASERUN_SB : int  NA 37 46 43 49 107 80 40 69 72 ...
##  $ TEAM_BASERUN_CS : int  NA 28 27 30 39 59 54 36 27 34 ...
##  $ TEAM_PITCHING_H : int  9364 1347 1377 1396 1297 1279 1244 1281 1391 1271 ...
##  $ TEAM_PITCHING_HR: int  84 191 137 97 102 92 122 116 114 96 ...
##  $ TEAM_PITCHING_BB: int  927 689 602 454 472 443 525 459 447 441 ...
##  $ TEAM_PITCHING_SO: int  5456 1082 917 928 920 973 1062 1033 922 827 ...
##  $ TEAM_FIELDING_E : int  1011 193 175 164 138 123 136 112 127 131 ...
##  $ TEAM_FIELDING_DP: int  NA 155 153 156 168 149 186 136 169 159 ...
```

For data imputation we looked at the columns with missing and use imputation on on those columns that have a rate 5% missing data.

```
Missing <- (colSums(is.na(Training_prep)) / 2276) * 100
print(Missing)
```

```
##      TARGET_WINS   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B
##         0.000000         0.000000         0.000000         0.000000
##  TEAM_BATTING_HR  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB
##         0.000000         0.000000         4.481547         5.755712
##  TEAM_BASERUN_CS  TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB
##        33.919156         0.000000         0.000000         0.000000
## TEAM_PITCHING_SO  TEAM_FIELDING_E TEAM_FIELDING_DP
##         4.481547         0.000000        12.565905
```

Used multiple imputation to impute the missing data using MICE predictive mean matching method.

```
Training_imp<-mice(Training_prep,
                   method = "pmm", #pmm=predictive mean matching
                   m=5,
                   maxit=5,
                   seed=10)|>
  complete()
```

```
##
##  iter imp variable
##   1   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   1   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   1   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   1   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   1   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   2   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   2   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   2   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   2   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   2   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   3   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   3   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   3   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   3   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   3   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   4   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   4   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   4   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   4   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   4   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   5   1  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   5   2  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   5   3  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   5   4  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
##   5   5  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS  TEAM_PITCHING_SO  TEAM_FIELDING_DP
```

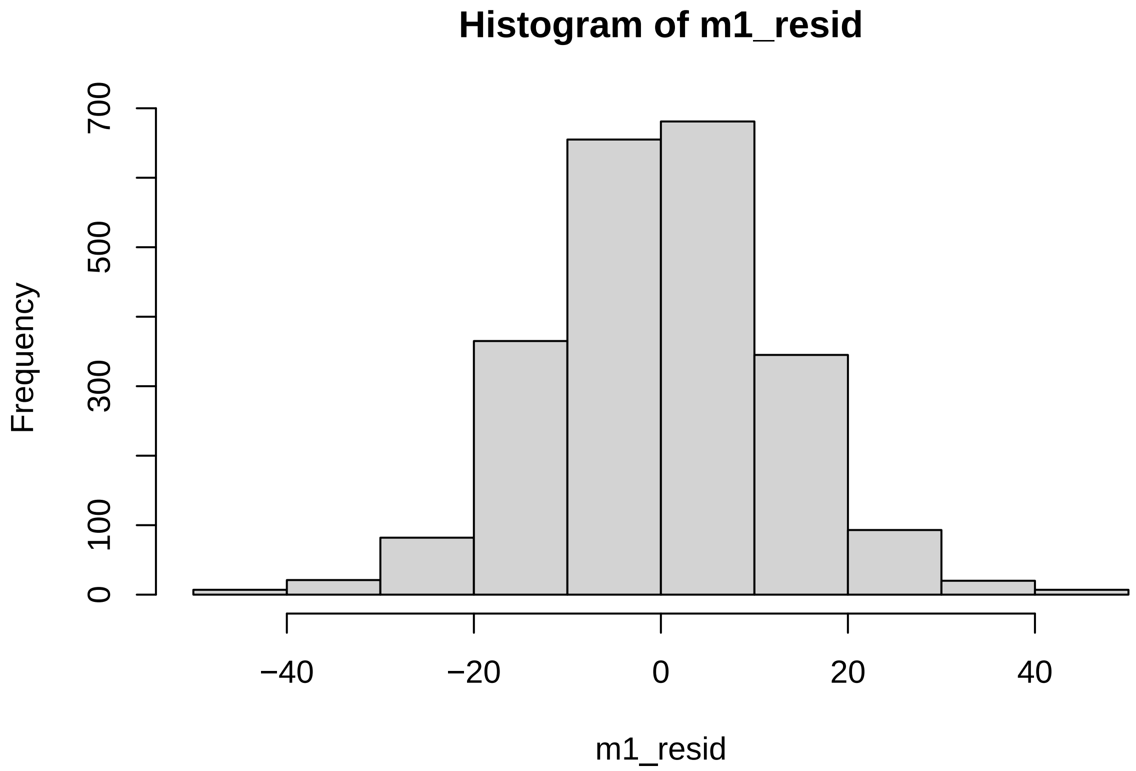
Multiple Linear Regression Models

Model 1: All Features

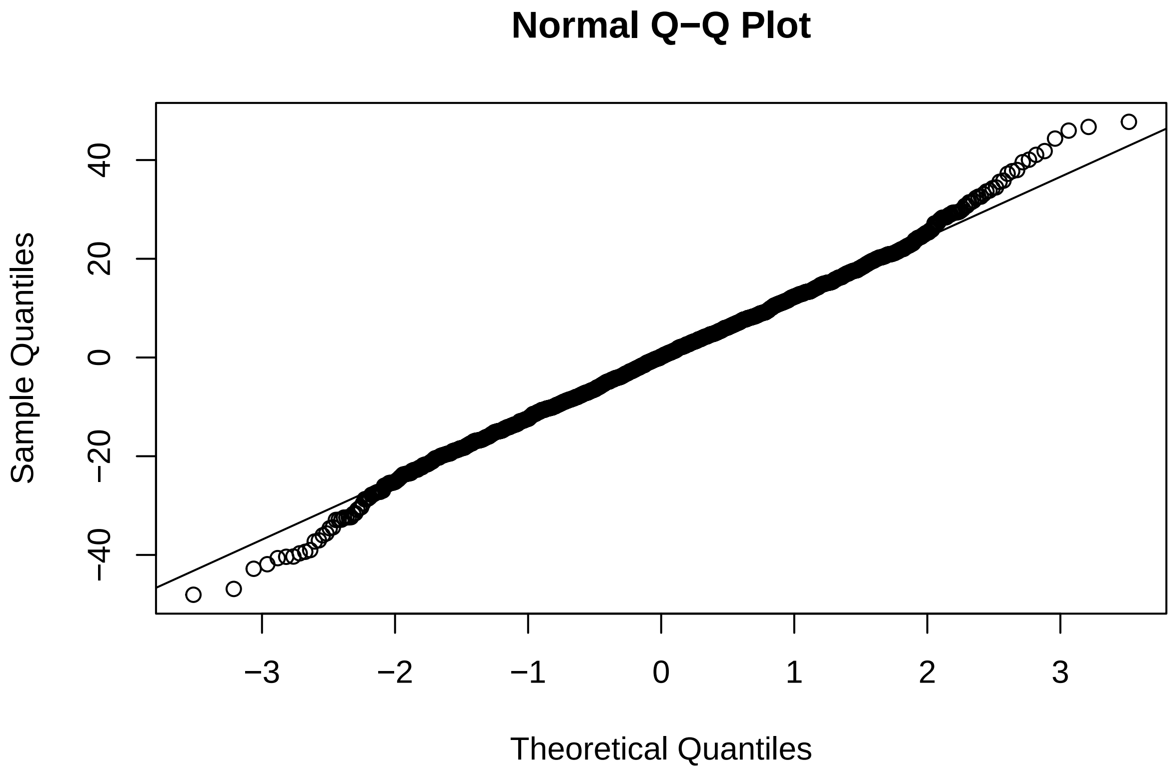
```r
model1 = lm(formula = TARGET_WINS ~
              TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B +
              TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO +
              TEAM_BASERUN_SB + TEAM_BASERUN_CS +
              TEAM_PITCHING_H + TEAM_PITCHING_HR + TEAM_PITCHING_BB + TEAM_PITCHING_SO +
              TEAM_FIELDING_E + TEAM_FIELDING_DP,
            data = Training_imp)

summary(model1)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B +
##     TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_BATTING_SO +
##     TEAM_BASERUN_SB + TEAM_BASERUN_CS + TEAM_PITCHING_H + TEAM_PITCHING_HR +
##     TEAM_PITCHING_BB + TEAM_PITCHING_SO + TEAM_FIELDING_E + TEAM_FIELDING_DP,
##     data = Training_imp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.066  -8.413   0.173   8.114  47.738
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      33.6652346  5.1731357   6.508 9.37e-11 ***
## TEAM_BATTING_H    0.0431257  0.0035895  12.014  < 2e-16 ***
## TEAM_BATTING_2B  -0.0199054  0.0088954  -2.238 0.025337 *
## TEAM_BATTING_3B   0.0412403  0.0164442   2.508 0.012215 *
## TEAM_BATTING_HR   0.0576471  0.0265424   2.172 0.029968 *
## TEAM_BATTING_BB   0.0130473  0.0056243   2.320 0.020440 *
## TEAM_BATTING_SO  -0.0150600  0.0024780  -6.077 1.43e-09 ***
## TEAM_BASERUN_SB   0.0494468  0.0054066   9.146  < 2e-16 ***
## TEAM_BASERUN_CS   0.0020950  0.0110596   0.189 0.849777
## TEAM_PITCHING_H   0.0013758  0.0003859   3.566 0.000371 ***
## TEAM_PITCHING_HR  0.0236405  0.0235842   1.002 0.316263
## TEAM_PITCHING_BB -0.0036554  0.0040041  -0.913 0.361385
## TEAM_PITCHING_SO  0.0015600  0.0008943   1.744 0.081220 .
## TEAM_FIELDING_E  -0.0415048  0.0027079 -15.327  < 2e-16 ***
## TEAM_FIELDING_DP -0.1119556  0.0124114  -9.020  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.66 on 2261 degrees of freedom
## Multiple R-squared:  0.358,  Adjusted R-squared:  0.354
## F-statistic: 90.06 on 14 and 2261 DF,  p-value: < 2.2e-16
```

```
# Residuals
m1_resid = model1$residuals
hist(m1_resid)
```



```
qqnorm(m1_resid)
qqline(m1_resid)
```



Model 2:

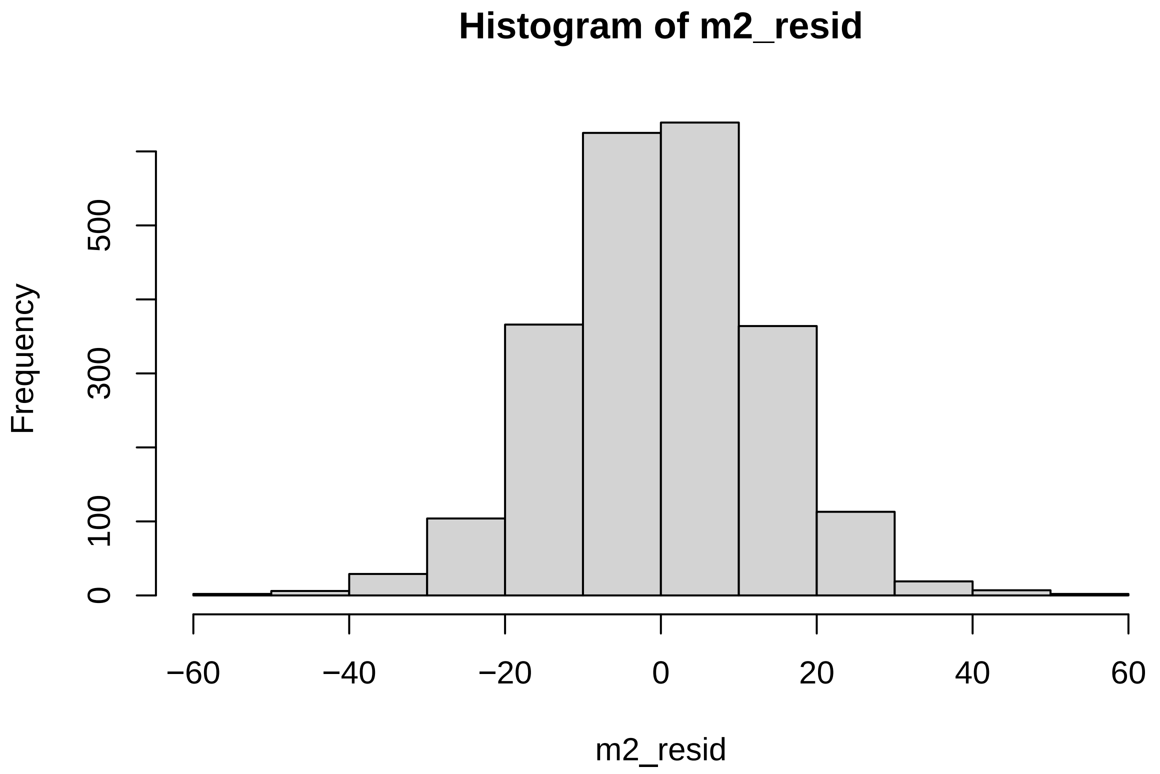
Drop: TEAM_PITCHING_HR for correlation with TEAM_BATTING_HR TEAM_BATTING_SO, TEAM_BASERUN_SB, TEAM_BASERUN_CS, TEAM_PITCHING_SO, TEAM_FIELDING_DP for missing values

```r
model2 = lm(formula = TARGET_WINS ~
              TEAM_BATTING_H + TEAM_BATTING_2B + TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB +
              TEAM_PITCHING_H + TEAM_PITCHING_BB + TEAM_FIELDING_E,
            data = Training_imp)

summary(model2)
```

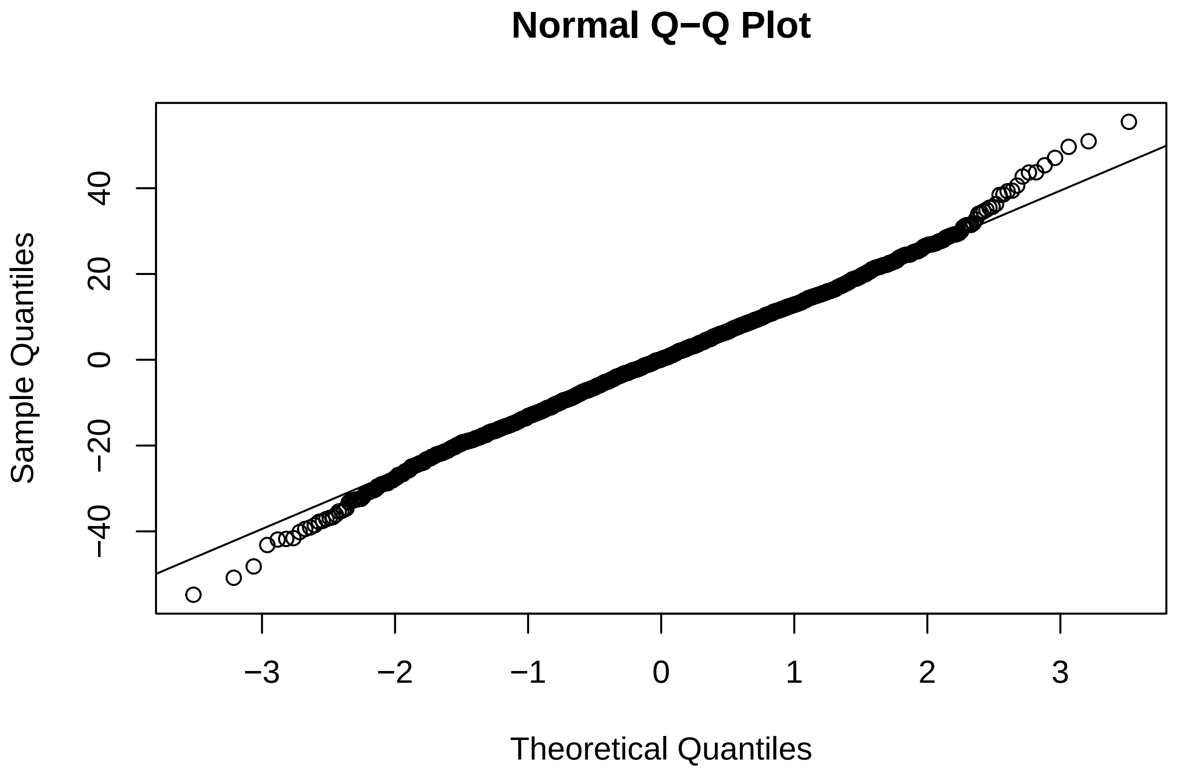
```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_2B +
##     TEAM_BATTING_3B + TEAM_BATTING_HR + TEAM_BATTING_BB + TEAM_PITCHING_H +
##     TEAM_PITCHING_BB + TEAM_FIELDING_E, data = Training_imp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.776  -8.875   0.097   8.860  55.466
##
## Coefficients:
```

```
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.290e+00  3.443e+00   2.117 0.034376 *
## TEAM_BATTING_H      4.848e-02  3.207e-03  15.118  < 2e-16 ***
## TEAM_BATTING_2B    -2.582e-02  9.057e-03  -2.851 0.004400 **
## TEAM_BATTING_3B     1.011e-01  1.665e-02   6.072 1.48e-09 ***
## TEAM_BATTING_HR     3.672e-02  7.749e-03   4.739 2.28e-06 ***
## TEAM_BATTING_BB    -7.926e-05  4.585e-03  -0.017 0.986208
## TEAM_PITCHING_H    -1.312e-03  3.683e-04  -3.561 0.000377 ***
## TEAM_PITCHING_BB    1.036e-02  2.802e-03   3.695 0.000225 ***
## TEAM_FIELDING_E    -1.664e-02  2.368e-03  -7.025 2.81e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.48 on 2267 degrees of freedom
## Multiple R-squared:   0.27,  Adjusted R-squared:  0.2675
## F-statistic: 104.8 on 8 and 2267 DF,  p-value: < 2.2e-16
```

```
# Residuals
m2_resid = model2$residuals
hist(m2_resid)
```



```
qqnorm(m2_resid)
qqline(m2_resid)
```

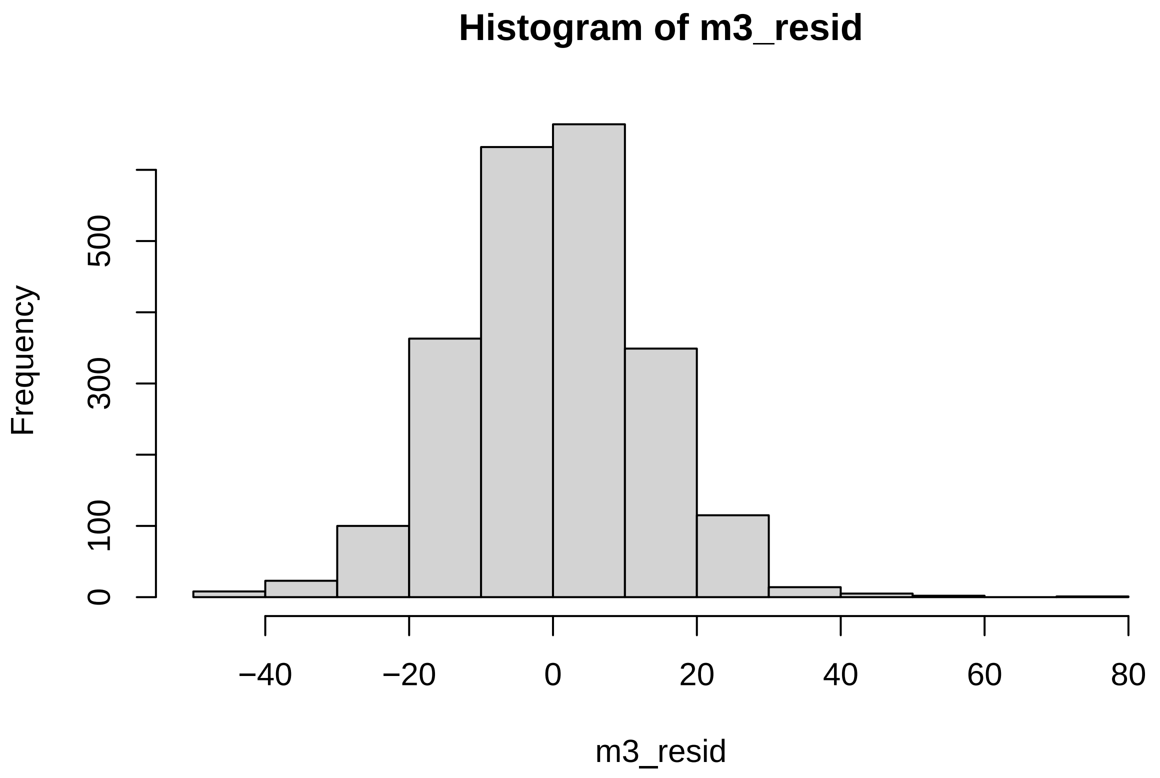


## Model 3: Only taking the high p-values in model 1 and model 2.

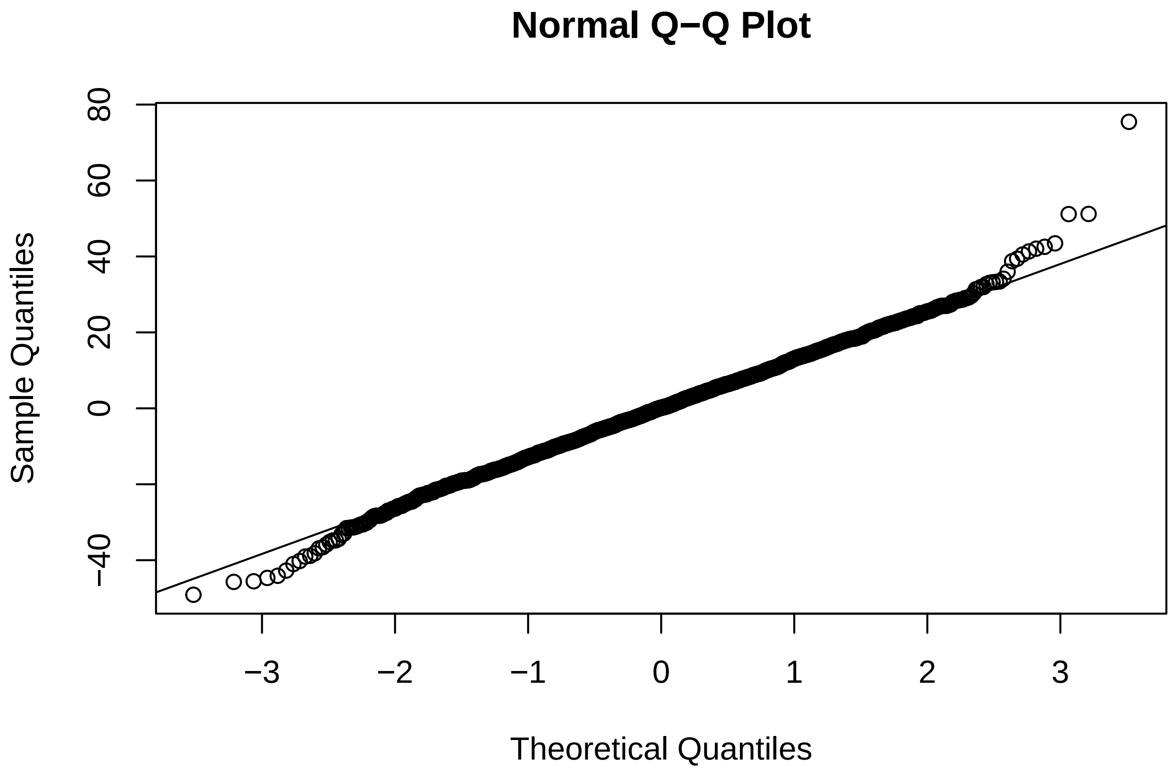
```
model3 = lm(formula = TARGET_WINS ~
              TEAM_BATTING_H + TEAM_BATTING_SO +
              TEAM_FIELDING_E + TEAM_FIELDING_DP +
              TEAM_BATTING_H + TEAM_BATTING_3B +
              TEAM_BATTING_HR + TEAM_FIELDING_E,
            data = Training_imp)

summary(model3)
```

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_SO +
##     TEAM_FIELDING_E + TEAM_FIELDING_DP + TEAM_BATTING_H + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_FIELDING_E, data = Training_imp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.101  -8.747   0.152   8.422  75.453
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      37.017739   4.502184   8.222 3.32e-16 ***
## TEAM_BATTING_H    0.043055   0.002725  15.798  < 2e-16 ***
## TEAM_BATTING_SO  -0.006218   0.002116  -2.938  0.00333 **
## TEAM_FIELDING_E  -0.026633   0.001649 -16.151  < 2e-16 ***
## TEAM_FIELDING_DP -0.140708   0.011571 -12.160  < 2e-16 ***
## TEAM_BATTING_3B   0.091049   0.015717   5.793 7.88e-09 ***
## TEAM_BATTING_HR   0.065951   0.009163   7.197 8.31e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.16 on 2269 degrees of freedom
## Multiple R-squared:  0.3035, Adjusted R-squared:  0.3016
## F-statistic: 164.7 on 6 and 2269 DF,  p-value: < 2.2e-16
```

```
# Residuals
m3_resid = model3$residuals
hist(m3_resid)
```



```
qqnorm(m3_resid)
qqline(m3_resid)
```



Select Models:

While Model 1 has higher multidisciplinary in certain predictors. But our analysis identified Model 1 as the strongest regression model. It achieved the lowest residual error (12.66) and the highest adjusted R² (0.354), making it the most accurate and reliable predictor of team wins.

```
result_table <- bind_rows(
  glance(model1) %>% mutate(Model = "Model 1"),
  glance(model2) %>% mutate(Model = "Model 2"),
  glance(model3) %>% mutate(Model = "Model 3")
) %>%
  transmute(
    Model,
    RSE         = sigma,
    Adj.R2      = adj.r.squared,
    F.Statistic = statistic
  )

result_table
```

```
## # A tibble: 3 x 4
##   Model     RSE Adj.R2 F.Statistic
##   <chr>   <dbl>  <dbl>       <dbl>
## 1 Model 1  12.7  0.354        90.1
## 2 Model 2  13.5  0.267       105.
## 3 Model 3  13.2  0.302       165.
```

‘

```
vif(model1)
```

```
##   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B  TEAM_BATTING_HR
##         3.823342         2.460052         2.995896        36.657149
##  TEAM_BATTING_BB  TEAM_BATTING_SO  TEAM_BASERUN_SB  TEAM_BASERUN_CS
##         6.756380         5.274069         4.349937         4.373084
##  TEAM_PITCHING_H TEAM_PITCHING_HR TEAM_PITCHING_BB TEAM_PITCHING_SO
##         4.182680        29.664612         6.297724         3.336076
##  TEAM_FIELDING_E TEAM_FIELDING_DP
##         5.399699         1.872039
```

```
vif(model2)
```

```
##   TEAM_BATTING_H  TEAM_BATTING_2B  TEAM_BATTING_3B  TEAM_BATTING_HR
##         2.691190         2.248967         2.707698         2.755238
##  TEAM_BATTING_BB  TEAM_PITCHING_H TEAM_PITCHING_BB  TEAM_FIELDING_E
##         3.958646         3.361075         2.720094         3.642208
```

```
vif(model3)
```

```
##   TEAM_BATTING_H  TEAM_BATTING_SO  TEAM_FIELDING_E TEAM_FIELDING_DP
##         2.038514         3.557484         1.852096         1.504965
##  TEAM_BATTING_3B  TEAM_BATTING_HR
##         2.531400         4.040915
```