**Baseball Dataset**

When looking at the baseball dataset, I initially found average double, triples, and homeruns per team, however, for submission, I grouped by year to see trends in the entire league. Values are rounded to nearest whole number.

|  |  |  |  |
| --- | --- | --- | --- |
| **Year** | **Avg # of Doubles** | **Avg # of Triples** | **Avg # of Home Runs** |
| 2018 | 275 | 28 | 186 |
| 2019 | 284 | 26 | 226 |
| 2021 | 262 | 22 | 198 |
| 2022 | 265 | 21 | 173 |
| 2023 | 274 | 24 | 196 |

From a historic perspective, the following are the expected five number summaries for doubles, triples, and home runs.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Min** | **Q1** | **Median** | **Mean** | **Q3** | **Max** |
| X2B | 213.0 | 252.0 | 271.0 | 272.2 | 290.8 | 355.0 |
| X3B | 8.0 | 18.0 | 24.0 | 24.4 | 29.0 | 50.0 |
| HR | 110.0 | 166.0 | 196.5 | 195.9 | 220.8 | 307.0 |

Moving forward in the analysis, we decided to run a linear regression. The goal of this regression was to use historical trends in the following variables in order to predict runs: walks, hits, doubles, triples, homeruns, stolen bases, and caught stealing. In my first linear regression attempt, the results had a RSE of 22.78, meaning that our predicted runs were off by about 22, on average, when being compared to actual runs scored in a given year by a given team. Multiple R -squared was strong at .927, indicating that a significant amount of the variance in the data was explained by the linear regression. The adjusted R-squared is also high at .924, indicating that the variables I chose were effective as well.

The only issue that arose, however, was the p-value associated with the variable triples. With a value of .17, we generally consider its significance too low to be helpful in predicting, and therefore, I deiced to remove it and re-run the model without it. The resulting summary of the model indicated a nearly identical accuracy metrics with a standard error of 22.85, multiple R-squared of .927, and adjusted R-squared of .924. Therefore, I chose to stick with the new model as my final model because it produced nearly the same scoring metrics, while also removing variables with less significance. Lastly, in this model, we can see the most impactful variables by considering the coefficients related to each variable. The three most significant were homeruns, doubles and caught stealing. These coefficients mean, that for every homerun, the model predicts .96 more runs per season, for every double, an additional .50 runs for the season, and for every time caught stealing, -.65 runs for the season.

Out of the linear regression modeling, we are able to compare runs and predicted runs. By doing so, we can see the five luckiest teams and the five unluckiest teams. Lucky teams will have worse scores in terms of variables, but a higher difference in predicted, and unlucky teams will have high scoring variables, but ultimately underperform in terms of runs.

The five luckiest:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Team** | **Year** | **Runs** | **Predicted Runs** | **Difference** |
| TBA | 2021 | 857 | 775 | 82 |
| BAL | 2023 | 807 | 760 | 47 |
| TBA | 2023 | 860 | 823 | 36 |
| CHN | 2023 | 819 | 783 | 36 |
| ARI | 2022 | 702 | 666 | 36 |

We can see here that at the greatest positive extreme, the linear regression is inaccurate by a little more than 10%, but within the top 5, the prediction is off by about 5%, meaning the model is substantially effective.

The five unluckiest:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Team** | **Year** | **Runs** | **Predicted Runs** | **Difference** |
| MIL | 2019 | 769 | 839 | -70 |
| DET | 2019 | 582 | 640 | -58 |
| SLN | 2023 | 719 | 773 | -54 |
| OAK | 2023 | 585 | 637 | -52 |
| BAL | 2018 | 622 | 665 | -43 |

The inaccuracy at the greatest negative extreme is similar here. The model at its worse is predicting close to a 10% error, but even by the 5th worst error, the error is as low as 7%.

**Appendix**

**Original Baseball Regression Model:**

*Including all input variables*

*model <- lm(R ~ Walks + H + X2B + X3B + HR + SB + CS, data = Teams)*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) -318.33701 39.37131 -8.086 2.47e-13 \*\*\**

*Walks 0.32727 0.03414 9.585 < 2e-16 \*\*\**

*H 0.38519 0.03954 9.743 < 2e-16 \*\*\**

*X2B 0.49956 0.09543 5.235 5.82e-07 \*\*\**

*X3B 0.34459 0.25319 1.361 0.175672*

*HR 0.98187 0.06939 14.150 < 2e-16 \*\*\**

*SB 0.27813 0.07270 3.826 0.000195 \*\*\**

*CS -0.74115 0.29384 -2.522 0.012763 \**

*---*

*Residual standard error:* ***22.78*** *on 142 degrees of freedom*

*Multiple R-squared****: 0.9277****, Adjusted R-squared****: 0.9241***

*F-statistic: 260.2 on 7 and 142 DF, p-value: < 2.2e-16*

**Adjusted Baseball Regression Model:**

*Including all input variables except for triples*

*new\_model <- lm(R ~ Walks + H + X2B + HR + SB + CS, data = Teams)*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept). -320.35056 39.46058 -8.118 1.99e-13 \*\*\**

*Walks 0.32654 0.03424 9.536 < 2e-16 \*\*\**

*H 0.39373 0.03915 10.057 < 2e-16 \*\*\**

*X2B 0.50091 0.09571 5.234 5.80e-07 \*\*\**

*HR 0.96032 0.06776 14.172 < 2e-16 \*\*\**

*SB 0.28728 0.07260 3.957 0.000119 \*\*\**

*CS -0.65646 0.28803 -2.279 0.024139 \**

*Residual standard error:* ***22.85*** *on* ***143*** *degrees of freedom*

*Multiple R-squared:* ***0.9267****, Adjusted R-squared****: 0.9237***

*F-statistic: 301.4 on 6 and 143 DF, p-value: < 2.2e-16*