Hitters Data Analysis

William Arnost, Daniel Crouthamel, Richard Palmer

# Introduction

The goal of this analysis is to determine which baseball performance metrics influence a player’s salary. We will conduct a modeling exercise to see which features are useful in predicting salary and test different modeling techniques. We will also perform a two way anova analysis to determine the effect of league and division on salary.

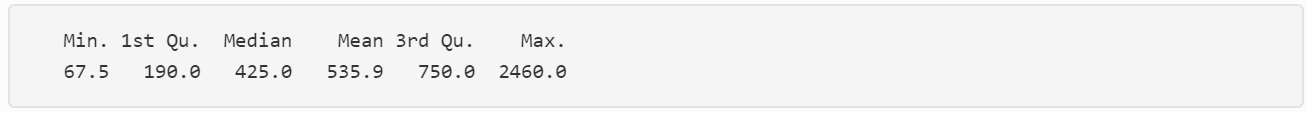
# Data Description

This data set contains salary information and career statistics for major league baseball players. The salary data was original from Sports Illustrated (April 20, 1987) while the career statistics are from the 1987 Baseball Encyclopedia Update published by Collier Books. We sourced the data from Kaggle (<https://www.kaggle.com/floser/hitters>) and it is also used in the ISLR R-package and is used in the textbook "An Introduction to Statistical Learning with applications in R" by G. James et al. (2013).

The data contains 322 observations, 1 per player, and 20 variables. The variables include salary and career statistics like Home Runs, times At Bat, Errors, and others.

# Exploratory Data Analysis

The salary variable will the response variable for modeling, so let’s look at that first. A quick summary shows that the minimum salary is $67.5K and the maximum is $2.46M. There are 59 NA values in the salary variable, so we will need to exclude those values for modeling (About 18% of the dataset).



The first two figures below are the histograms of Salary and logSalary respectively. We can see in the first histogram that Salary, untransformed, is very right skewed, with the maximum salary being almost 6x the median salary. We will need to transform the variable for modeling to deal with non-normality in the residuals.



In the second two graphs we compare the distribution of salary for two factors, League and Division. The distribution of salary in the American and National Leagues look similar, although the American league has a few outliers at the high end. There are more differences in the East and West divisions, with the east division having more salaries in the $1M+ range and the West division having more in the less than $250K range.

# Objective 1: Modeling

## Problem Statement

We would like to know which career statistics are indicative of a player’s salary. We will use a predictive model to assess which statistics, if any, are important. We will use variable selection techniques to help fit the best model.

## Model Selection

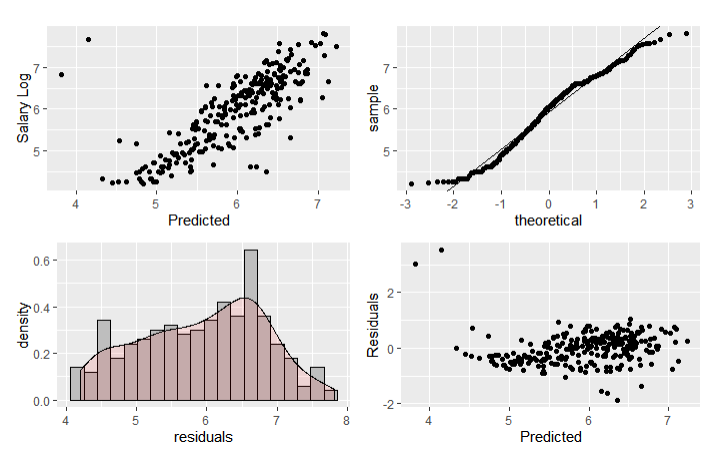
For model selection, we will compare four different possibilities in R, these being Least Squares, Ridge, Lasso and Elastic net. Additionally, we will perform an additional Lasso test in SAS for comparison. As mentioned above, logging salary seems to be appropriate. It also appears that logging the career attributes, those that start with C, will also be beneficial. Notice below that the scatter plots (far left column) is more linear with the right graph using log of C\* attributes. The residual plots using SAS in the last row show a more uniform cloud distribution.

|  |  |
| --- | --- |
|  |  |
|  |  |

## Assumption Checking

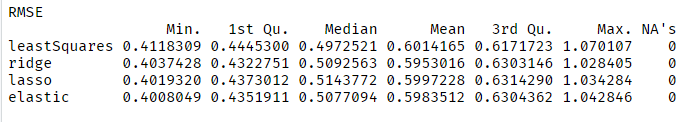
For the purpose of assumption checking, the Lasso model was used (see Comparing Competing Models). All models performed roughly the same, but since Lasso has fewer parameters, we opted for the simpler model.

The Residual vs Predicted (bottom right) is roughly a cloud shape around 0 and the QQ plot shows that the residuals are somewhat normally distributed, although there are few significant outliers.



## Compare Competing Models

We performed a comparison between 4 different models and the mean RMSE value was nearly the same all 4. Lasso was selected because it resulted in the least number of coefficients, thereby producing a simpler, easier to explain model.



<ADD IN SAS Analysis>

## Parameter Interpretation

### Interpretation

### Confidence Intervals

## Final Conclusions for Objective 1

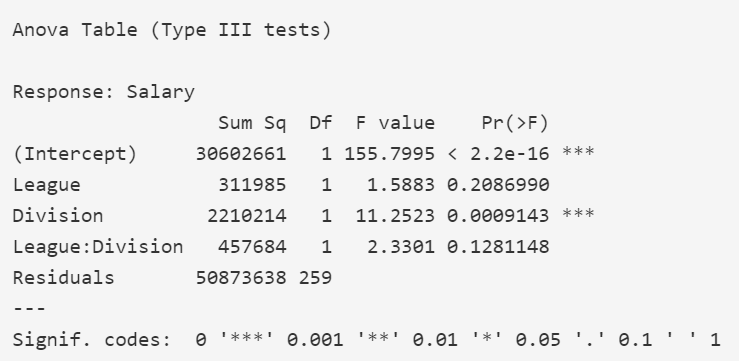
# Objective 2: Advanced Analysis

## State the method

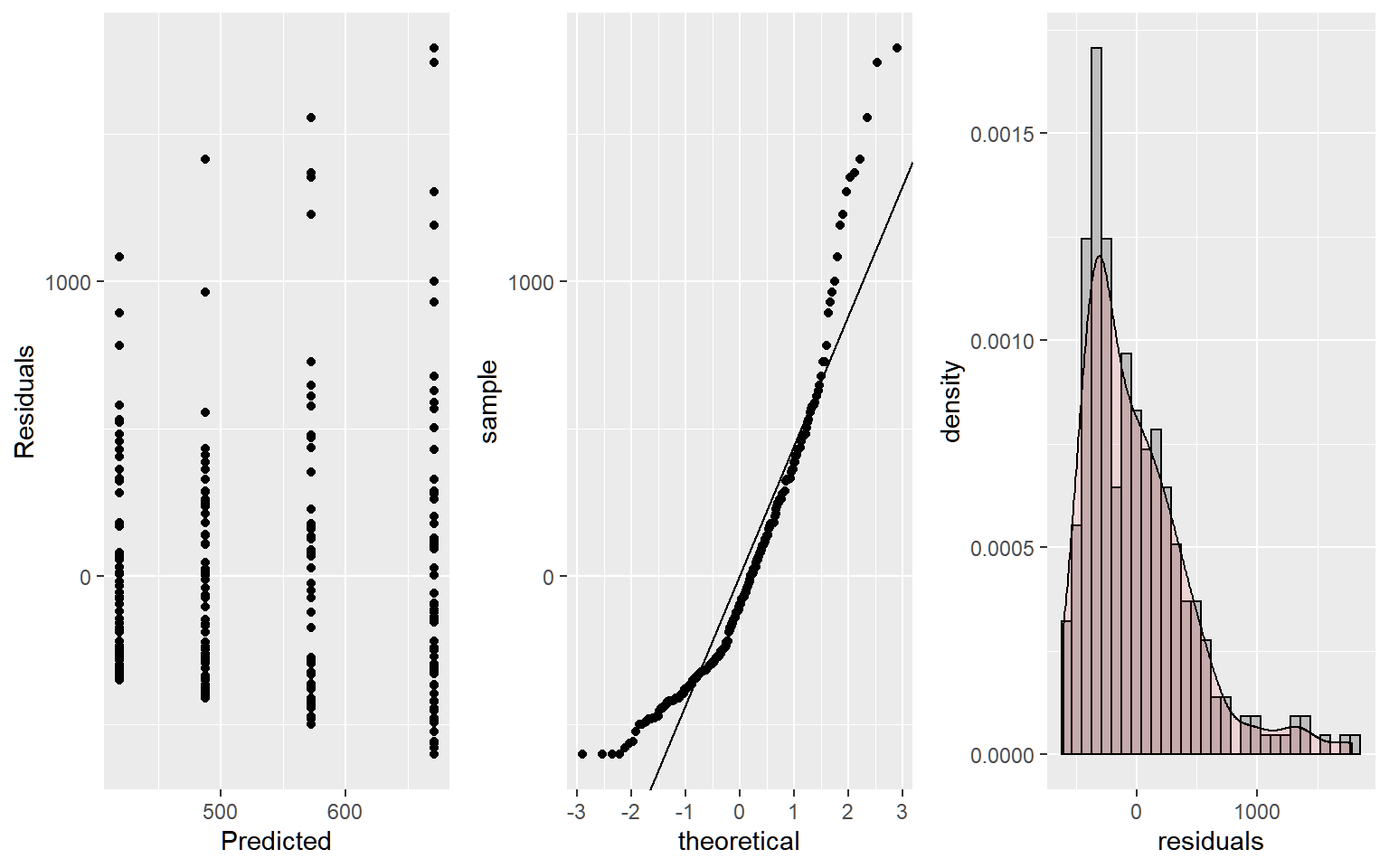
We will conduct a Two Way ANOVA test to see if there is a difference in mean salary when considering the League and Division of a player. Recall from Figure <x> the we didn’t see obvious differences in League but there appeared to be differences in Division

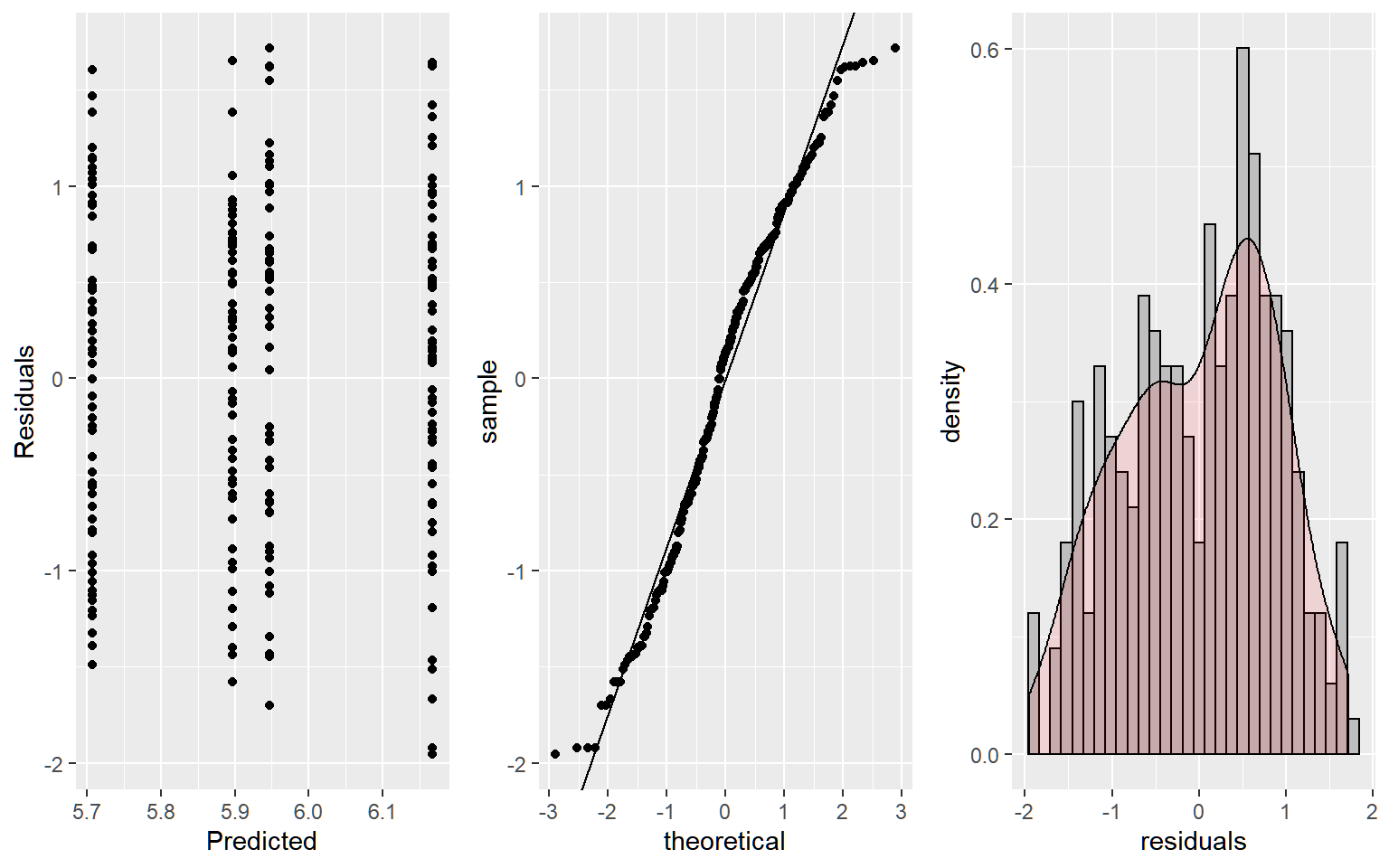
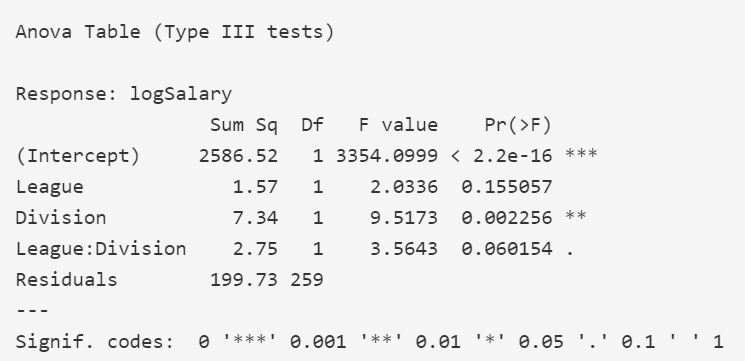
## Main Analysis Content

First, we looked at the means plot in figure <x>. We can see that the variances are clearly different by Division, but not by League. The mean of Salary is lower in the West Division for both Leagues, but neither League is consistently above the other. There might be a non-additive relationship here.



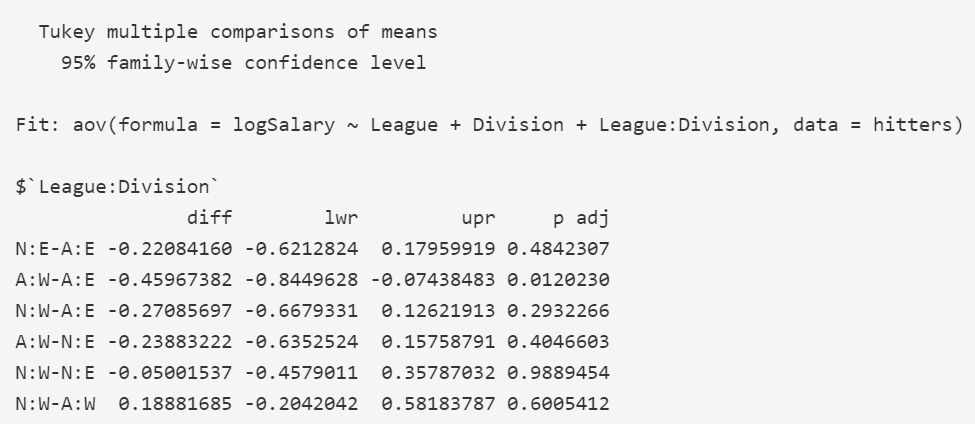
The F-Test for the interaction between League and Division does not appear significant, however looking at the diagnostics for this model we see that we have issues with normality and non-constant variance in the residuals. We hinted earlier that we would deal with this using a log transformation of Salary.





The diagnostics of the logSalary model look much better, the residuals are not perfectly normal but they have significantly improved, and the variance is much more consistent. Looking at the F-Tests, Division is still significant, and the interaction term just barely fails the F-Test. Because it is close, we may want to evaluate differences both with and without it.

## 



There is evidence that logSalary is different between the East and West Division (p-value 0.0145). If we look at both league and division, it seems the E-W difference is only present in the American League.

## Conclusion / Discussion

The East Division has a higher means salary.

# Appendix

## Well Documented Code

## Summary Graphics and Tables