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.

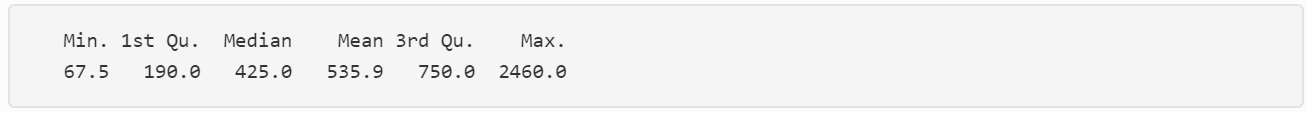
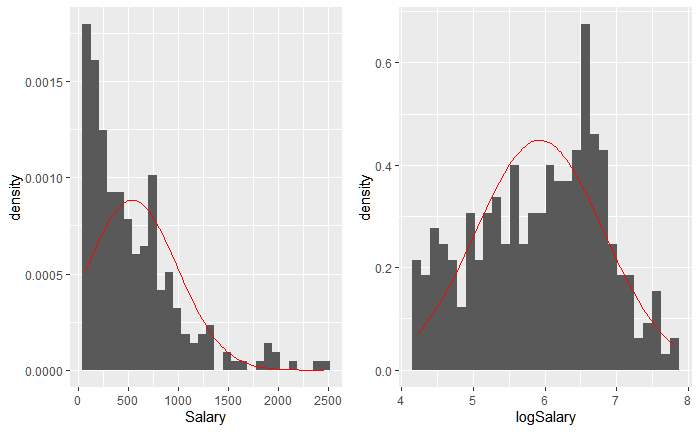
# Data Description

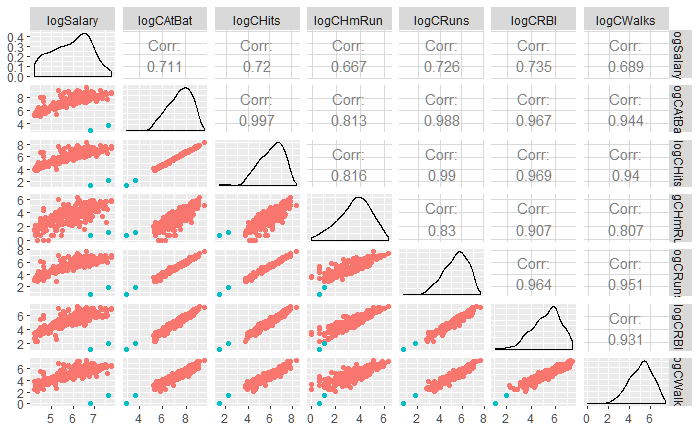
This data set contains salary information and career statistics for major league baseball players. The salary data was original from Sports Illustrated (April 20, 1987) while the career statistics are from the 1987 Baseball Encyclopedia Update published by Collier Books. We sourced the data from Kaggle (<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.

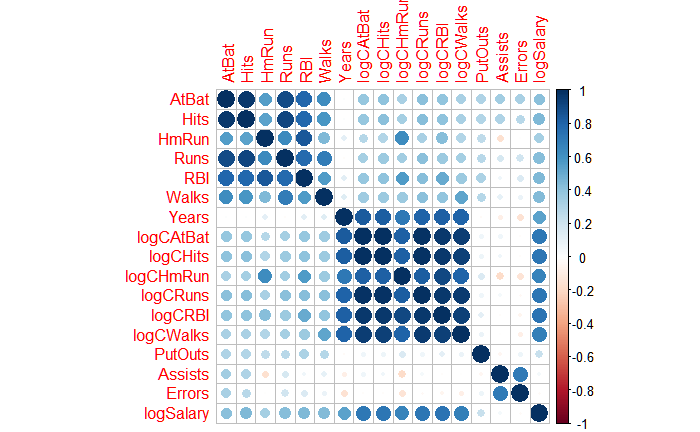
# Exploratory Data Analysis

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 two figures to the right 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.

Next, we look at a pairs plot to the left with logSalary and logged career statistics, which we transformed for similar reasons to salary. We see positive correlation with each of these variables compared to salary, and we can see that they have strong correlation with each other. We will need to check for multicollinearity in our model. We also identified two players who seem consistently out of place (colored in teal). They appear to be rookies with very few At Bats and other metrics.

The correlation matrix to the right shows how in season statistics and career statistics are related to each other strongly. However, in season and career statistics are not strongly correlated except for maybe Home Runs. Put Outs, Assists, and Errors do not seem to have a relationship with salary. We can see that the career statistics seem to have the strongest correlation with Salary. (In this plot, large dark blue circles indicate correlation close to 1).

# 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

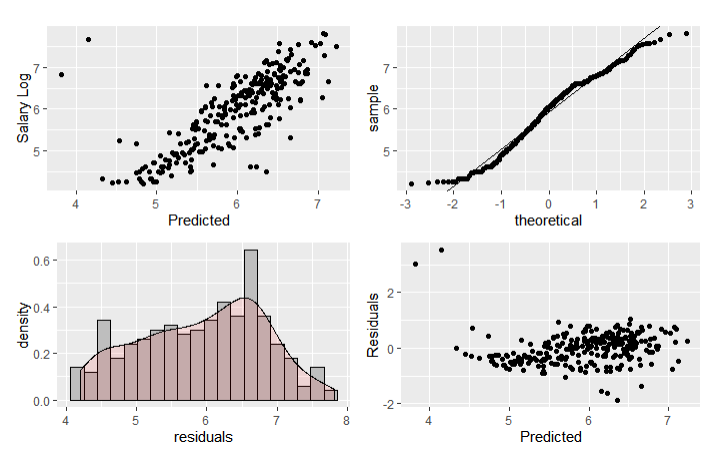
For model selection, we will compare four different possibilities in R, these being Least Squares, Ridge, Lasso and Elastic net. Additionally, we will perform an additional Lasso test in SAS for comparison. As mentioned above, logging salary seems to be appropriate. It also appears that logging the career 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

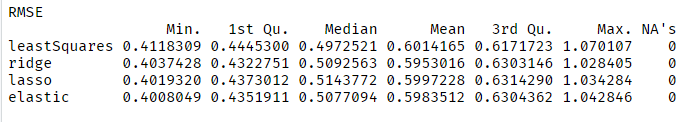
For the purpose of assumption checking, the Lasso model was used (see Comparing Competing Models). All models performed roughly the same, but since Lasso has fewer parameters, we opted for the simpler model.

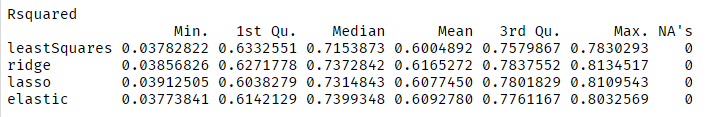
The Residual vs Predicted (bottom right) is roughly a cloud shape around 0 and the QQ plot shows that the residuals are somewhat normally distributed, although there are few significant outliers.



## Compare Competing Models

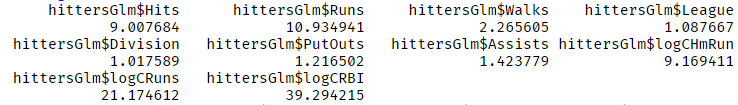
We performed a comparison between 4 different models and the mean RMSE value was nearly the same for all 4. We initially select Lasso as it resulted in the least number of coefficients, thereby producing a simpler, easier to explain model. The following graphic shows the various RMSE and Rsquared values computed during the 4 model selections, and the variability.



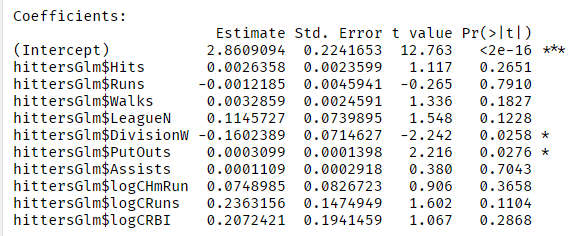


We then took the features selected by Lasso and created a model for it using the entire data set. In this case, the R2 for the model was nearly the same as the other test models, and the RMSE was a bit less too.

However, there still exists a lot of collinearity between the attributes selected. Using vif in the car library we found that there is some correlation between predictors, which makes sense. For example, hitting a home run will increase your Run and RBI attributes.



Additionally, after performing a summary on the glm object, we found the many of the parameters aren’t statistically significant. Similarly, when computing the confidence intervals, we found that many spanned 0.



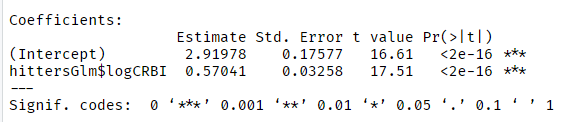
We then performed a final model using just one predictor, Career RBIs, since it probably represents the most productive attribute for a ball player. We found it’s RMSE to be .60, comparable to the mean RMSEs found while comparing the results from various models. If we remove the outliers, the RMSE goes to .56.

## Parameter Interpretation

### Interpretation & Confidence Intervals

Our final model includes just 1 predictor (Career RBIs) and the outliers found above during EDA. Note, we could make an argument to remove then. They are for rookies, those with very few At Bats. For the purpose of using existing MLB stats for making predictions, we could exclude them but they were left in the final model.

The estimated coefficient value for Career RBIs was found to be .5704 with a confidence interval of .5065 to .6343. Doubling the number of RBIs would result in a 2^(.5704) change in the mean salary, or approximately 1.5 times more dollars.





## Final Conclusions for Objective 1

We compared different selection techniques and found all of them preformed nearly the same. The Lasso model at first was chosen for simplicity and to possibly help reduce collinearity between the predictors. Although the final model chosen still shows evidence collinearity (based on VIF), the parameters chosen are still statistically significant and their estimates are contained within the computed confidence intervals. However, many of the confidence intervals spanned 0.

It should be noted that the model can be significantly changed based on what seed, number of folds, and partition (test vs train) sizes that are used. For example, using SAS to perform a lasso selection with glmselect produced a model with just Hits, Career Runs Log, Career RBIs Log, and Putouts. So this is even simpler, although not necessarily better. There is still a high correlation between Runs and RBIs, which makes sense, since things like RBIs, Runs, Hits, and Homeruns are all corelated with each other. Hitting a home run increase your hits, runs and RBI all at the same time.

We then tried a much simpler model, using Career RBIs. The test RMSE was found to be only slightly higher (.60 vs .57). We opted to go with using Career RBIs as it is easier to explain and is easily understood. The prediction power of using it vs the features selected by Lasso isn’t all that much different.

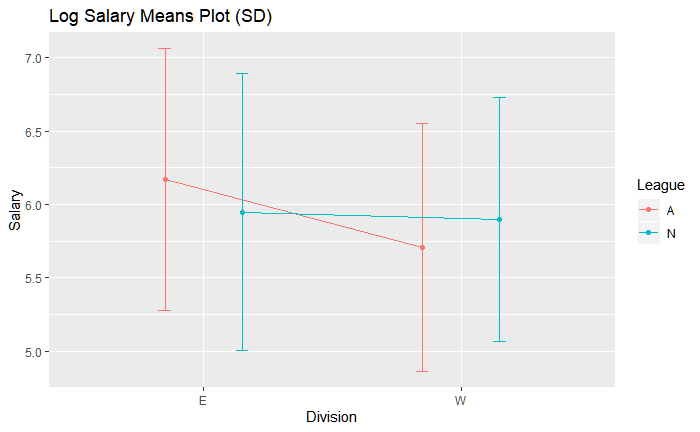
# Objective 2: Advanced Analysis

## State the method

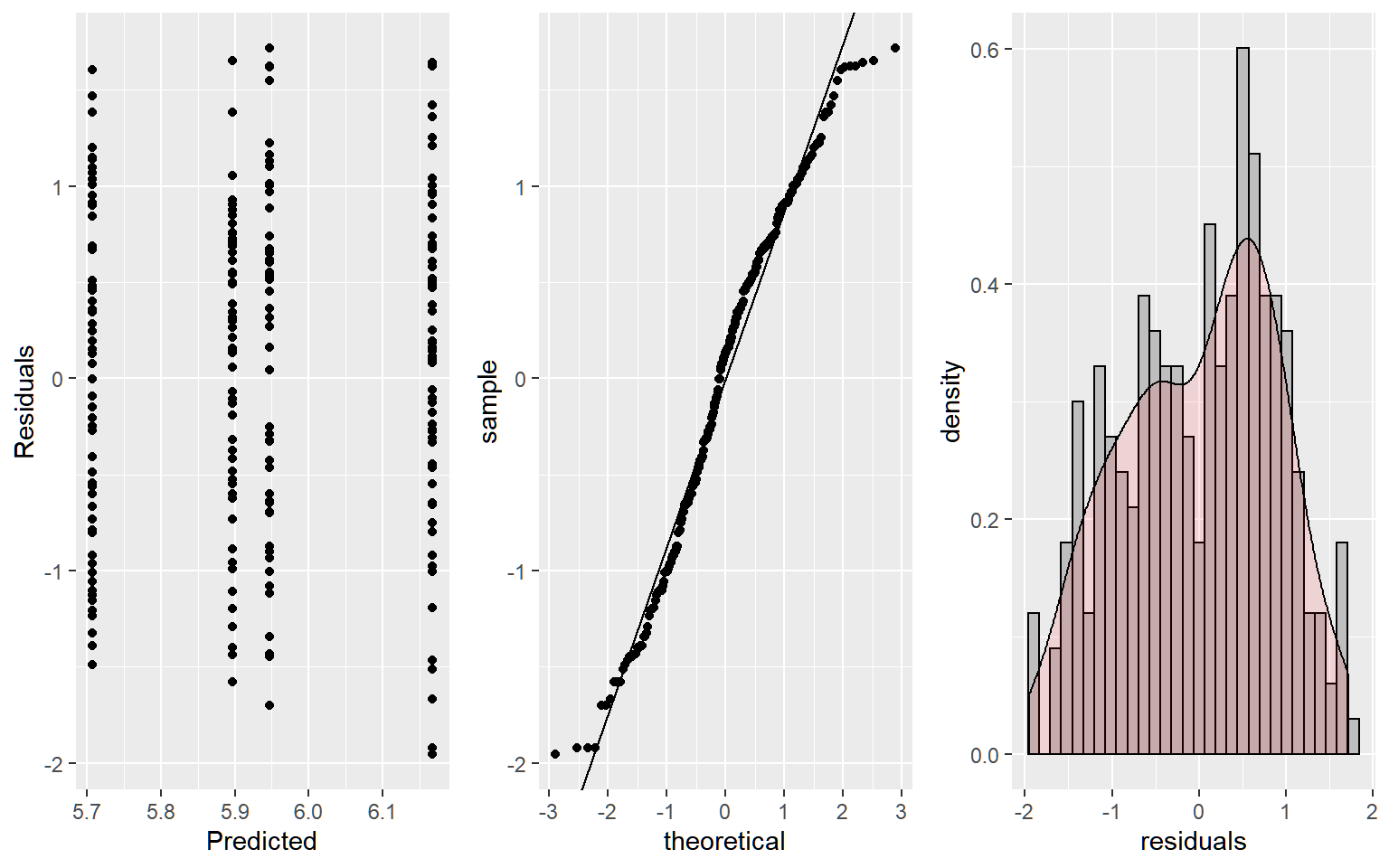
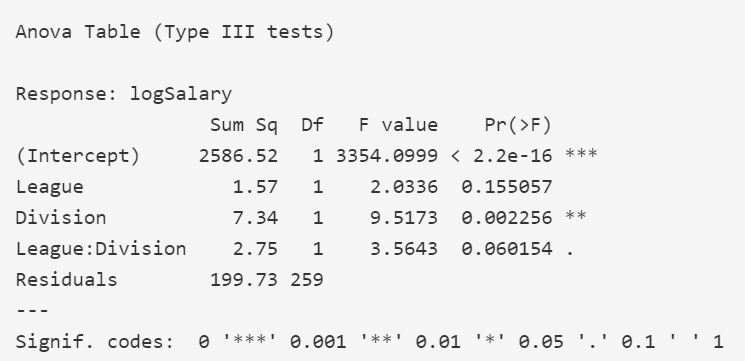
We will conduct a Two Way ANOVA test to see if there is a difference in mean salary when considering the League and Division of a player.

## Main Analysis Content

First, we looked at the means plot in the top left figure. The means are only slightly different for the National League across divisions. The American league shows a larger difference in log Salary, which is higher than the National league in the East division but below it in the West Division. There might be a non-additive relationship here.

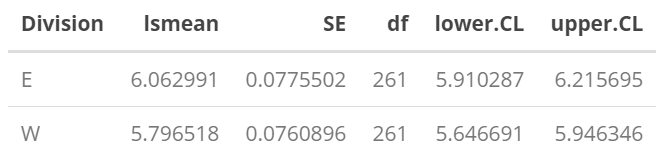


Two Way ANOVA Model



In the Anova table (left figure above), Division is significant, and the interaction term just barely fails the F-Test. League doesn’t appear significant either. As a result, we will focus on the differences in Division since it was the only significant factor. The diagnostics of the logSalary model look good (see the 3 plots in the right hand figure above), the residuals are mostly normal and variance appears constant. The observations are independent in that players in one League or Division cannot appear in another, but this deserves more discussion later.

This means table is for reference for the contrast to follow



In the contrast below, we can see there is evidence that logSalary is different between the East and West Division (p-value 0.0148). However, the estimated difference of -0.266 translates to a median salary difference of $766, with a confidence interval of $619 to $949, which doesn’t seem practically significant given the minimum salary in our dataset is $67.5K



## Conclusion / Discussion

Using a two way anova analysis we found that there was a difference in median salary between the east and west division, however the amount is unlikely to be practically significant. We did not find evidence that there was a difference in median salary between the American and National leagues, and that there was not enough evidence to support an interaction between league and division. I also ran a model removing the two rookie players, but the conclusions didn’t change.

Regarding independence of observations, there are some subtleties beyond whether players not being present in multiple divisions. If a team is losing, their innings end quicker, and everyone on that team probably takes a hit to statistics like at bats and hits. Also, salaries may be correlated among players on the same team. Unfortunately, we don’t have player names, positions, or teams in this data set to look if we are still ok to assume independence. We are assuming these don’t have a major impact on our tests.

# Appendix

Most of our code and graphics are documented in the attached .RMD file and on our github located here: <https://github.com/rmpalmer/DS6372_402_fall19_proj1>

We did some confirmatory checks in SAS, that code is found here:

SAS Code:

**proc** **import** datafile="C:\\Users\\dancr\\OneDrive\\Documents\\Data Science\\SMU\\MSDS 6372 - Applied Stats\\Project 1\\Hitters.csv"

dbms=dlm out=hitters replace;

delimeter=',';

getnames=yes;

**run**;

/\* Remove NAs. Create Numeric attribute for Salary \*/

**data** hitters;

set hitters;

If Salary = 'NA' Then Delete;

SalaryNumeric = input(Salary, best12.);

**run**;

/\* Log features \*/

**data** hitters;

set hitters;

SalaryLog = log(SalaryNumeric);

CRBILog = log(CRBI);

CHitsLog = log(CHits);

CAtBatLog = log(CAtBat);

CHmRunLog = log(CHmRun+**1**);

CRunsLog = log(CRuns);

CWalksLog = log(CWalks);

**run**;

/\* Compare standard vs log, for evidence of logging C\* attributes \*/

**proc** **reg** data=hitters plots(label)=(rstudentleverage cooksd);

model SalaryLog = CAtBat CHits CHmRun CRuns CRBI CWalks / VIF;

model SalaryLog = CAtBatLog CHitsLog CHmRunLog CRunsLog CRBILog CWalksLog / VIF;

**run**;

**quit**;

/\* SAS Model selection using Lasso \*/

**proc** **glmselect** data = hitters plots(stepaxis = number) = (criterionpanel ASEPlot) seed=**1**;

partition fraction(test = **.25**);

class Division League NewLeague;

model SalaryLog = AtBat Hits HmRun Runs RBI Walks

CAtBatLog CHitsLog CHmRunLog CRunsLog CRBILog CWalksLog

PutOuts Assists Errors Years Division League NewLeague

/ selection = lasso(choose=cv stop = AIC) CVDETAILS

CVMETHOD=RANDOM(**10**);

**run**;