Statistical Linear Models Project 1

Math 158, Linear Models, Spring 2018

Moira Dillon & Mike Adams

Due: Monday, February 5, 2018

We will be analyzing The Lahman Baseball Database, available on R. This database includes pitching, hitting, and fielding stats for Major League Baseball from 1871 through 2016. This data set is the largest available for baseball statistics. While this data set is available in R, we compiled the files together using a Python script and imported this csv into R (file and script on Github). We will be analyzing players salaries, awards, statistics for batting & fielding during regular and post season, along with their birth date, height, weight, playing hand, and birth country.

While there are various analyses we will be able to conduct, there are a few in particular we are looking forward to. We are interested in the relationship between playing statistics (batting & fielding) and recognition of success (awards & salaries), specifically how this may change over time. We are also interested in analyzing birth month & playing statistics, as Malcolm Gladwell suggests in *Outliers* that players born earlier in the calender year tend to be more likely to reach the professional level.

```
Lahman <- read.csv("Lahman.csv")
```

Our Github repository for this project is available here: https://github.com/moiradillon2/SLM_MAMD/

The observational unit for our data is the player, but there are possible ways to analyze the data by other observation units. For example, we may decide to explore the data using teams instead of players as the observational unit. The columns corresponding to year, stint (play with more than one team per season), team, league, games played, at bats, runs, hits, doubles, triples, home runs, RBIs, stolen bases, caught stealing (CS), walks, strikeout, intentional walks, hit by pitch, sacrifice hit, sacrifice fly, grounds into double play, salary, awards, award ties, birth month, birth country, weight, height, bats with R/L hand, throws with R/L hand, and full name.

For the quantitative variables, we can use skim() to summarize the vector, outputting the number of missing and complete entries, the mean, SD, and median, the variable type, and a histogram. Below is an example for number of runs.

```
skim(Lahman$R)
```

```
## Skim summary statistics
```

```
## Warning: package 'bindrcpp' was built under R version 3.2.5
```

```
##
## Variable type: integer
##  variable missing complete      n  mean    sd p0 p25 median p75 p100
##  Lahman$R       0   102816 102816 18.82 28.24  0   0      4  27  192
##      hist
##
```

Using skim(), we compiled the number of missing and complete entries along with the mean, SD, and median for all of the quantitative variables of interest in our data set.

```
summary_quant <- read.csv("summary.csv")
summary_quant
```

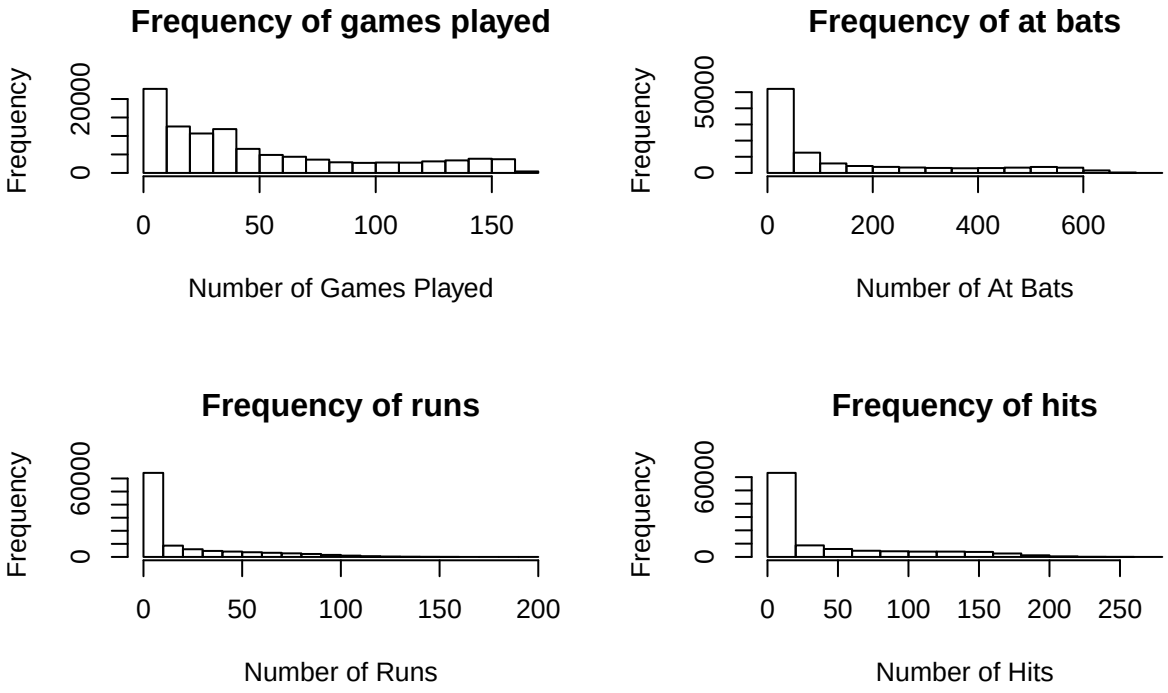
```
##    Variable missing complete       mean         sd median
## 1         G       0   102816      51.34      47.12     34
## 2        AB       0   102816     141.91     184.65     49
```

```
## 3       RBI       0   102816      18.82      28.24      4
## 4         H       0   102816      37.14      52.60      9
## 5      twoB       0   102816       6.29       9.66      1
## 6    threeB       0   102816       1.29       2.65      0
## 7        HR       0   102816       2.81       6.30      0
## 8       RBI     424   102392      17.00      26.35      3
## 9        SB    1300   101516       2.98       7.72      0
## 10       CS   23456    79360       1.23       2.75      0
## 11       BB       0   102816      13.07      20.75      3
## 12       SO    7838    94978      20.53      28.33      9
## 13      IBB   36565    66251       1.11       2.78      0
## 14      HBP    2810   100006       1.06       2.28      0
## 15       SH    6338    96478       2.30       4.24      0
## 16       SF   36034    66782       1.05       1.96      0
## 17     GIDP   26110    76706       2.98       4.74      0
## 18   salary   74704    28112 2120866.95 3444341.82 600000
## 19   weight    1184   101632     187.76      21.38    185
## 20   height    1125   101691      72.43       2.54     72
```

This allows us to compare mean, median and standard deviation for batting statistics, as well as salary, height, and weight, which provide important insight into our future analyses. For a majority of the variables, there is not an entry for every player in the data set, which will be important to keep in mind during future analyses.

We can create histograms to visualize the distribution for a variety of the batting statistics. Below we see plots for games played, at bats, runs, and hits.

```
par(mfrow=c(2,2))
hist(Lahman$G, xlab = "Number of Games Played", main = "Frequency of games played")
hist(Lahman$AB, xlab = "Number of At Bats", main = "Frequency of at bats")
hist(Lahman$R, xlab = "Number of Runs", main = "Frequency of runs")
hist(Lahman$H, xlab = "Number of Hits", main = "Frequency of hits")
```



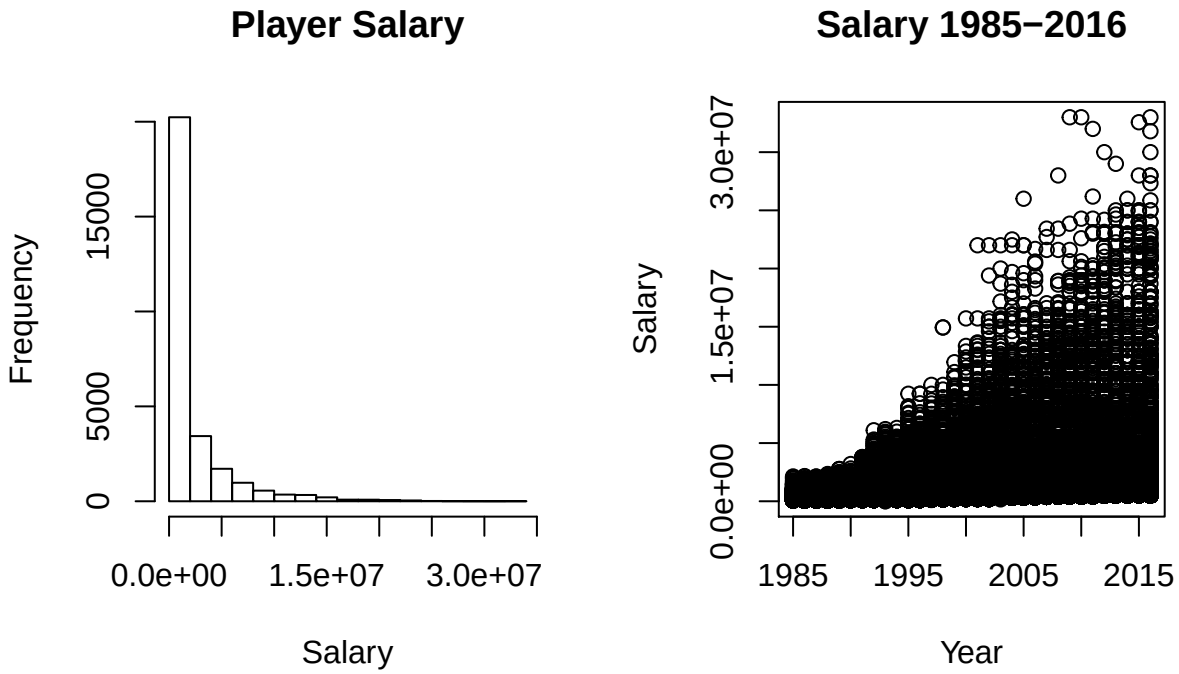
Note that this set of histograms does not include all of the quantitative variables in our data set.

For each variable plotted here, we observe that the data is skewed far to the left. For the number of games played per individual, we observe a slightly more symmetric distribution, although still skewed left.

It will be interesting to analyze how these frequency numbers change as we subset the data. For example, a question we might ask is "Do award winning baseball players have significantly more hits/runs/etc. than the entire professional baseball population?"

In addition to how playing statistics vary between award winners and the entire baseball population, we can analyze this data set based on salary.

```
par(mfrow=c(1,2))
hist(Lahman$salary, xlab = "Salary", main = "Player Salary")
plot(Lahman$yearID, Lahman$salary, xlim = c(1985,2016), xlab = "Year", ylab = "Salary", main = "Salary
```

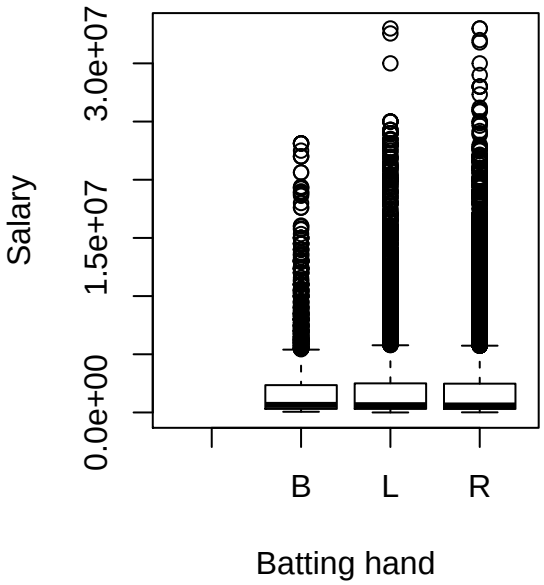


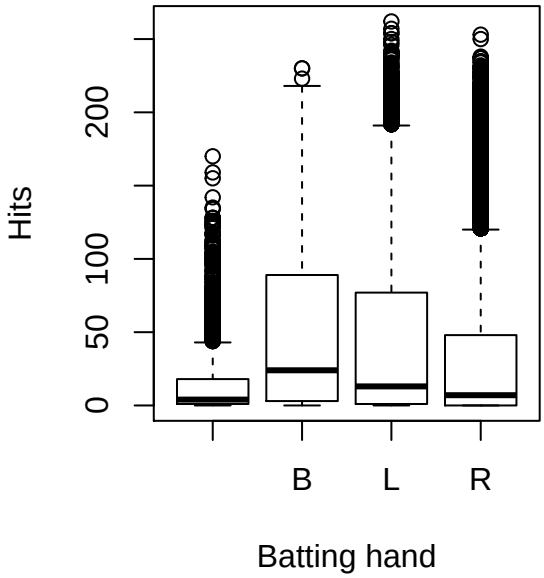
We observe that the data is also skewed far to the left. From the summary table above, we see that the mean salary is $2,120,866.95 while the median salary is $600,000. Based on the mean, median, and histogram, it is clear that some very highly paid players skew the mean upwards, while most players earned closer to the median value of $600,000.

Similar to how playing stats will likely vary significantly between award winners & all other baseball players, it will be interesting to analyze how salary changes for players based on their playing stats and their award records. By plotting salary over time (right), we observe that the spread of the distribution of salary has increased each year. (Note there is not salary data before 1985 in this data set).

In addition to the quantitative variables, there are a handful of categorical variables that will be very interesting to analyze, including batting and throwing hand (right or left), birth month, awards, and team and league. To begin, we can visualize how handedness may influence batting statistics and salary.

```
par(mfrow=c(1,2))
plot(Lahman$bats, Lahman$salary, xlab = "Batting hand", ylab = "Salary")
plot(Lahman$bats, Lahman$H, xlab = "Batting hand", ylab = "Hits")
```





Batting hand indicates which hand the player bats with - right, left, or both (switch-hitter). Note that for some players we do not have handedness but we do have hit data, which is why there is a box without a corresponding batting hand. While it appears there may not be a difference in salary, it appears there is more variance in number of hits based on batting hand.

Overall, we expect this to be a very interesting data set to explore with many interesting questions already emerging. Because our data set includes all professional baseball players, it is not necessarily a sample describing a larger population. However, we can subset our data and analyze how that subset compares to the entire baseball population. For example, as suggested above, we can analyze how playing statistics changes between those who have won awards versus the entire baseball population. There are similar questions we can ask with how salary changes with playing stats.