INFX 573 Lab: Simple Linear Regression

Pierre Augustamar

November 1st, 2016

Collaborators:

Don't forget to list the full names of your collaborators!

Instructions:

Before beginning this assignment, please ensure you have access to R and/or RStudio.

1. Download the `week6a_lab.Rmd` file from Canvas. Open `week6a_lab.Rmd` in RStudio (or your favorite editor) and supply your solutions to the assignment by editing `week6a_lab.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name.
3. Be sure to include code chucks, figures and written explanations as necessary. Any collaborators must be listed on the top of your assignment. Any figures should be clearly labeled and appropriately referenced within the text.
4. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit`, rename the R Markdown file to `YourLastName_YourFirstName_lab6a.Rmd`, and knit it into a PDF. Submit the compiled PDF on Canvas.

In this lab, you will need access to the following R packages:

```
# Load some helpful libraries
library(tidyverse)
```

Sports Statistics: Predicting Runs Scored in Baseball

Baseball is a played between two teams who take turns batting and fielding. A run is scored when a player advances around the bases and returns to home plate. The data we will use today is from all 30 Major League Baseball teams from the 2011 season. This data set is useful for examining the relationships between wins, runs scored in a season, and a number of other player statistics.

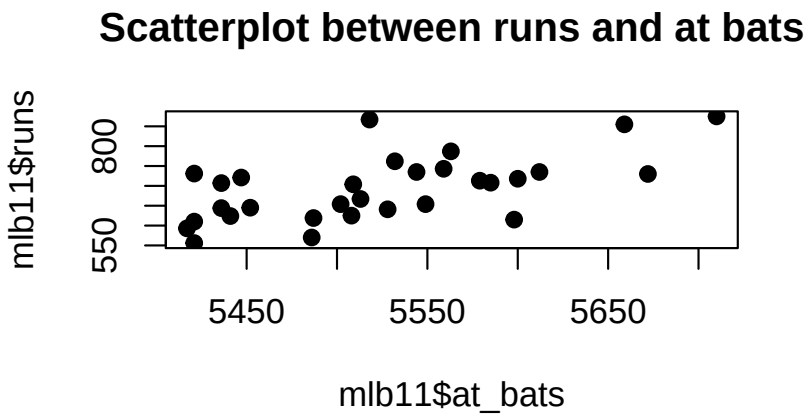
Note: More info on the data can be found here: https://www.openintro.org/stat/data/mlb11.php

```
# Download and load data
download.file("http://www.openintro.org/stat/data/mlb11.RData",
    destfile = "mlb11.RData")
load("mlb11.RData")
```

Use the baseball data to answer the following questions:

- Plot the relationship between runs and at bats. Does the relationship look linear? Describe the relationship between these two variables.

```
# plot relationiship between runs and bats
plot(mlb11$runs ~ mlb11$at_bats, main = "Scatterplot between runs and at bats",
    pch = 19)
```



Response - Does the relationship look linear? Describe the relationship between these two variables.

The graph shows a potenatial positive linear graph, but it is not strong nor weak. I would say it is roughly a moderate linear graph. The relationship between the two variables appear to be a positive correlation.

Response - If you knew a team's at bats, would you be comfortable using a linear model to predict the number of runs?

The scatter plot alone is not sufficient to determine this because it shows a moderate linear correlation between runs and at bats. I would need to calculate the corellation coefficient to have a better idea if a relationship were to exist between the two.

Response - If the relationship looks linear, quantify the strength of the relationship with the correlation coefficient. Discuss what you find.

```
# plot relationiship between runs and bats
cor(mlb11$runs, mlb11$at_bats)
```

```
## [1] 0.610627
```

The correlation coefficient of runs and ata bats is 0.610627. Since it is positive value then it is a uphill linear relationship, but it is well below 1. A value close to 1 would have projected a strong positive linear.Thus, it is a moderate huphill positive relationship. Also, it is worth nothing that there are few outliers.

- Use the `lm()` function to fit a simple linear model for runs as a function of at bats. Write down the formula for the model, filling in estimated coefficient values.

```
# plot relationiship between runs and bats
runAndAtBats = lm(mlb11$runs ~ mlb11$at_bats,
    data = mlb11)
summary(runAndAtBats)
```

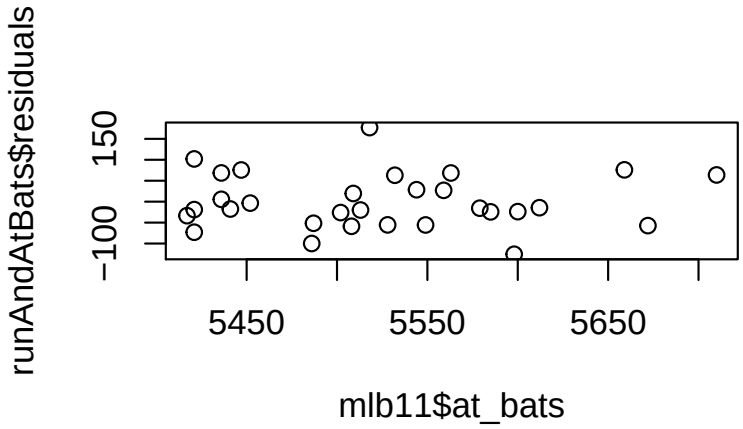
```
##
## Call:
## lm(formula = mlb11$runs ~ mlb11$at_bats, data = mlb11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -125.58  -47.05  -16.59   54.40  176.87
##
## Coefficients:
##                 Estimate Std. Error t value
## (Intercept)   -2789.2429   853.6957  -3.267
## mlb11$at_bats     0.6305     0.1545   4.080
##               Pr(>|t|)
## (Intercept)   0.002871 **
## mlb11$at_bats 0.000339 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## Residual standard error: 66.47 on 28 degrees of freedom
## Multiple R-squared:  0.3729, Adjusted R-squared:  0.3505
## F-statistic: 16.65 on 1 and 28 DF,  p-value: 0.0003388
```

Response - Describe in words the interpretation of β_1.

Y = bo + b1x implies tthat from the coefficients estimate then we have runs = -2789.2429 + 0.6305*at_bats. beta_1 represents the slope. It signifies that for every added run by a batter, the required at bats goes up by 0.6305.

- Make a plot of the residuals versus at bats. Is there any apparent pattern in the residuals plot?

```
# plot residuals versus at bats
plot(runAndAtBats$residuals ~ mlb11$at_bats)
```



Response

data are scattered all over the place. I do not recognize any apparent correlation between residuals and at bats.

- Comment of the fit of the model.