INFX 573 Lab: Simple Linear Regression

Shuyang Wu

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)
```

```
## Warning: package 'ggplot2' was built under R
## version 3.3.2
```

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.

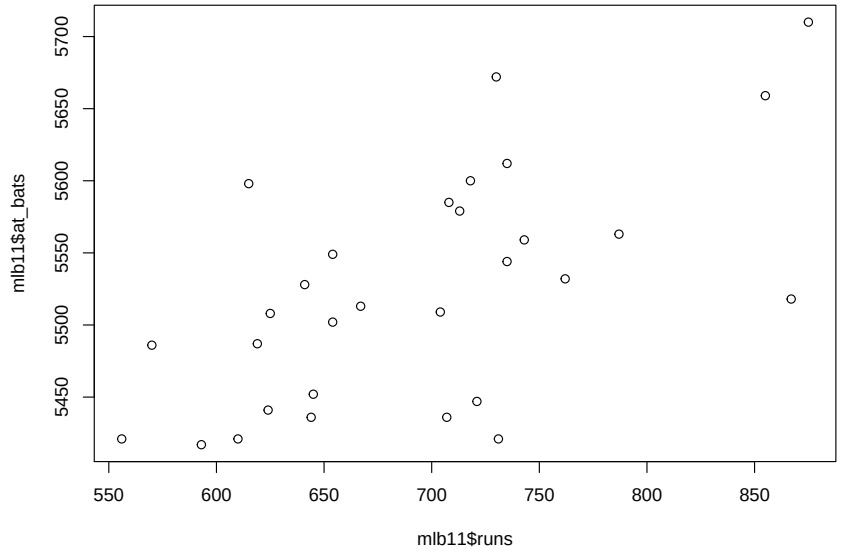
Note: More info on the data can be found here: https://www.openintro.org/stat/data/mlb11.php

```
# 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 runs and at bats
plot(mlb11$runs, mlb11$at_bats)
```




```
cor(mlb11$runs, mlb11$at_bats)
```

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

The two variables are positively correlated, the linear relationship is not very obvious because the points are spread out but I can see a trend.

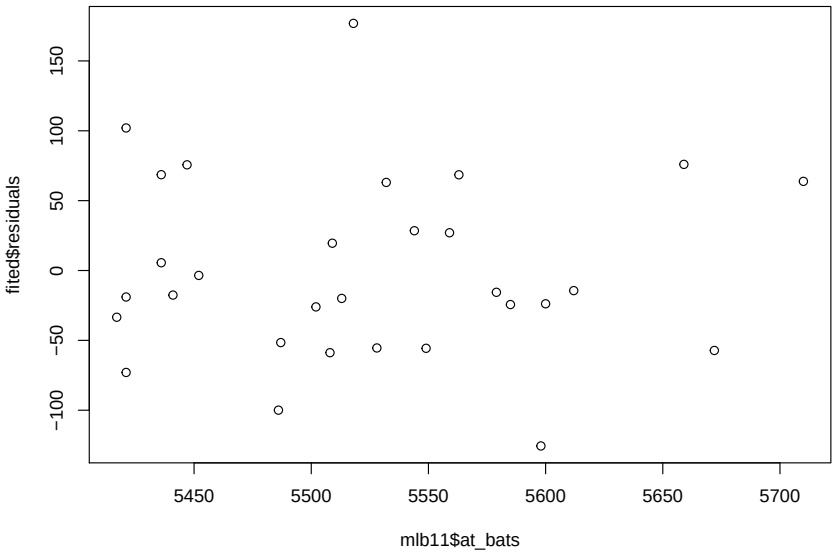
- If you knew a team's at bats, would you be comfortable using a linear model to predict the number of runs?

```
fited <- lm(runs ~ at_bats, data = mlb11)
summary(fited)
```

```
##
## Call:
```

```
## lm(formula = runs ~ 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
## at_bats         0.6305     0.1545   4.080
##             Pr(>|t|)
## (Intercept) 0.002871 **
## 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
```

```
plot(mlb11$at_bats, fited$residuals)
```



- If the relationship looks linear, quantify the strength of the relationship with the correlation coefficient. Discuss what you find. Correlation coefficient = 0.61, which means the two variables are somewhat postively correlated but not strong.
- 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.

runs = 0.6305 x at_bats - 2789.24

- Describe in words the interpretation of β_1. Change in one unit of X corresponds to change of beta_1 in Y.
- Make a plot of the residuals versus at bats. Is there any apparent pattern in the residuals plot? Most of the residuals centered around 0, but there are a few outliners.
- Comment of the fit of the model. R-square value is relatively small, which means some of the observations are a bit far from the fitted line.