# IMT 573 Lab: Linear Regression

*Miloni Desai*

*November 7th, 2019*

### Collaborators

### Objectives

Before beginning this assignment, please ensure you have access to R and RStudio; this can be on your own personal computer or on the IMT 573 R Studio Server.

1. Download the `lab6_linear_regression.rmd` file from Canvas or save a copy to your local directory on RStudio Server. Open `lab6_linear_regression.rmd` in RStudio and supply your solutions to the assignment by editing `lab6_linear_regression.rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures, and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do no need four different visualizations of the same pattern.
4. Collaboration on problem sets is fun and useful, and we encourage it, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment.
6. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the knitted PDF file to `lab6_YourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

### Setup

In this lab you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------- tidyverse 1.2.1 --
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

**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. #We can use a scatter plot to plot the relationship between runs and bats. The relationship does look linear. Since it is linear we could find the correlation coefficient to see the strength of the relationship which in this case does seem linear.

```
reg <- lm(mlb11$runs~mlb11$at_bats)
plot(mlb11$at_bats,mlb11$runs)
abline(reg)
```

#If you knew a team's at bats, would you be comfortableusing a linear model to predict the number of runs? #From the linear relationship that we found in the previous question and a strong correlation coefficient, we coud use a linear model to predict the number of runs using a teams at bats.

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

#If the relationship looks linear, quantify the strength of the relationship with the correlation coefficient. Discuss what you find. #We found the correlation coefficient and it is positive and 0.61 which is high and shows a strong relationship between at bats and runs.

#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. #The km function is a formula that takes the form y ~ . It can tells us that we want to make a linear model of runs as a function of at_bats. The second part specifies that R should take data, find variables in the mlb11 dataframe

#We use the summary function to get the estimated coefficient values.

#formula : $\hat{y}$ = -2789.2429 + 0.6305 *atbats

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

#Describe in words the interpretation of $\beta_1$. # Make a plot of the residuals versus at bats. Is there any apparent pattern in the residuals plot? # Comment of the fit of the model.