

# IMT 573: Module 7 Lab

## Regression

Jenny Skytta

Due: May 15, 2022

**Collaborators:** List collaborators here.

### Objectives

Regression is one of the fundamental and important data models we will encounter in this course. Our aim in this exercise is to see regression models in practice. Recall, our emphasis in class and here will also be on interpreting the results of these models in context, and always to evaluate the assumptions of the model to ensure valid and accurate statistical inference.

### Instructions

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 Cloud.

1. Open the `07_lab_regression.Rmd` and save a copy to your local directory. Supply your solutions to the assignment by editing `07_lab_regression.Rmd`.
2. First, replace the “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 chunks, 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 not need four different visualizations of the same pattern.
4. Collaboration on problem sets is fun and useful, and I 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. When the PDF report is generated rename the knitted PDF file to `lab7_YourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

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

```
# Load standard libraries
library(tidyverse)
library(knitr) # this will keep code on the page!
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

## Sports Statistics: Predicting Runs Scored in Baseball

Baseball is 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>

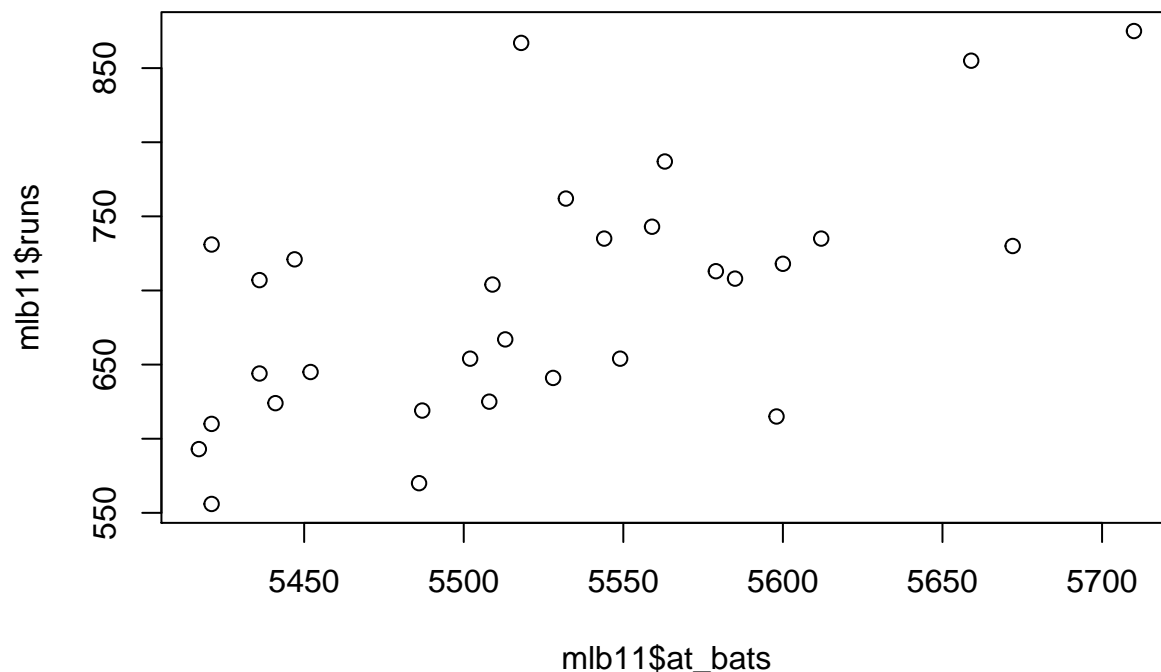
```
# Download and load data
download.file("http://www.openintro.org/stat/data/mlb11.RData", destfile = "data/mlb11.RData")
load("data/mlb11.RData")
```

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.

The relationship shown has a lot of variability and is dispersed but there still appears to be a slightly positively linear relationship. It seems to show that as at\_bats increases, runs increases as well.

```
# Simple plot of data
plot(mlb11$at_bats, mlb11$runs)
```



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

No, I think the disperement would cause some issues with prediction. The distance from the line of best fit is varied.

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

The correlation coefficient is  $\sim 0.61$  which denotes a moderate relationship. A stronger relationship would be closer to 1 or -1.

```
# Compute correlation
cor(mlb11$runs, mlb11$at_bats)
```

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

- 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 formula is as follows:  $\text{runs} = -2789.24 + 0.63 * \text{at\_bats}$   $y = b(\text{intercept}) + m(\text{slope}) * x(\text{independent variable})$

```
# Fit a simple linear model
m1 <- lm(runs ~ at_bats, data = mlb11)
summary(m1)
```

```
##
## 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 Pr(>|t|)
## (Intercept) -2789.2429   853.6957  -3.267 0.002871 **
## at_bats      0.6305     0.1545   4.080 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
```

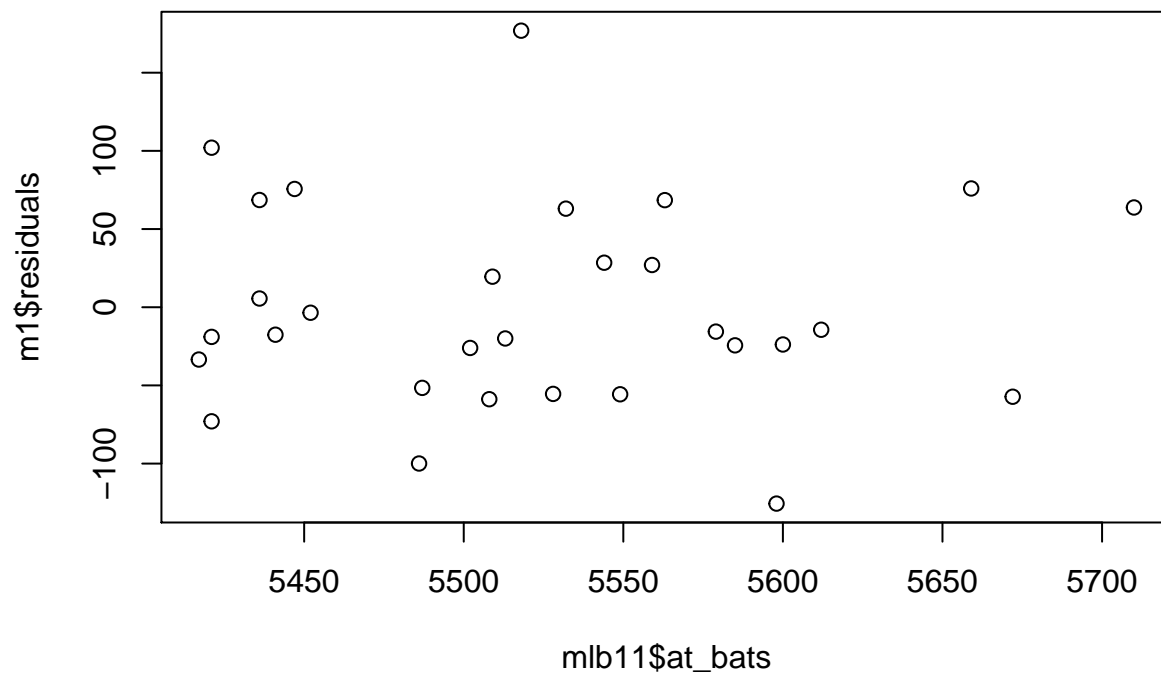
- Describe in words the interpretation of  $\beta_1$ .

Beta\_1 is represented by at\_bats variable. This variable does hold statistical significance but we have a low R-squared. Any study that attempts to predict human behavior will tend to have R-squared values less than 50% so this may not be too important. The p value suggests that our results are not random and that a correlation does exist between the variables.

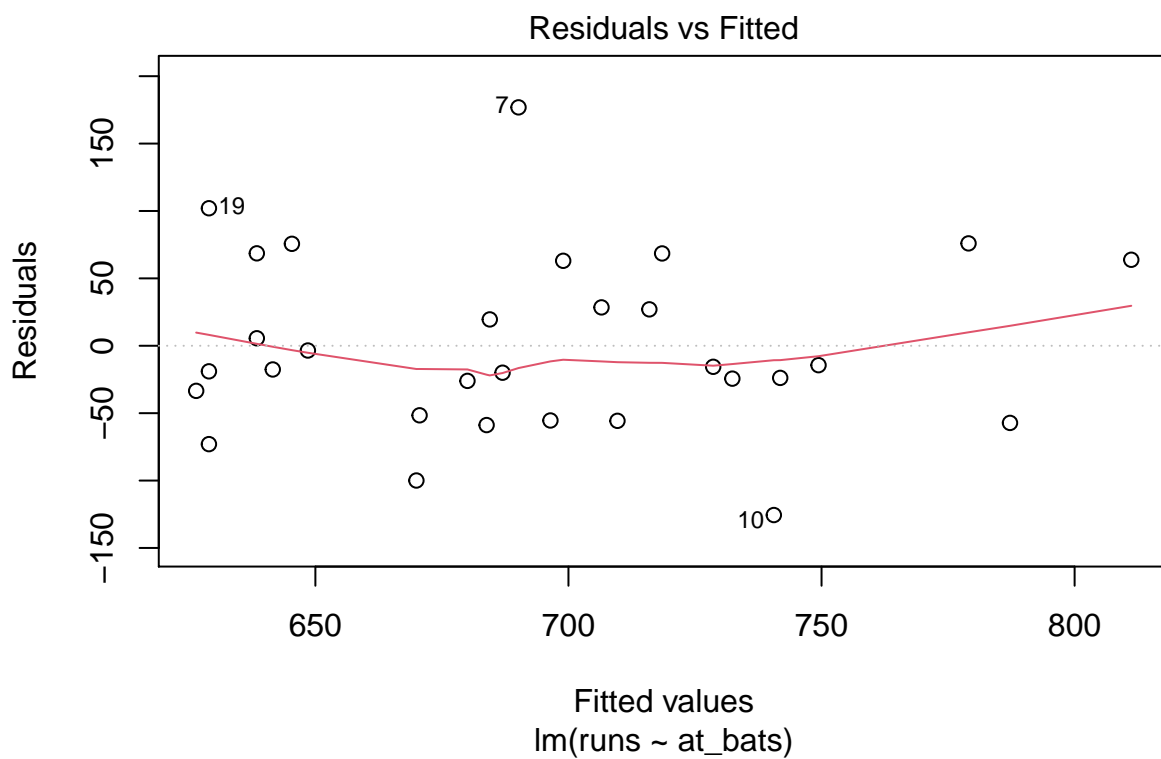
- Make a plot of the residuals versus at bats. Is there any apparent pattern in the residuals plot?

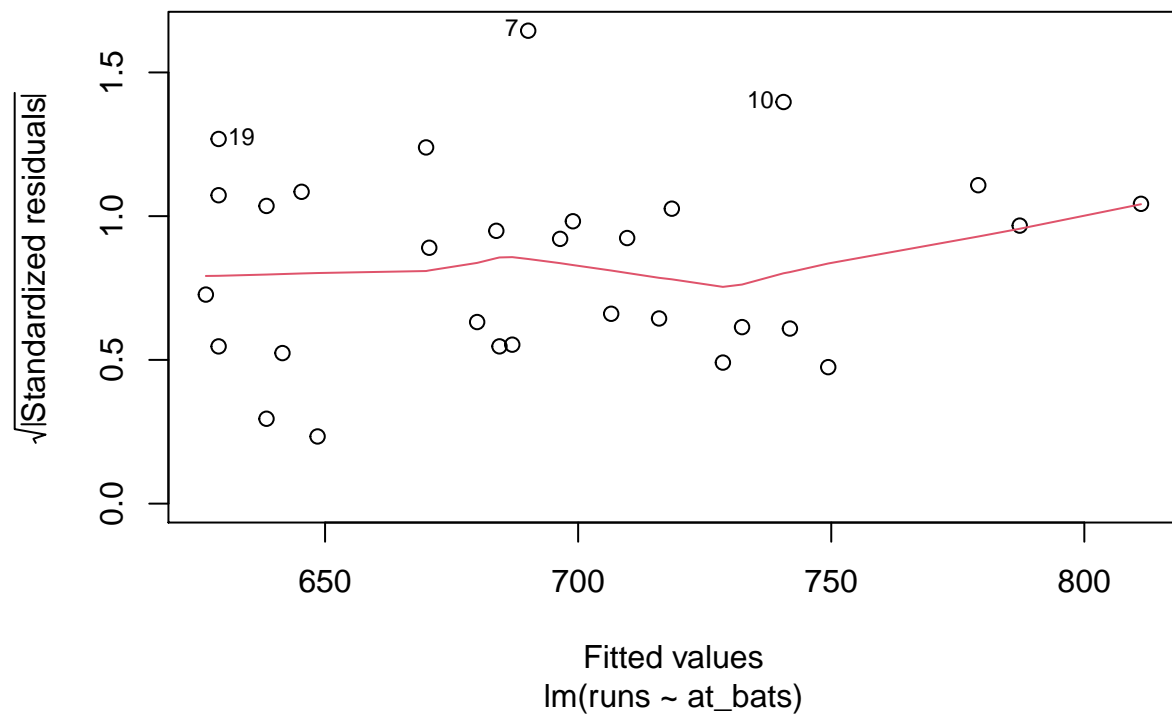
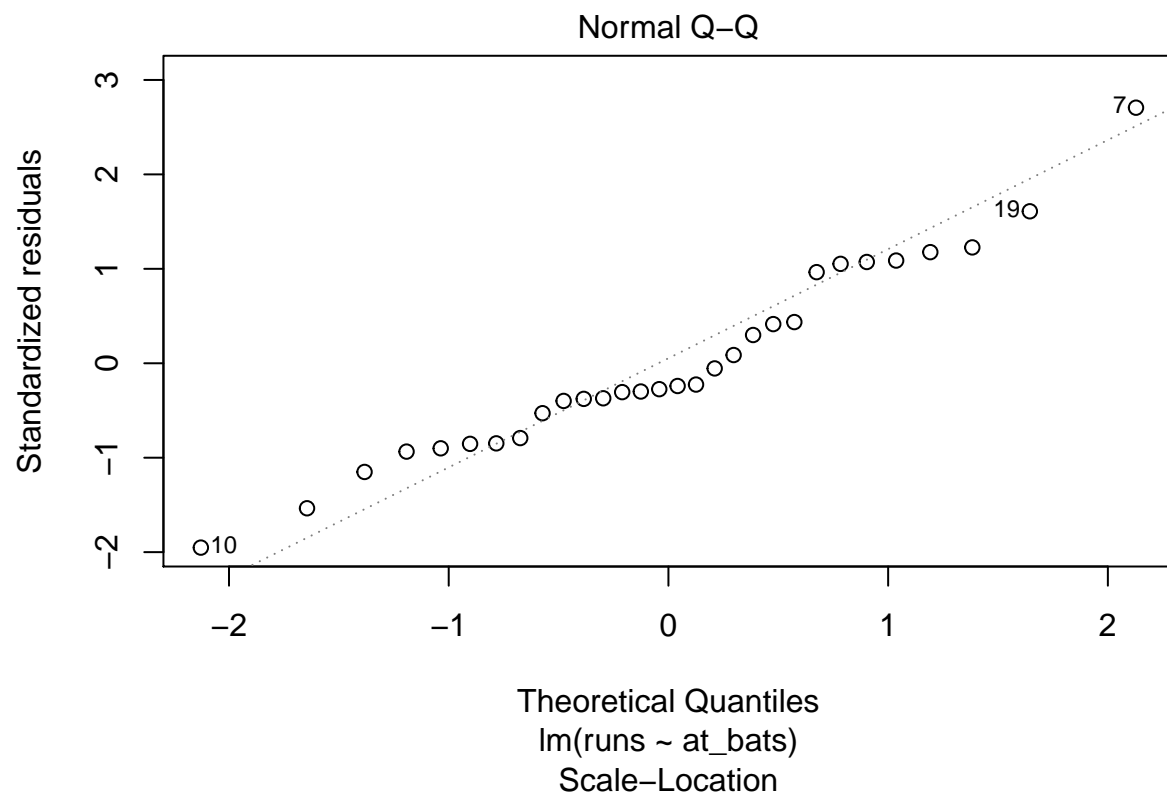
The residuals vs fitted plot is well behaved with the values loosely scattered around the 0 line. There is no discernible pattern. The Q-Q plot follows the line at a nearly 45 degree angle.

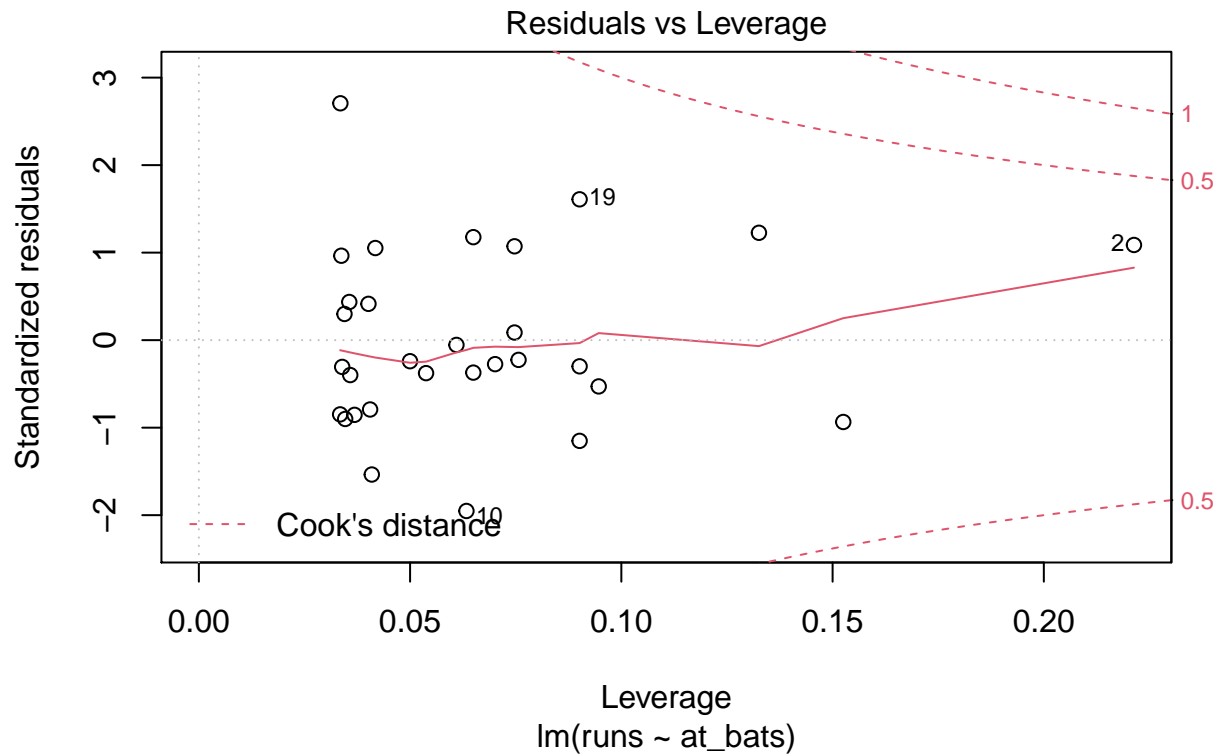
```
# Plot residuals
plot(m1$residuals ~ mlb11$at_bats)
```



```
plot(m1)
```







- Comment on the fit of the model.

Overall the fit of the model had predictive value. It may not be the best variable to form predictions but its not the worst.

### Citations

Code written above is from the previous course IMT 511 which used the below text to support class scripts.  
[https://www.google.com/books/edition/Programming\\_Skills\\_for\\_Data\\_Science/BnB6DwAAQBAJ?hl=en&gbpv=1&printsec=frontcover](https://www.google.com/books/edition/Programming_Skills_for_Data_Science/BnB6DwAAQBAJ?hl=en&gbpv=1&printsec=frontcover) —