IMT 573: Module 7 Lab

Regression - Solutions

YOUR NAME HERE

Due: August 06, 2021

Collaborators: List collaborators here.

Objectives

In this module, we have focused on exploring data. Visualization is a great way to do this. Let's play around with visualization in this lab.

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 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 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.

```r
# Load standard 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

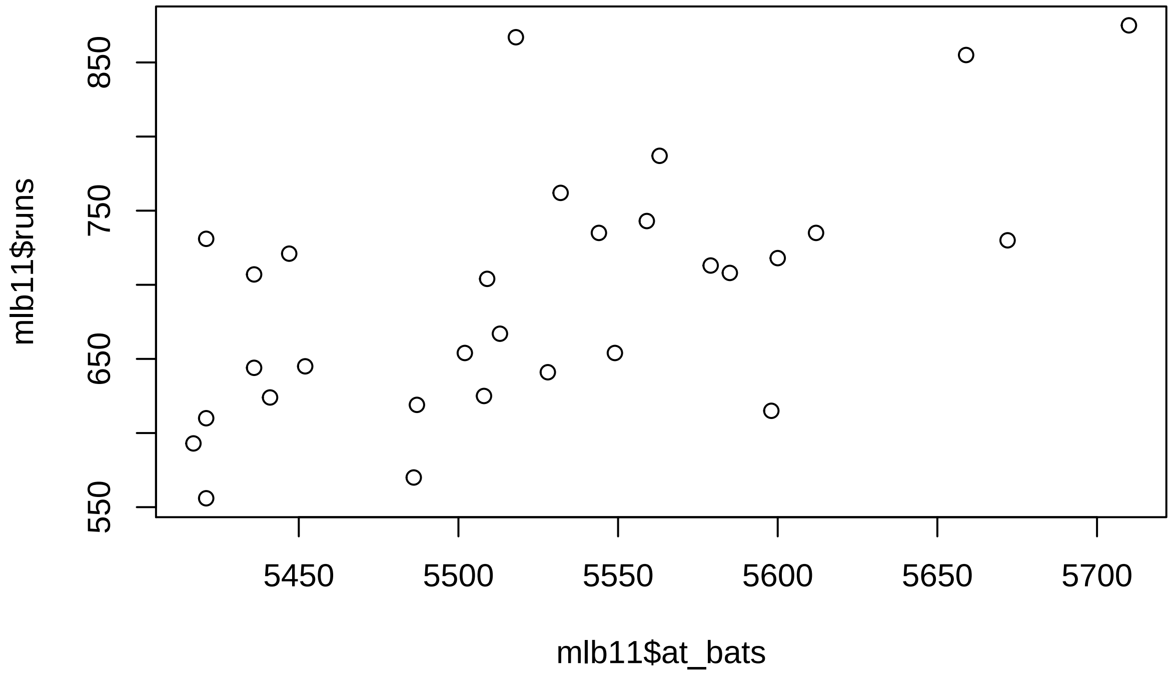
```r
# Download and load data
download.file("http://www.openintro.org/stat/data/mlb11.RData", destfile = "mlb11.RData")
load("mlb11.RData")
```

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.

Solution: After plotting the data we can see that there does seem to be a relationship between at bats and runs. We find that runs does increase as at bats increases. A linear model will likely work well, however, there does seem to be variability in the relationship as seen in the dispersed points.

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



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

Solution: While a linear model would likely be a good model for this data, there is variability. In predicting runs from at bats I would want to report an interval estimate, rather than a point estimate, to capture this uncertainty.

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

Solution: We can compute the correlation between these two variables. We find that the observed correlation in this dataset between runs and at bats is 0.61, which is a relatively strong relationship.

```r
# 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.

Solution: runs = -2789.24 + 0.6305 * at_bats

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

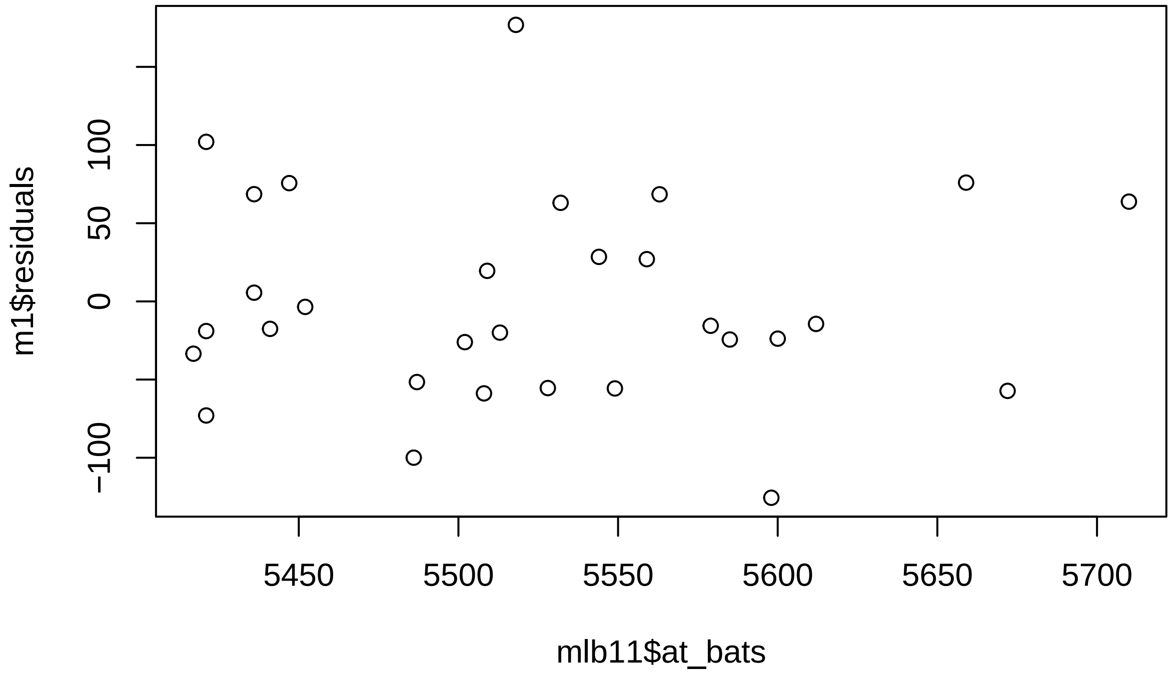
```
## 
## Call:
## lm(formula = runs ~ at_bats, data = mlb11)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -125.58  -47.05  -16.59   54.40  176.87 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -2789.2429   853.6957  -3.267 0.002871 ** 
## 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 β_1.

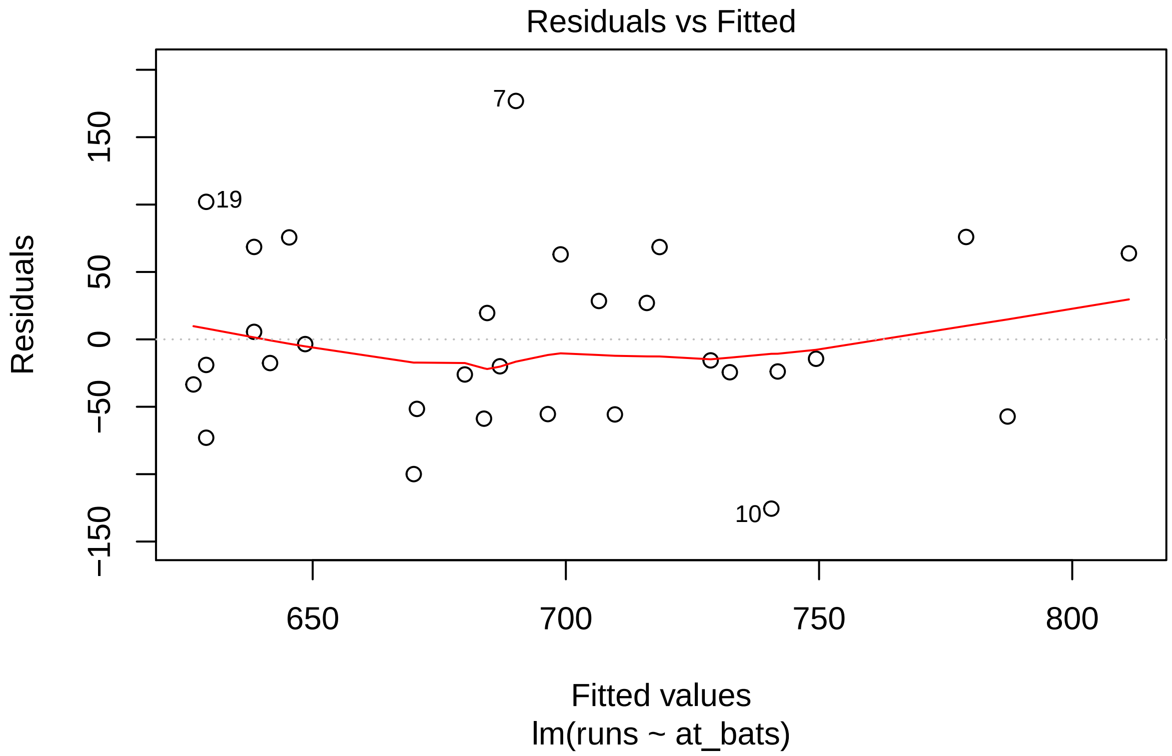
Solution: For an increase of 1 at bats there is a 0.6305 increase in runs.

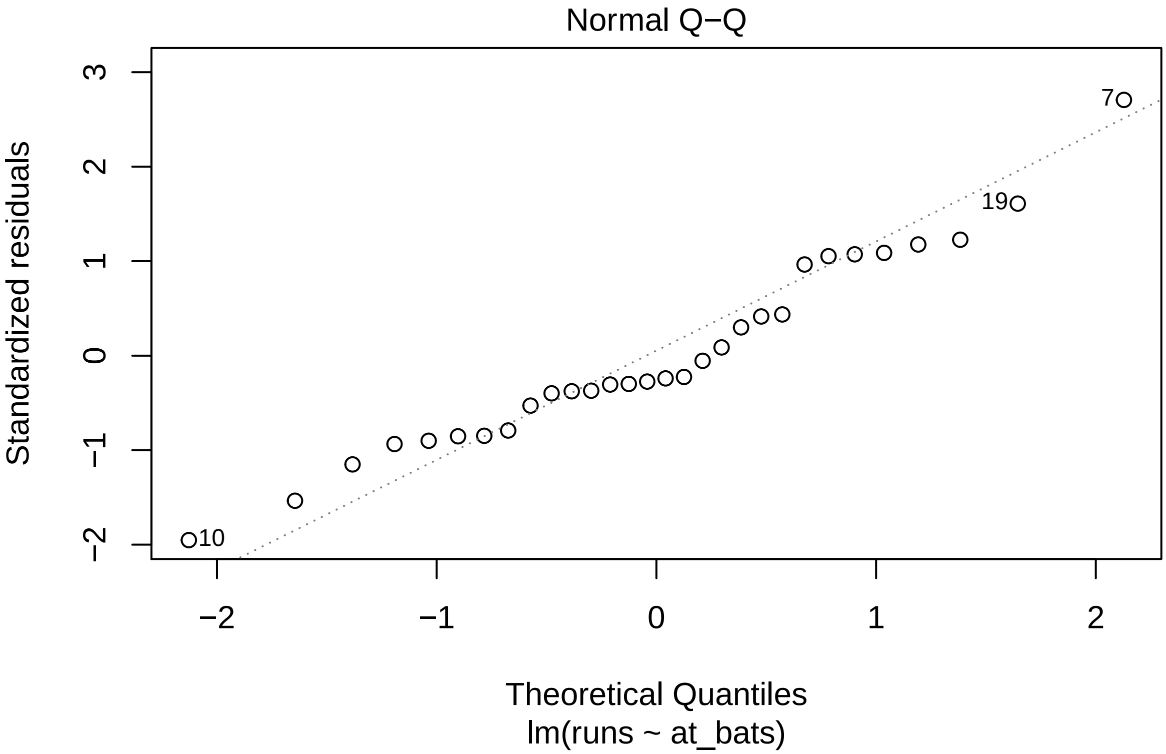
- Make a plot of the residuals versus at bats. Is there any apparent pattern in the residuals plot?

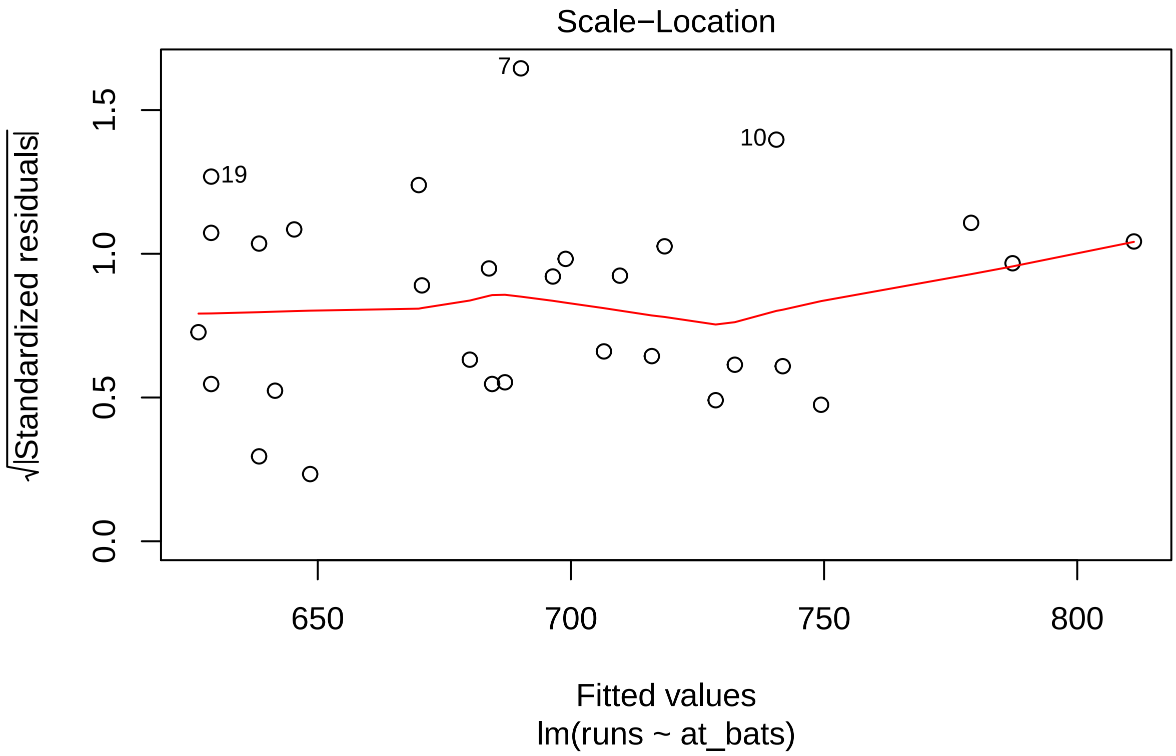
```r
# Plot residuals
plot(m1$residuals ~ mlb11$at_bats)
```

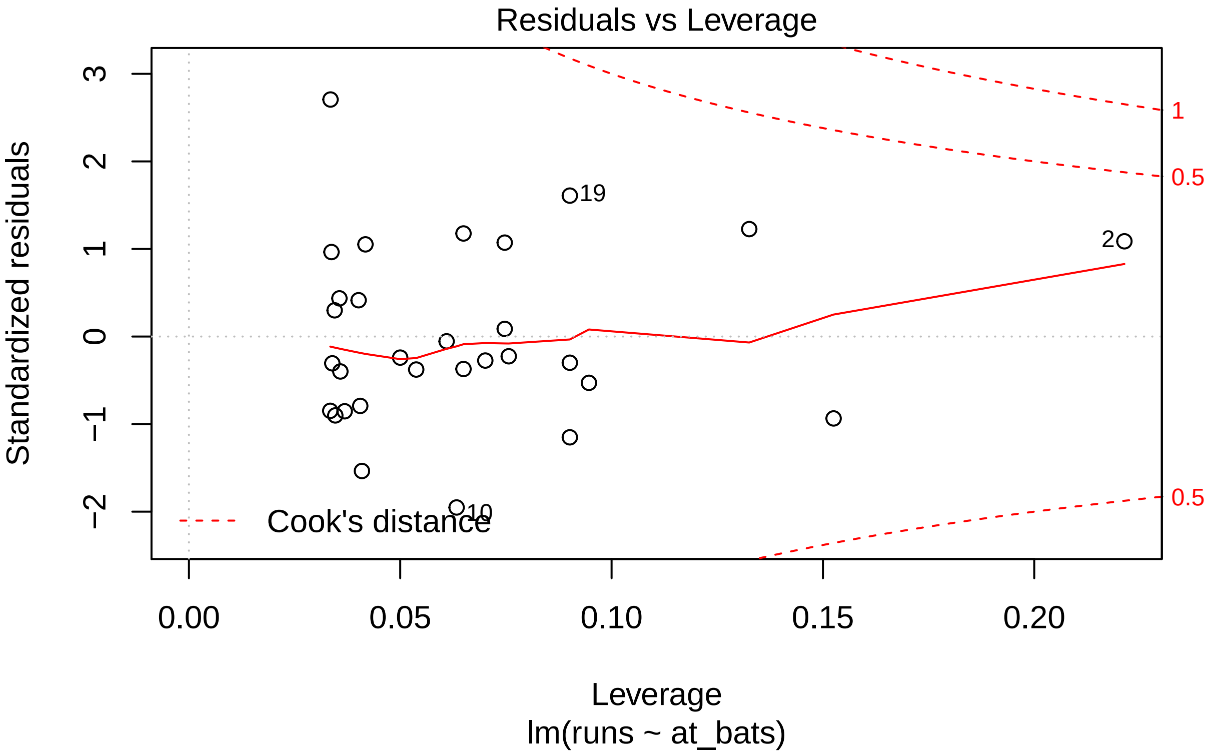

m1$residuals
−100
0
50
100
5450
5500
5550
5600
5650
5700
mlb11$at_bats

```r
plot(m1)
```


Residuals vs Fitted
Residuals
150
50
0
−50
−150
650
700
750
800
Fitted values
lm(runs ~ at_bats)
7
19
10


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
7
19
10
−2
−1
0
1
2
Theoretical Quantiles
lm(runs ~ at_bats)

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
650
700
750
800
Fitted values
lm(runs ~ at_bats)
7
10
19



Solution: There is no strong pattern in the residuals.

- Comment of the fit of the model.

Solution: The diagnostics plots of the linear model show that the residuals have approximately mean zero and constant variance. The residual plots are not perfect, but no immediate pattern is evident. The adjusted R^2 is 0.35, which is decent but not great.