Stat 50 - Lab 3

Correlation and Regression

Dr. Jorge Basilio

Feb 11, 2020

Getting Started

- Navigate to the Labs folder and find the Lab_3 folder. Inside you will find a Juptyper Notebook which can run and execute R code titled **"Stat50-Lab_3-YourLastName_YourFirstName-W20"**. Re-name the file with the obvious modifications.
- At the beginning of your Jupyter Notebook, double-click on the "Lab3" text. Replace the text "FirstName LastName" with your actual first and last name. Click "run cell" (looks like a play button) or hit `shift+enter`.
- By looking at this document, you are encouraged to copy and paste lines of code and modify them :-)

PART 1

In this lab we will explore how to study correlation and linear regression using R.

You will need to know the material from Labs 1 and 2. I recommend that you open them to serve as a reference in case you have forgotten how to do a certain task.

Opening Data

We will work with the Excel file "winter.xlsx" so let's load that up:

```
library(readxl)
winter_data <- read_excel("winter.xlsx")
```

Type (or copy) those lines of code in your lab notebook so you can also use the data in your lab.

Next, let's see the first few rows of

```
dim(winter_data) # quick glance how many rows and columns
```

```
## [1] 59  4
```

```
head(winter_data) # show the first few rows and all columns
```

```
## # A tibble: 6 x 4
##   City                          Mean_Jan_Temp_F Latitude Jan_degreesC
##   <chr>                                   <dbl>    <dbl>        <dbl>
## 1 Akron, OH                                  27     41.0        -2.78
## 2 Albany-Schenectady-Troy, NY                23     42.4        -5
## 3 Allentown, Bethlehem, PA-NJ                29     40.4        -1.67
```

```
## 4 Atlanta, GA                              45     33.4         7.22
## 5 Baltimore, MD                            35     39.2         1.67
## 6 Birmingham, AL                           45     33.3         7.22
```

```
names(winter_data) # names of columns/variabls
```

```
## [1] "City"            "Mean_Jan_Temp_F" "Latitude"        "Jan_degreesC"
```

Let's take a look at all the cities in the data set:

```
winter_data$City # show the full data set from "City"
```

```
##  [1] "Akron, OH"
##  [2] "Albany-Schenectady-Troy, NY"
##  [3] "Allentown, Bethlehem, PA-NJ"
##  [4] "Atlanta, GA"
##  [5] "Baltimore, MD"
##  [6] "Birmingham, AL"
##  [7] "Boston, MA"
##  [8] "Bridgeport-Milford, CT"
##  [9] "Buffalo, NY"
## [10] "Canton, OH"
## [11] "Chattanooga, TN-GA"
## [12] "Chicago, IL"
## [13] "Cincinnati, OH-KY-IN"
## [14] "Cleveland, OH"
## [15] "Columbus, OH"
## [16] "Dallas, TX"
## [17] "Dayton-Springfield, OH"
## [18] "Denver, CO"
## [19] "Detroit, MI"
## [20] "Flint, MI"
## [21] "Grand Rapids, MI"
## [22] "Greensboro-Winston-Salem-High Point, NC"
## [23] "Hartford, CT"
## [24] "Houston, TX"
## [25] "Indianapolis, IN"
## [26] "Kansas City, MO"
## [27] "Lancaster, PA"
## [28] "Los Angeles, Long Beach, CA"
## [29] "Louisville, KY-IN"
## [30] "Memphis, TN-AR-MS"
## [31] "Miami-Hialeah, FL"
## [32] "Milwaukee, WI"
## [33] "Minneapolis-St. Paul, MN-WI"
## [34] "Nashville, TN"
## [35] "New Haven-Meriden, CT"
## [36] "New Orleans, LA"
## [37] "New York, NY"
## [38] "Philadelphia, PA-NJ"
## [39] "Pittsburgh, PA"
## [40] "Portland, OR"
## [41] "Providence, RI"
## [42] "Reading, PA"
## [43] "Richmond-Petersburg, VA"
## [44] "Rochester, NY"
```

```
## [45] "St. Louis, MO-IL"
## [46] "San Diego, CA"
## [47] "San Francisco, CA"
## [48] "San Jose, CA"
## [49] "Seattle, WA"
## [50] "Springfield, MA"
## [51] "Syracuse, NY"
## [52] "Toledo, OH"
## [53] "Utica-Rome, NY"
## [54] "Washington, DC-MD-VA"
## [55] "Wichita, KS"
## [56] "Wilmington, DE-NJ-MD"
## [57] "Worcester, MA"
## [58] "York, PA"
## [59] "Youngstown-Warren, OH"
```

We can, of course, study all the descriptive statistics as we did in Lab 2 (mean, 5 number summary, box-plot, etc); however, we want to explore whether or not two variables are related via a **linear regression**. That is, we want to study the **correlation coefficient** between two variables.

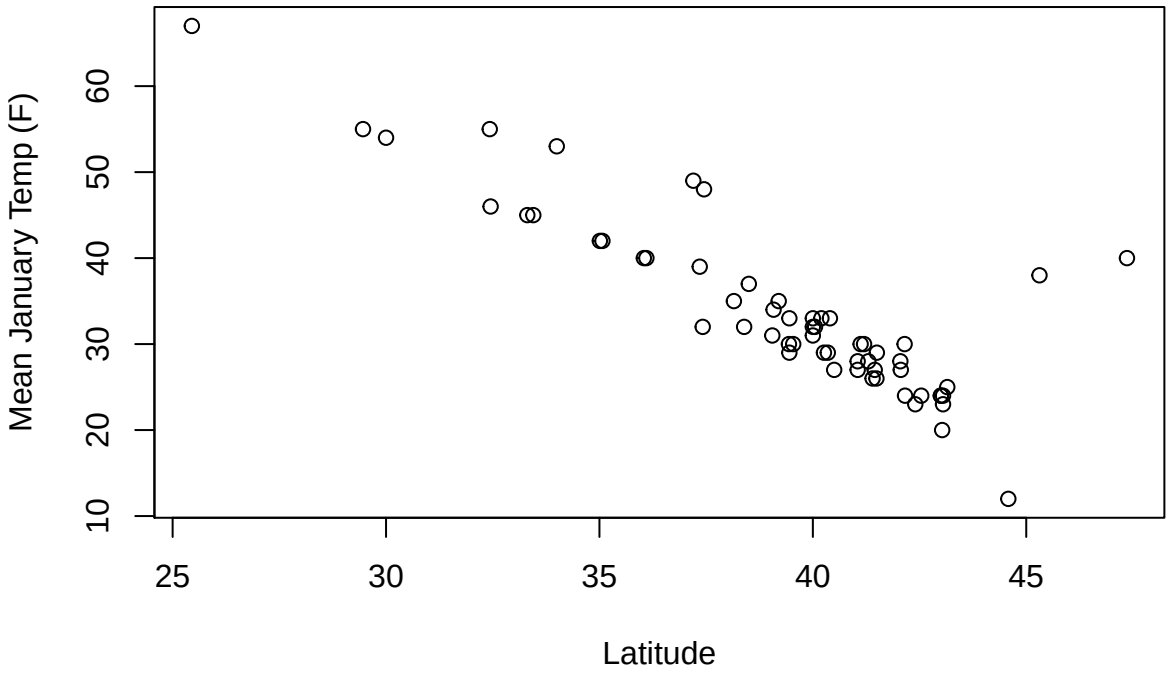
Scatterplots

Let's first create shorter names for the explanatory & response variables:

```
xvar <- winter_data$Latitude # be careful and spell it exactly as shown in above printout
yvar <- winter_data$Mean_Jan_Temp_F
```

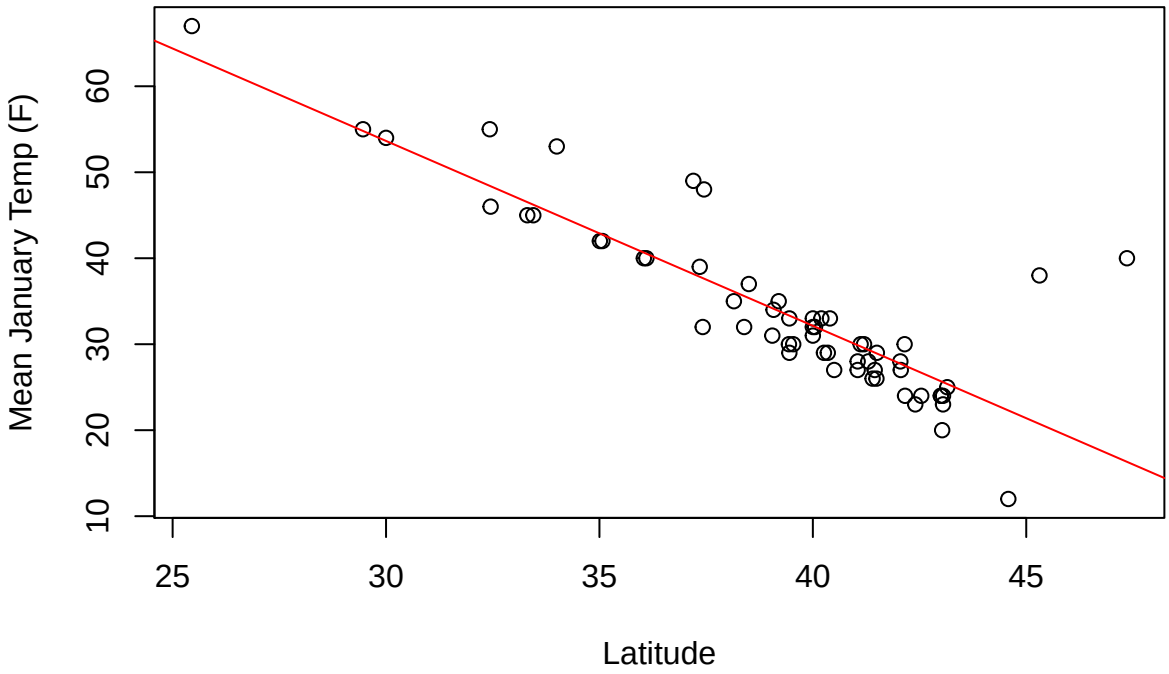
Next, make a **scatter plot**:

```
plot(xvar, yvar, xlab="Latitude", ylab="Mean January Temp (F)") # scatter plot with labels
```



How about we look at the **line of best fit**, or linear regression line:

```r
plot(xvar, yvar, xlab="Latitude", ylab="Mean January Temp (F)") # scatter plot with labels
abline(lm(yvar ~ xvar), col="Red") # draw regression line on scatter plot (with red color)
```



1. Based on the graph and the regression line, is there strong, moderate, weak, or no correlation between the Latitude and the Temperature (in degrees F) in January?

We can compute the **correlation coefficient**, r, as follows:

```r
corr_r <- cor(xvar,yvar)
cat("correlation coefficient is", corr_r) # this is one way to print the text and variables together
```

```
## correlation coefficient is -0.8573135
```

2. Based on the correlation coefficient, is there strong, moderate, weak, or no correlation between the Latitude and the Temperature (in degrees F) in January?

Standardizing the data

We mentioned in class that one of the benefits of studying the correlation coefficient was that it does not change if all of the values of either variable are converted to a different scale.

We will convert all the values of the data into z-scores. Recall the formula:

$$z = \frac{x - \bar{x}}{\sigma}$$

Instead of going into how to compute each z-score individually, R has a function that does this called `scale()`.

```r
# Use the "scale" function to standardize both variables as shown below.
# convert to z-score
# the "center" option subtracts the mean
# the "scale" option divides by the st. dev.
zscore_xvar = scale(xvar, center=TRUE, scale=TRUE)
zscore_yvar = scale(xvar, center=TRUE, scale=TRUE)
```

3. Create the scatter plot for the variables `zscore_xvar` and `zscore_yvar` and include the line of best fit.
4. Compute the correlation coefficient for the variables `zscore_xvar` and `zscore_yvar` and give it the name `corr_r_after`. Does it agree with the correlation coefficient we previously computed?

Line of Best Fit: Linear Regression Model

Now it is time to have R compute the linear regression line:

$$y = b_0 + b_1 x.$$

Recall that b_0 is called the **y-intercept** (where the line crosses the y-axis) and b_1 is called the **slope**. The interpreation of slope is very important: for each unit increase in the explanatory variable (x-variable), the response variable increases (if $b_1 > 0$) or decreases (if $b_1 < 0$) by b_1 units.

There's a few different ways to compute the line of best fit, or as R calls it the **linear regression model** (lm).

Method 1

In this method, we use the variables `xvar` and `yvar` which have been defined previously, and extracted from the original data set/excel file.

In R, `lm( response ~ explanatory)` is the general code we need to produce the line of best fit. Note that `lm` is short for *linear (regression) model*.

```
# Method 1: use the variables "xvar" for explanatory and "yvar" for the response
lm(yvar ~ xvar)
```

```
##
## Call:
## lm(formula = yvar ~ xvar)
##
## Coefficients:
## (Intercept)         xvar
##      118.14        -2.15
```

In the above, we see that the y-intercept is $b_0 = 118.14$ and the slope is $b_1 = -2.15$. Eyeballing the slope from (approximately) $(25, 65)$ and $(45, 20)$ we see that the slope is approximately

$$\frac{\Delta y}{\Delta x} = \frac{20 - 65}{45 - 25} = \frac{-45}{20}$$

which is approximately

```
-45/20
```

```
## [1] -2.25
```

This is reasonably close to the exact value of $b_1 = -2.15$ computed by R.

5. $(M \to E)$ Write a complete sentence that interprets the slope in the context of this scatterplot. Be sure to refer to the original variables (and the units!). As in the previous labs, create a new Markdown cell to write your sentence.
6. Use the regression line to **predict** the averate temperature in January at an altitude of 48 degrees. Write your answer in a complete sentece $(M \to E)$. As in the previous labs, create a new Markdown cell to write your sentence.

Method 2

In this method, we use the variables directly from the original data set/excel file. This is nicer since we can see directly the two variables being compared.

```
# Method 2: Direcly use variable from dataset/excel file
lm(Mean_Jan_Temp_F ~ Latitude, data=winter_data)
```

```
##
## Call:
## lm(formula = Mean_Jan_Temp_F ~ Latitude, data = winter_data)
##
## Coefficients:
## (Intercept)     Latitude
##      118.14        -2.15
```

We get the same results as before.

Method 3

In this method, we store the results of `lm()` in a new variable that we are going to call `lm_results`. Then we ask for the summary statistics of `lm_results`

```
# Method 3: store the linear model `lm()` into a variable
lm_results <- lm( Mean_Jan_Temp_F ~ Latitude, data=winter_data)
```

Recall: nothing happens! We only created the variable `lm_results` and stored it's results. To see what's there we need to type the name of the variable:

```
lm_results # show output of variable `lm_results`
```

```
##
## Call:
## lm(formula = Mean_Jan_Temp_F ~ Latitude, data = winter_data)
##
## Coefficients:
## (Intercept)     Latitude
##      118.14        -2.15
```

Again, we get exactly the same info as before. But now that it is stored as a new variable, we can do some extra things with it.

By using the `summary()` function, we get some new information:

```
summary(lm_results) # get extra info using `summary()`
```

```
##
## Call:
## lm(formula = Mean_Jan_Temp_F ~ Latitude, data = winter_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2978  -2.6353  -0.8719   0.3965  23.6789
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  118.139      6.743   17.52   <2e-16 ***
## Latitude      -2.150      0.171  -12.57   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.272 on 57 degrees of freedom
## Multiple R-squared:  0.735,  Adjusted R-squared:  0.7303
## F-statistic: 158.1 on 1 and 57 DF,  p-value: < 2.2e-16
```

There's a bunch of new stuff to look at now! Admittedly, we didn't study much of it so you can ignore it :-P

Typing math formulas into R

We have previously discussed how to create a "Markdown" cell where we can write paragraphs using normal english.

In a Markdown cell, there's special code for typing math formuals. There's two types: inline math formulas and displayed math formulas.

Displaying inline math uses dollar signs `$ ...math formula...$`. For example, we write the equation of the regression line as an inline displayed formula like this `$y=118.14 - 2.15x$` and it looks like this $y = 118.14 - 2.15x$. Notice how the styling is slightly different than the "plain text".

We can display the formula for the regression line in between a slash and a bracket `\[ ...math formula... \]`. So our regression line is displayed like this:

`\[  y=118.14 - 2.15x \]` and it looks like this displayed:

$$y = 118.14 - 2.15x$$

A new data set: Guns

Introduction

The U.S. Center for Disease Control and Prevention (CDC) publishes state by state data on mortality rates by different causes, including deaths by firearms. Using this in conjunction with gun ownership data in each state, we can explore the association, if any, between firearm deaths and gun ownership. The file "firearms2013.xlsx" contains these data for the year 2013.

7. Carry out the following tasks:
   [a] Read the file into R
   [b] Make a scatterplot of "deaths_per_100k" vs "gun_ownership_rate". Be sure to label your axes.
   [c] Compute the correlation coefficient between those two variables.
   [e] Plot the same two variables in standardized form
   [f] Construct a linear regression model to predict firearm deaths from gun ownership rate. Plot the model together with the original data.
   [g] Type the equation of the line of best fit. See the section above about "Typing Math formulas into R".
   [h] Write a short paragraph discussing the quality and appropriateness of the linear model, based on the scatter plot, correlation, etc. Are there any conclusions you can draw from the model?

Part 2

Moneyball

Acknoledgement: This part is adapted from a lab by the OpenIntro Stats textbook authors.

Batter up

The movie Moneyball focuses on the "quest for the secret of success in baseball". It follows a low-budget team, the Oakland Athletics, who believed that underused statistics, such as a player's ability to get on base, better predict the ability to score runs than typical statistics like home runs, RBIs (runs batted in), and batting average. Obtaining players who excelled in these underused statistics turned out to be much more affordable for the team.

In this lab we'll be looking at data from all 30 Major League Baseball teams and examining the linear relationship between runs scored in a season and a number of other player statistics. Our aim will be to summarize these relationships both graphically and numerically in order to find which variable, if any, helps us best predict a team's runs scored in a season.

The data

Let's load up the data for the 2011 season. The excel file is called `mlb11.xlsx`.

```
mlb11_data <- read_excel("mlb11.xlsx")
```

In addition to runs scored, there are seven traditionally used variables in the data set: at-bats, hits, home runs, batting average, strikeouts, stolen bases, and wins.

There are also three newer variables: on-base percentage, slugging percentage, and on-base plus slugging. For the first portion of the analysis we'll consider the seven traditional variables. At the end of the lab, you'll work with the newer variables on your own.

8. Carry out the following tasks:
   [a] Read the file into R.
   [b] Show the first few rows of the data.
   [c] Show all the variables in the data.

The Analysis

Runs vs At_bats

9. Carry out the following tasks:
   [a] Make a scatterplot of `runs` vs `at_bats`. Be sure to label your axes.
   [b] Compute the correlation coefficient between those two variables.
   [c] Construct a linear regression model to predict `runs` from `at_bats`. Plot the model together with the original data.
   [d] Type the equation of the line of best fit. See the section above about "Typing Math formulas into R".
   [e] ($M \rightarrow E$) Write a complete sentence that interprets the slope in the context of this scatterplot. Be sure to refer to the original variables (and the units!). As in the previous labs, create a new Markdown cell to write your sentence.
   [f] If you knew a team's `at_bats` was $5,475$ times, would you be comfortable using a linear model to predict the number of runs?

10. Write a short paragraph discussing the quality and appropriateness of the linear model, based on the scatter plot, correlation, etc. Are there any conclusions you can draw from the model?

Runs vs Homeruns

11. Carry out the following tasks:

[a] Make a scatterplot of Runs vs Homeruns. Be sure to label your axes.
[b] Compute the correlation coefficient between those two variables.
[c] Construct a linear regression model to predict Runs from Homeruns. Plot the model together with the original data.
[d] Type the equation of the line of best fit. See the section above about "Typing Math formulas into R".
[e] ($M \rightarrow E$) Write a complete sentence that interprets the slope in the context of this scatterplot. Be sure to refer to the original variables (and the units!). As in the previous labs, create a new Markdown cell to write your sentence.
[f] Which variable is correlated to `runs` better, `at_bats` or `homeruns`?

Conclusion

12. Now that you can summarize the linear relationship between two variables, investigate the relationships between runs and each of the other five traditional variables. Which variable best predicts runs? Support your conclusion using the graphical and numerical methods we've discussed (for the sake of conciseness, only include output for the best variable, not all five).

13. Now examine the three newer variables. These are the statistics used by the author of Moneyball to predict a teams success. In general, are they more or less effective at predicting runs that the old variables? Explain using appropriate graphical and numerical evidence. Of all ten variables we've analyzed, which seems to be the best predictor of runs? Using the limited (or not so limited) information you know about these baseball statistics, does your result make sense?

You've done it! You've now finished Lab_3! ;-)