Using R Projects to Explore Regression

John Ross, Southwestern U

SIGMAA SDS-ED, MathFest 2023

Introduction to Statistics at Southwestern University

- Southwestern University
  - Small Liberal Arts College
  - 1400 students tota
- Introduction to Statistics
  - 22-30 Students per class
  - Business(!), Biology, Psychology, Kinesiology
  - Not required for the math major
- Fall 2022: 3 sections, 77 students
  - **76% First-year students, 87% taking the class because it's required for their major**

My Favorite Statistics / Data Science Activity

Activity 1: In-class work on Linear Regression
Activity 2: Out-of-class work on Multiple Regression

~~My Favorite Statistics / Data Science Activity~~

My Favorite Statistics / Data Science Framework

Activity 1: In-class work on Linear Regression

Activity 2: Out-of-class work on Multiple Regression

My Favorite Statistics / Data Science Framework

- Loosely flipped (videos before each class)
- In-class active lecture and small **activities**
  - Ungraded, but themed
  - Not billed as a project; just part of the lesson
- Standard homework and medium-sized **Mini-Projects**

R is incorporated in all activities and mini-projects

Students get consistent and regular exposure

When using R, emphasis is on **interpretation/thought process**, not code

Regression, Fall 2022 Schedule (Days 11, 12, 13)

	Pre-class videos	Content Goals	R Commands
Wednesday, September 28	Video 11.1	Introducing the Linear Model	lm() command
Lecture 11: Intro Linear Regression	Video 11.2	Residuals: using LM as a model	ggplot() scatterplot and line
	Video 11.3	Checking residuals	resid(lm())
	Pre-class videos	**Content Goals**	**R Commands**
Monday, October 3	Video 12.1	Regression towards the mean	cor()
Lecture 12: More Linear Regression	Video 12.2	Checking SE	ggplot() for residuals
	Pre-class videos	**Content Goals**	**R Commands**
Wednesday, October 5	Video 13.1	Interpreting coefficients	lm(A$B+C)
Lecture 13: Multiple Regression		Indicator and Interaction var's	
Monday, October 10		**Mini-Project Due**	
Lecture 14		Hold their hand through one example	
		"Try your own" for a second example	

More Handholding

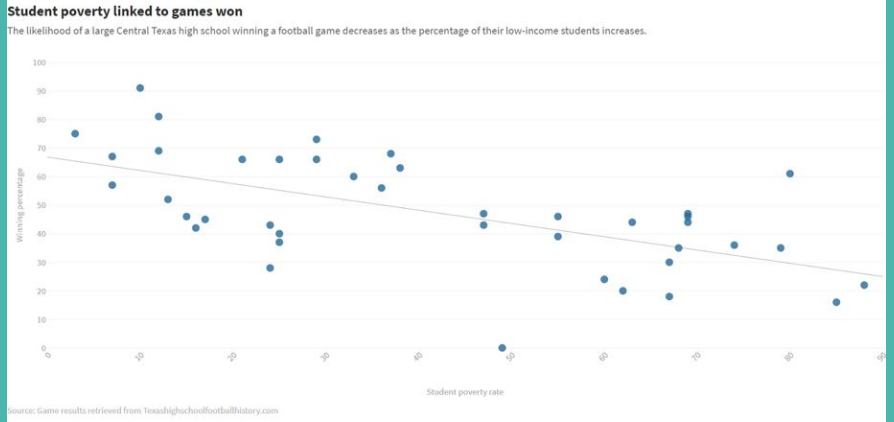
More Independence

Starting Class w/ Active Lecture

Student Poverty and Football Win %

"An American-Statesman analysis of the football records of 41… high schools in Central Texas over the past decade found a correlation between student poverty and performance on the football field: The wealthier the student body, the better the football team." – A.A.S., October 2019

Student Poverty and Football Win %



In-class R work on Linear Regression

Linear Regression – R material

Introducing the data set we'll look at

Importing the data set

Reminders of commands for basic data exploration

The `Longley` economic dataset

The **`Longley`** data frame is a well-known, small US-based macroeconomic data set that examines seven economic variables observed yearly over 16 years. It is already loaded into RStudio (under the **`datasets`** package, which should already be selected in RStudio), so we just need to give it a name and begin playing with it! Let's do that now. First, load our usual commands, and then create a new data set (which we will call **`LData`**) that captures the Longley data frame:

```
> library(ggplot2)
> library(dplyr)
> LData <- longley
```

(If you cannot load **`longley`**, make sure to include a command of the form **`library(datasets)`** to ensure that you have access the R's innate datasets.)

4. Use the commands **`names(LData)`** and **`head(LData)`** to get an initial sense of our data set. How many quantitative variables are there? How many categorical variables are there?

Linear Regression – R material

Graph the scatterplot

Make real-world connections

Create and plot the linear model

5. Our goal is to create several scatterplots, measure correlation, and (ultimately) create linear regressions. Let's begin by plotting a few time plots (scatterplots with x=time) to get a sense of the economic data over the 16 years in question.

   (a) Create a time plot with `x=Year` and `y=Population` to get a sense of population growth over time. Do this by entering the following ggplot:

   ```
   > ggplot(LData, aes(x=Year, y=Population)) + geom_point()
   ```

   What is the shape, direction, and form of this scatterplot? What do you notice?

   (b) Create more time plots; one with `y=Employed`, and one with `y=Armed.Forces`. What do you notice? [Bonus: can you connect this to any events in US history?]

6. The variable's `Year` and `Employed` seemed reasonably strongly correlated (that is, the linear relationship between them appeared to be quite strong). Intuitively, this means that we should have a large correlation coefficient r (close to 1.0). We should also get a line of best fit that "cuts through the data" nicely. Let's verify both of those!!

   (a) To calculate r, use the command `cor(LData$Year, LData$Employed)`. What is the r-value we get?

   (b) We can draw a line of best fit through the graph using the `geom_smooth` command. Try entering the following:

   ```
   > ggplot(LData, aes(x=Year, y=Employed))+
       geom_point()+
       geom_smooth(method = "lm", se = FALSE)
   ```

Linear Regression – R material

Find the regression coefficients

Interpret the coefficients using the equation of a line

Use the linear model to predict new data

Examine and interpret the residuals

7. In order to interact with the line of best fit, we need its equation! To find this, we can simply enter the `lm()` command. Try the following:

> lm(LData$ Employed ~ LData $ Year)

You should get the following information as output:

```
lm(formula = LData$Employed ~ LData$Year)

Coefficients:
(Intercept)   LData$Year
  -1335.1052       0.7165
```

This is telling us that the intercept of our equation is -1335.1052, and our slope (i.e. the coefficient attached to the explanatory variable `Year`) is equal to 0.7165. The equation of the regression line can be written in as $\hat{y} = 0.7165x - 1335.1052$ or also as $\widehat{\text{Emp}} = 0.7165(\text{Year}) - 1335.1052$.

(a) Interpret this slope in real-world language (i.e., in a sentence that includes years and employment numbers).

(b) Interpret this intercept in real-world language (i.e., in a sentence that includes years and employment numbers). Does such an interpretation make "real-world sense" in this case?

(c) Let's focus on the year 1958 for a moment. Using the work we've done so far, see if you can answer:

- What would we predict employment would be in 1958?
- What was the actual employment in 1958?
- What is the residual? How can we interpret this residual?

(d) If you had to use this model to guess what the employment numbers would be in the year 1965, and again in the year 2000. What would you guess for each of these years?

(e) The actual employment numbers in the US for 1965 and 2000 were 73.1 million and 129.7 million, respectively. How did your model do?

Linear Regression – R material

(Next day in class; a different data set, NHANES, is being used)

Remind students what we want to see in lines

Create the linear model, and examine the residuals

Interpret the residuals

Checking the appropriateness of a linear regression

3. Recall that the *residuals* of a regression are defined as

$$\text{residual} = \text{observed} - \text{predicted}.$$

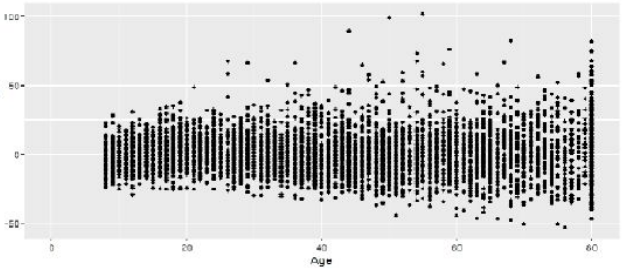
Writing this as a formula (where the letter e is used for the residual), we get

$$e = y - \hat{y}.$$

If our regression line is well-chosen and appropriate, we expect (or hope) that the residuals are relatively scattered, with no strong discernible patterns. For example, here is a "residual" plot for `Age` vs `BPSysAve`. We still plot age on the x-axis, but instead of plotting blood pressure on the y-axis, we plot the residuals on that axis (replacing each point's height with its *residual value*).

```
> lm.age.bp <- lm(ND$BPSysAve~ND$Age, na.action=na.exclude)
> resid.age.bp <- resid(lm.age.bp)
> ggplot(ND, aes(x=Age, y=resid.age.bp))+
+   geom_point()
```

What observations can we make about the residual plot? Is it boring? (hopefully!)



Linear Regression – R material

Open exploration!

Depending on the timing of the day, students will explore for 10-25 minutes.

If time allows, students will report out on anything interesting they found at the end of class.

Try your own!

9. Try finding your own interesting linear relationships between variables in either the `NHANES` or `longley` data set. In each case,

   - plot the two quantitative variables on a scatterplot.
   - calculate the linear regression coefficients (using the `lm()` command) and then graph the linear regression (using the `geom_smooth()` or the `geom_abline` commands)
   - Create two graphs of the residuals: the residual scatterplot, and also a histogram of the residual values. Verify whether the residuals are "appropriately boring."
   - Calculate the residual mean and standard deviation (i.e. the standard error).
   - Explain the relationship between the two variables you picked. What is the equation of the linear regression? How can you interpret the slope and intercept? Was the relationship suitably linear, and the residual plot suitably boring?

   Record your results here!

Mini Project on Multiple Regression

Multiple Regression – Mini-Project

An initial linear regression analysis is performed.

Students are reminded of commands, but interpretation of coefficients must be their own

2. Use your usual commands (like `names()`, `dim()`, and `head()`) to get a sense of the data set and its variables. Using the comments, list the 5 variables in this data set, and what kind of variable (quantitative or categorical) each variable is.

3. Let's begin with a simple linear regression, predicting Sepal Width as a function of Sepal Length.Try typing in the following commands:

```
> table(IrisData$Species)
> ggpplot(IrisData, aes(x=Sepal.Length, y=Sepal.Width)) +
      geom_point()+
      geom_smooth(method = "lm")
> lm.Width.Length < - lm(IrisData$Sepal.Width ~ IrisData$Sepal.Length)
> lm.Width.Length
```

4. In comments: Describe your linear regression. What is the equation of the line? (Write it in the form "y = mx + b".) What is the slope of your line, and what does it mean (in real-world terms)? What is the intercept of your line, and what does it mean (in real-world terms)?

Multiple Regression – Mini-Project

Color-coded scatterplot command is given

Expected plot output is shown

The paragraph at the bottom provides multiple regression motivation.

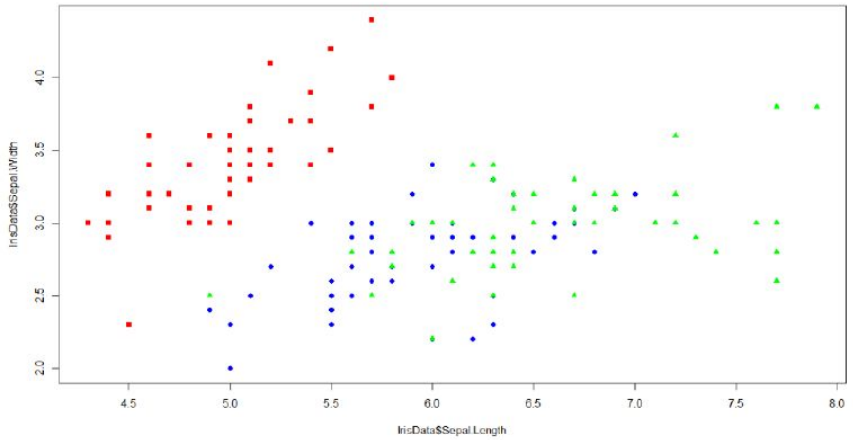
Motivating our decision to use Multiple Regression

To motivate our work for the rest of this sheet, we want to color-code the points in our scatterplot according to the species of iris. This will let us see that one type of iris has a much different pattern than the other two!

5. Try typing in the following command, which assigns colors to different species:

```
> ggpplot(IrisData, aes(x=Sepal.Length, y=Sepal.Width, color=Species)) +
      geom_point()
```

The scatterplot you just created should look (roughly) like this (the symbols might be different):



From our previous scatterplot, it's clear that one species (the species Setosa, represented by the red dots in the upper left) has a different width/length relationship than the other two species. **Our major goal for the rest of this sheet will be to design a multiple regression that predicts Sepal Width as a function of both the Sepal Length and the Iris species (whether or not it's Setosa).**

Multiple Regression – Mini-Project

Step-by-step commands for setting up and running the multiple regression analysis

ifelse() command and mutate() command are introduced

Creating an indicator and interaction variable for the Setosa species

To perform an appropriate regression analysis, we will create an indicator and a interaction variable for the Setosa species. To do this, we must add new variables (ie, new columns) to our dataset using the `mutate` command from the `dplyr` package. However, we also want to learn how to use an "if/else" command in RStudio.

6. We want our first new column to be our indicator variable, meaning we want an entry in this column to be a "1" if the plant species is Setosa, and a 0 otherwise, This is an excellent place to use the `ifelse` command in R. Type in the following command:

   > IrisData <- mutate(IrisData, SetosaIndicator = ifelse(Species == "setosa", 1, 0))

   This says that we assign a number of the variable SetosaIndicator as follows: *IF* the species of Iris is "setosa", we assign a 1. *ELSE* (or otherwise), we assign a 0. This is exactly what we were hoping to accomplish!

7. Next, we want to create our **interaction** variable, which we will name SetosaInteraction. As a reminder, an interaction term is of the form (indicator)*(predictor), and so we achieve this by creating a new variable and assigning it to be the indicator variable times the predictor variable (in this case, sepal length):

   IrisData <- mutate(IrisData, SetosaInteraction = SetosaIndicator*Sepal.Length)

Multiple Regression – Mini-Project

The command for the multiple regression is given

Students must interpret the coefficients

Students must set up two lines of best fit based on the multiple regression

The Multiple Regression Analysis

8. Finally, we can finish our multiple regression analysis! To do this, enter the command:

```
> mlm.width <- lm(IrisData$Sepal.Width~IrisData$Sepal.Length+
      IrisData$SetosaIndicator+IrisData$SetosaInteraction)

> mlm.width
```

9. What are our three coefficients and our intercept? (write your answer as a comment)
10. Use the coefficients and the intercept to write an *equation* describing the relationship between our four variables. (something of the form "predicted sepal width = ...")
11. Suppose we want to predict the sepal width of a virginica or a virsicolor iris flower based only on its sepal length. What is the equation we would use? (Write your answer as a comment, rounding your slope and intercept to two decimal places. Briefly explain your answer in a comment.)
12. Suppose we want to predict the sepal width of a setosa iris based only on its sepal length. What is the equation we would use? (Write your answer as a comment, rounding your slope and intercept to two decimal places. Briefly explain your answer in a comment.)
13. Add on two `geom_abline()` commands to your ggplot, with slope and intercept specified, to verify that your answers to the previous two problems are correct! If done correctly, you should get two lines of best fit: one that describes the setosa irises, and one that describes the virginica and virsicolor irises.

Multiple Regression – Mini-Project

At the end: open exploration!

Students must run the same type of analysis, but for new variables.

A recommendation for variables to use is given, but the final choice is left to the students

A Second Multiple Regression Analysis – this one is your choice!

Now comes the fun part: repeat this entire process, using other variables of your choice to do some multiple regression analysis! This means

14. Explicitly identify which quantitative variable you want to estimate, and which quantitative variable you want to predict.
15. Create a scatterplot that compares two these two quantitative variables, with the predictor on the x-axis and the response variable on the y-axis, and with different colors used to represent different `Species`.
16. Create an appropriate indicator variable and interaction variable.
17. Run a multiple regression using the `lm` command. In the comments, write down what your coefficients are.
18. Calculate the equation of each of your two lines of best fit (with one line depending on `Species` being a 1, the other with `Species` being a 0.
19. Graph your two lines using the `abline` command, to see if they line up appropriately in your scatterplot!

I recommend estimating `Sepal.Width` based on `Petal.Width` and `Species`, but you are free to explore and compare any `Species` with two quantitative variables you want!

Links to PDFs

This Slide Deck:

https://bit.ly/JohnRossMathFest2023

Direct Links to PDF Handouts:

Day 11 Activities

Day 12 Activities

Mini Project on Multiple Regression

rossjo@southwestern.edu