Intro to Stats, Day 11 – Linear Regression

Name: ______________________________________

As we have seen, **scatterplots** are useful tools for comparing two quantitative variables, determining whether they are associated in any way, and (in case the association is linear) we can come up with a number, r, to determine how correlated the data is.

In this worksheet, we aim to go a step further. Once we know that two variables are correlated (and, specifically, that their relationship is linear), we want to draw a line that best describes the pattern we see. The line we draw is called, interchangeably, a **least squares line**; a **line of best fit**; or a **regression line**. The line serves two dual purposes:

- we can view it as a **linear equation**, using our knowledge of the equations of lines to learn about the relationship between our variables.
- we can view it as a **linear model**, using this equation to model (or predict) where future y-values will fall based on their hypothetical x-value.

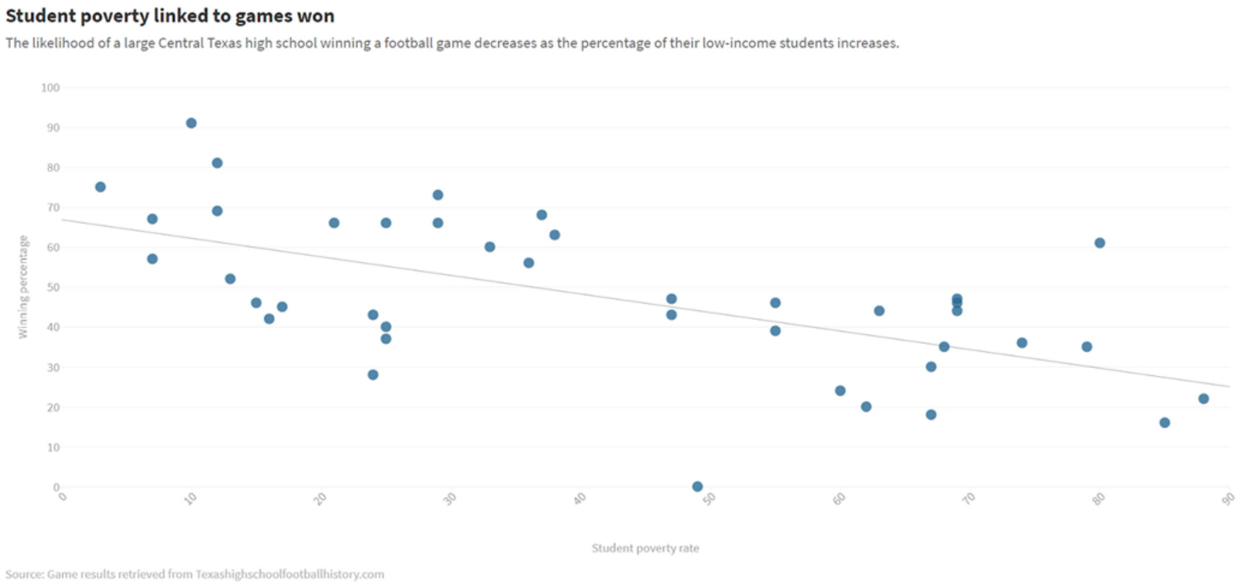
In other words, we want to be able to answer the question "given an x value of __________, what do I think the y value should be?" Being able to answer this question means creating (and agreeing upon) a line of best fit for the data, which we call a **regression line**.

By the end of this worksheet, you should be able to

- find a regression line for a scatterplot, given a basic summary of the statistics of the variables (mean, standard deviation, correlation).
- use a regression line to make predictions about the y variable of a set of data at a given x-value.
- determine the residual value of a data point, and explain what this means in terms of the actual data and the predicted data.
- examine the scatterplot of the residuals, and use that scatterplot as evidence of whether or not your regression line is a good fit.

Do Football Teams from High-Poverty Areas Lose More Football Games?

The Austin-American Statesman had an article in 2019 that explored the relationship between a high school's football team win percentage and their local poverty level. They recorded data from 40 high schools in the central Texas area, and plotted their poverty level against their football team's win percentage. The scatterplot they created is shown below:



Student poverty linked to games won

The likelihood of a large Central Texas high school winning a football game decreases as the percentage of their low-income students increases.

Source: Game results retrieved from Texashighschoolfootballhistory.com

1. Describe the trend you see in the scatterplot here, using the language of **direction**, **form**, and **linearity** that we looked at last class:

Since the scatterplot has a linear form, we might attempt to draw a line that best describes the pattern we see. The line is shown drawn on top of the graph above. Let's view it through the alternating lenses of a **linear equation** and a **linear model**.

2. First, let's examine the equation of the line, which is $\widehat{\text{FB}} = -0.5(\text{Poverty}) + 68$.

   (a) What is the **slope** of the line? How do we interpret the slope?

   (b) What is the **intercept** of the line? How do we interpret the intercept?

3. We can also talk about the line as a **model**, in which we can talk about predicting $\hat{y}$-values based on a corresponding x-value. If we have an actual y-value, we can compare it to its predicted $\hat{y}$-value to calculate a **residual**.

   (a) At $x = 14$, we claim that the predicted $\hat{y}$-value would equal 61. Write down what this means in real-world terms. Also, write down how we calculate this (where did the "61" come from?)

   (b) We had an actual point in our scatterplot with an x-value of 14 and a y-value of 46. What would the **residual** be at this point? What does that residual value mean, in real world terms?

   (c) Note that one of our points on the scatterplot was the point $(80, 70)$. Please answer the following:

   - Identify this point on the graph on the previous page.
   - What do the numbers 80 & 70 represent here in real-world terms?
   - What is the corresponding $\hat{y}$ value, an what does it represent?
   - What is the corresponding residual value, and what does it represent?

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?

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:

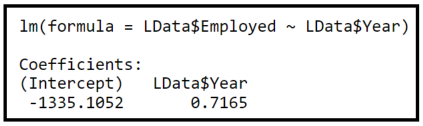
   ```
   >    ggplot(LData, aes(x=Year, y=Employed))+
             geom_point()+
             geom_smooth(method = “lm”, se = FALSE)
   ```

   Dissecting that last command: the `geom_smooth` command creates a smooth model for the data. The method we're using the **linear model** (that is the "lm" in the command). The `se=FALSE` piece keeps us from adding an "error" zone around the line – we'll discuss this at a later date – and, instead, only draws the line.

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 $\sim$ 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?

8. Try to perform a linear regression analysis that uses `GNP` (Gross National Product, a measurement of the monetary value of how much "stuff" the US produced in a year) and `Employed`. Let `GNP` be your explanatory variable, and let `Employed` be your response variable. Do the following:
    - Create the scatterplot and verify that a line of best fit is appropriate.
    - Plot the line of best fit on top of the scatterplot. Find the equation for this line.
    - Interpret the slope and the intercept. What do each of those values mean, in real world language?

Exploratory Try performing some linear regression analysis that is of interest to you! Use either the `longley`, `NHANES`[1], or `Institutions2014`[2] datasets. Find two quantitative variables that have a linear relationship, and explore! Show me what you find!

Checking the residual plot – BORING is BETTER

When our scatterplot is *truly* linear, we expect our line to cut through the point cloud neatly. This means we expect roughly the same number of positive and negative residuals, and we expect them to be randomly scattered around the line. To verify this, we often like to record, and then plot, the residual values!

9. We'll explore this using our original `longley` data, comparing `GNP` and `Employed`. After creating our scatterplot and creating our linear model, do the following:

```
>    Resid_Employed_GNP <- resid(lm(LData$Employed~LData$GNP))
>    ggplot(LData, aes(x=GNP, y=Resid_Employed_GNP )) + geom_point()
```

    The first command creates a list of the residual values. The second command creates a plot with *those* values as your y-values, and keeps our original x-value the same. If our linear regression is a good one, the result should be a boring plot!

10. Try doing the same thing, examining the residuals for a time plot comparing `Population` and `Year`. What do you notice? What does this lead you to conclude?

[1] ND <- NHANES   # make sure you have the NHANES package installed and activated first!

[2] ID2014 <- read.csv("http://bit.ly/RossStatsInstitutions2014")