Lab Simple Model

Abhi Thanvi, Paul Holaway

July 14th, 2022

Contents

Lab Simple Model

Welcome

Just like learning a new spoken language, you will not learn the language without practice. Labs are an important part of this course. Collaboration on labs is **extremely encouraged**. If you find yourself stuck for more than a few minutes, ask a neighbor or course staff for help. When you are giving help to your neighbor, explain the **idea and approach** to the problem without sharing the answer itself so they can figure it out on their own. This will be better for them and for you. For them because it will stick more and they will have a better understanding of the concept. For you because if you can explain it to other students, that means you understand it better too.

The Idea of this Lab

This is the last lab we have as a class, and this lab is gonna focus on a type of statistical model called Simple Linear Regression. Although, the model has "Simple" in it, be mindful that Linear models are very powerful because it makes your life easier and is one of the most powerful model in terms of accuracy and simplicity. It can be used in a lot of different areas and there are ways to make exponential relations into linear relations. So understanding this "simple" linear regression model and its intricacies is very necessary. This is the last lab and I have enjoyed every single minute with y'all. I hope this feeling is mutual, if yes feel free to put a thumbs up in the air lol. Before I cry on my laptop, let's get to the lab!

"Half of the time when companies say they want AI, what they really want is a Simple Linear Regression" - Expert Abhi

Problem 1: Modeling

Hey, this is Abhi! So I am actually an only child (no siblings), and usually I do not text too much either. I have a guess that people who have more siblings, most likely also text more. I believe it should be true because they have more family to text too, but also they are more likely to be extroverted. I am not sure if I can really justify this causation, but I think we can see if there is any association. Build a linear regression model and state whether this is a good model in this case.

Today we are not giving you any help to build this, because we want you to edit, code, and answer the problem. This simulates real-world scenario, and you are allowed to collaborate with group members and instructors are here to help!

Hint 1: Abhi is asking you to use the hello.csv

Hint 2: Make sure to be descriptive and feel free to add your analysis (as if you are talking to yourself) between your code chunks

Hint 3: Feel free to ask questions and do above and beyond... it might be helpful for your project ;). In real world you are rewarded with doing above expectations with $$$

Hint 4: If you see some questions, those should be answered! This is an exploratory simulation

Answer: (Work space for students)

```
library(tidyverse)
```

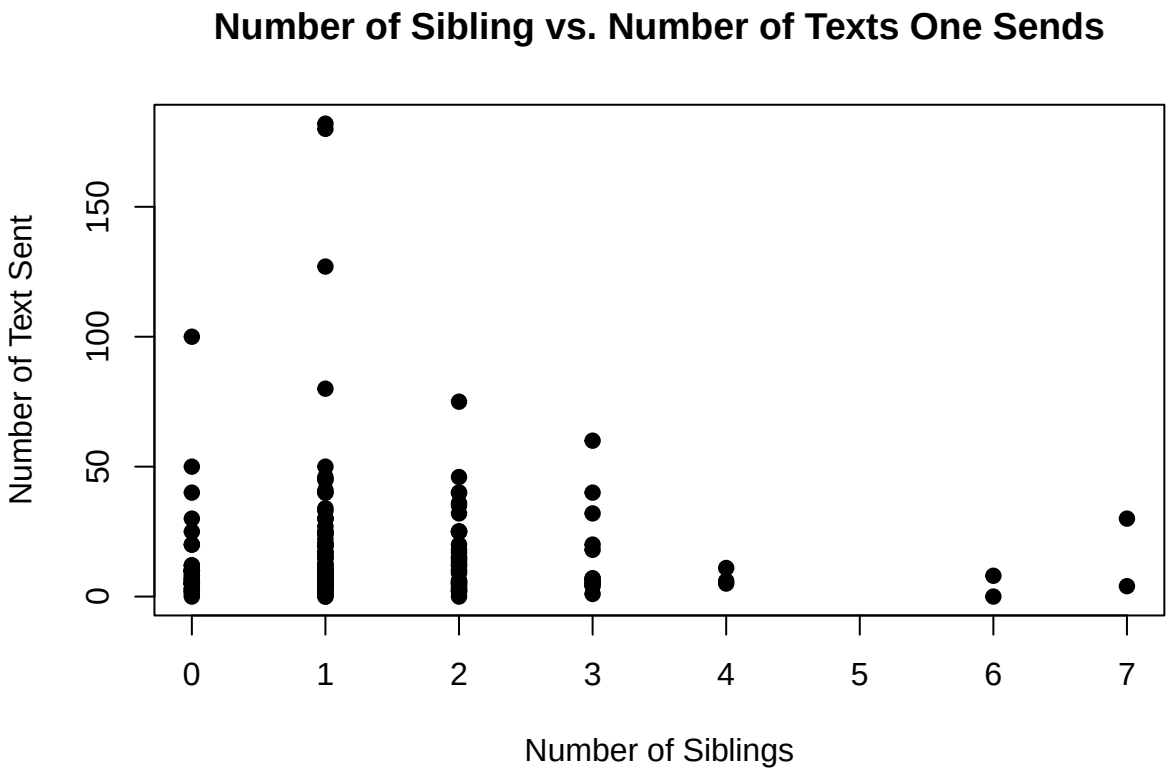
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
hello <- read.csv("~/Desktop/DPISu22/Data Sets/hello.csv", stringsAsFactors=TRUE)

plot(hello$Siblings, hello$Texts, xlab = "Number of Siblings", ylab = "Number of Text Sent",
     main = "Number of Sibling vs. Number of Texts One Sends", pch = 19)
```



Question: Do you feel a linear model would be a good fit? Why? Answer: (Student Response Expected) No, personally I feel the scatter plot isn't showing any linear relationship, if I am being very honest Abhi seems like he is completely wrong because the highest texts are sent by people with 1 or few siblings. :(Abhi needs to get better smh.

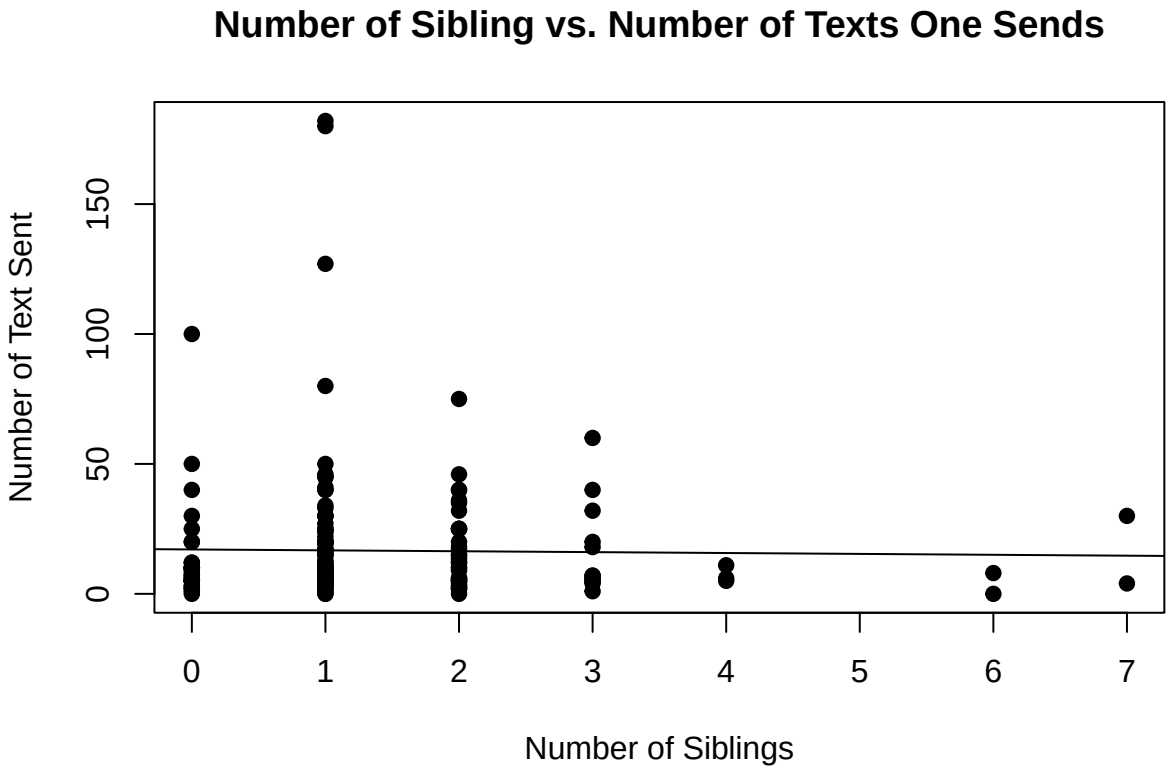
Continue to Build you model here and make sure you have a scatter plot with the line of best fit

```
hunch_model = lm(Texts ~ Siblings, data = hello)
summary(hunch_model)
```

```
##
```

```
## Call:
## lm(formula = Texts ~ Siblings, data = hello)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.080 -11.482  -8.739   3.261 165.261
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.0801     2.6446   6.458 8.64e-10 ***
## Siblings     -0.3415     1.5295  -0.223    0.824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.11 on 190 degrees of freedom
## Multiple R-squared:  0.0002623,  Adjusted R-squared:  -0.004999
## F-statistic: 0.04985 on 1 and 190 DF,  p-value: 0.8236
```

```
plot(hello$Siblings, hello$Texts, xlab = "Number of Siblings", ylab = "Number of Text Sent",
     main = "Number of Sibling vs. Number of Texts One Sends", pch = 19)
abline(hunch_model)
```



Question: Do you think a linear model is okay to use in this case? Why? (We are looking for a particular numerical measure)

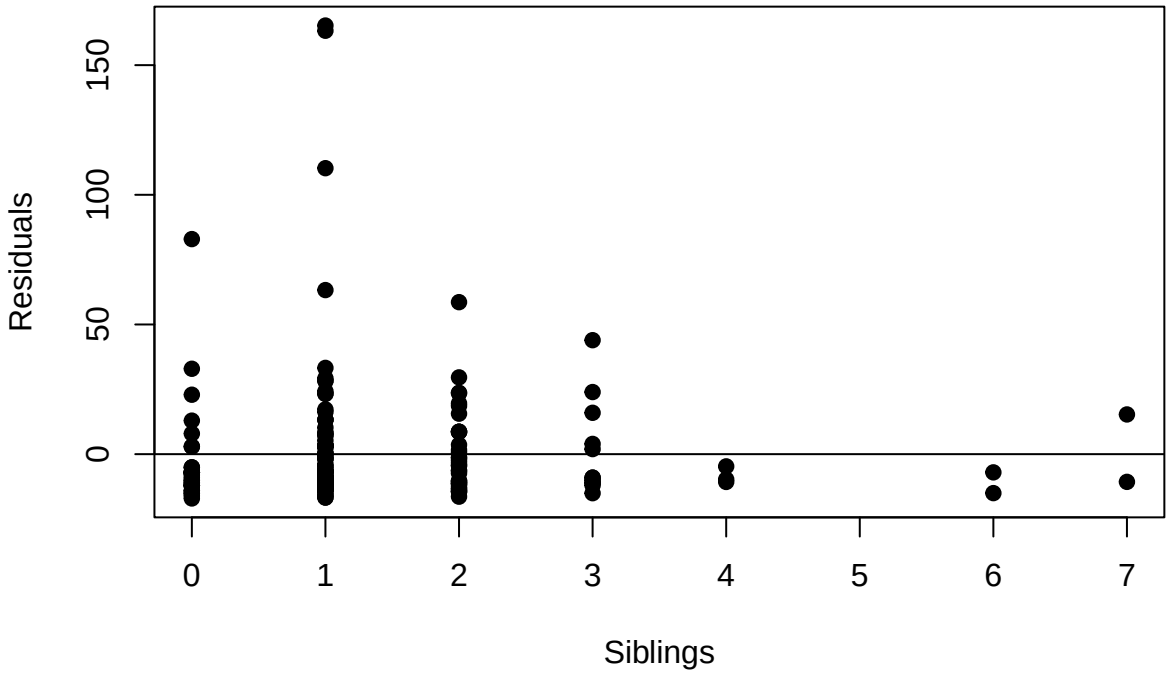
Answer: (Student Response Expected) Nope, R^2 values are literally very low and close to 0. This means

that linear model is not a good model for these variables. Also, the line is not looking like a good line of fit for the points.

Question 2: For some reason, Abhi doesn't believe you smh. Graph a residual plot to make sure Abhi understands we know what we are talking about. Use code and words (both) to convince Abhi that we are correct!

Answer:

```
#Code Section
plot(hello$Siblings, hunch_model$residuals, xlab= "Siblings", ylab="Residuals", pch = 19)
abline(0,0)
```



Word Section: Abhi like seriously try to understand that we are smart (at least more than you). The residual plot does not seem random and seems like its skewed to the left and there is definitely some pattern we can recognize. I understand the line of best fit looked linear, but it still doesn't mean a linear model would be a good fit for your question. If you are smart, you would understand what we are saying. ;)

Problem 2: Conceptual Modeling

Question 1: Which of the following is true about Residuals ?

A) Lower is better
B) Higher is better
C) A or B depend on the situation
D) None of these

Answer: (Student Response Here) A is correct, we want low residual values (this is not the r-squared value).

Question 2: Suppose that we have many independent variables (X1, X2, X3...) and dependent variable is Y. Now Imagine that you are applying linear regression by fitting the best fit line using least square error on this data.

You found that correlation coefficient for one of it's variable (Say X1) with Y is -0.93.

Which of the following is true for X1?

A) Relation between the X1 and Y is weak
B) Relation between the X1 and Y is strong
C) Relation between the X1 and Y is neutral
D) Correlation can't judge the relationship

Answer: (Student Response Here) B is correct, the 0.93 indicated a strong correlation or relation between X1 and Y. The negative just means that there is a negative correlation (imagine downward slope).

Question 3: Over-fitting is good because your model is perfectly predicting what it is supposed to?

A) TRUE
B) FALSE

Why?

Answer: (Student Response Here) B is correct (False), Over-fitting is not good because it means that your model is so good for your data set that it cannot predict anything else. Over-fitting is common issue, but a important one to be aware of.

Question 4: Bob calculated the correlation coefficient between Ice Cream Sales (X) and Temperature (Y). The coefficient turns out to be 0.72. Grace decided to switch it to see if the correlation is stronger with the flipped variables. What do you think happens with correlation coefficient?

Answer: (Student Response Here) The coefficient will stay the same since the value represents the association between two variables and it will be the same even if you flip your X variable and Y variable.

Project Questions

Feel free to work on your project if there is any time left after the labs. Paul and I are here to answer any questions during the second half of the lab times to answer mainly project related questions, but general questions are more than welcome too. Feel free to discuss among your group about any project ideas or help each other out. Remember collaboration is promoted, plagiarism is not! :)

Submission

Once you have finished your lab...

1. Go to the top left and click `File` and `Save`.
2. Click on the `Knit` button to convert this file to a PDF.
3. Submit **BOTH** the `.Rmd` file and `.pdf` file to Blackboard by 11:59 PM tonight.