# 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)

```
hello <- read.csv("~/Desktop/DPI 2022/hello.csv", stringsAsFactors=TRUE)
model = lm(Texts ~ Siblings, data = hello)
summary(model)
```

```
##
## Call:
```

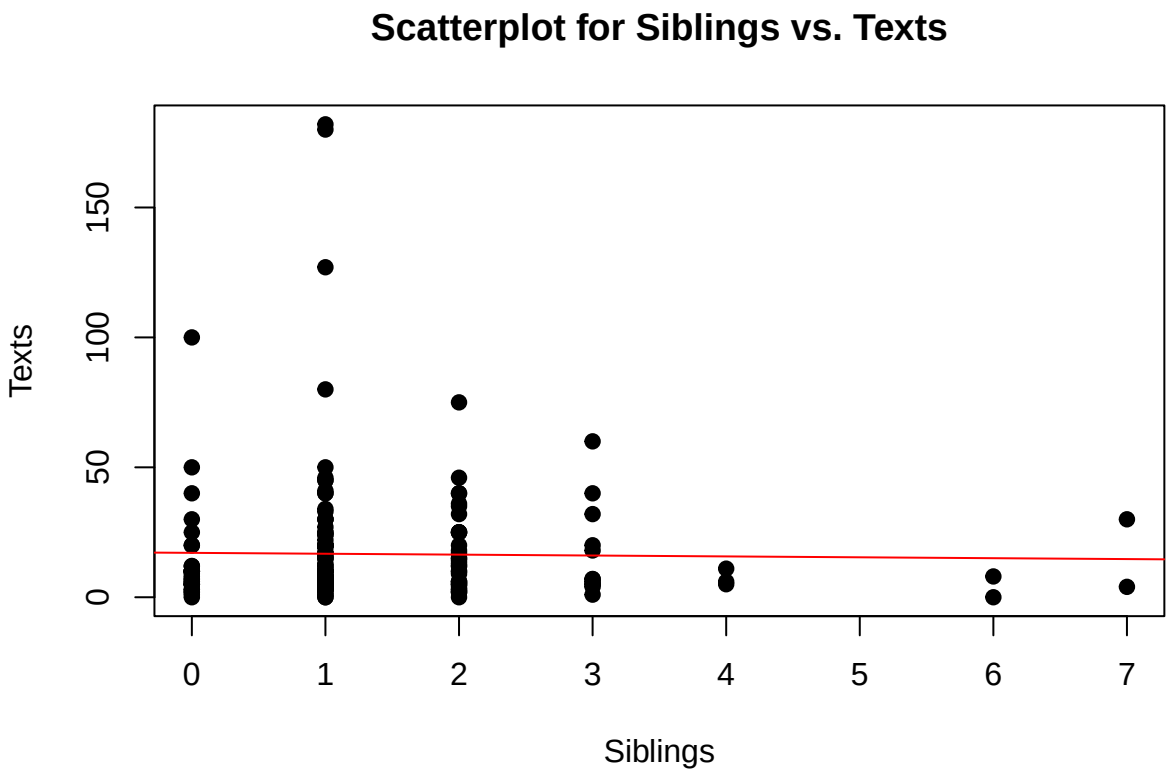
```
## lm(formula = Texts ~ Siblings, data = hello)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.080 -11.482  -8.739   3.261 165.261
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.0801     2.6446   6.458 8.64e-10 ***
## Siblings     -0.3415     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
```

**Question 1: Do you feel a linear model would be a good fit? Why?**

**Answer:** This is a terrible linear model, because the "Multiple R-squared" value is almost 0. This means that almost none of the variance in Y can be explained by X, which means the linear model explains pretty much none of the relation between texts and siblings.

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

```
plot(hello$Siblings, hello$Texts, xlab = "Siblings", ylab = "Texts", main = "Scatterplot for Siblings vs
abline(model, col = "red")
```



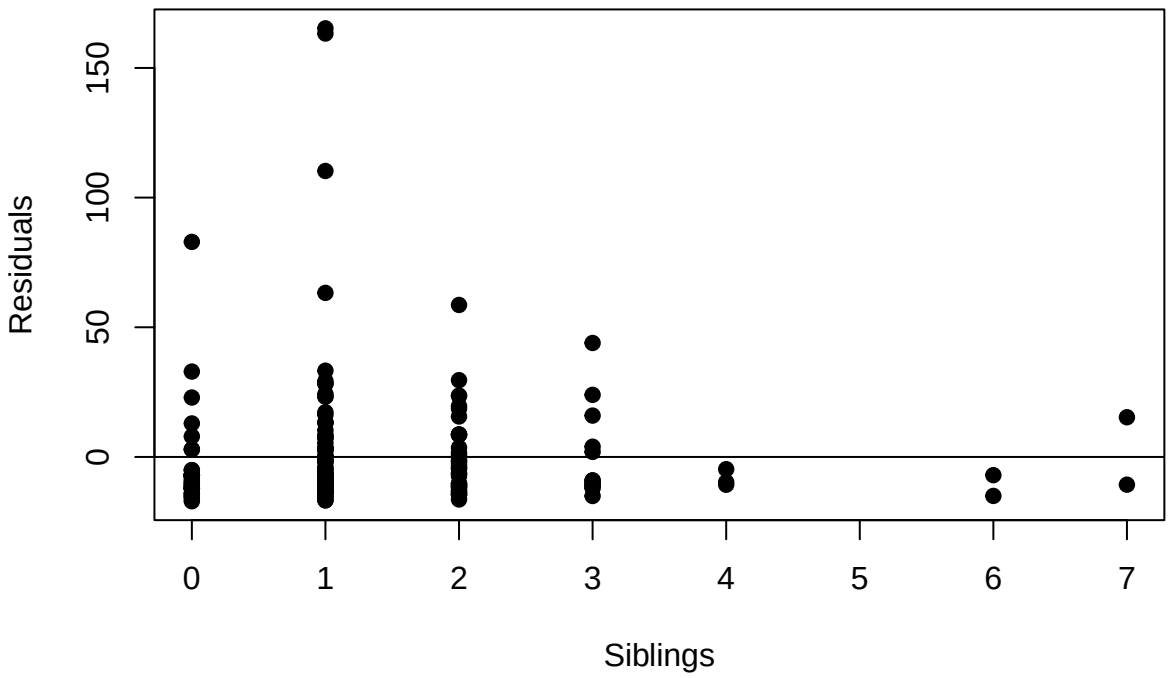

**Question 2: Do you think a linear model is okay to use in this case? Why? (We are looking for a particular numerical measure)**

**Answer:** A linear model is not good to use in this case, because "siblings" is not a continuous variable - it's discrete, since you can't have a non-integer amount of siblings. Linear models work best when comparing two continuous variables, and don't make much sense when comparing with discrete variables. The *particular numerical measure* that shows this is the $r^2$ value, because since it's super close to 0, we know that our linear model shows pretty much NONE of the relationship between siblings and texts.

**Question 3: 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:** When you look at this residual plot, you can see obvious patterns (lines); however, when you see a residual plot, you should see a jumbled, zero-correlation mess, but that's not the case here.

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




**Word Section:**

## 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:** C - If your residuals are too big, that means there's little correlation between my variables, but if your residuals are too small, that means your variables are over-fitted, and will be "too perfect" since it probably will end up fitting to data better than real-world values.

**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:** B - the relation between X1 and Y is strong, because a correlation coefficient close to -1 means that there's a strong negative (inverse) relationship between X1 and Y. This is because correlation coefficient ($r$) is the square root of the R-squared value ($r^2$). If you square your $r$ value of -0.93, you'd get an $r^2$ value of 0.8649, which means that a lot of the variance in Y can be attributed to X1, meaning that the relationship between them must be strong.

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

A) TRUE
B) FALSE

*Why?*

**Answer:** B - The answer is B, because overfitting can be bad. If you have overfitting, this likely means that your model works "too perfectly" for your specific sample data, and will probably end up having trouble coping with real data as a result.

**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:** The correlation coefficient will be the exact same, since a positive relation between two variables will still be positive if you flip them. What this means is that even though flipping the variables will cause the slope of the linear regression line to become the inverse of what it was before (for example, if the slope was 2 before, it would become 0.5), the points of the scatterplot would hug the line just as closely as they did before.

## 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.