2021 AlgLabII Final Project

Congratulations on making it to the final project! We have learned a wide variety of data analysis skills so far and now we will put them to the test. Our final project will allow us to practice our regression skills on public health problems you are interested in.

Step 1 : Logistics

- Final project will be in groups of 2-3 people.
- Every week for the remainder of class we will be going over an example of each step in the step by step outline of the project.
- Final project will be due December 3rd (the last day of class).
- Presentations will be the last class (Thurs 2nd/Fri 3rd).

Step 1 : Question

Our first step in the project is to formulate a scientific question. We have seen a number of scientific questions in our lab from "what is the association of drug dosage on survival time?" to "what is the association of poverty on education?".

You will have to formulate a scientific question that addresses the association between X and Y. Please chose a public health related X and Y that has at least one other variable Z.

Step 2 : Data

We have seen many examples of data sets, from simulated data to NHANES to COVID-19 data. In order to answer your question in Step 1.

Look for a dataset that has the variables X and Y that you are interested in from Step 1.

Step 3 : Repeat Step 1 and 2

Unfortunately, not all questions have data to answer them. You might have to iterate on Step 1 and Step 2 until you find a question that has data associated with it.

Step 4 : Plot/Summarize the data

As we have seen in lecture, plotting the data is key. This involves the following steps.

- Summarize X and Y.
- Plot X.

- Plot Y.
- Plot X against Y.

Step 5 : Formulate the Regression

Formulate your regression by investigating the type of your X,Y,Z. Are they numeric, categorical?

Specificy the correct regression in terms of your specific public health variables

$$Y \sim X + Z$$

Step 6 : Run / Interpret the Regression

For each coefficient (beta) interpret the results in the context of your question. We have seen this when answering questions related to "what is the effect of a unit increase in drug dosage".

Step 7 : Diagnose the regression

As we have seen in lecture, once we run a regression we must answer how well the regression fits the data.

- Plot the residuals
- Plot the fitted versus residuals
- Plot residuals versus leverage.
- Identify any outliers and re-run the regression with outliers removed.
- Identify any changes in the estimated coefficients.

Step 8 : Conclusion and Limitations

Conclude with some remarks about your estimated effect and what the limitations of your study might be.

Paper Outline

The write up should be no more than 3 pages.

- **Introduction** Introduce the scientific problem.
- **Data** Describe how you obtained the data and what the data represents.
- **Regression Analysis** Run/report the regression results and diagnostics.
- **Conclusion** Conclude with some remarks about your estimated effect and what the limitations of your study might be.

Presentation Outline

We will have presentations the last week of class. We will give each group 10-15 minutes to present their project. The presentation should follow the exact same setup as the paper with 1-2 slides for each section of the paper (Introduction, Data, Regression Analysis, Conclusion/Limitations).

Grading

- **Proposal** 20 points (Did you posit a specific scientific question and find a dataset that can answer this question)
- **Regression** 20 points (Did you forumlate a regression suitable to answer your scientific question and did you interpret it properly)
- **Write-Up** 35 points (Did you include each section Introduction, Data, Regression Analysis, Conclusion/Limitations with correct model specification, interpretation, and limitations).
- **Presentation** 25 points (Did each group member get a chance to speak. Did you cover each of the Introduction, Data, Regression Analysis, Conclusion/Limitations sections.)