STS 2300 – Project 2

Regression / Statistical Inference Report

# Background

The goal of this project will be for you to practice what you’ve learned about regression and statistical inference (confidence intervals and hypothesis tests) on a real-world issue of your choosing.

There are two options for doing this:

1. Use the data **project2\_data.csv** (posted in the data folder on our code repository), which includes information on dozens of variables from a random sample in 2022 of 30,000+ US residents. The data is from the National Health Interview Survey, which is conducted annually by the Centers for Disease Control and Prevention (CDC). I do **NOT** want you to use the full 30,000+ observations in the dataset. Instead, I want you to think of a question that interests you and use a subset of the data to investigate your question of interest question.
2. You may use another data set (or collect one) that can reasonably be considered a random sample from a population of interest. If you have an idea for this option, discuss it with me to get approval before moving forward.

After analyzing the data, you will present your results in a format of your choosing (e.g., paper, poster, or video).

# Step 1 – Choose Data Option and Question of Interest

Decide whether you will use the project2\_data.csv file or your own data. Then choose a question of interest that will let you do at least two of the following: linear regression, confidence interval, or hypothesis test. The two analyses you choose don’t have to answer the exact same question, but they should be related to a central question of interest.

If you use the project2\_data.csv, I encourage you to focus on a subset of the data as your sample (e.g., 18 – 22 year olds, people who smoke, people who work over 40 hours per week, etc.). At the end of this document, I have included some *possible* topics of interest, but I encourage you to choose something that interests you (whether it is from that list or not).

Once you have chosen your data and a question of interest, complete the project topic form and submit it.

# Step 2 – Analyze Your Data

You will have to do ***at least two of the following three*** things:

* Create a regression line related to your question of interest
* Calculate a confidence interval related to your question of interest
* Conduct a hypothesis test related to your question of interest

There is nothing to submit for this step, but it will prepare you for creating your final product.

# Step 3 – Create a Product Summarizing Your Results

This product can take several different forms. Below are some examples. You may think of something that is not listed here, which is great. Just check with me if you have another idea.

* Short paper using R Markdown
* Research poster (like someone might do for SURF Day)
* Video with slides

Whichever product type you choose, you will need the following sections and content. There is no strict page or length requirement as long as you thoroughly address each piece below. A rubric will be provided showing the points breakdown.

* Introduction
  + Introduce the reader/viewer to the data you used (where is it from, what variables does it include, etc.) and your question of interest. (Note: You only need to discuss variables relevant to your question of interest.)
* (If chosen as one of your two) Regression
  + Calculate and display your regression line equation
  + Interpret the y-intercept and (at least one) slope
  + Calculate and interpret the R2 value
  + Make a reasonable prediction with your equation
* (If chosen as one of your two) Confidence Interval
  + Calculate an estimate for your parameter and discuss why a confidence interval is needed (as opposed to just using your estimate).
  + Generate and graph a bootstrap distribution. Describe what it shows.
  + Use the bootstrap distribution to calculate a confidence interval. You can choose a confidence level.
  + Interpret your confidence interval in context of the problem.
* (If chosen as one of your two) Hypothesis Test
  + State your hypotheses. Define any parameters as needed. You can choose a significance level you will use.
  + Summarize your data (with an estimate and a graph).
  + Generate a null distribution with a graph that includes your p-value.
  + Calculate the p-value and write a sentence about what it means in context.
  + Provide a formal conclusion for your test.
* Summary
  + Summarize the main results and conclusions as they relate to your question of interest.
  + Discuss possible ramifications your results may have for someone interested in your topic.
  + Discuss any limitations to your findings. Think back to some of the things we discussed earlier this semester. Are the variables recorded the way you’d like them to be? Are there other variables that would have been useful?
  + Discuss future studies that could be done to further explore your topic. For example, is there other data someone could gather? Did what you found bring up a new question?
* References
  + Include citations if appropriate. If you used the project2\_data.csv file you should include:

Lynn A. Blewett, Julia A. Rivera Drew, Miriam L. King, Kari C.W. Williams, Annie Chen, Stephanie Richards, and Michael Westberry. IPUMS Health Surveys: National Health Interview Survey, Version 7.3 [dataset]. Minneapolis, MN: IPUMS, 2023. <https://doi.org/10.18128/D070.V7.3>

# Possible Topics of Interest

You are not required to choose one of these topics. These are only for people who need help settling on a topic.

* What proportion of people who have had depression take medication for it?
* Are people who work full time more likely to have had COVID-19?
* What proportion of college graduates are female?
* How fast do 10- to 18-year-olds grow taller? (looking at AGE vs. HEIGHT)
* How does health status differ among people with and without paid sick leave (or workplace health insurance)?
* Are sexual orientation rates different across different regions of the US?
* Are US citizens more or less likely to have graduated high school than non-US citizens?
* Are people who smoke also more likely to drink?