SDS 291: Case Study Guidelines

There will be four open-ended case studies assigned throughout SDS 291. These case studies give you a chance to *individually* grapple with a dataset, complete an appropriate analysis, and interpret your results. The focus of these assignments is equally on the correctness of your statistical analysis and on the clarity with which you communicate your procedure and findings. The final product for each case study is a short report of 1–2 pages, single spaced and excluding figures and tables, along with an accompanying R appendix with your code.

Format

Each case study will be comprised of two components: a short written report and a corresponding R code appendix.

- Your report should be written entirely in Quarto (or R Markdown / R Sweave), and your R code appendix should be included as a single code chunk at the end of your short report. I will provide a Quarto template for you to use for the first case study assignment, if you wish. You may also find the SDS 100 labs on Polishing Figures (Lab 5), Formatting in Quarto (Lab 8), and Referencing Figures and Tables (Lab 10) to be helpful resources as you work on your case study: https://smithcollege-sds.github.io/sds100/.
- Please include the following information at the top of your Quarto document:
  - your name,
  - the case study number (#1, #2, #3, or #4) and date, and
  - a descriptive title for your analysis.
- Your final submission to Moodle should include (i) a single PDF document containing both your short report and R code appendix and (ii) the corresponding .qmd file.

Content

Short Report: Your short report should be 1–2 pages of written analysis, single spaced, with the following four sections/components:

1. An introduction to the problem. Your introduction should provide sufficient background information to motivate and explain your analysis. In short, what question(s) are you trying to answer and why?
2. A description of the data you have on hand to answer your question. Your description should include a discussion of relevant summary statistics (EDA) in context, with at least one accompanying graph.
3. A description of the final model(s) explored and the relevant results. This description must include:
   - A mathematical description of the population model(s) used, with all relevant variables clearly defined. For example,

     $$E[\log(income)|education] = \beta_0 + \beta_1 education,$$

     where $\log(income)$ is the natural logarithm of an individual's net income (measured in US dollars in 2023) and *education* is the number of years of schooling the individual has completed.
   - The estimates of the model parameters (e.g., β_0 and β_1 in the above population model) and their accompanying standard errors.
   - Interpretations of the relevant model parameter(s) / estimated population mean(s) / prediction(s), with accompanying confidence interval(s) or prediction interval(s).

4. A final summary of your findings and conclusions, as well as a discussion of any limitations of your analysis and conclusions. When you think about possible limitations of your analysis, some questions to consider include: to what population(s) do your results generalize (and is this the population that you are actually most interested in)? Can the associations you describe be interpreted as cause-and-effect relationships? Are there any lingering assumption violations that you have concerns about?

Note, your written short report should be a *big-picture summary* of your analysis and conclusions. You should not give a step-by-step description of your complete analysis.

R Appendix: Your written short report *should not include any direct R output or code*, save for your plot(s) and nicely formatted tables. As such, all code chunks in the body of your written report should be set to either `# | echo: false` (if they produce a plot or a regression table) or `# | include: false` (for all other chunks). Instead:

- Copy and paste all code that you used to conduct your analysis (including additional regression models or assumption checks that you did not discuss in your write-up!) into one chunk at the end of your written report.
- Your code should be commented throughout, telling me what each section of code is doing.
- Please edit your code so that there are no typos or errors, and—while I do want to see the associated R output or plots from your code—please avoid printing entire (or partial!) datasets to the screen.

Collaboration

You are **not** allowed to work collaboratively on the case studies, and I ask that questions related to the case studies **not** be posted in the course Slack channel. However, you are welcome to reach out to me individually (preferably over Slack direct message) to discuss questions about the data set or your analysis of it!