Case Study 3 (F25)

Due Dates
Lectures 100/200: Monday, November 24th, at 8:00PM Ann Arbor Time via Gradescope
Lectures 300/400: Tuesday, November 25th, at 8:00PM Ann Arbor Time via Gradescope

Late Submission Policy
We offer a 1-hour grace period without any penalty. Submissions uploaded after the 1-hour grace period but within 24 hours of the deadlines listed above will be accepted but are subject to a 10 percent late penalty. No submissions will be accepted thereafter.

Paired Submission Policy
Only *active* lab approach students may submit with a partner. Your partner must be in the same lab section - and you both must be in attendance for the corresponding lab session. To receive the 5 points of extra credit, you must properly tag your partner to the submission via Gradescope. See the submission instructions at the bottom of the lab document (or ask your lab instructor) for additional help.

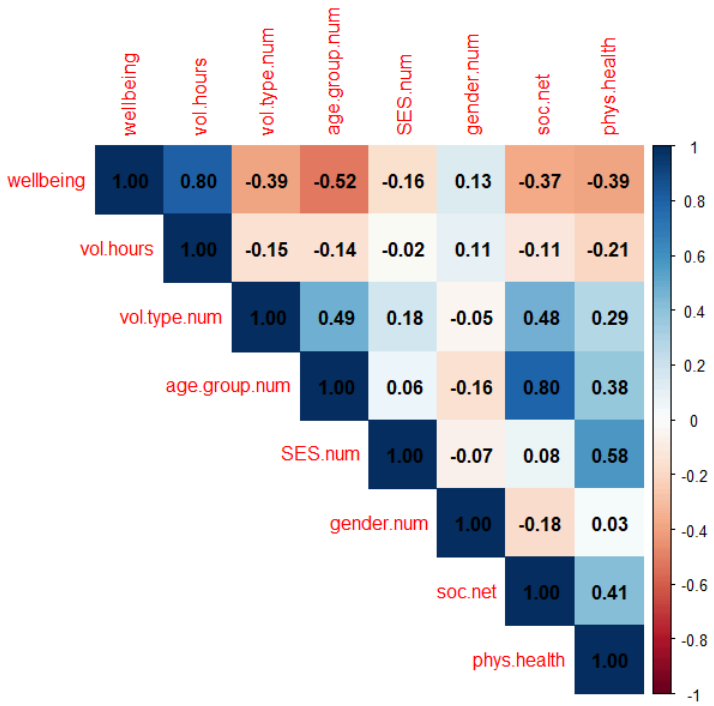
Purpose
The overall purpose of this case study is to build your skills in conducting an applied statistical analysis in a real-world context. Specifically, this assignment is designed to (1) improve your statistical writing skills; (2) assess your proficiency using testing and interval estimation methods to evaluate associations between variables; and (3) challenge you to make data-based arguments in a setting where decisions based on statistical analyses have substantial real-world implications.

Giving Back, Feeling Good: Does Volunteering Predict Psychological Well-Being?
The United States has long been known for its rich tradition of community service. Nearly two centuries ago, Alexis de Tocqueville (1835) commented on the American tendency to create or join voluntary associations in huge numbers and to invest their energies in uncompensated civic service. Although volunteer work is widely believed to be beneficial not only for the community but for the individuals who perform it, only a few studies have explored the effect of volunteer service on the psychological well-being of those who pursue it. In your third case study, we ask you to use what you've learned about linear regression modeling to explore the relationship between volunteer work and psychological well-being, accounting for factors that might influence this association.

Data
The *volunteer* dataset contains responses from n = 100 randomly selected members of *VolunteerMatch,* a national organization that focuses on connecting individuals with volunteer opportunities in their local communities. Each individual sampled volunteered at least once during the previous calendar year (2024). Although *VolunteerMatch* works to find volunteering opportunities for people of all ages, for the purposes of this assignment, we have restricted the data to volunteers who were college students and volunteers who were retirees. The variables recorded for these individuals include:

- **wellbeing:** self-reported well-being, measured on a scale from 0-100, with higher values indicating higher overall psychological well-being
- **vol.hours:** the total number of hours spent volunteering in 2024
- **vol.type:** an indicator of whether the individual's volunteering activities primarily addressed *social* causes (e.g., volunteering at a soup kitchen or safe house) or *environmental* causes (e.g., cleaning up trash in a community park)
- **age.group:** an indicator of whether the volunteer was a college student or a retiree
- **SES:** a variable indicating the volunteer's current socioeconomic status (low, medium, high), based on household income
- **gender:** an indicator of the volunteer's self-reported gender (female, male)
- **soc.net:** a variable describing the strength of the volunteer's social network, recorded as the average number of social interactions the volunteer had per week in the previous calendar year
- **phys.health:** a score describing the volunteer's overall health for their age, measured on a scale from 0-10, with higher values indicating better health

A correlogram of how these variables relate to each other is provided below:



Task

Your goal is to build a series of regression models that estimate the overall relationship between an individual's *well-being* (*wellbeing*) and the number of hours they commit to volunteer work (*vol.hours*). In developing your models, you will demonstrate what you've learned about simple and multiple linear regression by evaluating the fit of the models you estimate and discussing what factors are important to consider when analyzing how volunteering relates to a person's psychological well-being. Your analysis and written report will include the following main components:

1. *Simple Linear Regression Analysis:* Create a simple linear regression model (Model A) that predicts an individual's self-reported well-being (*wellbeing*) using *only* the number of hours spent volunteering (*vol.hours*). Use the results of this model to run a population-level test of linear association *and* construct a confidence interval that estimates the true but unknown population-level slope.
2. *Multiple Linear Regression Analyses:* Add additional predictors to your initial model to create three multiple linear regression models (Model B, C, and D). These models should satisfy the following criteria:
    B. Mode should include *vol.hours* and *at least* one additional *quantitative* predictor. This model should not include any categorical predictors, nor interactions.
    C. Model C should include *vol.hours* and *at least* one additional quantitative predictor. This model's set of predictors should differ from the previous model. This model should not include any categorical predictors, nor interactions.
    D. Model D should include *vol.hours* and *one categorical* predictor. This model should include an interaction between *vol.hours* and your categorical predictor. This model should not include any additional quantitative predictors.
3. *Discussion:* First, summarize your findings from (1). Then, explain how you selected additional predictors for the multiple linear regression models you created in (2). Summarize your findings from (2) and compare the results to (1). Provide a conclusion regarding whether the number of hours spent volunteering is linearly associated with an individual's well-being, and whether additional predictors influence this association.

These components should comprise a written report that demonstrates your ability to interpret the results of a regression model and your understanding of how to select helpful predictors.

Assignment

This case study is broken up into two parts: Part 1 (30 points), analyzing the data; and Part 2 (70 points), writing up your results and recommendation in a brief report.

Part 1 - Data Analysis Task (*Completed during lab - Worth 30 points, see below for rubric*)

Your analysis should be partitioned into three main components.

1. *Simple Linear Regression Analysis:* Create a simple linear regression model (Model A) that predicts an individual's self-reported well-being (*wellbeing*) using only the number of hours spent volunteering (*vol.hours*). Report the summary results of this model. Create a scatterplot that visualizes the estimated model. Create a 95% confidence interval for the population-level slope.
2. *Multiple Linear Regression Models:* Use the provided correlogram to determine which explanatory variable(s) might also assist in predicting an individual's self-reported well-being (*wellbeing*). Create *two* multiple linear regression models (Models B and C) that each contain *vol.hours* and *at least* one additional *quantitative* predictor. Report the summary results of these models. You do not need to provide graphical representations in this section.
3. *Multiple Linear Regression Model with an Interaction:* Create one last multiple linear regression model (Model D) that contains *vol.hours* and *one categorical* predictor. This model should include an interaction between *vol.hours* and the categorical predictor you included in your model. Report the summary results of this model. Create a scatterplot that visualizes the estimated interaction.

Part 2 - Writing Task (*Completed outside of lab - Worth 70 points, see below for rubric*)
Write a report summarizing your findings from the Data Analysis Task. Your report should provide a brief summary of the goals of your analysis, demonstrate a clear understanding of its results, and provide a conclusion regarding whether the number of hours spent volunteering is linearly associated with an individual's self-reported well-being, and whether additional predictors influence this association. Successful statistical reports should follow the structure recommended in the outline below and range between approximately 700 and 1000 words. **Submissions that are between 1100 and 1500 words will receive a 5-point penalty, and those that exceed 1500 words will receive a 10-point penalty.**

Introduction and Simple Linear Regression Model
Provide a summary of the goals of your analysis and describe the data. Report and summarize your findings from Model A. Highlight and interpret key results. A successful analysis will be separated into two paragraphs:

a. In the first paragraph, describe the purpose of the analysis and justify its importance. Summarize the data that have been collected. Successful summaries will state the observational units of the analysis, note the sample size, and describe the variables that were analyzed in the models you fit.
b. In the second paragraph, thoroughly define, interpret, and evaluate the results of your simple linear model. All definitions and interpretations should be provided *within the context of the analysis*.
    - Report and interpret the estimated slope coefficient
    - Report and interpret the RSE value
    - Report and interpret the R^2 value
    - Conduct a formal hypothesis test to evaluate evidence of a population-level linear relationship between *wellbeing* and *vol.hours*. Define the parameter of interest for this test. Report and define the associated test statistic. Report and define the corresponding *p*-value. Provide a written conclusion that explicitly identifies the level of evidence the data provide in favor of a linear relationship.
    - Report and define the 95% confidence interval for the population-level slope

Multiple Linear Regression Models
Report and summarize your findings from Models B and C. Highlight and interpret key results. A successful analysis will be separated into two paragraphs:

a. In the first paragraph, explain how you selected additional predictors for the three multiple linear regression models you created. This discussion should consider the strength of the association between a predictor and *wellbeing* as well as any potential problems of collinearity.
b. In the second paragraph, compare Models A, B, and C. First, compare the estimated slopes for the *vol.hours* predictor and comment on whether evidence of collinearity is found in any of the multiple linear regression models. You can but are not required to provide *interpretations* of the estimated slopes. Then, identify whether Model A, Model B, or Model C provides a superior fit to the collected data. Justify your opinion using output from the models.

Multiple Linear Regression Model with an Interaction and Conclusion
Report and summarize your findings from Model D. Highlight and interpret key results. A successful analysis will be separated into two paragraphs:

a. In the first paragraph, explore the interaction…
    - Identify the model's reference group and explicitly report the values of the estimated slopes associated with *each* group of your categorical variable. Interpret *at least* one of these slopes *within the context of the analysis*.
    - Use the summary output and graphical representation to explain whether the model provides evidence of an interaction between your model's quantitative and categorical predictors.
b. In the second paragraph, identify which of your four models provides a superior fit to the collected data. Justify your opinion using the output from the models. Then, based on the results of *all* your models (A, B, C, and D), provide concluding statements regarding how the relationship between *wellbeing* and *vol.hours* is (or is not) influenced by any additional predictors.

Criteria for Success:

Rubric for Data Analysis Task (30 points)

		Model A **(10 points)**	Models B and C **(10 points)**	Model D **(10 points)**
Statistical Analysis	**Exemplary** Correctly addresses all elements of the prompt, demonstrating expert understanding and interpretations of required concepts / skills.	10	10	10
	Proficient Correctly addresses all elements of the prompt, demonstrating understanding and interpretations of required concepts / skills with just one minor error.	8	8	8
	Emerging Correctly addresses most elements of the prompt, demonstrating solid understanding and interpretations of required concepts / skills. May include a few minor computational/conceptual mistakes.	6	6	6
	Needs Improvements Demonstrates limited understanding of statistical concepts / skills. Large errors or conceptual mistakes are present.	2	2	2
	Missing Completely missing any attempted analysis.	0	0	0

Criteria for Success:

Rubric for the Written Report (70 points)

		Overall Report **(5 points)**		
Presentation	**Proficient** Report flows in a logical order, containing minimal grammatical or spelling errors. Writing is clear, creative, and informative.	5		
	Emerging Report includes passages that are vague, unclear, or include substantial grammatical or spelling errors.	3		
	Needs Improvements Report displays little logical order, suffering from many grammatical or spelling errors.	1		
		Simple Linear Model **(30 points)**	Models B and C **(20 points)**	Model D and Recommendation **(15 points)**
Statistical Reasoning	**Exemplary** Correctly addresses all elements of the prompt, demonstrating expert understanding and interpretations of required concepts / skills.	30	20	15
	Proficient Correctly addresses all elements of the prompt, demonstrating solid understanding and interpretations of required concepts / skills. May include one or two minor conceptual mistakes.	27	18	13
	Fair Correctly addresses most elements of the prompt, demonstrating an understanding and interpretations of required concepts / skills. Includes several mistakes, one of which may be major.	24	15	11
	Emerging Correctly addresses some elements of the prompt, demonstrating an understanding and interpretations of required concepts / skills. May include more than two minor conceptual mistakes OR demonstrates major misunderstandings.	20	12	8
	Needs Improvements Demonstrates very limited understanding of statistical concepts / skills. Includes more than one major misunderstanding in core concepts.	10	8	5