BIOS 507 HOMEWORK 3

Due 3/3/2025 by 11:59pm

Directions: Complete all questions. Any R or SAS code used should be attached at the end of the homework. Collaboration is encouraged, but the final product must be your own work.

Problem 1

A research team was tasked with optimizing the synthesis of a novel organic compound. Their primary goal was to understand the reaction mechanism, particularly the role of a key intermediate species believed to be crucial for achieving high yields of the final product. The team hypothesized that higher concentrations of the intermediate species during the reaction would correlate with increased yields of the final product. Conducting this experiment required a significant amount of time each day, as it was difficult and highly technical. Because of this, the research team hired a new lab technician to conduct the experiment full time. Every day, the technician ran the experiment. First, they grew the intermediate species to the desired concentration, which was a difficult task that took most of the morning. Then, in the afternoon, they synthesized the novel organic compound. This synthesis process was also highly difficult, and took the technician the rest of the work day. Because each experimental run takes an entire day, the study had to be carried out over a long period of time, totaling 100 days. The data were logged in `catalyst_dataset.csv`, available on Canvas, and contain the following columns:

Day A column logging which day of the study the new technician performed the analysis.

Intermediate_Concentration A column logging the concentration of the intermediate generated during the morning phase of the analysis.

Yield A column logging yield of the novel compound obtained by the technician in the afternoon using the intermediate species obtained in the morning.

1. Ignore any potential "problems" in the data and answer the researcher team's questions using a simple linear regression model. Draw conclusions about the association between concentration of the intermediate species and total yield of the organic compound. Be sure to carry out all of the usual steps in the analysis.

2. Perform diagnostics. Is anything wrong? If so, what might be the reason for this?

Problem 2

A nutrition study was carried out to try to better understand the relationship between protein consumption at breakfast and energy levels throughout the day. A sample of size 20 was collected, and for each individual in the study the researchers measured:

- Protein consumption at breakfast on a continuous scale ranging from 1 to 25.
- A energy score (self-rated) ranging from 1 to 100

The data can be found on canvas under `nutrition_protein_dataset.xlsx`. Note that this is an excel file instead of a csv. It can be read into R using the `read_excel` function. This requires you to load the `readxl` library.

1. Analyze the data using a simple linear regression model and draw conclusions about the association between protein consumption and energy levels. Be sure to carry out all of the usual steps in the analysis.
2. Create a plot of the residuals vs fitted values. What do you notice?
3. How might this be affecting your conclusions? Specifically comment on the point estimate and the confidence interval from Part (1).
4. Interpret what you saw in the residuals vs fitted plot in terms of protein and energy levels (instead of residuals and fitted values). Provide some guess as to why this might happen. Note that there is no single "correct answer" here.

Problem 3

A student is performing a class project to assess the relationship between daily sodium intake and systolic blood pressure. The project requires them to design a study, collect the data, and analyze the results. The student goes door-to-door in their apartment building, and for families that are willing to participate, the student collects data on sodium intake and systolic blood pressure for each member of the family. The data are available in `sodium_SBP_data.csv` on Canvas. The data set contains the following columns:

family_id Which family each observation was collected from

sodium Self-report of daily sodium consumption (mg)

blood_pressure Measurement of blood pressure (mm Hg).

1. Investigate the association between sodium intake and blood pressure using a simple linear regression model. Be sure to carry out all of the usual steps in the analysis.

2. What is a potential issue with this study? How would it affect the quantities that you reported in Part 1? (HINT: when I created this dataset, I set the true association between sodium intake and blood pressure to be $\beta_1 = 0.01$.)

3. Generate a qqplot and a histogram of the residuals. What potential problem do you see? Why might this problem NOT be your first priority to fix.