STAT 705, Spring 2020, Assignment 1

This is a graded assignment, worth 25 points.

Directions

Two datasets are provided for this assignment. For each dataset, you will need to write the SAS code to answer the following questions. Enter your answers directly into Canvas.

In addition to answering these questions, you will need to to upload your SAS program file. This is the file with the "sas" extension, and it contains the statements that appear on the Code tab in SAS Studio. The complete code for both datasets needs to be in one file. You can use any text editor (like Notepad or Wordpad) to copy and paste the code for both datasets into one file.

Please do not upload your SAS output file, the SAS log file or the HTML file. I need your original program file so that I can execute it and generate the output for myself. This is what I use to assign partial credit, in the event your answers are incorrect.

Dataset #1

Global warming is of particular concern to some researchers, and the temperature of the world's oceans is thought to be related to global warming. The dataset consists of measurements taken at the surface of the ocean at 100 locations around the globe. At each location, the surface temperature (in degrees Celsius) and the salinity of the water (in g/kg, grams of salt per kilogram of water) was measured. We want to develop a simple linear regression model that will use the salinity to estimate the temperature.

Use SAS to fit a simple linear regression model that will estimate the temperature from the salinity. As a result of the analysis, you will need to answer the following questions. Use SAS to perform the calculations, and perform hand calculations only when necessary. You will submit your answers directly into Canvas using the Quiz tool.

Note: We are using "Salinity" to predict "Temperature", so please be careful when choosing X and Y for the regression analysis. You can lose a lot of points if X and Y are backward in your analysis.

1. Do any of the model assumptions appear to be violated? Explain.
2. Does the model seem to adequately fit the data? Explain.
3. What is the estimated regression equation?
4. Provide a point estimate for the temperature at a single location in which the salinity is 33.495 g/kg. (Note: This value for salinity is not in the provided dataset. It is incorrect to change this value.)
5. Provide an interval estimate for the temperature at a single location in which the salinity is 33.495 g/kg.
6. Provide an interval estimate for the average temperature for all locations that have salinity 33.495 g/kg.

(Questions for dataset #1 continue on the next page.)

7. Interpret the estimated slope on salinity.
8. Explain why it would not be appropriate to use the estimated regression equation to estimate the temperature for a location that has salinity 28.125 g/kg.
9. What proportion of the variability in temperature is explained by this regression model?

Dataset #2

For the provided (X, Y) pairs, fit a linear regression model using Y as the response variable and X as the predictor variable. **As with the first dataset, be careful when designating X and Y in your code. You can lose a lot of points if X and Y are backward.**

10. Identify the violation(s) of the assumptions for this model. Explain what you see in the SAS output that makes you think an assumption has been violated.
11. Specify one transformation that should alleviate the violation. (Examples: take the log of Y, take the square root of X, take 1/Y).
12. Re-fit the model using the transformed variable you identified in question 11. Did the transformation alleviate the violation?