ST 513 Mini-Project 2

In this project you will create a report and a SAS program. These should both be uploaded to Moodle via the assignment link.

Notes about the project:

- You'll be working alone for this project. You can discuss the project with others but cannot share code.
- This project is pretty straightforward. I'm asking you to take the concepts and principles we've discussed so far and apply it to some data.
- I'm going to push you to learn some SAS stuff we didn't do in class.
- As this is relatively straightforward, there will be no resubmissions.
- Be sure that your SAS program adheres to the SAS program file submission guidelines (available on Moodle in the "Resources and Information" Section).

Datasets:

The dataset USEuropeCars.csv has information about different cars that were on the road for at least 100,000 miles. The goal is to use this sample of cars to relate to the population of all European and US made cars.
The variables are:

- observation: the observation number
- horsepower: the horsepower of the car
- region: the region -- USA, Europe, Asia -- where the car is from
- mpg_before: the average miles per gallon before the car reached 75,000 miles
- mpg_after: the average miles per gallon after the car reached 75,000 miles

SAS Programs:

You should be documenting all of the code you end up using with corresponding comments in a .sas program (or notebook). Things to do below:

Data things:

- Read the data into a permanent library
- Remove the observation variable.
- Remove the observation that has Asia as the region.
- There is an obvious outlier in one of the other variables, remove this observation as well.
- Create an average gas mileage variable.

Analysis items:

- Create histograms for horsepower, mpg_before, mpg_after, and average mpg. Include different coloring for each region. Overlay smoothed density plots.
- Create scatterplots to assess the relationship between the numeric variable. Color the points by the region variable (check SGSCATTER procedure).

- We discussed fitting parameters of distributions using either MOM estimators or MLEs. We can easily fit MLEs using the UNIVARIATE procedure.
- We also saw that we could use a qqplot to assess normality. qqlots are more general than that. We can use a qqplot to compare any fitted distribution to a theoretical one. In all cases, we are looking for a straight-line. If we have one, this indicates roughly that the model is a good fit for the data.
    - Use PROC UNIVARIATE and the CDFPLOT and QQPLOT statements to fit a gamma distribution to the horsepower variable.
    - Use PROC UNIVARIATE and the CDFPLOT and QQPLOT statements to fit a gamma distribution to the mpg_before variable.
    - Use PROC UNIVARIATE and the CDFPLOT and QQPLOT statements to fit a weibull distribution to the mpg_before variable.
    - Note: you can use EST in your qqplot options to specify the estimated value for a parameter like alpha.
- Use SAS to create a 90% confidence interval for the mean horsepower for all cars (regardless of region).
- Use SAS to create a 95% confidence interval to inspect the relationship between the average overall gas mileage between European and US cars.
- Use SAS to create a 99% confidence interval to investigate the average of the differences between after and before gas mileage.

Report:

You should write a report to an audience that has strong quantitative literacy skills but doesn't necessarily know a lot about statistics. The report should include the following:

- Intro:
    - Intro describing the data and the goals of the project (which you should glean from the programming tasks).
    - Discussion of the variables led by your plots created (these plots should show up in the text and should be described).
- Estimation:
    - Describe the idea of maximum likelihood and why we want to use it to estimate parameters. This should be clear and concise!
    - Show the plots corresponding to each estimation above. State the fitted distributions (i.e. The estimated curve is a Gamma with shape parameter …) and discuss the qqplots. State whether you think the fitted curves are a good model for the variable.
- Inference:
    - Take a paragraph or two and describe the purpose of confidence intervals and how to interpret them. Also describe the idea of confidence.
    - Have a subsection for each of the confidence intervals you were asked to create. Describe the procedure briefly (including stating the interval formula).
        - State the assumptions you are making for each interval and comment on the assumptions based on the data (if appropriate).
        - Interpret each interval in the context of the problem. Make any

obvious conclusions that would be important (such as with the last interval asked for).

Rubric:

Item	Points	Notes
Intro	10	Worth either 0, 4, 7, or 10
Data read in and appropriately modified	10	Worth either 0, 4, 7, or 10
Graphical summaries correctly created and described	10	Worth either 0, 4, 7, or 10
Estimation correctly done and described	20	Worth either 0, 5, 10, 15, or 20
Confidence intervals created correctly	20	Worth either 0, 5, 10, 15, or 20
Confidence intervals discussed appropriately (both idea and interpretations)	20	Worth either 0, 5, 10, 15, or 20
Assumptions	10	Worth either 0, 4, 7, or 10

Notes on grading:

- For each item in the rubric, your grade will be lowered one level for each error (syntax, logical, or other) in the code or for each required description that is missing or lacking. The descriptions describe the coding part but also correspond to the relevant part of the write-up.
- You should use Good Programming Practices. If you do not follow GPP you can lose up to 25 points on the project.