Statistics 108, Project 1, *Due: November 13th by 5pm*

Turn in the report in electronic form (word or pdf) through Canvas

Instruction: This project is to analyze a dataset, from start to finish, based on the simple linear regression model. It is an *individual* project. Students could discuss with each other to get better understanding of the project. Copying solutions or computing codes from other students or other sources is plagiarism. At a minimum, all students involved will receive a 0 on this project for any type of academic dishonesty.
R codes: Attach the entire R codes you used to analyze the data at the end of the report.

Data description: The data in the file "UN.txt" contains PPgdp, the 2001 gross national product per person in US dollars, and Fertility, the birth rate per 1000 femals in the population in the year 2000. The data are for 184 localities, mostly UN member countries, but also other areas such as Hong Kong that are not independent countries. In this problem, we study the relationship between Fertility and PPgdp.

Data visualization and pre-processing.

1. Draw the scatterplot of Fertility on the vertical axis versus PPgdp on the horizontal axis and summarize the information in this graph. Does a simple linear regression model seem to be a plausible for a summary of this graph?

2. In order to get a better fit, we seek to transform the variables. What transformations you would take so that a simple linear regression model is proper? State why you choose these transformations. Draw the scatter plot of the transformed variables. Comment on the plot.

Model fitting and diagnostic.

3. Fit the simple linear model on the transformed data through three ways. Report the least square estimates for the coefficients and R^2. Add the fitted line to the scatter plot on the transformed data and comment on the fit.

   (a) Plain coding (not using the 'lm' function or matrix manipulation)

   (b) Using the 'lm' function

   (c) Through matrix manipulation

4. Draw the diagnostic plots and comment.

Making inferences based on the model.

5. Test whether there is a linear relationship between the transformed variables at 0.05 significance level.

6. Provide a 99% confidence interval on the expected Fertility for a region with PPgdp 20,000 US dollars in 2001.

7. Provide a 95% confidence band for the relation between the expected Fertility and PPgdp. Add the bands to the scatter plot of the original data.

8. Assuming that the same relationship between Fertility and PPgdp holds, give a 99% prediction interval on Fertility for a region with PPgdp 25,000 US dollars in 2018[1].

9. Based on the diagnostic plots in Part 4, do you have any concern on the above hypothesis testing and inferences? If so, what are the concerns?

[1] In reality, we would need to consider inflation, but we simplify the problem here.