Class Project (Due Date: December 2 at 11:59 PM

Problem Description

You are working for a statistical consulting firm and your clients are two towns in New York State that want you to estimate their expenditures for the year 2030. The file `ny.csv` contains data on 914 towns in New York. The variables are the following:

Variable	Description
`expend`	The expenditure per person in the town.
`wealth`	Wealth per person in the town in terms of real estate holding.
`pop`	The town's population.
`pgs`	Percentage of revenue from state and federal grants/subsidies.
`density`	The population per area of the town.
`income`	Mean income per person in the town.

Your main objective is to construct a multiple linear regression model to predict the expenditures for the two towns using 95% prediction intervals.

Modeling Strategy

This is a real data set that will require you to expand upon the techniques we have used in class. Some additional ones you should use include the following:

- You should make histograms of the initial predictor variables. If they are extremely right-skewed then a transformation such as the log (natural log) or square root transformation should be used. The function `mutate` will allow you to do this or you can use the $I()$ function to create these variables when fitting the linear model with the `lm` function.
- You should make a histogram of the response variable. If it is extremely right-skewed then a transformation such as the log (natural log) or square root transformation should be used. The function `mutate` will allow you to do this.
- The function `hist.data.frame` in the `Hmisc` library will create histograms for all the variables in the data set at one time.
- To obtain an appropriate model it may be necessary to include higher polynomial terms for the predictor variables. You should try using third degree polynomials for each predictor variable/transformed variable.
- There will be lots of predictor variables that are all not needed in the final model. You should employ a model building technique called stepwise forward regression that I will discuss in a video.

Important: There is not a correct answer for the model. There are several possibilities that are plausible.

Prediction Intervals

The towns in question have the following projected data for the year 2030.

Town	Population	Wealth	PGS	Density	Income
Warwick	310333	89000	26	325	20000
Tuxedo	292246	115000	7	656	25000

You should make 95% prediction intervals for the response variable based on these projections. If you make a log transformation on Y or a square root transformation on Y then it will be necessary to convert that interval into its original variable (`expend`).

- If your prediction interval is $(3, 6)$ and you make a log transformation then it may seem that $(e^3,\ e^6)$ is the correct way to convert it into the original variable. This is incorrect. If we assume that the estimate of the variance is $\hat{\sigma^2}$ then the correct transformation is

  $$(e^{3+\hat{\sigma}^2/2},\ e^{6+\hat{\sigma}^2/2}).$$

  In general if the prediction interval for $\log(Y)$ is $(a,\ b)$ then the transformed interval is $(e^{a+\hat{\sigma}^2/2},\ e^{b+\hat{\sigma}^2/2}))$.

- If your prediction interval is $(3, 6)$ and you make a square root transformation then it may seem that $(3^2 = 9,\ 6^2 = 36)$ is the correct way to convert it into the original variable. This is incorrect. If we assume that the estimate of the variance is $\hat{\sigma^2}$ then the correct transformation is

  $$(3^2 + \hat{\sigma}^2,\ 6^2 + \hat{\sigma}^2).$$

  In general if the prediction interval for $\sqrt{Y}$ is $(a,\ b)$ then the transformed interval is $(a^2 + \hat{\sigma}^2,\ b^2 + \hat{\sigma}^2)$.

Written Report

A very important part of conducting a statistical analysis is your ability to communicate the results to a person that may know very little about statistics. As part of the project you will need to write a report on the model and its findings. It will include the following:

- The beginning of the report should include a title, the date, and your name. The title should be concise and be in terms of the project.
- An executive summary that describes the results of the study in one or two sentences.
- An introduction section that starts with a statement that describes the overall goal of the study. It will also include the following:
  - A subsection that describes the design of the study.
  - A subsection that describes the variables in the study.

- A methodology section that describes the statistical analysis techniques used to analyze the data. This should include the following: exploratory analysis, linear modeling, stepwise forward regression, and prediction intervals.
- A results section that provides the major results from the statistical analysis. Tables and graphics should be placed in an Appendix and referenced in this section.
- A conclusions section that provides the interpretation of the results in terms of the problem (predicting expenditures for the towns in 2030).
- An appendix that has all of the tables and graphics. It should not include your R code.