STOR 557: Fall 2025: Homework 1

Due date: Thursday, August 28, 2025

This is an exercise in standard linear regression, designed to test your knowledge and recall of STOR 455 (or if anyone here didn't take that course, whatever previous course you have had in linear regression). Like many exercises in statistics, there is no single "right answer": the homework will be graded based on how logically you lay out your analysis, and your ability to describe what you did in plain English. The prerequisites for this exercise are a basic knowledge of R, including in particular the "lm" function, and a knowledge of linear regression itself, including such aspects as variable selection and transformation of variables.

The dataset is from the 1970s, and is an assessment of the cost of building a nuclear power plant at a time when many Americans believed the future of power generation was nuclear. In R, upload the data using the command

nukes=
read.table('https://rls.sites.oasis.unc.edu/faculty/rs/source/Data/nukes.dat',header=T)

This creates a data matrix "nukes" with 32 rows corresponding to 32 power plants, and 12 variables, one of which ("num") you will ignore because it is just a counting variable. The variables of interest to us are given in the following table:

C	Cost in dollars $\times 10^{-6}$, adjusted to 1976 base
D	Date construction permit issued
T_1	Time between application for and issue of permit
T_2	Time between issue of operating license and construction permit
S	Power plant net capacity (MWe)
PR	Prior existence of an LWR on same site $(= 1)$
NE	Plant constructed in north-east region of USA $(= 1)$
CT	Use of cooling tower $(= 1)$
BW	Nuclear steam supply system manufactured by Babcock-Wilcox $(= 1)$
N	Number of power plants constructed by architect/engineer
PT	Partial turnkey plant $(= 1)$

The objective is to predict cost (C) as a function of the other ten variables. However, it is to be expected that not all ten variables will be statistically significant and some subset will likely give the best predictions. Therefore, variable selection will be an important part of your answer.

Part 1 (10 points): For an initial analysis, assume that each of the variables C, T1, T2, S and N is replaced by its natural logarithm, denoted LC, LT1, etc. Then, find the best regression model for predicting LC as a function of the other ten variables. You may use any standard method(s) for variable selection, but be sure to explain what you did and why.

Part 2 (5 points): Suppose a new nuclear plant is to be built according to the following specs: D=68.92, T1=12, T2=65, S=800, PR=0, NE=0, CT=0, BW=0, N=13, PT=0. Obtain (a) a 95% confidence interval, and (b) a 95% prediction interval, for the cost of the new plant, and explain the distinction between the two types of interval.

Part 3 (5 points): Are the transformations made in part 1 appropriate? Write a few sentences to describe verbally why such transformations might be appropriate (or not, as the case may be). If you have time, try a few alternative transformations (e.g. identity, square root) and compare them with the logarithmic transformations. However, I don't want you to consider a very large number of alternative analyses: a few alternatives, just to show how you would make such comparisons, will suffice. (Also, you should be aware that the standard theory of linear regression treats such transformations differently for the response and the explanatory variables. In this context, C is the response and all the other variables are explanatory.)

Your answer may be written in R Markdown or any alternative such as latex or a Word document (however, R Markdown is popular with students in this course and fully acceptable to me – however, please convert to a pdf file for submission). Your answer should not be excessively long (suggest 5 pages, definitely not more than 10). If you use R Markdown, I strongly suggest that you delete output produced by R Markdown that is not relevant to your answer (e.g. long tables of numerical output: just include the parts that are relevant to your answer). Submission will be through gradescope and I will make sure the submission page is open well before the due date.