Homework 4: Introduction to Statistical Computing Using R

STAT 42100: Modern Statistical Modeling Using R and SAS
STAT 52100: Statistical Computing

Instruction: Collaboration on homework assignments is acceptable, but all write-ups must be done independently and clearly indicate the submitter's understanding of the material. Unclear or disorganized homework may have points removed, even if the content is correct. **No hardcopy of the typed report needs to be turned in. The pdf file should be submitted electronically on canvas.** An assignment handed in after the deadline is late, and may or may not be accepted. My solutions to the assignment questions will be available when everyone has handed in their assignment.

Edit the program(s) and output together into a single document, showing the lines of code and relevant output produced by SAS/R. Your answers must be easy for the grader to find. A simple structure is, for each part of each question in order, to put these three things:

- Your code
- Your output
- Your answers and explanation

For SAS, the code and output are naturally separate, so this order is good; for R, the code and output can be intermingled, and that is OK. If your assignment is disorganized or otherwise difficult for the grader to deal with, you can expect to lose marks.

You are reminded that work handed in with your name on it must be entirely your own work. It is as if you have signed your name under it. If it was done wholly or partly by someone else, you have committed an academic offence, and you can expect to be asked to explain yourself. The same applies if you allow someone else to copy your work. The graders will be watching out for assignments that look suspiciously similar to each other (or to my solutions). Besides which, if you do not do your own assignments, you will do badly on the exams.

Problem 1: Permutation The `driving.txt` data set on canvas is from a study which examined the driving habits of illegal drug users as compared to non-illegal drug users. The outcome we will look at is following distance. It may be hypothesized that drug users like to engage in risky behavior and follow at closer speeds than other drivers.

1. Test the null hypothesis that the mean following distance of drug users is the same as that of non-illegal drug users using a t-test.
2. Test the same null hypothesis as above using a permutation test with test statistic:
$$T = \left| \frac{\sum_{i=1}^{n} g_i x_i}{\sum_{i=1}^{n} g_i} - \frac{\sum_{i=1}^{n} (1-g_i) x_i}{\sum_{i=1}^{n} (1-g_i)} \right|$$
where g_i is a 0-1 indicator of group membership.
3. Test the same null hypothesis using a test statistic that compares the absolute difference of medians of the two groups.
4. Briefly comment on the results of the three tests.

Problem 2: Bootstrapping Design a simulation experiment as follows, specifying the details of your parameter set as you go along. Set up a program to sample from the mixture of normals distribution

$$Y = pN(0, \sigma_1^2) + (1-p)N(0, \sigma_2^2)$$

where p is a probability to draw Y from $N(0, \sigma_1^2)$, assumed to be close to 1 and $\sigma_1 << \sigma_2$, so that the distribution has a small probability of producing a large draw. Repeat the following steps for $p = .99$ and $p = .95$ with $\sigma_2/\sigma_1 = 5$:

- Draw a large sample from Y.
- Estimate the mean of Y under the assumption that Y is normal (i.e., not a mixture distribution).
- Write down a (i) 95% confidence interval for the mean, (ii) the standard error for the estimate of the mean, both based on the assumption that Y is normal.
- Create a bootstrap estimate for the 95% confidence interval and standard error. Compare the result to those arrived at assuming that Y was normal.

What are your general conclusions about the impact of contamination on the standard error?

Problem 3: Optimization Let X be the observed number of responders out of n patients entered on a phase II cancer clinical trial. Suppose $X \sim$

Binomial$(n; p)$. Having observed $X = r$, the 'exact' $1 - \alpha$ upper confidence limit on p is defined as the value p_u satisfying

$$\alpha = \mathbb{P}(X \leq r|p_u) = \sum_{i=0}^{r} \binom{n}{i} p_u^i (1 - p_u)^{n-i},$$

and the $1-\alpha$ exact lower confidence limit is defined as the value p_l of p satisfying

$$\alpha = \mathbb{P}(X \geq r|p_l) = 1 - \mathbb{P}(X < r|p_l).$$

We also define $p_l = 0$ if $r = 0$ and $p_u = 1$ if $r = n$. Please use `uniroot()` function to calculate the lower an upper confidence limit for p for any given n and r.

Problem 4: MLE for Poisson Regression A random variable Y is said to have a Poisson distribution with parameter $\mu > 0$ if it takes integer values $y = 0, 1, 2, \cdots$ with probability

$$\mathbb{P}(Y = y) = \frac{e^{-\mu}\mu^y}{y!}.$$

The mean and variance of this distribution can be shown to be $\mathbb{E}(Y) = \text{var}(Y) = \mu$. Use Ornstein's data to model the logarithm (natural base) of the expected number of interlocking directorates (interlocks) as a linear function of level of "assets". You can load this data set from `library(car)` by `data(Ornstein)`. Design the log-likelihood function for this Poisson count model. Optimize your log-likelihood function over the coefficients of your inputs.