STAT/BIOSTAT 571: Homework 1

To be handed in on Weds January 13th, in class. Please see 'Chapter 0' of the slides for a summary of how to answer questions appropriately, and the guidelines from 570. Where solutions require use of `R`, summarize your findings in a written answer, and append your **annotated** code, to show what you did. For each question, write up your solution on your own, using **full sentences**.

1. [**Review – Sandwich estimates**] Consider exercise 2.5 from Jon's book, which implements exponential regression for data on 15 rats.

   (a) With your own coding (i.e. not using any R packages) implement sandwich estimates for the variance of the MLEs described in this question, and compare the corresponding 95% confidence intervals with the likelihood-based approaches in the question.

   (b) Using `family=Gamma`, and negating both the intercept and other covariates, it is possible to get `glm()` to fit this regression. Using this approach, does R's `sandwich` package (using the `HC0` variety of sandwiches) match what you got in part a), or not? [Hint: run some checks of your coding in part a), so that any discrepancy with `sandwich` coding is not due just to coding error]

2. [**Review – MLEs with many parameters**] This question considers behavior similar to the Neyman-Scott problem that you have seen before. Here, the setting is the commonly-used 'pair-matched' study design, where we collect outcomes Y on pairs of people whose covariate values X differ, for some X of interest, but whose other covariates values (age, sex, etc) are identical, or negligibly different.

   The simplest pair-matched design has binary Y and binary X. A commonly-used model for such data assumes that

$$\begin{aligned} X_{ij} &= j \\ Y_{ij}|X_{ij} = x_{ij} &\overset{ind}{\sim} Bern(p_{ij}) \\ p_{ij} &= \text{expit}(a_i + \beta x_{ij}), \end{aligned}$$

   for $j = 0, 1$ and $1 \leq i \leq n$. The parameter β is therefore a log-odds ratio describing the association between Y and X, which is assumed identical across all pairs. The parameters a_i are pair-specific intercept parameters; differences between the a_i account for the differences in the mean outcome due to the other covariates, that were used in the matching process.

   (a) Give a formula for the MLE for β. Hint: note that each vector outcome $\{Y_{i0}, Y_{i1}\}$ has only four different possible values. Additionally, first try to maximize with respect to all the a_i; when can you do this without using calculus?

   (b) For what value is this MLE $\hat{\beta}$ consistent? (i.e. what is the limiting value of $\hat{\beta}$?)

   (c) Using as little technical language as possible (or a diagram) explain *why* the asymptotic bias in $\hat{\beta}$ is in the direction you found. [Note: this part is hard!]

3. [**Motivation – Impact of ignoring correlation**] First, review the Gauss-Markov theorem (from e.g. 533). In this question we will assume that outcomes are multivariate Normal, specifically that

$$\mathbf{Y_{n\times 1}}|\mathbf{x}_{n\times p} \sim N(\mathbf{x}_{n\times p}\boldsymbol{\beta}_{p\times 1}, \boldsymbol{\Sigma}_{n\times n})$$

   where the subscripts denote the dimensions of the vector and matrix quantities.

(a) Assuming that $\mathbf{x}$ and Σ are of full rank, give a formula for the OLS estimate of $\boldsymbol{\beta}$, and its covariance.

(b) In the situation where $\Sigma = \sigma^2 \mathbf{R}$ for scalar σ^2 and correlation matrix $\mathbf{R}$, give a formula for the expectation of

$$S^2 = \frac{1}{n-p} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})^2,$$

i.e. the usual estimate of σ^2 when the Y_i are uncorrelated and errors are homoskedastic.

(c) Using both results above, describe why confidence intervals which (correctly) assume homoskedasticity can still give invalid coverage when outcomes are correlated.

(d) Give examples of your answer from c) using the data from R's built-in `cars` dataset for simple linear regression. This data has $n = 50$ and simple linear regression has $p = 2$ (i.e. intercept and slope). Using `dist` (stopping distance) as the covariate, you should generate outcomes using values of σ and $\boldsymbol{\beta}$ of your choice, and using the correlation matrix $\mathbf{R}$ where $\mathbf{R}_{ij} = \rho^{|i-j|}$ for a value $\rho \in (-1, 1)$ you should also choose. For what values of ρ does the coverage of nominal 95% intervals seriously concern you? When is it exactly 95%? How do the values of σ and $\boldsymbol{\beta}$ affect the results?

4. [**Motivation – Exploring vector outcomes**] The class site contains annual data on self-reported health, from a sample of 1,798 people followed over 31 years. The responses are coded from 5 to 0, representing Excellent, Very Good, Good, Fair, Poor and Dead. (Naturally, Dead is not truly 'self-report' status, nor does anyone report feeling better after being reported Dead). There are many analyses one can do with this data, but ideally the graphs we draw of the data should reflect the analysis of interest. Explaining your choice of graphing method, make plots of the data that illustrate;

(a) The proportion of people moving between pairs of health states, over time

(b) The absolute number of people in different health states, over time

(c) The variety and relative frequencies of different trajectories taken by different participants, over the entire course of the study

Note: there is no uniquely 'correct' answers to any of these, and you are free to use any graphing techniques you think are helpful. In your explanation of what *you* did, briefly describe strengths and weaknesses in your plotting methods.