Homework Assignment 1

STAT 544

Due on February 12, 2021

1. Suppose you standardize (i.e., subtract the sample means and then divide by the sample standard deviations) the quantitative explanatory variables in a regression model. Why might it be desirable to do this? Explain how to interpret the resulting standardized regression coefficients.

2. Recall that for the OLM we typically use the variance estimator

$$\hat{\sigma}^2 = \boldsymbol{e}'\boldsymbol{e}/(n-p),$$

   where $\boldsymbol{e} = (e_1, \ldots, e_n)'$ are the residuals, n is the sample size, and $p = \dim(\boldsymbol{\beta})$. Compare and contrast this estimator with the maximum likelihood estimator

$$\hat{\sigma}^2_{\text{mle}} = \boldsymbol{e}'\boldsymbol{e}/n.$$

   Why do we usually use the first estimator instead of the MLE?

   Also, the commonly used notation $\hat{\sigma}^2$ is not quite correct. Explain.

3. On our course website you will find dataset `anorexia.dat`. These data were collected for 72 young girls who were afflicted with anorexia. The girls were randomly assigned to receive one of three types of therapy during the study period: cognitive behavioral therapy ('b'), family therapy ('f'), or control ('c'). Let the response be the change in weight during the study, and fit an ordinary linear model to the data. Use the control group as the baseline. Perform a complete analysis, including model assessment, and present your results concisely.

4. Analyze the anorexia data once more, this time fitting an ANCOVA model. Which model is better, the ANOVA model or the ANCOVA model? Justify your answer.

5. Write an R function called `rcat` that produces pseudorandom draws from a categorical distribution. The function should take exactly two arguments: `n`, the number of values to return, and `probs`, a vector of k probabilities. You may assume that the elements of `probs` sum to 1.

6. Recall that for any GLM that employs the canonical link function, the likelihood equations are $\mathbf{X}'(\boldsymbol{Y} - \boldsymbol{\mu}) = \mathbf{0}$, where $\mathbf{X}_{n\times p}$ is the model matrix,

$\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ are the observations, and $\boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta}) = \mathbb{E}\boldsymbol{Y}$. State and prove a central limit theorem for $\mathbf{X}'\boldsymbol{Y}$, the sufficient statistic for the regression coefficients $\boldsymbol{\beta}$.

7. Code a simple simulation study to confirm that, for a sufficiently large sample size, $\mathbf{X}'\boldsymbol{Y}$ is approximately Gaussian distributed. Consider only one (discrete) member of the exponential family.