Please provide complete and well-written solutions to the following exercises.

Due November 3, 9AM, to be submitted in blackboard, under the Assignments tab.

# Homework 5

**Exercise 1.** Let $n > m$ be integers. Let $A$ be an $n \times m$ real matrix of known (deterministic) constants. Let $\beta \in \mathbf{R}^m$ be an unknown vector of (deterministic) constants. And let $\varepsilon \in \mathbf{R}^n$ be a random vector with $\mathbf{E}\varepsilon = 0$ and such that $\varepsilon$ is a vector of i.i.d. random variables. Define $Y \in \mathbf{R}^n$ by $Y = A\beta + \varepsilon$. Assume that $A^T A$ is invertible. Define $Z := (A^T A)^{-1} A^T Y$.

Show that the estimator

$$\Big(\frac{1}{n-m}\sum_{i=1}^{n}(Y_i - (AZ)_i)^2\Big)(A^T A)^{-1}$$

is an unbiased estimator of the covariance matrix of $Z := (A^T A)^{-1} A^T Y$.

**Exercise 2.** Assume the one-way ANOVA assumptions. Consider the null hypothesis $H_0$ that $\beta_1 = \cdots = \beta_p$. Recall that, under this assumption, the $F$ statistic takes the form

$$F = \frac{1}{S^2}\sum_{j=1}^{p} n_i(\overline{Y_j} - \overline{Y})^2.$$

The alternative hypothesis $H_1$ is that $\beta_i \neq \beta_j$ for some $1 \leq i < j \leq p$. We can therefore reject $H_0$ when $F$ is large.

Show that the generalized likelihood ratio test of $H_0$ versus $H_1$ coincides with the hypothesis test we just described. (The likelihood function should just use the Gaussian assumptions for the random variables $Y_1, Y_2, \ldots$.) (Also, you should assume that $\sigma > 0$ is unknown.) (When you form the generalized likelihood ratio, the exponential terms from the Gaussian distribution should eventually become constants.)

**Exercise 3.** In statistics and other applications, we can be presented with data points $(x_1, y_1), \ldots, (x_n, y_n)$. We would like to find the line $y = mx + b$ which lies "closest" to all of these data points. Such a line is known as a linear regression. There are many ways to define the "closest" such line. The standard method is to use least squares minimization. A line which lies close to all of the data points should make the quantities $(y_i - mx_i - b)$ all very small. We would like to find numbers $m, b$ such that the following quantity is minimized:

$$f(m, b) = \sum_{i=1}^{n}(y_i - mx_i - b)^2.$$

Using the second derivative test, show that the minimum value of $f$ is achieved when

$$m = \frac{(\sum_{i=1}^{n} x_i)\left(\sum_{j=1}^{n} y_j\right) - n\left(\sum_{k=1}^{n} x_k y_k\right)}{(\sum_{i=1}^{n} x_i)^2 - n\left(\sum_{j=1}^{n} x_j^2\right)} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{j=1}^{n}(x_j - \overline{x})^2}.$$

$$b = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - m\sum_{j=1}^{n} x_j\right) = \overline{y} - m\overline{x}.$$

Briefly explain why this is actually the minimum value of $f(m, b)$. (You are allowed to use the inequality $(\sum_{i=1}^{n} x_i)^2 \leq n(\sum_{i=1}^{n} x_i^2)$.)

**Exercise 4.** Let

$$h(x) := \frac{1}{1 + e^{-x}}, \qquad \forall\, x \in \mathbf{R}.$$

Fix $x \in \mathbf{R}$ and $y \in [0, 1]$. Define $t\colon \mathbf{R}^2 \to \mathbf{R}$ by

$$t(a, b) := \log\Big([h(ax + b)]^y[1 - h(ax + b)]^{1-y}\Big), \qquad \forall\, a, b \in \mathbf{R}.$$

Show that $t$ is concave. Conclude that $t$ has at most one global maximum.