Statistical Modeling and Methods: Homework 2

Due by 5pm on April 3, online through Blackboard

Homework format: all homework must be written in latex. You must turn in both your tex and pdf files. Attach your code and computer output if there is any programming.

1. Let X_i be i.i.d. from $N(\mu_i, 1)$, $i = 1, \dots, n$. Find the explicit distributions of $\bar{X}_i = n^{-1}\sum_{i=1}^n X_i$ and $\sum_{i=1}^n (X_i - \bar{X})^2$, and show that they are independent of each other.

2. Let $X \sim N_p(\mu, \Sigma)$, where $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, Σ_{11} is $r \times r$, $r \le p$. Let $X = (X_1', X_2')'$ with X_1 being $r \times 1$.

   (a) Show that X_1 and X_2 are independent if and only if $\Sigma_{12} = 0$.

   (b) Obtain the explicit conditional distribution of X_1 given $X_2 = x_2$.

3. (a) Let $X_1, \dots, X_n$ be i.i.d. $N(0, 1)$, for any nonzero $a \in \Re^n$, find the conditional distribution of $\sum_{i=1}^n X_i^2$ given $\sum_{i=1}^n a_i X_i = 0$.

   (b) Show that the range $R_n = X_{(n)} - X_{(1)}$ is independent of the sample mean $\bar{X}_n$, where $X_{(i)}$ is the ith order statistic.

4. Consider a linear model with p parameters, fit by ordinary least squares to a set of training data $(x_1, y_1), \dots, (x_n, y_n)$ with the OLS estimate $\hat{\beta}_{OLS}$. Suppose we have some test data $(\tilde{x}_1, \tilde{y}_1), \dots, (\tilde{x}_m, \tilde{y}_m)$ drawn at random from the same population as the training data. Denote $R_{tr}(\beta) = n^{-1}\sum_{i=1}^n (y_i - \beta^T x_i)^2$ and $R_{te}(\beta) = m^{-1}\sum_{i=1}^m (\tilde{y}_i - \beta^T \tilde{x}_i)^2$, show that $E\{R_{tr}(\hat{\beta}_{OLS})\} \le E\{R_{te}(\hat{\beta}_{OLS})\}$, where the expectations are taken over all random quantities (including x_i's and $\tilde{x}_i$'s).

5. Consider the linear regression model $Y = X\beta + \epsilon$ with $E(\epsilon) = 0$ and $\mathrm{var}(\epsilon) = \sigma^2 I_n$, where X is $n \times p$, β is $p \times 1$, and I_n is the $n \times n$ identity matrix.

   (a) Suppose that a linear function $l'Y$ is an unbiased estimator of $p'\beta$, show that $l'Y$ is BLUE (best linear unbiased estimator) for $p'\beta$ if and only if $\mathrm{cov}(l'Y, m'y) = 0$ for all m such that $E(m'Y) = 0$.

(b) Find p such that $p'\beta$ is estimable and $\text{var}(p'\hat{\beta})/\|p\|^2$ is minimum (or maximum), where $\hat{\beta}$ is a solution to the normal equation.

6. Consider the following analysis of covariance model in which m_1 experimental animals are fed with diet 1, m_2 animals with diet 2, and so on. The initial weights x_{ij}'s and weights y_{ij}'s after 10 days on the experimental animals were recorded. The model below is fit for this dataset

$$y_{ij} = \mu + rx_{ij} + \epsilon_{ij}, \quad j = 1, \ldots, m_i, i = 1, \ldots, k,$$

where (x_{ij}, y_{ij}) are the weights of the jth animal on the i diet.

(a) Obtain explicitly the least squares estimates of μ_i's and r.

(b) Assume that ϵ_{ij}'s are independent with $E(\epsilon_{ij}) = 0$ and $\text{var}(\epsilon_{ij}) = \sigma^2$, obtain the variance of the LS estimates of μ_i's and r.

7. Let $A =$ be a $p \times p$ positive definite matrix.

(a) Show that $a_{ii} > 0$, where a_{ij} denotes the element on the ith row and jth column.

(b) Show that $a_{ii}a^{ii} \geq 1$ with equality holding when and only when $a_{ij} = 0$ for all $i \neq j$, where a^{ij} are the element on the ith row and j column of A^{-1}.

(c) Show that $a^{ii}a_{ii} = 1$ for all i implies that $a_{ij} = 0$ for all $i \neq j$.

8. Use the facts in Question 7 to answer the following problem of weighing design. Suppose that there are m objects whose individual weights have to be determined. One method is to weigh each object r times and take the average value as the estimate of its weight. This procedure needs a total of mr weighings and the precision of each estimated weight is σ^2/r, where σ^2 is the error variance of an individual observation. Another method is to weigh the objects in combinations. Each operation consists in placing some of the objects in one balanced pan and others in the other pan, and placing weights to achieve equilibrium. This method results in an observational equation of the type

$$y = x_1w_1 + \ldots + x_mw_m + \epsilon,$$

where $w_1, \ldots, w_m$ are the hypothetical weights of the objects, $x_i = 0, 1$ or -1 corresponds to the situations that the ith object is not used, placed in the left pan or in the right pan, and y is the weight required for equilibrium. The n operations described as using different combinations of the objects for the left and right pans yield n observations equations, from which the unknown weights may be estimated by the least squares. Here the $n \times n$ design matrix of the observational equations have entries $-1, 0$ or 1. The problem of interest is to choose the design matrix in such a way that the precision of the individual estimates is as accurate as possible.

(a) Show that $\mathrm{var}(\hat{w}_j) \geq \sigma^2/n$, where $\hat{w}_j$'s are the LS estimates, $j = 1, \ldots, m$.

(b) Show that the maximum (or minimum) variance is achieved for all the m estimates if and only if all the entries are -1 or 1, and the columns of X are orthogonal.

(c) Show that, to achieve the best precision of σ^2/n by weighing the objects individually (i.e., method 1), the number of weighting operation is mn.

9. Consider a one-way ANOVA model:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad j = 1, \ldots, m_i, \; i = 1, \ldots, k, \; \epsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

Suppose that you want to test $H_0 : \alpha_1 = \ldots = \alpha_\ell, \alpha_{\ell+1} = \ldots = \alpha_k$, for some $\ell < k$. This hypothesis is meaningful if you suppose that the true means of the k groups are in two clusters.

(a) Express the hypothesis in the form of $H'\beta = \xi$, and show that the column spaces satisfy $\mathrm{C}(H) \subseteq C(X')$.

(b) Find explicitly an appropriate statistic for testing H_0.

(c) Obtain the noncentrality parameter associated with this testing problem, and check that the noncentrality parameter is zero if and only if H_0 is true.