00101756: Modern Statistical Modeling
Homework 1 *Due*: May 9, 2023

1. Let $\mathbf{X}$ be an $n \times p$ matrix of rank r, and $m = \min(n, p)$. For any $k < r$, denote

$$\mathbf{X}_k = \sum_{j=1}^{k} d_j \mathbf{u}_j \mathbf{v}_j^T,$$

where $d_1 \geq d_2 \geq \cdots \geq d_m \geq 0$ are the singular values of $\mathbf{X}$, and $\mathbf{u}_j$ and $\mathbf{v}_j$ the left and right singular vectors corresponding to d_j, respectively.

   (a) Show that $\mathbf{X}_k$ is the best rank-k approximation to $\mathbf{X}$ in the sense that

$$\min_{\text{rank}(\mathbf{Y})=k} \|\mathbf{X} - \mathbf{Y}\|_2 = \|\mathbf{X} - \mathbf{X}_k\|_2 \quad (= d_{k+1}).$$

   (b) Show that the statement is still true if the matrix 2-norm is replaced by the Frobenius norm. Give an expression for $\|\mathbf{X} - \mathbf{X}_k\|_F$.

2. Let $X_1, \ldots, X_n$ be a random sample from the uniform distribution over the unit ℓ_2-ball in $\mathbb{R}^p$.

   (a) Find the median distance M from the origin to the closest data point. What are the values of M for a sample of size 10^6 and $p = 1, \ldots, 15$?

   (b) Find the mean distance D from the origin to the closest data point. What are the values of D for a sample of size 10^6 and $p = 1, \ldots, 15$?

3. Prove Theorem 19.5 in UML by following Exercises 19.1–19.4.

4. Consider a linear model with p parameters, fit to a training sample $(x_1, y_1), \ldots, (x_n, y_n)$ with the OLS estimate $\widehat{\boldsymbol{\beta}}$. Suppose we have a test sample $(x_{n+1}, y_{n+1}), \ldots, (x_{n+m}, y_{n+m})$ from the same population. Denote $R_{\text{tr}}(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2$ and $R_{\text{te}}(\boldsymbol{\beta}) = m^{-1} \sum_{i=1}^{m} (y_{n+i} - \boldsymbol{\beta}^T \mathbf{x}_{n+i})^2$. Show that

$$E\{R_{\text{tr}}(\widehat{\boldsymbol{\beta}})\} \leq E\{R_{\text{te}}(\widehat{\boldsymbol{\beta}})\},$$

where the expectations are taken over all (x_i, y_i).

5. Consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $E\boldsymbol{\varepsilon} = \mathbf{0}$ and $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$.

   (a) Show that a linear estimator $\mathbf{c}^T \mathbf{y}$ is the BLUE of $\mathbf{h}^T \boldsymbol{\beta}$ for some $\mathbf{h}$ if and only if $\text{Cov}(\mathbf{c}^T \mathbf{y}, \mathbf{d}^T \mathbf{y}) = 0$ for all $\mathbf{d}$ with $E(\mathbf{d}^T \mathbf{y}) = 0$.

   (b) Find $\mathbf{h}$ such that $\mathbf{h}^T \boldsymbol{\beta}$ is estimable and $\text{Var}(\mathbf{h}^T \widehat{\boldsymbol{\beta}}) / \|\mathbf{h}\|_2^2$ is minimized or maximized, where $\widehat{\boldsymbol{\beta}}$ is a solution to the normal equation.

6. Consider the simple linear model $y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1, \ldots, n$, where x_i are fixed and ε_i are i.i.d. with $E\varepsilon_i = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$. Find a sufficient condition using the Lindeberg–Feller central limit theorem such that the OLS estimate $\widehat{\beta}$ is asymptotically normal. Give the normalized form of $\widehat{\beta}$ and its limiting distribution.

7. Consider the ANCOVA model

$$y_{ij} = \mu + \alpha_i + \gamma x_{ij} + \varepsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, k,$$

where ε_{ij} are i.i.d. $N(0, \sigma^2)$.

(a) Suppose you want to test $H_0 : \alpha_1 = \cdots = \alpha_k$. Express H_0 in the form of $\mathbf{H}^T \boldsymbol{\beta} = \boldsymbol{\xi}$ and show that $\mathcal{C}(\mathbf{H}) \subset \mathcal{C}(\mathbf{X}^T)$, where $\mathbf{X} = (x_{ij})$ is the design matrix.

(b) Explicitly obtain a test statistic for testing H_0.

(c) Find the distribution of the test statistic in part (b) when H_0 is true and when H_0 is not true. Check that the noncentrality parameter is zero if and only if H_0 is true.

8. Consider the two-way ANOVA model

$$y_{ij} = \mu + \alpha_i + \gamma_j + \varepsilon_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, m,$$

where ε_{ij} are i.i.d. $N(0, \sigma^2)$.

(a) Find a necessary and sufficient condition for $c_0\mu + \sum_{i=1}^k c_i\alpha_i + \sum_{j=1}^m c_{k+j}\gamma_j$ to be estimable.

(b) Use Scheffé's method to obtain simultaneous confidence intervals of level $1-\alpha$ for $\sum_{i=1}^k c_i\alpha_i$ with $\sum_{i=1}^k c_i = 0$.

(c) Use Tukey's method to obtain simultaneous confidence intervals of level $1-\alpha$ for $\gamma_j - \gamma_{j'}$, $j \neq j'$.