Statistical Modelling and Methods: Homework 3

Due by 5pm on April 30, 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. Consider the ANCOVA model, where x_{ij} are known covariate,

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

   (a) Suppose that you want to test $H_0 : \alpha_1 = \ldots = \alpha_k$. Express this hypothesis in the form of $H'\beta = \xi$ and show that $\text{Col}(H) \subseteq \text{Col}(X')$.

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

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

2. This refers to the same ANCOVA model in Question 1. Suppose that you want to test $H_0 : \gamma = 0$. Repeat parts (b) and (c) as in Question 1.

3. Consider a two-factor ANOVA model with one observations per cell

$$y_{ij} = \mu + \alpha_i + \gamma_j + \epsilon_{ij}, \quad i = 1, \ldots, r, \quad j = 1, \ldots, s, \quad \epsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

   (a) Find a necessary and sufficient condition for $p_0\mu + p_1\alpha_1 + \ldots + p_r\alpha_r + p_{r+1}\gamma_1 + \ldots + p_{r+s}\gamma_s$ to be estimable.

   (b) Use Scheffe's method to obtain simultaneous confidence intervals with level $1-\alpha$ of all contrasts for factor 1, i.e., $\sum_{i=1}^r p_i\alpha_i$ with $\sum_{i=1}^r p_i = 0$.

   (c) Use Tukey's method to obtain simultaneous confidence intervals with level $1-\alpha$ for $\gamma_j - \gamma_{j'}$, $j \neq j'$ (assume that the "studentized" distribution of Q is known).

4. Consider the following simplified linear model selection: $Y = X\beta + \epsilon$, $\epsilon \sim N_n(0, \sigma I_n)$ and $X'X = nI_m$, where I_k is the $k \times k$ identify matrix. Also assume that $\beta_j \neq 0$ for $j = 1, \ldots, p*$ and $\beta_j = 0$, $j = 1, \ldots, m$. Design matrices for the pth model to be considered sequentially consist of the first p columns of X. If one uses the mean square error, $MSE_p = SSE_p/(n-p)$

as a model selection criterion, where SSE_p is the residual sum of squares of the pth model, show that the MSE_p is *not* an appropriate model selection criterion quantifying the following probabilities, as $n \to \infty$: $P(MSE_{p*} > MSE_{p*-1})$ and $P(MSE_{p*} > MSE_{p*+1})$.

5. Consider a simple linear model $Y_i = \alpha + \beta x_i + \epsilon_i, i = 1, \ldots, n$, where x_i are known covariate values and ϵ_i are i.i.d. with $E\epsilon_i = 0$ and $E\epsilon_i^2 = \sigma^2$. Find a sufficient condition using Linderberg-Feller's Central Limit Theorem such that the least square estimate $\hat{\beta}$ is asymptotically normal, and give the normalized form of $\hat{\beta}$ and its limit distribution.

6. Consider a linear regression model $y_i = x_i'\beta + \epsilon_i$, where $\beta \in R^p$, ϵ_i are i.i.d. with mean zero and variance σ^2, $i = 1, \ldots, n$. Denote the residuals $r_i = y_i - x_i'\hat{\beta}$, where $\hat{\beta}$ is the solution to the normal equation of the whole sample, and the deleted residuals $r_i^{(-i)} = y_i - x_i\hat{\beta}^{(-i)}$, where $\hat{\beta}^{(-i)}$ is the solution to the normal equation based on the sample with the ith pair $\{x_i, y_i\}$ removed. Show that

$$r_i^{(-i)} = \frac{y_i - x_i'\hat{\beta}}{1 - h_{ii}},$$

where h_{ii} is the ith diagonal element of the hat matrix H.

7. (a) Recall that $E_\theta\{U(\theta, Y)\} = 0$ and $E_\theta\{(\partial/\partial\theta)U(\theta, Y)\} + \text{var}\{U(\theta, Y)\} = 0$, where $U(\theta, Y)$ is the score of a random variable Y that has a density function $f(y; \theta)$. These are usually called the first and second order *Bartlett's identities*. Derive the third order Bartlett's identity.

   (b) Partition a $p \times 1$ parameter vector θ as $\theta = \begin{pmatrix} \psi \\ \lambda \end{pmatrix}$, where ψ with dimension $p_1 < p$ is the parameter of primary interest, and λ with dimension $(p - p_1)$ is the nuisance parameter (*not of interest, but still need to be estimated*). Denote the information matrix of $\hat{\theta}$ obtained from an independent sample $\{Y_1, \ldots, Y_n\}$ by

$$I_n(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix}.$$

   Find the appropriate information matrix of $\hat{\psi}$ obtained from this sample.