STAT5030 Linear Models (Final Exam 2020-2021)

3 May 2021

1. (**15 marks**) Let $\boldsymbol{x} = (X_1, \ldots, X_k)^\top \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where μ is a $k \times 1$ constant vector and $rank(\boldsymbol{\Sigma}) = k$.

   (a) What is the distribution of $U = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$?

   (b) Let $\boldsymbol{A} = \boldsymbol{\Sigma^{-1}} - (\boldsymbol{\Sigma^{-1} 1_k 1_k^\top \Sigma^{-1}})/(\boldsymbol{1_k^\top \Sigma^{-1} 1_k})$. Here $\mathbf{1}_k$ is a $k \times 1$ vector with all elements being 1. Find the distribution of $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$.

2. (**15 marks**) In the one-way ANOVA model

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, 2, 3, \quad j = 1, \ldots, n,$$

   where ϵ_{ij} are independently distributed as $N(0, \sigma^2)$, and $\tau_i's$ are fixed but unknown.

   (a) Consider a hypothesis $H_{01} : \mu + \tau_1 = 2(\mu + \tau_2) = 3(\mu + \tau_3)$. Is H_{01} testable? If yes, derive a test for testing H_{01}.

   (b) Is $H_{02} : \tau_2 = (\tau_1 + \tau_3)/2$ testable? If yes, derive a test for testing H_{02}.

3. (**30 marks**) Consider a linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

   where $\boldsymbol{Y}$ is $n \times 1$, $\boldsymbol{X}$ is an $n \times (p+1)$ fixed design matrix, $\boldsymbol{\beta}$ is a $(p+1)$-vector of regression coefficient and $\boldsymbol{\epsilon}$ has mean $\mathbf{0}$ and known positive definite covariance matrix $\boldsymbol{V}$.

   (a) When $\boldsymbol{X}$ is of full rank, find the best linear unbiased estimates (BLUE) of $\boldsymbol{p}^\top \boldsymbol{\beta}$, where $\boldsymbol{p} \in \mathbb{R}^{p+1}$ is a constant vector. (Students are required to show the detailed proof of the Gauss-Markov theorem.)

(b) When $\boldsymbol{X}$ is not of full rank, find a sufficient and necessary condition for $\boldsymbol{c}^\top\boldsymbol{\beta}$ to be estimable, where $\boldsymbol{c} \in \mathbb{R}^{p+1}$.

(c) For model (5), suppose that $\boldsymbol{\beta}$ is constrained by $\boldsymbol{R}\boldsymbol{\beta} = r$, where $\boldsymbol{R}$ is a full-rank $m \times (p+1)$ matrix with $m < p+1$. Let $\boldsymbol{V} = \sigma^2\boldsymbol{I}$. Derive the constrained least square estimate of $\boldsymbol{\beta}$ and show that it is better than the ordinary least square estimate of $\boldsymbol{\beta}$(without the constraint) in the sense of Gauss-Markov theorem.

4. (**20 marks)** Consider a linear regression model

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots x_{ip}\beta_p + \epsilon_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \quad i = 1, \ldots, n. \tag{2}$$

By convention, the response and covariates are centered and standardized. Model (2) can be written in matrix form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{X}_{n\times p} = (x_1, \ldots, x_p)$ is orthogonal design and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$.

(a) The *non-negative lasso* is defined to minimize

$$\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \text{ subject to } \beta_j \geq 0, j = 1, \ldots, p, \tag{3}$$

over $\boldsymbol{\beta}$. Here $\|\cdot\|_2$ and $\|\cdot\|_1$ are the L_2 norm and the L_1 norm, respectively. Find the solution to the non-negative lasso problem. Explain the difference between lasso and the non-negative lasso by comparing the solutions. (Students are required to provide detailed steps.)

(b) Zou and Hastie (2005) introduced the *elastic-net penalty*

$$\lambda\sum_{j=1}^{p}(\alpha\beta_j^2 + (1-\alpha)|\beta_j|),$$

a different compromise between ridge and lasso with $0 \leq \alpha \leq 1$. Consider the elastic-net optimization problem

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda[\alpha\|\boldsymbol{\beta}\|_2^2 + (1-\alpha)\|\boldsymbol{\beta}\|_1],$$

where $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)^\top$, $\boldsymbol{X} = (x_1, x_2, \ldots, x_p)_{n\times p}$. Show how one can turn this into a lasso problem.

5. (**20 marks)** Consider binary outcome $Y \in \{0, 1\}$, where $Y = 0$ represents the control (non-disease) and $Y = 1$ represents the case (disease). The logistic regression model assumes that Y is associated with the $(p+1)$-dimensional predictor X via the logistic link function

$$P(Y = 1|X = x) = \frac{e^{x^\top \beta}}{1 + e^{x^\top \beta}}, \tag{4}$$

where $\beta \in R^{p+1}$ including an intercept. A latent variable formulation of model (4) is as follows: suppose that there is an unobserved continuous random variable $\tilde{Y}$ such that $Y = 1$ if and only if $\tilde{Y} > \theta$, where θ is some unknown constant.

(a) Show that θ is not identifiable under model (4).

(b) For identifiability, we fix $\theta = 0$ without loss of generality. Assume that the latent $\tilde{Y}$ depends on X via a linear regression model

$$\tilde{Y} = X^\top \beta + U, \tag{5}$$

where U is the error term. Show that when U follows the standard logistic distribution, model (5) is the logistic regression model in (4).

Hint: The density function of the standard logistic distribution is

$$f(x) = \frac{e^x}{(e^x + 1)^2}, \quad x \in R.$$

(c*) (**Optional question: 10 bonus marks**) For the logistic regression model (4), the observations are $(Y_i, X_i), i = 1, \ldots, n$, a random sample of (Y, X) with size n. Let $\hat{\beta}_n$ be the maximum likelihood estimator of β. Derive the asymptotic distribution of $\sqrt{n}(\hat{\beta}_n - \beta_0)$ as $n \to \infty$, where β_0 is the true value of β in model (4). If it is possible, please provide the technical assumptions needed to establish the asymptotic distribution.