STAT5030 Linear Models (Qualifying exam 2019-2020)

1. Consider a linear regression model

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \dots x_{ip}\beta_p + \epsilon_i, \quad i = 1, \dots, n.$$

The ridge regression is to apply squared penalty on the least squares estimate by minimizing

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\lambda \geq 0$ is a tuning parameter. By convention, the response is centered and the covariates are standardized. The error term ϵ has zero mean. The resulting estimate is denoted by $\hat{\boldsymbol{\beta}}^{\text{ridge}}$.

   i. Denote the design matrix by $\boldsymbol{X}_{n\times p} = (x_1, \dots, x_p)$. Derive the explicit expression of $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ in detailed steps.

   ii. Show the details how to compute the ridge solution via the singular value decomposition (SVD).

   iii. Show that there always exists a λ such that the mean squared error (MSE) of $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ is less than the MSE of $\hat{\boldsymbol{\beta}}^{\text{ols}}$, the ordinary least square estimate. (Please provide detailed derivation of each step).

2. In the following, $\boldsymbol{I}_m$ is an $m \times m$ identity matrix, $\boldsymbol{0}_m$ is an $m \times 1$ vector of zero elements, and $\boldsymbol{J}_m = \boldsymbol{1}_m \boldsymbol{1}_m'$, where $\boldsymbol{1}_m$ is an $m \times 1$ vector of 1's. You may use, without proof, the fact that

$$[\boldsymbol{I}_m + \phi \boldsymbol{J}_m]^{-1} = \left[\boldsymbol{I}_m - \frac{\phi}{1 + m\phi} \boldsymbol{J}_m\right].$$

   i. Consider the following linear model:

$$\begin{aligned} Y_{ijt} &= \gamma_i + \tau_j + \epsilon_{ijt}, \\ \epsilon_{ijt} &\sim N(0, \sigma_E^2), \gamma_i \sim N(0, \sigma_\gamma^2), i = 1, 2; j = 1, 2; t = 1, 2; \end{aligned} \tag{1}$$

   where all random variables on the right hand side of the model are mutually independent. Write the model as $\boldsymbol{Y} = \boldsymbol{Z\gamma} + \boldsymbol{X\tau} + \boldsymbol{\epsilon}$, where

$$\boldsymbol{Y} = [Y_{111}, Y_{112}, Y_{121}, Y_{122}, Y_{211}, Y_{212}, Y_{221}, Y_{222}], \boldsymbol{\gamma} = [\gamma_1, \gamma_2], \boldsymbol{\tau} = [\tau_1, \tau_2]$$

   and find $\boldsymbol{Z}, \boldsymbol{X}$. Next, find the variance-covariance matrix of $\boldsymbol{Z\gamma} + \boldsymbol{\epsilon}$.

   ii. State the distribution of $\boldsymbol{Y}$ and find the best linear unbiased estimator of $\boldsymbol{\tau}$ in part (a). Give a condition for $\boldsymbol{C'\tau}$ to be estimable under model (1), where C' is $q \times p$ of rank q (and $q \geq 1$). Justify your answer.

   iii. For given constant vector $\boldsymbol{d}$ and estimable set of functions $\boldsymbol{C'\tau}$, state a test statistic for testing

$$H_0 : \boldsymbol{C'\tau} = \boldsymbol{d} \quad \text{versus} \quad H_1 : \boldsymbol{C'\tau} \neq \boldsymbol{d},$$

   where $\boldsymbol{C'}$ is $q \times p$ of rank q (and $q \geq 1$). Find the expected value of the numerator of the test statistic.

   iv. Let $\phi = \sigma_\gamma^2 / \sigma_E^2$ and let $\boldsymbol{C'} = [1, -1]$. Assuming that the distribution of your test statistic in part(c) is non-central F, does the power of this test depend on the value of σ_γ? If so, in which way?