# 國立臺灣大學數學系
# 九十六學年度博士班入學考試試題
# 科目：迴歸分析

2007.05.04

1. Consider the multiple linear regression model $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$, $i = 1, \cdots, n$, where $\epsilon_i$'s are uncorrelated with zero mean and constant variance $\sigma^2$. Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p)^T$ be the least squares estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$.

(1a) (10%) Let $e_{i(j)}$'s and $d_{ij}$'s denote separately the residuals computed based on the regression models $Y_i = \beta_{0j} + \sum_{l\neq j} \beta_{lj} x_{il} + \eta_i$ and $x_{ij} = \gamma_0 + \sum_{l\neq j} \gamma_l x_{il} + \xi_i$, $i = 1, \cdots, n; j = 1, \cdots, p$. Show that $Var(\widehat{\beta}_j)$ increases as $R_j^2 = 1 - (\sum_{i=1}^n d_{ij}^2)/(\sum_{i=1}^n e_{i(j)}^2)$ increases.

(1b) (10%) Let $R^2$ be the coefficient of determination and $r$ be the sample correlation coefficient of $(Y_1, \widehat{Y}_1), \cdots, (Y_n, \widehat{Y}_n)$, where $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip}, i = 1, \cdots, n$. Show that $r^2 = R^2$.

(1c) (10%) Assume that $\epsilon_1, \cdots, \epsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$. Derive the distribution of the studentized residual of the first observation $r_1^* = e_1/\sqrt{MSE_{(1)}(1 - h_{11})}$, where $e_1$ and $h_{11}$ are separately the residual and the leverage point of the first case and $MSE_{(1)}$ is the mean squared error obtained after deleting the first case.

2. Consider the linear model $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 I_n)$ and $\mathbb{X}$ is a $n \times p$ full rank matrix. Let $\widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{e}$ be separately the maximum likelihood estimator and the residual vector.

(2a) (10%) Show that $\widetilde{\boldsymbol{\beta}}$ and $SSE = \boldsymbol{e}^T\boldsymbol{e}$ are independent.

(2b) (7%) (8%) Construct a $1-\alpha$, $0 < \alpha < 1$, confidence interval for $\boldsymbol{x}_0^T\boldsymbol{\beta}$ and a $1-\alpha$ prediction interval for the future independent run $Y_0 = \boldsymbol{x}_0^T\boldsymbol{\beta} + \epsilon_0$ with $\epsilon_0 \sim N(0, \sigma^2)$.

3. (5%) (5%) Consider the growth curve model $\boldsymbol{Y}_i^{(t)} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i^{(t)}$, $i = 1, \cdots, n$, where $\mathbb{X}$ is a $m \times p$ full rank matrix with $p < m$ and $\boldsymbol{\varepsilon}_1^{(t)}, \cdots, \boldsymbol{\varepsilon}_n^{(t)} \overset{i.i.d}{\sim} (\boldsymbol{0}, \Omega^{(t)})$. Find the estimated generalized least squares estimator of $\boldsymbol{\beta}$ and an unbiased estimator of $\Omega^{(t)}$.

4. (5%) (5%) (5%) Consider the model $y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_p x_{tp} + \varepsilon_t$, $t = 1, \cdots, n$, where $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$ with $u_t$'s $\overset{i.i.d.}{\sim} N(0, \sigma_u^2)$. Find the consistent estimators of $\rho$ and $\sigma_u^2$, and the estimated generalized least squares estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$.

5. Consider the centered and scaled regression model $\boldsymbol{Y} = \tilde{\mathbb{Z}}_{(s)}\boldsymbol{\beta}_{(s)} + \boldsymbol{\varepsilon}$, where $\tilde{\mathbb{Z}}_{(s)} = (\boldsymbol{1}, \mathbb{Z}_{(s)})$ with $\mathbb{Z}_{(s)}$ being a $n \times p$ standardized covariate matrix, $\boldsymbol{\beta}_{(s)} = \begin{pmatrix} \gamma_0 \\ \boldsymbol{\beta}_{(0)} \end{pmatrix}$, and $\boldsymbol{\varepsilon} \sim (0, \sigma^2 I_n)$. Assume that an approximate linear relationship exists among the column vectors of $\mathbb{Z}_{(s)}$.

(5a) (10%) Show that $(R^{-1})_{jj} = VIF_j$, $j = 1, \cdots, p$, where $R = (\mathbb{Z}_{(s)}^T \mathbb{Z}_{(s)})$ and $VIF_j$ is the variance inflation factor associates with the $j$th covariate.

(5b) (10%) Let $\widehat{\boldsymbol{\beta}}_{(0)}$ and $\widehat{\boldsymbol{\beta}}_{(0)R}$ denote separately the least squares estimator and the ridge regression estimator of $\boldsymbol{\beta}_{(0)}$. Show that $Var(\widehat{\boldsymbol{\beta}}_{(0)}) - Var(\widehat{\boldsymbol{\beta}}_{0R})$ is at least positive semi-definite.