國立臺灣大學數學系
九十五學年度博士班資格考試試題
科目：迴歸分析

2007.06.01

1. (25 pts) Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad 1 \le i \le n,$$

where

$$\epsilon_1, \cdots, \epsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

Define

$$C_k = \frac{SSE_k}{\sigma^2} - n + 2k$$

and

$$C_{k\sigma^2} = \frac{SSE_k}{\hat{\sigma}^2} - n + 2k$$

where SSE_k, $1 \le k \le p$ is the residual sum of squares after the least squares fitting based on

$$Y_i = a_0 + a_1 x_{i1} + \cdots + a_k x_{ik} + e_i, \quad 1 \le i \le n.$$

Here $\hat{\sigma}^2 = SSE_p/(n-p-1)$.

(a) (7 pts) Show that $E(SSE_p) = (n-p-1)\sigma^2$.

(b) (6 pts) When $\beta_{k+1} = \cdots = \beta_p = 0$, show that $E(C_k) = k+1$.

(c) (12 pts) When $\beta_{k+1} = \cdots = \beta_p = 0$, show that

$$C_{k\sigma^2} = (p-k)F + 2k - p,$$

where $F \sim F_{k-p+1, n-k-1}$, the F-distribution with $k-p+1$ and $n-k-1$ degrees of freedom.

2. (25 pts) Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad 1 \le i \le n,$$

where

$$\epsilon_1, \cdots, \epsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

Write $\mathbf{X} = (x_{ij})_{n \times p}$, and $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$. Let w denote the linear space spanned by $\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_p$ where $\mathbf{1} = (1, \ldots, 1)^T$. It is known that the dimension of w is $p+1$. Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be an orthonormal basis for R^n such that $\mathbf{v}_1, \ldots, \mathbf{v}_{p+1}$ span w. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$. Write $\mathbf{Y} = \sum_{i=1}^n a_i \mathbf{v}_i$, $\beta_0 \mathbf{1} + \sum_{j=1}^p \beta_j \mathbf{x}_j = \sum_{i=1}^n b_i \mathbf{v}_i$, and $(\epsilon_1, \ldots, \epsilon_n)^T = \sum_{i=1}^n c_i \mathbf{v}_i$.

(a) (6 pts) Determine a_i's, b_i's, and c_i's.

(b) (9 pts) Let $\mathbf{e}$ denote the residual vector after least-squares fit. Denote $SSE = \mathbf{e}^T\mathbf{e}$. Derive the distribution of SSE.

(c) (10 pts) Let $\hat{\beta}_1$ denote the least squares estimate of β_1. Derive the distribution of $\sqrt{n}(\hat{\beta}_1 - \beta_1)/\sqrt{SSE/(n-p)}$.

3. (25 pts) Assume

$$Y_{in} = x_{in}^2 + \epsilon_i, \quad 1 \le i \le n, \quad x_{in} = \frac{i}{n},$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. random variables with mean 0 and variance σ^2. Suppose that $Y_{in} = \beta_0 + \beta_1 x_{in} + \epsilon_i$ is used to model the relationship between x_{in} and Y_{in}. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the least-squares estimates of β_0 and β_1, respectively.

(a) (10 pts) As $n \to \infty$, does $\hat{\beta}_1$ converge to a constant? If it does, determine that constant.

(b) (8 pts) Determine the asymptotic distribution of $\hat{\beta}_1$.

(c) (7 pts) Determine $\lim_{n\to\infty} Ee_{in}$ when $i = [n/2]$. Here e_{in} denote the residual of the ith observation.

4. (25 pts) Consider a simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where ϵ_i's, $1 \le i \le n$, are iid random variables with mean zero and variance σ^2. Since x_i's cannot be observed, Z_i's are observed instead. It is known that

$$Z_i = x_i + \eta_i,$$

where η_i's, $1 \le i \le n$, are iid random variables with mean zero and variance τ^2. Moreover, ϵ_i's and η_i's are independent. Suppose that a simple linear regression model

$$Y_i = \alpha_0 + \alpha_1 Z_i + e_i$$

is used to fit (Z_i, Y_i) by the least-squares method. The resulting estimates are denoted by $\hat{\alpha}_0$ and $\hat{\alpha}_1$.

(a) (10 pts) When $x_i = i/n$, does $\hat{\alpha}_1$ converge to a constant as $n \to \infty$? If it is, determine that constant.

(b) (15 pts) For each x_i, we observe

$$Z_{ij} = x_i + \eta_{ij}, \quad j = 1, 2.$$

Is it possible to get a consistent estimator of β_1? If it does, please give such a consistent estimator.