臺灣大學數學系
100 學年度下學期博士班資格考試題
科目：迴歸分析

2012.02.24

1. (10%) Let β_1, β_2, and β_3 be the interior angles of a triangle, so that $\beta_1+\beta_2+\beta_3 = 180$ degrees. Suppose we have available estimates Y_1, Y_2, Y_3 of $\beta_1, \beta_2, \beta_3$, respectively. We assume that $Y_i \sim N(\beta_i, \sigma^2)$, $i = 1, 2, 3$ (σ^2 is unknown) and that the Y_i's are independent. Give the F-test for testing the null hypothesis that all three sides of triangle are equal?

2. (20%) Consider the quadratic forms, $Q_1 = \mathbf{Y}^T\mathbf{A}\mathbf{Y}$ and $Q_2 = \mathbf{Y}^T\mathbf{B}\mathbf{Y}$, where $\mathbf{Y} \sim N(\mathbf{0}, \Sigma)$ and $\mathbf{A}$ and $\mathbf{B}$ are nonnegative definite matrices of order $n \times n$.

   (a) (5%) Find the covariance of $\mathbf{y}^T\mathbf{A}\mathbf{y}$ and $\mathbf{y}^T\mathbf{B}\mathbf{y}$.

   (b) (5%) Show that if Q_1 and Q_2 are uncorrelated, they are also independent.

   (c) (5%) Show that $E(\mathbf{y}^T\mathbf{A}\mathbf{y}) \leq \lambda_{\max} \sum_{i=1}^n \sigma_{ii}$, where $\lambda_{\max}$ is the largest eigenvalue of $\mathbf{A}$ and σ_{ii} is the ith diagonal element of Σ.

   (d) (5%) If Σ is known, can you compute the exact probability $P(\mathbf{y}^T\mathbf{A}\mathbf{y} > \mathbf{y}^T\mathbf{B}\mathbf{y})$? Please explain.

3. (20%) Consider the multiple regression models

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad 1 \leq i \leq n,$$

   where $\sum_{i=1}^n X_{i1} = \sum_{i=1}^n X_{i2} = 0$ and $\sum_{i=1}^n X_{i1}X_{i2} = n\rho$.

   (a) (10%) Derive the least squares estimate of $(\beta_0, \beta_1, \beta_2)$, $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$, satisfying the constraint $|\beta_1| + |\beta_2| \leq 1$.

   (b) (10%) When $\beta_1 = 0$, $\beta_2 = 1$, $\sum_{i=1}^n X_{i1}^2 = \sum_{i=1}^n X_{i2}^2 = n$ and ϵ_i's are iid with distribution $N(0, 1)$, derive the joint distributions of $(\hat{\beta}_1, \hat{\beta}_2)$.

4. (15%) Assume that (Y_i, X_i), $1 \leq i \leq n$, $Y_i = X_i + \epsilon_i$ where $X_i = i/n$ and ϵ_i are i.i.d. $N(0, 1)$ random variables. The following three polynomial regression models have been proposed to identify the regression model of (X_i, Y_i)

$$M_1: \ Y_i = \beta_{10} + \epsilon_i, \qquad M_2: \ Y_i = \beta_{20} + \beta_{21}X_i + \epsilon_i,$$
$$M_3: \ Y_i = \beta_{30} + \beta_{31}X_i + \beta_{32}X_i^2 + \epsilon_i$$

   by finding M_j which minimizes Mallows' C_p. For model M, $C_p(M)$ is defined as $RSS(M) + 2 \times dim(M) - n$. Here $RSS(M)$ denotes the residual sum of squares when model M is used, and $dim(M)$ denotes the number of unknown parameters in model M. We now evaluate the operating characteristics of the just mentioned model selection algorithm.

   (a) (2%) When $RSS(M_1) = 138.53$, $RSS(M_2) = 102.93$, and $RSS(M_3) = 102.89$, which model will be chosen?

   (b) (3%) Describe the distributions of $RSS(M_2)$ and $RSS(M_3)$.

   (c) (10%) Give a good estimate of the probability of model M_3 is chosen when n goes to infinity.

5. (15%) Let

$$Y = \begin{cases} 1, & \text{if a person has a disease,} \\ 0, & \text{otherwise,} \end{cases}$$

and suppose that presence of the disease depends on covariates x and on a separate exposure variable, $\mathcal{E}$, with levels $j = 1, \ldots, k$, through the model

$$\log \frac{P(Y = 1|x, \mathcal{E} = j)}{P(Y = 0|x, \mathcal{E} = j)} = a(x) + \lambda_j, \quad j = 1, \ldots, k,$$

for some function $a(x)$ (possibly unknown).

(a) (5%) Controlling for x, what is the odds ratio comparing the odds of having the disease in two different exposure categories, j and j'. (Note that the odds in favor of an event or a proposition is the ratio of the probability that an event will happen to the probability that it will not happen and the odds ratio is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group.)

(b) (10%) Suppose that the data on exposure is collected retrospectively; that is, diseased and nondiseased individuals are sampled and their exposure categories are determined. Let

$$Z = \begin{cases} 1, & \text{if a person is sampled,} \\ 0, & \text{otherwise,} \end{cases}$$

and suppose that

$$P(Z = 1|Y, x, \mathcal{E}) = P(Z = 1|Y, x),$$

i.e., the selection probability for an individual given disease status, Y, and covariates, x, is conditionally independent of exposure, $\mathcal{E}$.
Derive an expression for $P(Y = 1|Z = 1, x, \mathcal{E} = j)$, and show that, controlling for x, the odds-ratio comparing the odds of disease for sampled individuals in two different exposure categories, j and j', is the same as that derived in part (a).

6. (20%) Suppose the pairs (Y_i, X_i), $i = 1, \ldots, n$, are iid realization of a random vector (Y, X) with a distribution π on R^2, and that Y and X both have a finite second moment and $\mu_X = E(X) = 0$. We wish to estimate μ_Y which is $E(Y)$. For any $\beta \in R$, the random variable $Y - \beta X$ has expectation μ_Y, and hence

$$\bar{Y} - \beta \bar{X} \tag{1}$$

is an unbiased estimate μ_Y.

(a) (5%) Find the value of β, call it β_{opt}, for which the estimate (1) has smallest variance.

(b) (5%) Show that, if $corr(Y, X) \neq 0$, then if we use β_{opt}, the variance of estimate (1) is strictly smaller than the variance of $\bar{Y}$.

(c) (10%) In general β_{opt} is not known, and must be estimated. Let $\hat{\alpha}$ and $\hat{\beta}$ be the usual estimates of the coefficients when we do simple linear regression of Y on X. Note that we do not assume that a linear regression model of the form

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad i = 1, \ldots, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

is true: we do not assume that the errors are homoscedastic, that they are normally distributed, or even that they have mean 0. In fact we do not assume anything except that the pairs (Y_i, X_i) are iid from π. The values $\hat{\alpha}$ and $\hat{\beta}$ are just the usual expressions that arise when we do simple linear regression; they are simply functions of the data (Y_i, X_i), $i = 1, \ldots, n$.
Show that $\hat{\alpha}$ is asymptotically equivalent to (1) with the optimal β, i.e. show that as n goes to ∞, the two quantities $n^{1/2}(\hat{\alpha} - \mu_Y)$ and $n^{1/2}(\bar{Y} - \beta_{opt}\bar{X} - \mu_Y)$ have the same limiting distribution.