Qualifying Exam: CAS MA575, Linear Models

Boston University, Fall 2015

1. In clinical trials or experimental designs, it is often of interest to compare two groups of subjects, for example the control group and the treatment group, with the same set of explanatory variables. Consider running two separate linear regression models for the two groups, namely

$$\boldsymbol{Y}^{(k)} = \boldsymbol{X}^{(k)}\boldsymbol{\beta}^{(k)} + \boldsymbol{e}^{(k)},$$

where $\boldsymbol{\beta}^{(k)} = [\beta_1^{(k)}, \dots, \beta_p^{(k)}]^\top$ is the regression coefficient, $\boldsymbol{e}^{(k)}$ is a vector of independent normal random variables with mean zero and (unknown) variance $\{\sigma^{(k)}\}^2$, and $k \in \{\mathrm{C}, \mathrm{T}\}$ indicates the group (control or treatment). Note that in certain cases the original data set is considered to be confidential, and only a group of (summary) statistics will be made available to the general public. Suppose for each group $k \in \{\mathrm{C}, \mathrm{T}\}$, we were given the following information:

Table 1

$n^{(k)}$	:	group size
$\hat{\boldsymbol{\beta}}^{(k)}$	:	ordinary least squares estimate
$\hat{\sigma}^{(k)}$	:	standard error of regression
$\mathrm{se}(\hat{\beta}_j^{(k)})$	:	standard error for each component of $\hat{\boldsymbol{\beta}}^{(k)}$

Throughout this problem, assume for simplicity that $n^{(\mathrm{C})} = n^{(\mathrm{T})} = n$ and that the two groups are independent with deterministic design matrices.

(a) Find the distribution of $\hat{\beta}_1^{(\mathrm{T})} - \hat{\beta}_1^{(\mathrm{C})}$. Can it be fully determined by quantities in Table 1?

(b) Suppose one is interested in testing the null hypothesis

$$H_0: \ \beta_1^{(\mathrm{T})} = \beta_1^{(\mathrm{C})} = \beta_1 \text{ for some } \beta_1 \in \mathbb{R}.$$

Based only on quantities provided in Table 1, provide a statistical test for the above null hypothesis. You need to specify the test statistic and the distribution, either asymptotic or not, that you will be using to obtain the cut-off value.

(c) Under the null hypothesis described in part (b), a better estimate can be obtained for the regression coefficient β_1. In particular, consider the weighted average

$$\hat{\beta}_1(w) = w\hat{\beta}_1^{(\mathrm{C})} + (1 - w)\hat{\beta}_1^{(\mathrm{T})}.$$

Find the weight w that minimizes the variance of $\hat{\beta}_1(w)$. Can this optimal weight be fully determined by quantities in Table 1? If not, how would you choose the weight w in practice if you only have access to quantities in Table 1?

(d) For this part only, suppose you have access to the original data set. Under the constraint that $\boldsymbol{\beta}^{(\mathrm{C})} = \boldsymbol{\beta}^{(\mathrm{T})} = \boldsymbol{\beta}$, find the least squares estimate of $\boldsymbol{\beta}$ as a function of $\{\boldsymbol{Y}^{(k)}, \boldsymbol{X}^{(k)}\}_{k\in\{\mathrm{C},\mathrm{T}\}}$. Can it be fully determined by quantities in Table 1?

(e) Suppose you were given an additional piece of information that the two groups share the same design matrix, namely $\boldsymbol{X}^{(\mathrm{C})} = \boldsymbol{X}^{(\mathrm{T})}$. In this case, under the constraint that $\boldsymbol{\beta}^{(\mathrm{C})} = \boldsymbol{\beta}^{(\mathrm{T})} = \boldsymbol{\beta}$, is it possible to use quantities in Table 1 to fully determine the least squares estimate of $\boldsymbol{\beta}$ as in part (d)? If so, specify the formula.

2. This problem concerns the situation where doubts are casted on the stability assumption of the regression coefficient and the independence assumption of the errors. Consider the following change-point model:

$$y_i = \mu_i + e_i, \quad i = 1, \ldots, n,$$

where

$$\mu_i = \begin{cases} \beta_1, & \text{if } i \le n/2; \\ \beta_2, & \text{otherwise.} \end{cases}$$

Suppose you only observe $y_1, \ldots, y_n$. For simplicity, assume that the sample size is even, namely $n = 2m$ for some integer $m > 0$.

(a) Find the least squares estimate $(\hat{\beta}_1, \hat{\beta}_2)$ for (β_1, β_2).

Assume that the errors satisfy

$$e_i = \epsilon_i - a\epsilon_{i-1}, \quad i = 1, \ldots, n,$$

where $a \in \mathbb{R}$ is a parameter controlling the dependence strength and ϵ_k, $k \in \mathbb{Z}$, are independent normal random variables with mean zero and variance $\sigma^2 > 0$.

(b) For parts (b) and (c) only, assume that $a = 0$. Find the joint distribution of $(\hat{\beta}_1, \hat{\beta}_2)$. Are $\hat{\beta}_1$ and $\hat{\beta}_2$ independent in this case? How does the variance of $\hat{\beta}_1$ change when $n \to \infty$ (for example whether it decreases to zero linearly in n, quadratically or at some other rate)?

(c) Following part (b), find an unbiased estimate $\hat{\sigma}^2$ of σ^2 and devise a statistically valid test for the null hypothesis:

$$H_0 : \ \beta_1 - \beta_2 = 0.$$

You need to specify the test statistic and its distribution under the null hypothesis.

(d) Now suppose that $a = 1$, find the joint distribution of $(\hat{\beta}_1, \hat{\beta}_2)$. Are $\hat{\beta}_1$ and $\hat{\beta}_2$ independent in this case? How does the variance of $\hat{\beta}_1$ change when $n \to \infty$ in this case (for example whether it decreases to zero linearly in n, quadratically or at some other rate)? Compare your result with the one in part (b) and comment on the effect of dependence among the errors. Is dependence always a "bad" thing?

(e) Now suppose that $0 < a < 1$. Find the distribution of $\hat{\beta}_1$. How does the variance of $\hat{\beta}_1$ change when $n \to \infty$ (for example whether it decreases to zero linearly in n, quadratically or at some other rate)? Compare your result with the ones in parts (b) and (d) and comment.