Qualifying Exam: CAS MA575, Linear Models

Boston University, Fall 2017

1. Consider fitting the following second-order model to a given data set:

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j}^{k} \beta_{ij} x_i x_j + e \quad . \tag{1}$$

Here $Y, x_i, x_j, e, \beta_0, \beta_i,$ and β_{ij} are all scalars. Let $z_1, z_2, \ldots, z_k$ denote shifted and rescaled versions of $x_1, x_2, \ldots, x_k$, respectively, such that

$$z_i = \frac{x_i - a_i}{b_i}, \quad i = 1, 2, \ldots, k,$$

where a_i and b_i are known constants. Applying this transformation to model (1), we obtain

$$Y = \gamma_0 + \sum_{i=1}^{k} \gamma_i z_i + \sum_{i<j}^{k} \gamma_{ij} z_i z_j + \epsilon, \tag{2}$$

where the γ_i's and γ_{ij}'s are now the unknown parameters. For a collection of n measurements, models (1) and (2) can be expressed in matrix form as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{3}$$

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}. \tag{4}$$

(a) Show that the column spaces of $\boldsymbol{X}$ and $\boldsymbol{Z}$ are identical.

(b) Show that $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{Z}\hat{\boldsymbol{\gamma}}$, where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$, $\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$.

(c) Show that the regression sum of squares and the error sum of squares are the same for models (3) and (4).

2. Suppose we want to study the effect of variables x_1 (with three levels `L, M, H`) and x_2 (with two levels `O, P`) on a response variable Y. A common approach is to choose n combinations of values of (x_1, x_2) and to obtain measurements on the response Y at each of those combinations. For this, suppose one uses the so-called *balanced design* where each treatment combination has an equal number of k observations, and collects the following data:

x_1	x_2	Y
L	O	$a_1, \ldots, a_k$
M	O	$b_1, \ldots, b_k$
H	O	$c_1, \ldots, c_k$
L	P	$d_1, \ldots, d_k$
M	P	$e_1, \ldots, e_k$
H	P	$f_1, \ldots, f_k$

Consider the linear regression model:

$$\texttt{lm}(\texttt{Y} \sim \texttt{x}_1 + \texttt{x}_2) \tag{5}$$

(a) Find the least squares estimate of model (5) for the data provided above.

(b) Construct a test for the null hypothesis that x_2 is an irrelevant variable.

(c) Construct a test for the null hypothesis that x_1 is an irrelevant variable. (Note that here x_1 has three levels and, as a result, the conventional t-test is not directly applicable.)