UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS

Generalised Regression Models

GRM: Problem Sheet 3 **Semester 1, 2022–2023**

Work on Questions 1, 2 and 3 in the workshop.

1. The observations y_1, y_2 and y_3 were taken on the random variables Y_1, Y_2 and Y_3, where

$$Y_1 = \theta + e_1 \quad Y_2 = 2\theta - \phi + e_2 \quad Y_3 = \theta + 2\phi + e_3$$

$$\mathrm{E}(e_i) = 0, \quad \mathrm{var}(e_i) = \sigma^2 \ \ (i = 1, 2, 3), \quad \mathrm{cov}(e_i, e_j) = 0 \ \ (i \neq j).$$

Formulate this model in matrix notation, and thus find the vector of least squares estimates of the regression parameter vector $\begin{pmatrix} \theta \\ \phi \end{pmatrix}$. Obtain the covariance matrix of the least squares estimators, and use it to determine estimated standard errors of these estimators.

2. If X is an $n \times p$ matrix of rank p, verify that the matrix $X(X^T X)^{-1} X^T$ is symmetric, idempotent and of rank p.

3. The file Births.txt contains the number of normal births in a hospital recorded for each hour of the day, as described in Numerical Example 1.8. Denoting the number for hour t by y_t, fit a periodic regression of the number of births on hour in the form

$$\mathrm{E}(Y_t \,|\, t) = \beta_0 + \beta_1 \cos\left(\frac{\pi t}{12}\right) + \beta_2 \sin\left(\frac{\pi t}{12}\right) \qquad (t = 1, 2, \ldots, 24).$$

   (a) Calculate the least squares estimates of β_0, β_1 and β_2.

   (b) Hence obtain an estimate of the time at which the expected number is a maximum.

4. In the following cases, responses Y_i are assumed to be uncorrelated and have common standard deviation σ given the values of any explanatory variables. For each model described below:

   - express the model as a special case of the linear model defined in Section 3.3 of the lecture notes, $\mathrm{E}(\mathbf{Y} \,|\, \mathbf{X}) = \mathbf{X}\beta$, $\mathrm{var}(\mathbf{Y} \,|\, \mathbf{X}) = \sigma^2 \mathbf{I}_n$, giving appropriate forms for the matrix $\mathbf{X}$ and the vector β of unknown parameters;
   - obtain formulae for $\mathbf{X}^T\mathbf{X}$, $\mathbf{X}^T\mathbf{y}$ and the least squares estimate $\widehat{\beta}$ of β;
   - try to simplify the formulae for $\widehat{\beta}$.
   - state the degrees of freedom for the residual sum of squares.

   (a) The first m responses have expectation μ and the remaining $n - m$ have expectation $\mu + \delta$, where μ and δ are unknown parameters (rather than μ and $\mu + \delta$).

   (b) The first n_1 responses have expectation μ_1, the next n_2 have expectation μ_2 and the final n_3 have expectation μ_3.

   (c) For $i = 1, 2, \ldots, m$, $\mathrm{E}(Y_i \,|\, x_i) = \alpha_1 + \beta x_i$; for $i = m+1, \ldots, n$, $\mathrm{E}(Y_i \,|\, x_i) = \alpha_2 + \beta x_i$. Thus the first m responses satisfy one regression equation and the remaining $n - m$ satisfy another having the *same* slope but a different intercept.

5. Conduct the same exercise as in Question 4 for the following two models but do not attempt to simplify the formulae for $\widehat{\beta}$.

   (a) Responses have expectations $\beta_0 + \beta_1 x_i + \beta_2 x_i^2$ $(i = 1, 2, \ldots, n)$ which are quadratic in x.

   (b) For $i = 1, 2, \ldots, m$, $\mathrm{E}(Y_i \mid x_i) = \alpha + \beta_1 x_i$; for $i = m+1, \ldots, n$, $\mathrm{E}(Y_i \mid x_i) = \alpha + \beta_2 x_i$. So the first m and the remaining $n - m$ expectations satisfy regression equations with the *same* intercept but a different slope.

6. Suppose that a and c are scalars and $\mathbf{b}$ is a p-vector. Find simple expressions for the determinant and inverse of a $(p+1) \times (p+1)$ matrix of the form

$$\mathbf{A} = \begin{pmatrix} a & \mathbf{b}^T \\ \mathbf{b} & c\mathbf{I}_p \end{pmatrix}.$$

   For what values of a, $\mathbf{b}$ and c is $\mathbf{A}$ singular?

7. A mechanical engineer is developing a method for studying the pattern of flow in a gas by photographing tiny light-scattering particles carried along with the gas. A transparent section in a tube containing the gas is used to illuminate the particles and photograph them at frequent intervals with a very short exposure. Suppose that the flow is effectively two-dimensional and that we can identify the particles appearing in successive photographs. Let the coordinates of the ith of n such particles be (x_{i1}, x_{i2}) in one photograph and (y_{i1}, y_{i2}) in the next. If the movement of the set of particles can be approximated by a translation plus a rotation and an enlargement then y_{i1} and y_{i2} should be given approximately by

$$\begin{aligned} y_{i1} &\simeq \alpha_1 + \beta_1 x_{i1} + \beta_2 x_{i2}, \\ y_{i2} &\simeq \alpha_2 + \beta_1 x_{i2} - \beta_2 x_{i1}, \end{aligned}$$

   where α_1, α_2, β_1 and β_2 are unknown parameters. The engineer proposes to derive estimates $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\beta}_1$ and $\widehat{\beta}_2$ of these parameters by minimizing

$$\sum_i \left\{ (y_{i1} - \alpha_1 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 + (y_{i2} - \alpha_2 - \beta_1 x_{i2} + \beta_2 x_{i1})^2 \right\}.$$

   Use the theory of least squares estimation for the model $\mathrm{E}(\mathbf{Y} \mid \mathbf{X}) = \mathbf{X}\beta$ to derive the normal equations satisfied by $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\beta}_1$ and $\widehat{\beta}_2$ in terms of the following statistics:

$$\bar{x}_1,\ \bar{x}_2,\ \bar{y}_1,\ \bar{y}_2,\ t = n^{-1}\sum_i (x_{i1}^2 + x_{i2}^2),\ u = \sum_i (x_{i1} y_{i1} + x_{i2} y_{i2}),\ v = \sum_i (x_{i2} y_{i1} - x_{i1} y_{i2}).$$

   Note that $\mathbf{X}$ is $2n \times 4$, and that it is convenient to write the vector of responses as

$$\mathbf{y} = \begin{pmatrix} y_{11} & \cdots & y_{n1} & y_{12} & \cdots & y_{n2} \end{pmatrix}^T.$$

   [The intention of the study is that fitting this model to the data should help to identify any errors in an initial attempt to match the particles between successive photographs.]