UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS

Generalised Regression Models

GRM: Problem Sheet 4 **Semester 1, 2022–2023**

Work on Questions 1 and 2 in the workshop.

1. Pairs of observations (x_i, y_i) $(i = 1, \dots, n)$ are made on an explanatory variable x and a response variable y. The first m values of x_i are less than some specified value z, and the remainder are greater than z. Hypothesis H_0 states that (given the x_i) the responses Y_i are independent and follow a simple linear regression model, so that Y_i has distribution $N(\alpha + \beta x_i, \sigma^2)$, where α, β and σ are unknown; the alternative hypothesis H_1 states that the responses are Normally distributed with common unknown variance σ^2 and

$$\mathrm{E}(Y_i \,|\, x_i) = \alpha + \beta x_i \qquad (i = 1, \dots, m),$$

$$\mathrm{E}(Y_i \,|\, x_i) = \alpha + \beta x_i + \delta(x_i - z) \quad (i = m+1, \dots, n),$$

where α, β and δ are unknown. Show that the expected response given x does not change at z but that its slope does. [This is called a ‘segmented’ or ‘broken-stick’ regression.]

Show how a Student-t statistic can be used to test H_0 against H_1. What columns of explanatory variables would be used to estimate α, β and δ?

2. The data shown below (and given in Barometer.txt) comprise 10 readings of an aneroid barometer (in mm) and corresponding readings of a mercury barometer (mm) and the temperature (°C) and humidity (%) at the time the two barometer readings were taken.

Aneroid barometer	**Mercury barometer**	**Temperature**	**Humidity**
749.0	744.4	10.0	69.1
746.0	741.3	6.2	48.3
756.0	752.7	6.3	50.0
758.9	754.7	5.3	62.7
751.7	747.8	4.8	60.0
757.5	754.0	3.8	31.3
752.4	747.8	17.1	71.4
752.5	748.6	22.2	25.6
752.2	747.7	20.8	30.7
759.5	755.6	21.0	40.2

The mercury barometer measures air pressure by the height of a column of mercury. The aneroid barometer uses the movement of the elastic top of a metal box containing a vacuum, but its scale is graduated in mm of mercury. A relation between their readings (possibly affected by other atmospheric variables) has been suggested based on fitting a regression model of the form

$$\mathrm{E}(Y \,|\, x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \,, \tag{1}$$

where y, x_1, x_2 and x_3 denote the four variables in the order given above, and β_0, β_1, β_2 and β_3 are unknown parameters.

NOTE: The residual sum of squares for the model is $RSS_{full} = 0.675$.

(a) Calculate the residual sum of squares for a simpler model which states that the expected aneroid barometer reading is *equal* to the mercury barometer reading,

$$E(Y_i | x_{i1}, x_{i2}, x_{i3}) = x_{i1},$$

(no parameters to estimate), i.e. calculate $RSS_{simple} = \sum_i (y_i - \widehat{y}_i)^2 = \sum_i (y_i - x_{i1})^2$.

(b) Using the extra sum of squares ($RSS_{simple} - RSS_{full}$) to compare the two models, examine whether the data are consistent with the model in which the expected aneroid barometer reading is equal to the mercury barometer reading.

3. It is believed that the response Y can be modelled in terms of an explanatory variable x by

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i \quad (i = 1, \ldots, n)$$

where the ε_i are independently distributed as $N(0, \sigma^2)$.

(a) By fitting the model,

$$E(Y_i) = \gamma_0 + \gamma_1 \phi_1(x) + \gamma_2 \phi_2(x) + \gamma_3 \phi_3(x),$$

with explanatory variables $\phi_0(x) = 1$, $\phi_1(x) = x$, $\phi_2(x) = x^2 - 4$, $\phi_3(x) = x^3 - 7x$ (orthogonal polynomials), estimate the regression coefficients of the cubic model from the following data.

x	-3	-2	-1	0	1	2	3
y	8	5	2	0	2	7	7

(b) Use analysis of variance to suggest the most appropriate model for these data given that it is not more complex than the cubic model.

4. The tensile strength of steel is believed to depend upon the amounts present of three substances, A, B, and C. In a exploratory experiment, each of the substances is present at concentrations of either 4 parts per million (ppm) or 6 ppm. Two replicates were made of each treatment combination. The amounts present are denoted in the table by x_1, x_2 and x_3, with the tensile strength being denoted by y.

x_1	6	6	4	4	4	6	6	4
x_2	6	4	6	6	4	6	4	4
x_3	6	6	4	6	6	4	4	4
y	36 42	23 18	25 23	27 21	17 18	32 39	20 24	12 12

The 16 observations are independent of one another, and the model of interest is

$$E(Y) = \alpha + \beta(x_1 - 5) + \gamma(x_2 - 5) + \delta(x_3 - 5).$$

(a) Obtain the design matrix for these data and hence obtain estimates of the parameters.

(b) Provide a complete analysis of variance table for these data and determine whether there is any significant lack of fit to the proposed model.

(c) Can any of the explanatory variables be omitted from the model?

(d) Estimate the mean Y-value corresponding to the combination $x_1 = 5, x_2 = 6, x_3 = 7$, giving an estimate of the standard error of the prediction.

FURTHER EXERCISES (WEIGHTED LEAST SQUARES):

5. Suppose that, given the values x_i of an explanatory variable x, responses Y_i are uncorrelated and have expectations $\beta_0 + \beta_1 x_i$ $(i = 1, 2, \ldots, n)$. The variances of the Y_i, which need not be equal, may conveniently be denoted by w_i^{-1}. A suitable method for estimating the parameters β_0 and β_1 under these assumptions is *weighted least squares*, which minimizes the weighted sum of squares $\sum_i w_i (y_i - \beta_0 - \beta_1 x_i)^2$ with respect to β_0 and β_1.

   Show that the weighted least squares estimates $\widehat{\beta}_0$, $\widehat{\beta}_1$ satisfy the two equations

$$\begin{aligned} \sum_i w_i \widehat{\beta}_0 + \sum_i w_i x_i \widehat{\beta}_1 &= \sum_i w_i y_i , \\ \sum_i w_i x_i \widehat{\beta}_0 + \sum_i w_i x_i^2 \widehat{\beta}_1 &= \sum_i w_i x_i y_i . \end{aligned}$$

6. Suppose that T denotes a statistic (such as a sample mean) which is based on n random variables and has a Normal distribution (at least approximately) with expectation μ and variance σ^2/n. If g is a differentiable function and the distribution of $g(T)$ is required, then in principle this can be derived analytically, but it is often simpler to have a Normal approximation to it. Suppose that $g(t)$ has a Taylor series expansion about μ in which the first two terms are given by

$$g(t) = g(\mu) + (t - \mu)\, g'(\mu) + \ldots .$$

   If the quadratic and higher-order terms in this expansion can be ignored (for large enough n), then $g(T)$ is approximately equal to $g(\mu) + (T - \mu)\, g'(\mu)$, a linear function of T. Thus we might expect the distribution of $g(T)$ to be roughly Normal with expectation $g(\mu)$ and variance $\{g'(\mu)\}^2 \sigma^2/n$.

   Let Y have the Binomial distribution Bi(n, θ) and T be the corresponding proportion of 'successes', equal to Y/n. Apply the above argument to show that the logit function $\ln\left(\frac{T}{1-T}\right)$ of T has the approximate distribution $N\left(\ln\left(\frac{\theta}{1-\theta}\right), \frac{1}{n\theta(1-\theta)}\right)$.

7. For Example 1.6 of the Notes, assume that the probability θ_i of death at dose d_i has the logistic form

$$\theta_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} ,$$

   where $x_i = \ln d_i$, which may be expressed as

$$\text{logit}(\theta_i) = \beta_0 + \beta_1 x_i$$

   where $\text{logit}(\theta_i) = \ln\left(\frac{\theta_i}{1-\theta_i}\right)$.

   Let r_i denote the number dead out of the n_i beetles given dose d_i, $p_i = \frac{r_i + \frac{1}{2}}{n_i + 1}$ (the proportion dead, modified to avoid values of 0 and 1) and

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) .$$

   Then, from Question 6, $\text{logit}(p_i)$ has approximate expectation $\text{logit}(\theta_i) = \beta_0 + \beta_1 x_i$ and variance $\frac{1}{n\theta_i(1-\theta_i)}$.

   Use the weighted least squares method outlined in Question 5 in a linear regression of $\text{logit}(p_i)$ on log dose to estimate β_0 and β_1 for the data in the file Beetles.txt, approximating the variance of the logits by w_i^{-1}, where $w_i = [n_i\, p_i(1 - p_i)]$.

8. Consider a simple linear regression model in which several values $y_{j1}, y_{j2}, \ldots$ of the response are recorded at the jth value of an explanatory variable. Thus n_j responses y_{jk} correspond to one value x_j of the explanatory variable x $(j = 1, 2, \ldots, g;\, k = 1, 2, \ldots, n_j)$ and there are $n = \sum_j n_j$ responses in all. The corresponding random variables Y_{jk} are assumed to be uncorrelated and to satisfy

$$\mathrm{E}\left(Y_{jk}\,|\,x_j\right) = \beta_0 + \beta_1 x_j\,,\ \mathrm{var}\left(Y_{jk}\,|\,x_j\right) = \sigma^2\,.$$

Show that the least squares estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ of β_0 and β_1 for this model satisfy the equations

$$\begin{aligned}\sum_{j=1}^{g} n_j\,\widehat{\beta}_0 + \sum_{j=1}^{g} n_j x_j\,\widehat{\beta}_1 &= \sum_{j=1}^{g} n_j \bar{y}_j\,,\\ \sum_{j=1}^{g} n_j x_j\,\widehat{\beta}_0 + \sum_{j=1}^{g} n_j x_j^2\,\widehat{\beta}_1 &= \sum_{j=1}^{g} n_j x_j \bar{y}_j\,,\end{aligned}$$

where $\bar{y}_j$ denotes the mean $n_j^{-1}\sum_k y_{jk}$ of the jth set of responses. Hence show that these estimates are equivalent to weighted least squares estimates as defined in Question 5, but with the responses y_i and weights w_i replaced by $\bar{y}_j$ and n_j respectively.