UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS

Generalised Regression Models

GRM: Solutions 4 **Semester 1, 2022–2023**

---

1. The expected response at $x = z$ under both formulae is $\alpha + \beta z$, but the slope changes from β to $\beta + \delta$. Defining
$$u_i = 0\ (i = 1, \ldots, m), \quad u_i = x_i - z\ (i = m+1, \ldots, n),$$
we obtain
$$\mathrm{E}(Y_i \,|\, x_i) = \alpha + \beta x_i + \delta u_i \quad (i = 1, \ldots, n).$$
To test the hypothesis H_0 that $\delta = 0$, we fit the regression of y on x and u and compare the value of the t-statistic
$$\hat{\delta}/(\text{estimated standard error of } \hat{\delta})$$
with the distribution $t(n-3)$. The formulae for $\mathbf{X}$ and $\mathbf{X}^T\mathbf{y}$ are
$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_m & \mathbf{x}_1 & \mathbf{0}_m \\ \mathbf{1}_{n-m} & \mathbf{x}_2 & \mathbf{x}_2 - z\mathbf{1}_{n-m} \end{pmatrix}, \quad \mathbf{X}^T\mathbf{y} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \\ \sum\limits_{i=m+1}^{n} (x_i - z)\, y_i \end{pmatrix}.$$
The second and third columns of $\mathbf{X}$ would have to be defined and used with the **lm** function in R. The t-statistic for the latter column would be used to test the hypothesis that the slope of the line is constant.

2. The residual SS under model (1)
$$\mathrm{E}(Y_i | x_{i1}, x_{i2}, x_{i3}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3},$$
is
$$RSS_{full} = 0.675 \ \ \text{(with 6 degrees of freedom)},$$
and the residual SS under the model
$$\mathrm{E}(Y_i \,|\, x_{i1}, x_{i2}, x_{i3}) = x_{i1}$$
is
$$RSS_{simple} = \sum_i (y_i - \widehat{y}_i)^2 = \sum_i (y_i - x_{i1})^2 = 171.07 \text{ (with 10 degrees of freedom)},$$
The extra SS and the corresponding MS relative to model (1) are 170.395 (on $10 - 6 = 4$ degrees of freedom) and 42.60, and the F-statistic is
$$F = \frac{\frac{RSS_{simple} - RSS_{full}}{4-0}}{\frac{RSS_{full}}{6}} = \frac{\frac{171.07 - 0.675}{4}}{\frac{0.675}{6}} = \frac{\frac{170.395}{4}}{0.113} = \frac{42.60}{0.113} = 379.$$
Comparison with $F(4,6)$ provides very strong evidence against the simpler model ($F(4,6)(5\%) = 4.534$ and $F(4,6)(1\%) = 9.148$).

3. Cubic model is:
$$\mathrm{E}(Y) = \gamma_0 + \gamma_1\phi_1(x) + \gamma_2\phi_2(x) + \gamma_3\phi_3(x),$$
where $\phi_0(x) = 1,\ \phi_1(x) = x,\ \phi_2(x) = x^2 - 4$ and $\phi_3(x) = x^3 - 7x$.

(a) Using
$\widehat{\beta} = (X^TX)^{-1}X^T\mathbf{y} = \text{diag}(\sum\phi_0^2(x), \sum\phi_1^2(x), \sum\phi_2^2(x), \sum\phi_3^2(x))^{-1}(\sum y\phi_0(x), \sum y\phi_1(x), \sum y\phi_2(x), \sum y\phi_3(x))$
we obtain the estimates:

$$\begin{aligned}
\widehat{\gamma}_0 &= \frac{\sum y_i\phi_0(x_i)}{\sum\phi_0^2(x_i)} = \frac{8(1)+5(1)+2(1)+0(1)+2(1)+7(1)+7(1)}{7} = \frac{31}{7} \\
\widehat{\gamma}_1 &= \frac{\sum y_i\phi_1(x_i)}{\sum\phi_1^2(x_i)} = \frac{8(-3)+5(-2)+2(-1)+0(0)+2(1)+7(2)+7(3)}{28} = \frac{1}{28} \\
\widehat{\gamma}_2 &= \frac{\sum y_i\phi_2(x_i)}{\sum\phi_2^2(x_i)} = \frac{8(5)+5(0)+2(-3)+0(-4)+2(-3)+7(0)+7(5)}{84} = \frac{63}{84} \\
\widehat{\gamma}_3 &= \frac{\sum y_i\phi_3(x_i)}{\sum\phi_3^2(x_i)} = \frac{8(-6)+5(6)+2(6)+0(0)+2(-6)+7(-6)+7(6)}{216} = -\frac{18}{216} = -\frac{1}{12}
\end{aligned}$$

Thus, the fitted cubic regression equation is:

$$\begin{aligned}
\widehat{y} &= \widehat{\gamma}_0 + \widehat{\gamma}_1\phi_1(x) + \widehat{\gamma}_2\phi_2(x) + \widehat{\gamma}_3\phi_3(x) \\
&= \frac{31}{7} + \frac{1}{28}x + \frac{63}{84}(x^2-4) - \frac{1}{12}(x^3-7x) \\
&= 1.429 + 0.619x + 0.75x^2 - 0.083x^3
\end{aligned}$$

(b) As $\phi_1(x_i)$, $\phi_2(x_i)$ and $\phi_3(x_i)$ are orthogonal, the **extra** sums of squares are given by $\frac{[\sum y_i\phi_1(x_i)]^2}{\sum\phi_1^2(x_i)}$, $\frac{[\sum y_i\phi_2(x_i)]^2}{\sum\phi_2^2(x_i)}$ and $\frac{[\sum y_i\phi_3(x_i)]^2}{\sum\phi_3^2(x_i)}$ for linear, quadratic and cubic terms respectively.
ANOVA table:

Source	SS	df	MS	F
Linear	$\frac{1^2}{28} = 0.036$	1		
Quadratic	$\frac{63^2}{84} = 47.250$	1	47.25	15.86
Cubic	$\frac{18^2}{216}$=1.5000	1	1.5	0.5
Residual	8.928	3	2.98	
Total	$S_{yy} = 57.714$	6		

Considering the coefficient of $\phi_3(x)$ first we test $H_0: \gamma_3 = 0$. As $0.5 < F_{1,3}(5\%) = 10.13$ we do not reject the null hypothesis at the 5% level, and conclude that $\gamma_3 = 0$.
Next, considering the coefficient of $\phi_2(x)$ we test $H_0: \gamma_2 = 0$. As $15.86 > F_{1,3}(5\%) = 10.13$ we reject this null hypothesis at the 5% level, and conclude that $\gamma_2 \neq 0$.
As the quadratic term is required in the regression we do not test lower order terms.
The model is quadratic, and the fitted value is given by

$$\begin{aligned}
\widehat{y} &= \widehat{\gamma}_0 + \widehat{\gamma}_1\phi_1(x) + \widehat{\gamma}_2\phi_2(x) \\
&= \frac{31}{7} + \frac{1}{28}x + \frac{63}{84}(x^2-4) \\
&= 1.429 + 0.036x + 0.75x^2
\end{aligned}$$

Note: We do not need to refit the model to re-estimate the γ coefficients as the explanatory variables $\phi_0(x) = 1,\ \phi_1(x) = x,\ \phi_2(x) = x^2-4$ and $\phi_3(x) = x^3-7x$ are orthogonal. If they were not orthogonal then we would need to re-estimate the parameters γ_0, γ_1 and γ_2.

4. Model:

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

(a) Writing in matrix notation gives:

$$\mathbf{y}=\begin{pmatrix}36\\42\\23\\18\\25\\23\\27\\21\\17\\18\\32\\39\\20\\24\\12\\12\end{pmatrix},\quad X=\begin{pmatrix}1&1&1&1\\1&1&1&1\\1&1&-1&1\\1&1&-1&1\\1&-1&1&-1\\1&-1&1&-1\\1&-1&1&1\\1&-1&1&1\\1&-1&-1&1\\1&-1&-1&1\\1&1&1&-1\\1&1&1&-1\\1&1&-1&-1\\1&1&-1&-1\\1&-1&-1&-1\\1&-1&-1&-1\end{pmatrix},\quad \beta=\begin{pmatrix}\alpha\\\beta\\\gamma\\\delta\end{pmatrix}$$

Thus,

$$X^TX=\begin{pmatrix}16&0&0&0\\0&16&0&0\\0&0&16&0\\0&0&0&16\end{pmatrix},\quad X^T\mathbf{y}=\begin{pmatrix}389\\79\\101\\15\end{pmatrix},$$

and the least squares estimates are given by

$$\widehat{\beta}=(X^TX)^{-1}X^T\mathbf{y}=\begin{pmatrix}16&0&0&0\\0&16&0&0\\0&0&16&0\\0&0&0&16\end{pmatrix}^{-1}\begin{pmatrix}389\\79\\101\\15\end{pmatrix}=\begin{pmatrix}\frac{389}{16}\\\frac{79}{16}\\\frac{101}{16}\\\frac{15}{16}\end{pmatrix}$$

(b) Total (corrected) SS: $S_{yy}=\sum y^2-\frac{(\sum y)^2}{16}=1201.4375$
Pure error SS:

$$SS_E=\text{within SS for the pairs of obsns}=\sum_{\text{all pairs}}[y_1^2+y_2^2-\frac{1}{2}(y_1+y_2)^2]=\frac{1}{2}(6^2+5^2+\cdots+0^2)=83.5$$

using $y_1^2+y_2^2-\frac{1}{2}(y_1+y_2)^2=\frac{1}{2}(y_1-y_2)^2$ [This is the error SS from a one-way ANOVA.]
ANOVA table:

Source	SS	df	MS	F
x_1	$\frac{79^2}{16}=390.0625$	1	390.0625	37.37
x_2	$\frac{101^2}{16}=637.5625$	1	637.5625	61.08
x_3	$\frac{15^2}{16}=14.0625$	1	14.0625	1.35
Lack of fit	76.250	4	19.0625	1.83
Pure error	83.5000	8	10.4375	
Total	$S_{yy}=1201.4375$	15		

To test lack of fit compare 1.83 with $F_{4,8}$. Therefore, as $F_{4,8}(5\%)=3.838$, there is no evidence of lack of fit (at the 5% level).

(c) To test each of the coefficients of x_1, x_2 and x_3, compare F statistics value with $F_{1,8}$. Therefore, as $F_{1,8}(5\%)=5.318$, x_1 (with $F=37.37$) and x_2 (with $F=61.08$) should be retained in the model. However, x_3 can be omitted from the model as its F statistic value of $F=1.35$ is below $F_{1,8}(5\%)=5.318$ (using a 5% level test).

(d) An estimate is required for the **expected** response for $x_1=5$, $x_2=6$, $x_3=7$ (rather than the future response for Y). Using $\mathbf{c}^T\beta$ with $\mathbf{c}^T=(1,0,1)$, the estimate is given by:

$$\widehat{\mathrm{E}(Y)}=\mathrm{E}(\mathbf{c}^T\widehat{\beta})=\mathrm{E}((1,0,1)\widehat{\beta})=\frac{389}{16}+\frac{79}{16}(0)+\frac{101}{16}(1)=\frac{490}{16}=30.625$$

if using the model with x_3 omitted, and the regression coefficients $\beta = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}$.

The variance of the estimator of the **expected** response, using $\mathbf{c}^T\widehat{\beta}$ with $\mathbf{c}^T = (1,0,1)$ is given by

$$\mathrm{var}(\widehat{\mathrm{E}(Y)}) = \mathbf{c}^T (X^TX)^{-1}\mathbf{c}\sigma^2 = (1)^2\frac{\sigma^2}{16} + (0)^2\frac{\sigma^2}{16} + (1)^2\left(\frac{\sigma^2}{16}\right) = \frac{\sigma^2}{8}$$

and thus, the estimated standard error is given by

$$ESE(\widehat{\mathrm{E}(Y)}) = \sqrt{\text{Estimate of var}(\widehat{\mathrm{E}(Y)})} = \sqrt{\frac{\widehat{\sigma}^2}{8}} = \sqrt{\frac{13.37}{8}} = 1.29$$

using the estimate of σ^2 (combining SS for pure error, lack of fit and x_3 in the ANOVA table),

$$\widehat{\sigma}^2 = \frac{RSS}{\text{df for RSS}} = \frac{83.5 + 76.25 + 14.0625}{8+4+1} = \frac{173.8125}{13} = 13.37.$$

5. The derivatives of the weighted sum of squares, Q, with respect to β_0 and β_1 are respectively

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_i w_i\,(y_i - \beta_0 - \beta_1 x_i) \qquad \text{and} \qquad \frac{\partial Q}{\partial \beta_1} = -2\sum_i w_i x_i\,(y_i - \beta_0 - \beta_1 x_i)\,.$$

Equating each of these to zero gives the following *normal equations* for determining the least squares estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

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

6. Using the Normal approximation to the Binomial, Y is approximately $N(n\theta, n\theta(1-\theta))$ for large n (and θ not too near 0 or 1). Hence $T = Y/n$ is approximately $N(\theta, \sigma^2/n)$ with $\sigma^2 = \theta(1-\theta)$. For $g(t) = \ln\left(\frac{t}{1-t}\right)$ we have

$$g'(t) = \frac{1}{t(1-t)}\,,$$

so that the logistic transformation has approximately the Normal distribution with expectation $\ln\left(\frac{\theta}{1-\theta}\right)$ and variance $\frac{1}{\{\theta(1-\theta)\}^2}\frac{\theta(1-\theta)}{n} = \frac{1}{n\theta(1-\theta)}$.

7. If we assume R_i (the number dead at dose d_i) to have the Binomial distribution $\mathrm{Bi}(n_i, \theta_i)$ with n_i large and θ_i not too close to 0 or 1, then, from Question 6, P_i and logit(P_i) have approximate distributions

$$N\left(\theta_i, \frac{\theta_i(1-\theta_i)}{n_i}\right), \quad N\left(\ln\left(\frac{\theta_i}{1-\theta_i}\right), \frac{1}{n_i\theta_i(1-\theta_i)}\right).$$

Under the model proposed, $\ln\{\theta_i/(1-\theta_i)\}$ has the form $\beta_0 + \beta_1 x_i$ with x_i equal to $\ln d_i$. The variance of logit(P_i) is approximately $\{n_i\,\theta_i(1-\theta_i)\}^{-1}$, so for weighted least squares estimation of β_0 and β_1 we may approximate w_i by $n_i\,p_i(1-p_i)$.

For the data of Beetles.txt, the (modified) proportions $p_i = \frac{r_i + \frac{1}{2}}{n_i + 1}$ dead are

p_i :	0.221	0.294	0.500	0.820	0.892	0.976	0.992,

the weights (given by $w_i = n_i\,p_i(1-p_i)$) are

w_i :	10.34	12.86	14.00	9.29	5.70	1.44	0.49

and the logits of the proportions are

logit(p_i) :	−1.26	−0.88	0.00	1.52	2.11	3.71	4.80.

Note that the weights are highest when the proportions are close to 0.5 (if the n_i are equal).

Writing x_i and y_i for $\ln d_i$ and $\text{logit}(p_i)$, the sums required are:

$$\sum_i w_i = 54.114\,,\ \sum_i w_i x_i = 221.668\,,\ \sum_i w_i x_i^2 = 908.454\,,$$
$$\sum_i w_i y_i = 9.509\,,\ \sum_i w_i x_i y_i = 45.436\,.$$

Solving the equations given in Question 5 for $\widehat{\beta}_0$ and $\widehat{\beta}_1$, e.g. using

$$\begin{aligned}\begin{pmatrix}\widehat{\beta}_0\\ \widehat{\beta}_1\end{pmatrix} &= \begin{pmatrix}\sum_i w_i & \sum_i w_i x_i\\ \sum_i w_i x_i & \sum_i w_i x_i^2\end{pmatrix}^{-1}\begin{pmatrix}\sum_i w_i y_i\\ \sum_i w_i x_i y_i\end{pmatrix}\\ &= \frac{1}{\sum_i w_i \sum_i w_i x_i^2 - (\sum_i w_i x_i)^2}\begin{pmatrix}\sum_i w_i x_i^2 & -\sum_i w_i x_i\\ -\sum_i w_i x_i & \sum_i w_i\end{pmatrix}\begin{pmatrix}\sum_i w_i y_i\\ \sum_i w_i x_i y_i\end{pmatrix},\end{aligned}$$

gives the *weighted least squares* estimates $\widehat{\beta}_0 = -60.6$ and $\widehat{\beta}_1 = 14.8$.

[Note that the equations for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are rather ill-conditioned because the range of the log-doses is small: a better formulation of the logistic model would be

$$\text{logit}(\theta_i) = \gamma + \beta_1(x_i - \bar{x})\,.]$$

8. To obtain the least squares estimates, minimize

$$Q = \sum_{j=1}^{g}\sum_{k=1}^{n_j}(y_{jk} - \widehat{\beta}_0 - \widehat{\beta}_1 x_j)^2.$$

Differentiating Q with respect to β_0 and β_1 gives

$$\begin{aligned}\frac{\partial Q}{\partial \beta_0} &= -2\sum_j\sum_k(y_{jk} - \widehat{\beta}_0 - \widehat{\beta}_1 x_j)\\ \frac{\partial Q}{\partial \beta_1} &= -2\sum_j\sum_k(y_{jk} - \widehat{\beta}_0 - \widehat{\beta}_1 x_j)x_j\,.\end{aligned}$$

Using $\sum_k y_{jk} = n_j \bar{y}_j$, in the equations $\frac{\partial Q}{\partial \beta_0} = \frac{\partial Q}{\partial \beta_1} = 0$ gives

$$\begin{aligned}\sum_j n_j(\bar{y}_j - \widehat{\beta}_0 - \widehat{\beta}_1 x_j) &= 0\\ \sum_j n_j(\bar{y}_j - \widehat{\beta}_0 - \widehat{\beta}_1 x_j)x_j &= 0\,.\end{aligned}$$

These normal equations may be written as

$$\begin{aligned}\sum_j n_j\,\widehat{\beta}_0 + \sum_j n_j x_j\,\widehat{\beta}_1 &= \sum_j n_j \bar{y}_j\\ \sum_j n_j x_j\,\widehat{\beta}_0 + \sum_j n_j x_j^2\,\widehat{\beta}_1 &= \sum_j n_j x_j \bar{y}_j\,.\end{aligned}$$

From Question 5, these are the equations satisfied by the weighted least squares estimates of β_0 and β_1 when the responses are taken as $\bar{y}_j$ and the weights equal n_j $(j = 1,\ldots,g)$.