Regression modelling - Model solution 2013/2014.

1(a) (i) linear

(ii) linear

(iii) non linear

1(b) (i) $y_i = \alpha + \beta(x_i - \bar{x}) + \varepsilon_i$

$$\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \underline{\theta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad A = \begin{bmatrix} 1 & x_1 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{bmatrix}$$

(ii) $y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i$

$$\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \underline{\theta} = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}, \quad A = \begin{bmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}$$

(iii) $y_{ij} = \alpha_i + \beta_i(x_{ij} - \bar{x}_{i\cdot}) + \varepsilon_{ij}$

$$\underline{y} = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{pmatrix}, \quad \theta = \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{pmatrix}, \quad A = \begin{bmatrix} 1 & (x_{11} - \bar{x}_{1\cdot}) & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (x_{1n_1} - \bar{x}_{1\cdot}) & 0 & 0 \\ 0 & 0 & 1 & (x_{21} - \bar{x}_2) \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & (x_{2n} - \bar{x}_2) \end{bmatrix}$$

1(b)

(iv) $\hat{\underline{\theta}} = (A^TA)^{-1}A^Ty$

$$(A^TA) = \begin{bmatrix} n & \sum(x_i-\bar{x}) \\ \sum(x_i-\bar{x}) & \sum(x_i-\bar{x})^2 \end{bmatrix}$$

$$(A^TA)^{-1} = \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & 1/\sum(x_i-\bar{x})^2 \end{bmatrix}$$

$$A^Ty = \begin{pmatrix} \sum y_i \\ \sum y_i(x_i-\bar{x}) \end{pmatrix}$$

$$\hat{\underline{\theta}} = \begin{pmatrix} \bar{y} \\ \frac{\sum y_i(x_i-\bar{x})}{\sum(x_i-\bar{x})^2} \end{pmatrix}$$

1 (c) Gauss Markov Thm states that the OLS are unbiased and also have the minimum variance of all linear ~~estimators~~, so the estimators are best linear unbiased estimators

(d) Benefits of centring the covariates are to simplify (A^TA) and hence make the inverse easier to calculate (a diagonal matrix)., Parameters are asymptotically independent

(e) R^2 is the % of variability in the response explained by the covariates. R^2 (adjusted) takes account of the number of covariates included in the model. We would aim for $\geq 70\%$

2 (a)

(i) $y_i = \alpha + \beta x_{1i} + \gamma x_{2i} + \varepsilon_i$

$$\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \underline{\theta} = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}, \quad A = \begin{bmatrix} 1 & x_{11} & x_{21} \\ \vdots & | & | \\ 1 & x_{1n} & x_{2n} \end{bmatrix}$$

(ii) $\underline{b}^T \hat{\underline{\theta}} \pm t(n-k, \frac{1+c}{2}) \sqrt{\frac{r}{n-k} \underline{b}^T (A^T A)^{-1} \underline{b}}$ $\quad \frac{r}{n-k} = \frac{120.22}{27}$

$\hat{\beta} = -0.5322$, $\sqrt{0.04165 \times 4.452} = 0.4306$

$\hat{\gamma} = 3.2535$, $\sqrt{0.04863 \times 4.452} = 0.4652$

$t(27, 0.975) = 2.0518$

β lies within $\hat{\beta} \pm 2.0518 \times 0.4306$, -0.5322 ± 0.8835

γ " " $\hat{\gamma} \pm 2.0518 \times 0.4652$, 3.2535 ± 0.9546

The interval for β includes 0, therefore x_1 could be dropped from the model, but not x_2 since interval for γ does not include 0

(iii) $R^2 = \frac{337.89 - 120.22}{337.89} = 64.4\%$
% variation explained in fat by the APPs

(b) , the residual $r_i = y_i - \hat{y}_i$

standardised $r_i / \sqrt{var(r_i)}$

(i) it would be usual to produce a probability plot of the residuals, should be a straight line

(ii) Usually a scatterplot of r_i vs $\hat{y}_i$, a random distribution of points around 0, with no fanning out.

2(c) Forward stepwise selection.

(i) the algorithm starts with no explanatory variables only an intercept

(ii) compute F statistics of all models with one parameter more

(iii) add the variable with the largest F value

loop over steps (ii) and (iii) until F value of the model meets a threshold or all the variables have been added.

In terms of stopping rules R^2 and R^2_{adj} are often used, occasionally a dummy variable (randomly generated is used), if this is entered into the model, this would suggest over-selection.

3 (a)

95% CI for β

$$\hat{\beta} \pm t\left(n-k, \tfrac{1+c}{2}\right)\sqrt{\frac{r}{n-k}\, b^T(A^TA)b}$$

$\hat{\beta} = 191.737$, $\sqrt{\frac{r}{n-k}} = 8.587$

$b^T(A^TA)^{-1}b = 5.09076$

$t(47, 0.975) = 2.0110$

CI is 191.737 ± 38.960, interval does not include 0, therefore diffusivity and porosity are related.

(b) 95% PI

$$\hat{\alpha} + 0.25\hat{\beta} \pm t\left(n-k, \tfrac{1+c}{2}\right)\sqrt{\frac{r}{n-k}\left(1 + b^T(A^TA)^{-1}b\right)}$$

$\underline{b} = (1 \;\; 0.25)$

$$b^T(A^TA)^{-1}b = (1 \;\; 0.25)\begin{bmatrix} 0.70511 & -1.86690 \\ -1.86690 & 5.09076 \end{bmatrix}\begin{pmatrix} 1 \\ 0.25 \end{pmatrix}$$

$\hat{\alpha} + 0.25\hat{\beta} = -36.092 + 0.25 \times 191.737 = 11.7505 \pm 5.1517.$

these

$$\begin{pmatrix} 0.70511 - 0.25 \times 1.8669 \\ -1.86690 + 0.25 \times 5.09076 \end{pmatrix}^T \begin{pmatrix} 1 \\ 0.25 \end{pmatrix} = (0.2376 \;\; -0.5943)\begin{pmatrix} 1 \\ 0.25 \end{pmatrix}$$

$= 0.2376 - 0.1485$

$= 0.0890$

PI is 11.7505 ± 5.1517, so $\underline{(6.6 \text{ to } 16.7)}$.

3(b) a prediction interval is always wider than the confidence interval since it includes the variation within the population.

3(c) R^2 is ~67% so this is moderately high, around 67% of the variability in porosity is explained by diffusivity.

3(d) The residual vs fitted plot shows a random scatter round zero, suggesting the model performs reasonably well; No evidence of non-constant variability.
The QQ plot shows some deviation from normality but not sufficiently to challenge the distributional assumptions.

3(e) the leverage $h_i = H_{ii}$, $H = X(X^TX)^{-1}X^T$
a large leverage means that $var(\hat{\varepsilon}_i)$ is small,
so the fit is forced close to y_i
an average value for h_i is p/n, so rule of thumb is that values $> 2p/n$ should be investigated.

3(f) $\hat{\rho} = 0.579$

$z(r) = 0.6610$

$z(r) + \frac{2}{\sqrt{n-3}} = 0.6610 + 0.2119 = 0.8729$ $n = 92$.

$z(r) - \frac{2}{\sqrt{n-3}} = 0.6610 - 0.2119 = 0.4491.$

transforming back:

$r(z_L) = 0.7014.$

$r(z_u) = 0.4219.$

approx 95% CI for ρ is $(0.4219, 0.7014)$

which is statistically significant, since does not include 0.

4 (a)

(i) $y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, 3.$

normal equations.

$$(A^TA)\hat{\underline{\theta}} = A^T \underline{y}$$

$$\underline{y} = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ y_{31} \\ \vdots \\ y_{3n_3} \end{pmatrix}$$

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & 1 & 0 & 0 \\ \vdots & 0 & 1 & \vdots \\ \vdots & \vdots & \vdots & 0 \\ \vdots & \vdots & 0 & \vdots \\ 1 & 0 & \vdots & 1 \end{bmatrix}$$

$$\underline{\theta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$$

$$A^T \underline{y} = \begin{pmatrix} \sum\sum y_{ij} \\ \sum y_{1j} \\ \sum y_{2j} \\ \sum y_{3j} \end{pmatrix} \qquad (A^TA) = \begin{bmatrix} n & n_1 & n_2 & n_3 \\ n_1 & n_1 & 0 & 0 \\ n_2 & 0 & n_2 & 0 \\ n_3 & 0 & 0 & n_3 \end{bmatrix}$$

$$(A^TA)\hat{\theta} = \begin{pmatrix} n\hat{\mu} + n_1\hat{\alpha}_1 + n_2\hat{\alpha}_2 + n_3\hat{\alpha} \\ n_1\mu + n_1\hat{\alpha}_1 \\ n_2\hat{\mu} + n_2\hat{\alpha}_2 \\ n_3\hat{\mu} + n_3\hat{\alpha}_3 \end{pmatrix}$$

4a(ii) So we do not have 4 independent equations, for the 4 unknown parameters, so therefore we must impose a constraint.

There are several possible constraints such as $\hat{\alpha}_1 = 0$, we typically use $\sum_{i=1}^{3} n_i \hat{\alpha}_i = 0$.

(iii) Using this latter constraint, plugging into the equations

$$n\hat{\mu} = \sum_i \sum_j y_{ij}, \quad \hat{\mu} = \bar{y}_{\cdot\cdot}$$

$$\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$$

in the standard class notation.

4(b)

(i) $y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij}$

$$y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{pmatrix}, \quad \theta = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta \end{pmatrix}, \quad A = \begin{bmatrix} 1 & 1 & 0 & x_{11} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & 1 & 0 & x_{1n_1} \\ \vdots & 0 & 1 & x_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & x_{2n_2} \end{bmatrix},$$

$$A^TA = \begin{bmatrix} n & n_1 & n_2 & \sum_i\sum_j x_{ij} \\ n_1 & n_1 & 0 & \sum_j x_{1j} \\ n_2 & 0 & n_2 & \sum_j x_{2j} \\ \sum\sum x_{ij} & \sum_j x_{1j} & \sum_j x_{2j} & \sum_i\sum_j x_{ij}^2 \end{bmatrix}$$

This matrix A is not of full rank, since first column is the sum of columns two and three

One possible constraint is to set $\alpha_1 = 0$

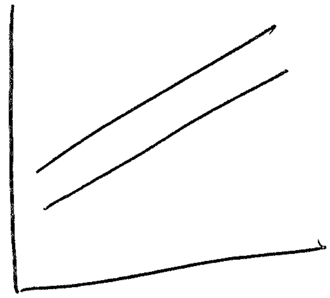
then:

$$A = \begin{bmatrix} 1 & 0 & x_{11} \\ \vdots & \vdots & \vdots \\ \vdots & 0 & x_{1n_1} \\ & 1 & x_{21} \\ 1 & \vdots & \vdots \\ 1 & 1. & x_{2n_2} \end{bmatrix}$$

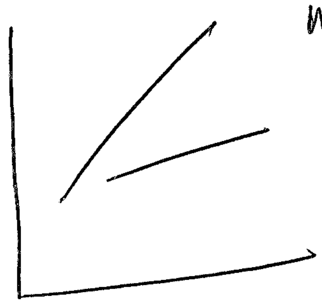
4(b)(ii) the proposed model allows for a difference in intercept but a common slope

A more complex model would allow the slopes also to be different.

model 1.



model 2.



$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij}$

$y_{ij} = \mu + \alpha_i + \beta_i x_{ij} + \varepsilon_{ij}$