Regression Methods: Examination

31 January 2024

Instructions: The time allotted for the examination is 180 minutes. You may answer in either English or French. No written material may be brought into the examination, but a simple calculator may be used. Full marks may be obtained with complete answers to four questions. The final mark will be based on the best four solutions.

Notation: 1_n, 0_n and I_n respectively denote the $n \times 1$ vectors of ones and of zeros, and the $n \times n$ identity matrix; $A_{r\times s}$ means that A is an $r \times s$ matrix; $X \sim \mathcal{N}_p(\mu, \Omega)$ means that X has a p-dimensional multivariate normal distribution with mean vector $\mu_{p\times 1}$ and variance matrix $\Omega_{p\times p}$; and $X_{p\times 1} \sim (\mu, \Omega)$ means that $\mathrm{E}(X) = \mu_{p\times 1}$ and $\mathrm{var}(X) = \Omega_{p\times p}$.

First name:

Last name:

SCIPER:

Exercise	Points	Indicative marks
1		/10 points
2		/10 points
3		/10 points
4		/10 points
5		/10 points
Total:		/40 points

Problem 1 If $X_{n\times p}$ contains real numbers and is of rank p, where $p < n$, then the linear subspace of $\mathbb{R}^n$ spanned by the columns of X is denoted $\text{Span}(X)$.

(a) Find the $p \times 1$ vector $\hat{\beta}$ that minimises the Euclidean distance $\|y - X\beta\|$ between a vector $y \in \mathbb{R}^n$ and $\text{Span}(X)$.

(b) What are the *second-order distributional assumptions* for the linear model? Show that under these assumptions the estimator $\hat{\beta}$ is unbiased and obtain its variance matrix.

(c) Let $\theta = c^{\mathrm{T}}\beta$, where $c \neq 0$ is a $p \times 1$ vector of constants. Show that under second-order assumptions the estimator $\hat{\theta} = c^{\mathrm{T}}\hat{\beta}$ has the smallest variance among all linear unbiased estimators of θ.

Problem 2

(a) Explain the meaning of the italicised words and phrases in the phrase 'the interpretation of an *analysis of variance table* is greatly simplified if the *terms* are *orthogonal*'.

(b) The linear model $y_{rc} \overset{\text{ind}}{\sim} (\eta_{rc}, \sigma^2)$ for a 3×3 Latin square with treatments A, B and C,

	Column		
Row	1	2	3
1	A	B	C
2	C	A	B
3	B	C	A

expresses the mean response in the rth row and cth column as

$$\eta_{rc} = \mu + \alpha_r + \beta_c + \gamma_{t(r,c)}, \qquad r, c \in \{1, 2, 3\}, \quad t \in \{A, B, C\},$$

where the treatment $t(r, c)$ depends on the row and column, and

$$\begin{pmatrix} \eta_{11} \\ \eta_{12} \\ \eta_{13} \\ \eta_{21} \\ \eta_{22} \\ \eta_{23} \\ \eta_{31} \\ \eta_{32} \\ \eta_{33} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}}_{X} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \gamma_A \\ \gamma_B \\ \gamma_C \end{pmatrix}.$$

Identify the terms of the design matrix X corresponding to rows, columns and treatments.

(c) Let $a = 1/3$. Giving a careful argument, explain how the analysis of variance table when the model in (b) is fitted to data is related to that for a model with design matrix

$$X_* = \begin{pmatrix} 1 & -a & -a & -a & -a & -a & -a \\ 1 & -a & -a & 2a & -a & 2a & -a \\ 1 & -a & -a & -a & 2a & -a & 2a \\ 1 & 2a & -a & -a & -a & -a & 2a \\ 1 & 2a & -a & 2a & -a & -a & -a \\ 1 & 2a & -a & -a & 2a & 2a & -a \\ 1 & -a & 2a & -a & -a & 2a & -a \\ 1 & -a & 2a & 2a & -a & -a & 2a \\ 1 & -a & 2a & -a & 2a & -a & -a \end{pmatrix}.$$

How many degrees of freedom are there for row, column and treatment effects? How many degrees of freedom, if any, are available for estimation of σ^2?

(d) How does the analysis of variance in (c) change if the order of the terms changes?

Problem 3

(a) An observation Y has probability function of the form

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y;\phi)\right\}, \quad y \in \mathcal{Y}, \theta \in \Theta, \phi > 0,$$

where $\mathcal{Y}, \Theta \subset \mathbb{R}$. Find the cumulant-generating function of Y and hence show that $\mathrm{E}(Y) = b'(\theta)$ and $\mathrm{var}(Y) = \phi b''(\theta)$.

(b) Show that the Poisson distribution can be written in the above form, and hence find the relation between its mean and variance function.

(c) Explain the terms *generalized linear model* and *canonical link function*. Derive the canonical link function for the Poisson distribution.

(d) Independent Poisson variables $Y_1, \ldots, Y_n$ have means $\mu_1, \ldots, \mu_n$, where $\mu_j = \lambda_0 + x_j\lambda_1 + z_j\lambda_2$, with x_j and z_j being constants. Can this be written in the form of a generalized linear model? If so, give the corresponding link function.

Problem 4

(a) Give two definitions of residuals in a general regression model, and discuss how they might be used to assess the fit of the model to data.

(b) Figure 1 shows residuals for fits of generalized linear models to four sets of $n = 200$ observations. In each case (i) say what you think the model is, giving reasons, (ii) say whether the fit seems adequate, giving reasons, and (iii) suggest what steps you would take to deal with any model failure suggested by the plot.

Problem 5 Consider a regression model $y_{n\times 1} \sim (\mu_{n\times 1}, \sigma^2 I_n)$, where the jth element of μ equals $\mu(x_j)$, with $\mu(x)$ twice continuously differentiable for $x \in [a,b]$ and $a < x_1 < \cdots < x_n < b$.

(a) Explain why it may be desirable to estimate μ by minimising the penalised sum of squares

$$\sum_{j=1}^{n} \{y_j - \mu(x_j)\}^2 + \lambda \int_a^b \mu''(x)^2 \,\mathrm{d}x$$

for some $\lambda > 0$, and discuss the cases $\lambda \to 0$ and $\lambda \to \infty$.

(b) If one can write

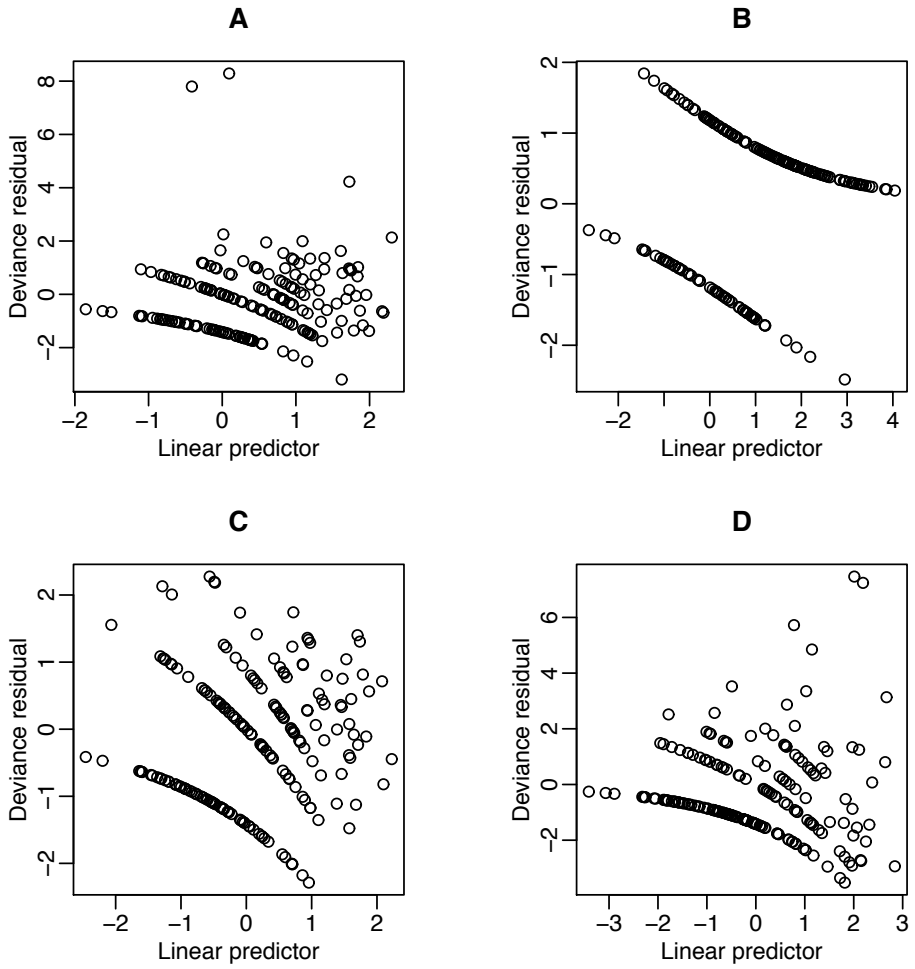
$$\int_a^b \mu''(x)^2 \,\mathrm{d}x = \mu^{\mathrm{T}} \Delta \mu,$$

for some symmetric positive semi-definite matrix $\Delta_{n\times n}$ of rank $n-2$, show that the vector of fitted values $\hat{\mu}_{n\times 1}$ may be written as $H_\lambda y$, where H_λ should be given. Show also that

$$\mathrm{tr}(H_\lambda) = \sum_{j=1}^{n} \frac{1}{1+\lambda d_j},$$

for some non-negative $d_1, \ldots, d_n$, and discuss how $\mathrm{tr}(H_\lambda)$ varies as a function of λ.

Figure 1: Diagnostic plots for generalized linear model fits.



(c) By writing $\mu - \hat{\mu} = (I_n - H_\lambda)\mu + H_\lambda(\mu - y)$, or otherwise, show that

$$\mathrm{E}\left\{(\hat{\mu} - \mu)^{\mathrm{T}}(\hat{\mu} - \mu)\right\} = \sigma^2 \mathrm{tr}(H_\lambda^{\mathrm{T}} H_\lambda) + \|(I - H_\lambda)\mu\|^2,$$

and discuss the interpretation of the terms on the right-hand side of this expression.

(d) Show that $\mathrm{tr}(H_\lambda^{\mathrm{T}} H_\lambda) < \mathrm{tr}(H_\lambda)$ for any $\lambda > 0$. What does this imply about the corresponding fitted models?

———————— END OF THE EXAM PAPER ————————