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

Solution 1

(a) [2, seen] Minimising $\|y - X\beta\|$ with respect to β is equivalent to minimising

$$\|y - X\beta\|^2 = (y - X\beta)^{\mathrm{T}}(y - X\beta) = y^{\mathrm{T}}y - 2y^{\mathrm{T}}X\beta + \beta X^{\mathrm{T}}X\beta,$$

and differentiation and setting the result equal to zero gives

$$-2X^{\mathrm{T}}y + 2X^{\mathrm{T}}X\beta = 0,$$

and, since the $p \times p$ matrix $X^{\mathrm{T}}X$ has the same rank as X and therefore is invertible, this gives

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y.$$

(b) [3, seen] The second-order assumptions are that $\mathrm{E}(y) = X\beta$ and $\mathrm{var}(y) = \sigma^2 I_n$. Under these assumptions

$$\mathrm{E}(\hat{\beta}) = \mathrm{E}\left\{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y\right\} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{E}(y) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X\beta = \beta,$$

and, writing $A = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, we have

$$\mathrm{var}(\hat{\beta}) = \mathrm{var}(Ay) = A\mathrm{var}(y)A^{\mathrm{T}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\sigma^2 I_n\{X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\}^{\mathrm{T}} = \sigma^2(X^{\mathrm{T}}X)^{-1}.$$

(c) [5, seen/unseen] This is a version of the Gauss–Markov theorem. A linear estimator of θ must be of the form $\tilde{\theta} = b^{\mathrm{T}}y$ for some constant $n \times 1$ vector b. If $\tilde{\theta}$ is unbiased then $\mathrm{E}(\tilde{\theta}) = b^{\mathrm{T}}X\beta = \theta = a^{\mathrm{T}}\beta$ for all β, i.e., $b^{\mathrm{T}}X = a^{\mathrm{T}}$, and $\mathrm{var}(\tilde{\theta}) = b^{\mathrm{T}}\mathrm{var}(y)b = \sigma^2 b^{\mathrm{T}}b$.

The estimator $\hat{\theta}$ has variance $\sigma^2 a^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}a$, so

$$\mathrm{var}(\tilde{\theta})-\mathrm{var}(\hat{\theta}) = \sigma^2 b^{\mathrm{T}}b-\sigma^2 a^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}a \propto b^{\mathrm{T}}b-b^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}b = b^{\mathrm{T}}\{I_n-X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\}b.$$

It is readily checked that $P = I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is symmetric and idempotent, so its eigenvalues are either 0 or 1. The spectral theorem gives $P = UDU^{\mathrm{T}}$, where U is orthogonal and D is a diagonal matrix containing the eigenvalues of P, so

$$\mathrm{var}(\tilde{\theta}) - \mathrm{var}(\hat{\theta}) = \sigma^2 b^{\mathrm{T}}Pb = \sigma^2 b^{\mathrm{T}}UDU^{\mathrm{T}}b = \sigma^2 d^{\mathrm{T}}Dd \geq 0,$$

where $d = U^{\mathrm{T}}b \neq 0$. Hence $\hat{\theta}$ has minimum variance among linear unbiased estimators of θ.

Solution 2

(a) [3, seen] See Slides 8–12, 39–40.

(b) [2, unseen] The terms are (i) the first column, corresponding to a grand mean, (ii) columns 2–4, corresponding to rows, (iii) columns 5–7, corresponding to columns, and (iv) columns 8–10, corresponding to the treatments.

(c) [4, unseen] The matrix X is 9×10, so it is clearly rank-deficient; in fact the sum of the columns of X for rows is the grand mean, and likewise for columns and treatments. Hence the rank of X is at most $10 - 3 = 7$. One way to obtain a full-rank matrix is to drop the first column of each term, giving a 9×7 matrix, and then to orthogonalise the remaining columns with respect to the grand mean. This process gives the matrix X_*, and it is easy to check that its columns are all orthogonal, so it has rank 7.

The number of degrees of freedom for rows, columns and treatments is 2 each, so there are $9 - (1 + 2 + 2 + 2) = 2$ left for estimation of σ^2.

(d) [1, unseen] The ANOVA is unchanged if the order of the terms changes, because of the orthogonality.

Solution 3

(a) [3, seen] Slide 119

(b) [2, seen] Example 22

(c) [3, seen] Slides 116–118

(d) [2, unseen] Yes, this is a GLM with Poisson errors and identity link function (which might give negative fitted values).

Solution 4

(a) [4, seen] Slides 108–111.

(b) [6, unseen] A: the residuals seem to be from a Poisson regression model, which have this characteristic banding pattern (corresponding to values 0, 1, 2, etc.) working from the bottom to the top in the figure), but there are two or three outliers, shown by the residuals larger than 3 or so. Drop these observations and try to fit again.

B: Residuals from a binary regression model, since there are negative residuals corresponding to the 0s and positive ones corresponding to the 1s. There seem to be no obvious problems here.

C: These look like residuals from a Poisson regression, for the reasons given in A, with no obvious problems.

D: These are residuals for count data, for the reasons given above, but they seem to be overdispersed (they are spread from -2.5 to $+6$, without any obvious outliers, unlike in A), so perhaps fitting using a quasilikelihood is indicated.

Solution 5

(a) [2, seen/unseen] Slide 249; $\lambda \to 0$ gives no penalty; $\lambda \to \infty$ gives a straight line.

(b) [4, seen] We aim to minimise

$$(y-\mu)^{\mathrm{T}}(y-\mu) + \lambda\mu^{\mathrm{T}}\Delta\mu = y^{\mathrm{T}}y - 2y^{\mathrm{T}}\mu + \mu^{\mathrm{T}}(I_n + \lambda\Delta)\mu,$$

and this gives $\hat{\mu} = (I_n + \lambda\Delta)^{-1}y = H_\lambda y$, say. The spectral theorem gives $\Delta = UDU^{\mathrm{T}}$, where $D = \mathrm{diag}(d_1, \ldots, d_n)$ with $d_1 = d_2 = 0 < d_3 < \cdots < d_n$ and U orthogonal, so

$$H_\lambda = (UU^{\mathrm{T}} + \lambda UDU^{\mathrm{T}})^{-1} = U(I_n + \lambda D)^{-1}U^{\mathrm{T}}.$$

Therefore

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

which decreases monotonically as a function of λ, with $\mathrm{tr}(H_0) = n$ and $\mathrm{tr}(H_\infty) = 2$, corresponding to a line that passes through all the points and a straight line. Its interpretation is as equivalent degrees of freedom for the fitted model.

(c) [2, seen] Writing $\mu - \hat{\mu} = (I_n - H_\lambda)\mu + H_\lambda(\mu - y)$ gives

$$(\hat{\mu}-\mu)^{\mathrm{T}}(\hat{\mu}-\mu) = (y-\mu)^{\mathrm{T}} H_\lambda^{\mathrm{T}} H_\lambda (y-\mu) + 2(y-\mu)^{\mathrm{T}} H_\lambda^{\mathrm{T}} (H_\lambda - I_n)\mu + \mu^{\mathrm{T}}(H_\lambda - I_n)^{\mathrm{T}}(H_\lambda - I_n)\mu.$$

The last term here is constant and the second has expectation zero, while the first equals

$$\mathrm{tr}\left\{(y-\mu)(y-\mu)^{\mathrm{T}} H_\lambda^{\mathrm{T}} H_\lambda\right\}$$

and therefore has expectation $\sigma^2 \mathrm{tr}(H_\lambda^{\mathrm{T}} H_\lambda)$, because $\mathrm{E}\{(y-\mu)(y-\mu)^{\mathrm{T}}\} = \sigma^2 I_n$. Hence

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

as required.

The interpretation is in terms of variance $\sigma^2 \mathrm{tr}(H_\lambda^{\mathrm{T}} H_\lambda)$ corresponding to 'double smoothing' and squared bias $\|(I - H_\lambda)\mu\|^2$. For the bias term

$$(I_n - H_\lambda)\mu = (I_n + \lambda\Delta)^{-1}(I_n + \lambda\Delta - I_n)\mu = \lambda H_\lambda \Delta \mu,$$

which equals zero only if (i) $\lambda = 0$ (no penalty) or (ii) $\Delta\mu = 0$ (μ is in the kernel of Δ, i.e., $\mu''(x) = 0$, i.e., $\mu(x)$ is a straight line).

(d) [2, unseen] Since H_λ is symmetric and semi-positive definite, the previous spectral decomposition gives

$$\begin{aligned}
\mathrm{tr}(H_\lambda^{\mathrm{T}} H_\lambda) &= \mathrm{tr}\{U(I_n + \lambda D)^{-1} U^{\mathrm{T}} U (I_n + \lambda D)^{-1} U^{\mathrm{T}}\} \\
&= \mathrm{tr}\{U^{\mathrm{T}} U (I_n + \lambda D)^{-2}\} \\
&= \sum_{j=1}^{n} (1 + \lambda d_j)^{-2} \\
&< \sum_{j=1}^{n} (1 + \lambda d_j)^{-1} \\
&= \mathrm{tr}(H_\lambda),
\end{aligned}$$

since $1 + d_j\lambda \geq 1$, with some terms strictly larger than unity. This implies that the fitted model with $H_\lambda^{\mathrm{T}} H_\lambda$ will be smoother, because its equivalent degrees of freedom are smaller.

——————— END OF THE EXAM PAPER ———————