Regression Methods: Examination 2023

1 February 2023

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 Consider a linear model for a response vector $y_{n\times 1}$ that satisfies

$$\mathrm{E}(y) = X\beta, \quad \mathrm{var}(y) = \sigma^2 I_n,$$

where the known matrix $X_{n\times p}$ is of rank $p < n$, $\beta_{p\times 1}$ is to be estimated, and $\sigma^2 > 0$.

(a) Find the estimator $\hat{\beta}$ that minimises the Euclidean distance $\|y - X\beta\|$.

(b) Show that the fitted values $\hat{y} = X\hat{\beta}$ may be written as Hy, where the $n \times n$ matrix H is symmetric and idempotent, and interpret this in terms of the geometry of $\mathbb{R}^n$.

(c) If y has a normal distribution, show that $\hat{y}$ is independent of $e = y - \hat{y}$ and by writing $y = \hat{y} + e$ show that $\hat{\beta}$ is independent of $e^{\mathrm{T}}e$.

(d) Derive the distributions of $\hat{\beta}$ and of $e^{\mathrm{T}}e$ when y has a normal distribution.

Hint: For (d) it may be helpful to write $I_n - H = U_{n\times n}D_{n\times n}U^{\mathrm{T}}_{n\times n}$, where the properties of D and U should be given.

Problem 2 In an investigation on the teaching of arithmetic, 45 pupils were divided at random into five groups of nine. Groups A and B were taught in separate classes by the usual method. Groups C, D, and E were taught together for a number of days. On each day C were praised publicly for their work, D were publicly reproved and E were ignored. At the end of the period all pupils took a standard test, leading to the table below.

Group	Test result y									Average	Variance
A (Usual)	17	14	24	20	24	23	16	15	24	19.67	17.75
B (Usual)	21	23	13	19	13	19	20	21	16	18.33	12.75
C (Praised)	28	30	29	24	27	30	28	28	23	27.44	6.03
D (Reproved)	19	28	26	26	19	24	24	23	22	23.44	9.53
E (Ignored)	21	14	13	19	15	15	10	18	20	16.11	13.11

(a) Let y_{jg} denote the test result for pupil $j \in \{1, \ldots, 9\}$ in group $g \in \{A, \ldots, E\}$. Express the distributional specification $y_{gj} \overset{\text{ind}}{\sim} \mathcal{N}(\alpha + \beta_g, \sigma^2)$ as a linear model, giving the form of the 45×6 design matrix X. Describe problems that might arise in fitting this model, and suggest potential solutions.

There is no need to write out the entire design matrix.

(b) The output from a fit of this linear model is given below:

```
Call:  lm(formula = y ~ group, data = arithmetic)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1111 -2.6667  0.5556  2.5556  4.8889

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   19.667      1.147  17.151  < 2e-16 ***
groupB        -1.333      1.622  -0.822   0.4158
groupC         7.778      1.622   4.796 2.26e-05 ***
groupD         3.778      1.622   2.330   0.0250 *
groupE        -3.556      1.622  -2.193   0.0342 *
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 3.44 on 40 degrees of freedom
Multiple R-squared:  0.6042,Adjusted R-squared:  0.5647
F-statistic: 15.27 on 4 and 40 DF,  p-value: 1.163e-07
```

(i) Explain what is estimated in the `Estimate` column.

(ii) What does the output suggest about the relative success of the teaching methods?

(iii) Suggest how the model might be changed to increase the precision of estimation.

Problem 3

(a) What do you understand by the phrase *degrees of freedom of a regression model fit*?

(b) A regression model fit to independent response variables $y^{\mathrm{T}} = (y_1, \ldots, y_n)$ with equal variances σ^2 provides fitted values $\hat{y}^{\mathrm{T}} = (\hat{y}_1, \ldots, \hat{y}_n)$. Let $m(y, \hat{y}) = \sum_{j=1}^{n} \mathrm{cov}(y_j, \hat{y}_j)/\sigma^2$. Discuss the situations where y_j and $\hat{y}_j$ are independent and where $\hat{y}_j = y_j$.

(c) Find $m(y, \hat{y})$ for a linear model with p linearly independent covariates.

(d) A penalized least squares fit sets $\hat{y} = Sy$, where $S = X(X^{\mathrm{T}}X + \lambda D)^{-1}X^{\mathrm{T}}$ and $\lambda > 0$. Under suitable conditions on $X_{n\times p}$ and $D_{p\times p}$, show that the corresponding function $m(y, \hat{y})$ is monotonic decreasing in λ.

Hint: You may use without proof the result that if $A_{q\times q}$ and $B_{q\times q}$ are positive semidefinite, $A+\lambda B$ is invertible for some $\lambda > 0$, η is an eigenvalue of $(A+\lambda B)^{-1}A$, and (i) if B is invertible, then $\eta = \eta'/(\lambda + \eta')$, where η' is an eigenvalue of $B^{-1/2}AB^{-1/2}$, (ii) if A is invertible, then $\eta = 1/(1 + \lambda\eta'')$, where η'' is an eigenvalue of $A^{-1/2}BA^{-1/2}$.

Problem 4 Independent Bernoulli variables $Y_1, \ldots, Y_n$ satisfy

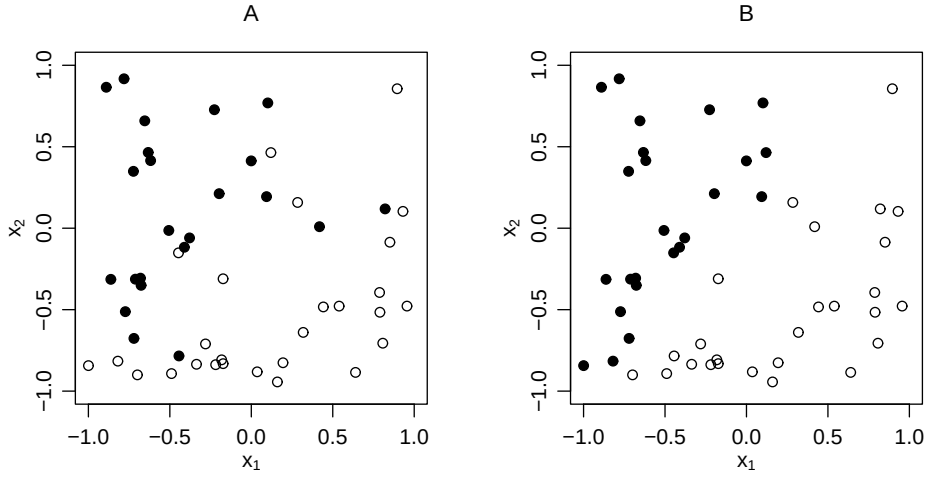
$$\Pr(Y_j = 1) = 1 - \Pr(Y_j = 0) = \frac{1}{1 + \exp(-x_j^{\mathrm{T}}\beta)},$$

where $x_1, \ldots, x_n$ are known $p \times 1$ vectors of constants and the unknown parameter $\beta \in \mathbb{R}^p$ is to be estimated from the observed values $y_1, \ldots, y_n$ of $Y_1, \ldots, Y_n$.

(a) Find the log likelihood $\ell(\beta)$ and show that it is never positive.

(b) If there exists a vector γ such that $x_j^{\mathrm{T}}\gamma > 0$ when $y_j = 1$ and $x_j^{\mathrm{T}}\gamma < 0$ when $y_j = 0$ $(j = 1, \ldots, n)$, then show by considering $\ell(t\gamma)$, where t is scalar, that the maximum likelihood estimate has at least one component that equals $\pm\infty$. What is then the value of the log likelihood?

(c) The figure below shows two sets of observed responses as solid black ($y_j = 1$) or as a circle ($y_j = 0$) for data with $n = 50$ and $p = 2$. In each case say whether you expect to have difficulties with likelihood estimation, and explain what you would expect when fitting a model.



Problem 5 What problems might arise in fitting regression models with Poisson response distributions to over-dispersed count data? How would they manifest themselves, and what solutions might you propose?

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