Modern Regression: Examination 2022

5 July 2022

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 number:

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

Problem 1

(a) State the assumptions underlying the application of a linear model to data. Which of these assumptions are primary, and which are secondary?

(b) A linear model with mean vector $X_{n\times p}\beta_{p\times 1}$, with $n > p$ and X of rank p, is fitted to a response vector $y_{n\times 1}$. Explain the role of the vector $e = y - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ in checking the model assumptions, and by considering its mean and variance suggest a preferable basis for model checking.

(c) It is suspected that the mean of y is in fact $X\beta + Z_{n\times q}\gamma_{q\times 1}$. Explain what might be learned by plotting e against the columns of the matrix $Q = \{I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\}Z$.

(d) The figure below shows such plots for three columns of such a matrix Q. How does y depend on the corresponding columns of Z?

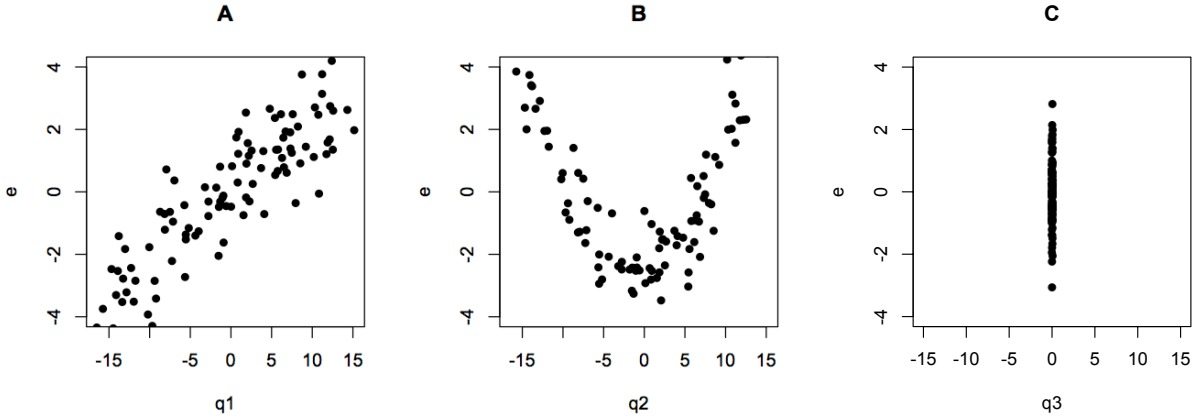

Figure 1: Plots for part (d) of question 1.

Problem 2

(a) Under what circumstances might you use a penalised linear regression? What changes to the response vector and design matrix would you make first?

(b) Use the singular value decomposition of the design matrix X to show that the value of β that minimises the expression

$$(y - X\beta)^{\mathrm{T}}(y - X\beta) + \lambda\beta^{\mathrm{T}}\beta$$

can be written in the form

$$\hat{\beta}_\lambda = \sum_{j:d_j>0} v_j \times \frac{d_j}{d_j^2 + \lambda} u_j^{\mathrm{T}} y,$$

and discuss how the parameter estimate $\hat{\beta}_\lambda$ and corresponding fitted values $\hat{y}_\lambda$ vary with λ.

(c) Find the 'hat matrix' of this regression, and show that the corresponding equivalent degrees of freedom are monotonic decreasing in λ.

Hint: Recall the singular value decomposition $U_{n\times n}D_{n\times p}V_{p\times p}^{\mathrm{T}}$ of a matrix $X_{n\times p}$, with $U = (u_1, \ldots, u_n)$ and $V = (v_1, \ldots, v_p)$ orthogonal and D (rectangular) diagonal with elements $d_1 \geq \cdots \geq d_m \geq 0$, where $m = \min(n, p)$.

Problem 3

(a) The table below shows data from a survey on the prevalence of upper respiratory tract infection. The response is the number of swabs positive for pneumococcus during a certain period. Observations were made on 18 families, each consisting of a father, mother and three children, the youngest of whom was always a pre-school child; the children are numbered in decreasing order of age. The families were randomly selected from those living in 'overcrowded' conditions, in 'crowded' conditions and in 'uncrowded' conditions.

Table 1: Number of swabs positive for pneumococcus during fixed periods

Crowding category	Family number	Family status					
				Child			
		Father	Mother	1	2	3	Total
Overcrowded	1	5	7	6	25	19	62
	2	11	8	11	33	35	98
	3	3	12	19	6	21	61
	4	3	19	12	17	17	68
	5	10	9	15	11	17	62
	6	9	0	6	9	5	29
		41	55	69	101	114	380
Crowded	7	11	7	7	15	13	53
	8	10	5	8	13	17	53
	9	5	4	3	18	10	40
	10	1	9	4	16	8	38
	11	5	5	10	16	20	56
	12	7	3	13	17	18	58
		39	33	45	95	86	298
Uncrowded	13	6	3	5	7	3	24
	14	9	6	6	14	10	45
	15	2	2	6	15	8	33
	16	0	2	10	16	21	49
	17	3	2	0	3	14	22
	18	6	2	4	7	20	39
		26	17	31	62	76	212
Total		106	105	145	258	276	890

Which of the terms in the linear model

$$y_{fs} = \eta_f + \alpha_s + \beta_c + \gamma_{cs} + \varepsilon_{fs}, \quad f = 1, \ldots, 18, s = 1, \ldots, 5, c = 1, 2, 3,$$

where $c \equiv c(f)$ denotes the Crowding category for Family f and s denotes Status, would you treat as random effects, and which as fixed effects? Justify your reasoning.

(b) The table below shows an analysis of variance for the data above. Explain the structure of the table and give a careful interpretation of it. Do Crowding and Status affect the response? Why are they compared to different lines marked 'Residual'?

	Sum of squares	Degrees of freedom	Mean square	Variance ratio against	
				a	b
Between families	1146.09	17			
Crowding	470.49	2	235.24		5.22*
Residual	675.60	15	45.04[b]	1.78	1.00
Within families	3122.80	72			
Status	1533.67	4	383.42	15.17†	
Status×Crowding	72.40	8	9.05	0.36	
Residual	1516.73	60	25.28[a]	1.00	
Total	4268.89	89			

$*$: $0.01 < P < 0.05$
† : $P < 0.01$

(c) The averages for the different family members are

Father	Mother	Child		
		1	2	3
5.9	5.8	8.1	14.3	15.3

Explain why the standard error for differences of these averages is $\sqrt{2 \times 25.28/18}$, and briefly comment on the Status effect.

Problem 4

(a) The table below shows an extract from data on the numbers of traffic accidents with personal injuries reported to the police on Swedish roads on 92 days in 1961 and 92 matching days in 1962. On some of these days a general speed limit was imposed. The goal of the experiment was to assess whether imposing a limit affected the number of such accidents. Discuss the suitability of treating the numbers of accidents on day j in year $196i$ as independent Poisson variables with means

$$\mu_{ij} = \exp(\alpha_i + \beta_j + \gamma I_{ij}), \quad i = 1, 2, \quad j = 1, \ldots, 92,$$

where the indicator variable I_{ij} equals 1 when a speed limit is imposed on day j of year i and otherwise equals 0. How should the parameters in this model be interpreted?

Day j	I_{1j}	I_{2j}	y_{1j}	y_{2j}
1	0	0	9	9
2	0	0	11	20
		⋮		
13	0	1	28	16
14	0	1	17	20
		⋮		
26	1	1	15	22
27	1	1	18	24
		⋮		
44	1	0	16	24
45	0	0	17	18
		⋮		
91	0	1	8	15
92	0	1	21	9

(b) Find the conditional distribution of y_{1j} given that $y_{1j} + y_{2j} = m_j$.

(c) The following output summarises the fit of a generalized linear model to these data.

```
Call:
glm(formula = cbind(y1, y2) ~ I(lim1 - lim2), family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4060  -0.7515   0.0306   0.7735   2.8227

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
Intercept     0.02904    0.03461   0.839    0.401
lim1 - lim2  -0.29169    0.04307  -6.772 1.27e-11

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 154.24  on 91  degrees of freedom
Residual deviance: 107.95  on 90  degrees of freedom
```

Explain how this fit is related to the computation in (b), and give an interpretation of the output. In particular: (i) discuss whether there is evidence that $\alpha_1 \neq \alpha_2$; (ii) whether imposing the speed limit has an effect; (iii) give 95% confidence limits for any effect of the speed limit on the mean number of accidents; and (iv) discuss the fit of the model.

Problem 5 The log likelihood function for data $y_1, \ldots, y_n$ believed to come from a parametric statistical model that is regular for maximum likelihood estimation is of the form

$$\ell(\beta) = \sum_{j=1}^{n} \log f\{y_j; \eta_j(\beta)\},$$

where $f(y_j; \eta_j)$ denotes the density function for y_j and β is a $p \times 1$ vector of unknown real-valued parameters.

(a) If the maximum likelihood estimate $\hat{\beta}$ is known to satisfy the score equation

$$\frac{\partial \ell(\hat{\beta})}{\partial \beta} = 0,$$

derive the iterative weighted least squares algorithm to obtain $\hat{\beta}$.

(b) Find the components of the algorithm when $y_1, \ldots, y_n$ are independent with densities of the form

$$f(y; \eta) = \exp\{y\eta - \kappa(\eta)\}, \quad y \in \mathcal{Y}, \quad \eta \in \Theta,$$

for suitable sets $\mathcal{Y}$ and Θ, and $\eta_j = x_j^{\mathrm{T}} \beta$ for vectors x_j of explanatory variables.

(c) Briefly discuss any potential problems when applying the algorithm derived in (b), and say how you might overcome them.