Modern Regression: Examination 2020

26 August 2020

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: $a_+ = \max(a, 0)$ for $a \in \mathbb{R}$; $A_{r\times s}$ 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

1. The penalized log likelihood function for independent observations $y_1, \ldots, y_n$ believed to come from a parametric statistical model that is regular for likelihood inference may be written in the form

$$\ell_{\mathrm{p}}(\beta) = \sum_{j=1}^{n} \ell_j\{\eta_j(\beta)\} - \tfrac{1}{2}\beta^{\mathrm{T}} D_\lambda \beta,$$

where β is a $p \times 1$ vector of unknown parameters and the $p \times p$ positive semi-definite symmetric matrix D_λ is known.

   (a) Give two examples of circumstances under which it would be appropriate to include a penalty term such as $\frac{1}{2}\beta^{\mathrm{T}} D_\lambda \beta$ when maximising a log likelihood.

   (b) Derive a penalized iterative weighted least squares algorithm to maximise $\ell_{\mathrm{p}}(\beta)$ with respect to β.

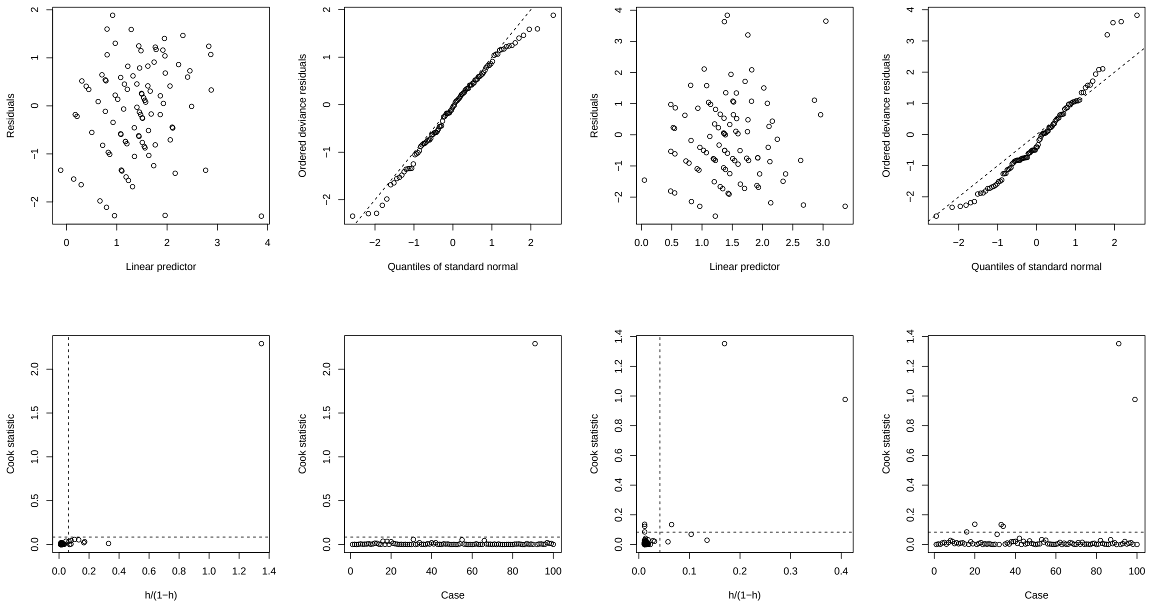
   (c) If $\ell_j(\eta_j) = -\frac{1}{2}(y_j - \eta_j)^2$ and $\eta_j = x_j^{\mathrm{T}}\beta$, show that your algorithm converges in one step. In what way(s) is this setting unrealistic?

2. The plots below show diagnostic quantities computed for two log-linear models fitted to independent observations $y_1, \ldots, y_n$, using a Poisson response distribution.

   (a) Describe the quantities on the axes of each of the four panels of (a), and explain how these plots are used in detecting problems with the fitted model.

   (b) Which of the two models do you think fits the data better? Justify your conclusions, giving your full reasoning.

   (c) What further plots would you wish to see, to determine if your reasoning in (b) is correct?



(a) Linear predictor $x_1 + x_2$

(b) Linear predictor x_1

Figure 1: Diagnostic plots for two different log-linear models fitted to Poisson response data.

3. Independent continuous observations $Y_1, \ldots, Y_n$ arise from a regression model

$$Y_j = x_j^{\mathrm{T}}\gamma + \sigma\varepsilon_j, \quad \varepsilon_j \overset{\text{iid}}{\sim} F, \quad j = 1, \ldots, n,$$

but only the $p \times 1$ covariate vectors $x_1, \dots, x_n$ and the values z_j of the indicator variables $Z_j = I(Y_j > 0)$ are observed.

(a) Find the probability that $Z_j = 1$. Are the parameters γ and σ identifiable? Explain your reasoning.

(b) Derive the likelihood based on $z_1, \dots, z_n$, and put it into generalized linear model form, identifying the response distribution and the link function in the special cases where (i) $\varepsilon_j \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, (ii) the ε_j have distribution function $\exp\{-\exp(-u)\}$, for $u \in \mathbb{R}$.

(c) Now suppose that the observed variables are the indicators that $Y_j \in \mathcal{I}_k$, where

$$\mathcal{I}_1 = (-\infty, \zeta_1], \quad \mathcal{I}_2 = (\zeta_1, \zeta_2], \quad \dots, \quad \mathcal{I}_K = (\zeta_{K-1}, \infty), \quad K > 2,$$

and $\zeta_1 < \cdots < \zeta_{K-1}$ are unknown. In what circumstances would such a model be useful? Give an expression for the corresponding likelihood function. If $x_j^{\mathrm{T}}\gamma = \gamma_0 + \gamma_1 x_j$, which of the parameters are estimable? Explain.

4. In a cooperative trial, research groups from different institutions work together in order to establish best practice in their field. The Analytical Methods Committee of the UK Royal Society of Chemistry has the broad aim of establishing a comprehensive framework for appropriate quality in chemical measurement. In order to do so, it conducted a cooperative trial in which seven specimens were sent to six laboratories in three separate batches at one-month intervals, and two replicate analyses were performed on each occasion. The response is the concentration of (unspecified) analyte in g/kg. The purpose of the study was to assess components of variation in such trials, and for this purpose, the laboratories and batches are regarded as random.

(a) A possible model for the response for specimen s, laboratory l, batch b and replicate r is

$$y_{slbr} = \mu_s + \alpha_{sl} + \beta_{slb} + \varepsilon_{slbr}, \quad s = 1, \dots, S, l = 1, \dots, L, b = 1, \dots, B, r = 1, \dots, R.$$

Give an interpretation for each of the terms on the right-hand side of this expression, explain which of them you would treat as random, and suggest suitable distributions for them.

(b) In such trials the data for each specimen are typically analysed separately, so the index s can be dropped and the analysis of variance table for a single specimen may be written as

Source of variation	df	Sum of squares	Mean square, MS	E(MS)
Between laboratories	$L-1$	$BR\sum_{l=1}^{L}(\bar{y}_{l\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$	MS_L	$\sigma^2 + R\sigma_B^2 + BR\sigma_L^2$
A	$L(B-1)$	$R\sum_{l=1}^{L}\sum_{b=1}^{B}(\bar{y}_{lb\cdot} - \bar{y}_{l\cdot\cdot})^2$	MS_B	B
Between replicates within batches	C	$\sum_{l=1}^{L}\sum_{b=1}^{B}\sum_{r=1}^{R}(y_{lbr} - \bar{y}_{lb\cdot})^2$	MS_R	D

Give the missing parts A, B, C, D of the table. How would you estimate the components of variance?

(c) In the trial mentioned above, it was found that for Specimen 1, $\mathrm{MS}_L = 0.378$, $\mathrm{MS}_B = 0.017$ and $\mathrm{MS}_R = 0.0063$. Give a numerical estimate of the variance for a single additional measurement on this specimen for a randomly chosen laboratory and batch.

5. A basic epidemiological quantity is the instantaneous reproduction number R_t for a population, defined as the average number of secondary cases that each infected individual would infect if the conditions remained as they were on day t. A simple model for this is that if the number of cases on day t is Y_t, then

$$Y_{t+1} \mid Y_t = y_t, Y_{t-1} = y_{t-1}, \ldots \sim \mathrm{Pois}\left(\mu_{t+1}\right), \quad \mu_{t+1} = R_t \sum_{s=0}^{S} w_s y_{t-s}, \quad t = 1, \ldots, n-1,$$

where w_s denotes the infectivity of an infected person s days after their infection, and $\sum_{s=0}^{S} w_s = 1$.

(a) Show that if $\mu_{t+1} > 0$ then we can write $\log \mu_{t+1} = x_t + \beta_t$ where the x_t can be treated as known. Hence show that for some t_0 the log likelihood for $\beta_{t_0}, \ldots, \beta_{n-1}$ may be written as

$$\ell(\beta_{t_0}, \ldots, \beta_{n-1}) = \sum_{t=t_0}^{n-1} \left\{ y_{t+1}(x_t + \beta_t) - \exp(x_t + \beta_t) \right\}.$$

Find the maximum likelihood estimators of the β_t, and discuss their drawbacks.

(b) Figure 2 shows the fit of a *generalized additive model* with *Poisson response distribution* and 8.9 *equivalent degrees of freedom* to data on Covid-19, using a known infectivity function. Explain the meaning of the terms in italics in the previous sentence. How might such a model improve on the approach in (a)? Do you find the figure credible? How might the approach be improved?

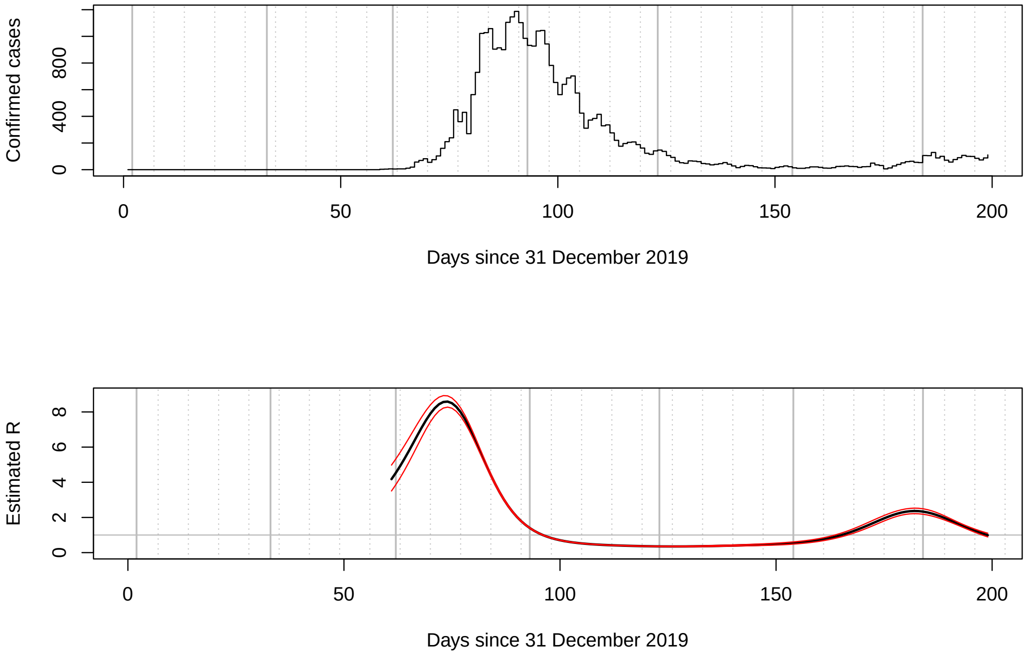


Figure 2: Top panel: daily confirmed cases of Covid-19 for Switzerland to mid-July 2020. The solid grey vertical lines mark the first day of a month, the dotted grey vertical lines mark Mondays. Lower panel: estimated value of R_t, with 95% pointwise confidence intervals. The grey horizontal line marks $R_t = 1$, above which the expected number of cases increases, and below which the expected number of cases decreases.