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

Solution 1

(a) [2, seen] We expect some subset of the material on slides 18, 19, 36. Primary aspects concern the question being asked; in this case this is the linearity of the model, ie., $\mathrm{E}(y) = X\beta$. Secondary aspects concern how one gets an answer, so these are variance formulation, further distributional assumptions (e.g., for $\mathrm{var}(y)$).

(b) [3, seen] e is the vector of raw residuals, used for all kinds of model checking (e.g., slide 37; a good reply will list the plots mentioned there). Under the second order assumptions we have $e = (I_n - H)y$ and if $y \sim (X\beta, \sigma^2 I_n)$, then in terms of the hat matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ we find that $\mathrm{E}(e) = 0$ and $\mathrm{var}(e) = \sigma^2(I_n - H)$, so (in the usual notation) the standardised residuals $r_j = e_j/\sqrt{s^2(1 - h_{jj})}$ are preferable, as they have constant variance.

(c) [2, unseen] Fitting the model $y \sim (X\beta, \sigma^2 V)$ will give residuals $e = (I_n - H)y$ whose expected values will be $(I_n - H)E(y) = (I_n - H)(X\beta + Z\gamma) = (I_n - H)Z\gamma = Q\gamma$. The columns of Q represent the residual variation in Z after removing any linear dependence on the columns of X, so plotting e against the columns of Q shows how y varies with Z, after removing their mutual linear dependence on X.

(d) [3, unseen] Let z_r denote the rth column of Z. Panel A shows positive linear dependence on z_1 after allowing for X. Panel B shows quadratic dependence on z_2 after allowing for X. Panel C shows that the third column of Q is essentially constant, so z_3 must be almost a linear function of the columns of X, i.e., this shows collinearity between X and z_3. Hence nothing more can be explained by adding z_3 to the regression.

Solution 2

(a) [3, seen] We use penalties to improve prediction and/or stabilise estimates when explanatory variables X may be (almost) collinear (e.g., when $p > n$ so we have a 'wide' regression), for variable selection (e.g., with the lasso), as a result of a Bayesian analysis (the penalty stems from the prior), or in a mixed model.

When using the lasso and/or ridge methods we usually centre (and sometimes scale) the response and design matrices, i.e., replace y and X by $y - \bar{y}1_n$ and $X - 1_n\bar{x}$, where $\bar{y}$ and $\bar{x}$ contain the row means of y and X.

(b) [4, seen] See slides 88 and 93; note that we do not write $D^{\mathrm{T}}D = D^2$ because D is not necessarily square. The given expression shows that as $\lambda \to \infty$, $\hat{\beta}_\lambda \to 0$, and we have

$$\hat{y}_\lambda = X\hat{\beta}_\lambda = UDV^{\mathrm{T}} \times V(D^{\mathrm{T}}D + \lambda I_p)^{-1}D^{\mathrm{T}}U^{\mathrm{T}}y = \sum_{j:d_j>0} u_j \times \frac{1}{1 + \lambda/d_j^2} u_j^{\mathrm{T}} y \to 0.$$

(c) [3, seen] The hat matrix is $H_\lambda = UD(D^{\mathrm{T}}D + \lambda I_p)^{-1}D^{\mathrm{T}}U^{\mathrm{T}}$ and the usual definition of equivalent degrees of freedom is

$$\mathrm{tr}(H_\lambda) = \sum_j \frac{d_j^2}{d_j^2 + \lambda},$$

which is monotone decreasing in λ.

Solution 3

(a) [3, unseen] The description makes it clear that families are randomly selected, so the η_f are random, say $\mathcal{N}(0, \sigma_f^2)$, and the 'errors' ε_{fs} are also random, say $\mathcal{N}(0, \sigma^2)$. Status and crowding are not selected at random, so the parameters α_s, β_c and the interaction γ_{cs} are fixed.

Judging from the table there are strong effects of Status and Crowding (the marginal averages vary a lot), but it is not so clear whether the effect of Status depends on the level of Crowding (i.e., whether the γ_{cs} are non-zero).

(b) [5, unseen] Variation between families is based on the averages

$$\bar{y}_{f\cdot} = \eta_f + \bar{\alpha}_{\cdot} + \beta_c + \bar{\gamma}_{c\cdot} + \bar{\varepsilon}_{f\cdot}, \quad f = 1, \ldots, 18,$$

which are independent $\mathcal{N}(\bar{\alpha}_{\cdot} + \beta_c + \bar{\gamma}_{c\cdot}, \sigma_f^2 + \sigma^2/5)$ variables, so the three Category averages

$$\bar{y}_{c\cdot}^{C} = \frac{1}{6} \sum_{f=6(c-1)+1}^{6c} \bar{y}_{f\cdot} \overset{\text{ind}}{\sim} \mathcal{N}(\bar{\alpha}_{\cdot} + \beta_c + \bar{\gamma}_{c\cdot}, \sigma^2/6_f + \sigma^2/30), \quad c = 1, 2, 3$$

have different means, but (for example) the sum of squares for families in Category 1 satisfies

$$\sum_{s=1}^{5} \sum_{f=1}^{6} (\bar{y}_{f\cdot} - \bar{y}_{1\cdot}^{C})^2 = 5 \sum_{f=1}^{6} (\bar{y}_{f\cdot} - \bar{y}_{1\cdot}^{C})^2 \overset{\text{D}}{=} 5\chi_5^2(\sigma_f^2 + \sigma^2/5) \overset{\text{D}}{=} (5\sigma_f^2 + \sigma^2)\chi_5^2,$$

and the sum of these for the three categories therefore has 15 df and its mean square will estimate $5\sigma_f^2 + \sigma^2$. This is the residual at the first level, and it is the basis for testing hypotheses about differences between Crowding categories (significant at between 0.01 and 0.05, according to the ANOVA).

Comparisons within families (e.g., on status) involve differences such as

$$\bar{y}_{\cdot 1} - \bar{y}_{\cdot 2},$$

from which the family effects η_f disappear, so the comparison will be based on the residual sum of squares within families, which can be used to estimate σ^2. This is the basis for assessing the presence of β_s (which show a highly significant effect of Status) and γ_{fs} (which show no significant variation, i.e., no interaction between Status and Crowding).

(c) [2, unseen] The variance of the difference of the averages for the father and mother (for example),

$$\bar{y}_{\cdot 1} - \bar{y}_{\cdot 2},$$

is $\sigma^2/18 + \sigma^2/18$ (recall that the η_f disappear from this difference, which is comparing members of the same family), which gives the stated standard error if we repace σ^2 by the mean square 25.28. The value of this is 1.676, so there are clearly no significant differences between the parents and the first child, or between the last two children, but equally clearly Children 2 and 3 have higher averages than the rest of the family.

Solution 4

(a) [2, seen/unseen] Treating accident numbers as Poisson variables is quite common (though one needs to watch out for excess zeros and/or overdispersion). The parameters α_i correspond to year effects, the β_j to day effects, and γ to an effect of the speed limit that is the same for both years.

(b) [3, seen/unseen] See slides 271–272 and/or problem 39. Here we expect the binomial distribution to be derived for full points (the question says 'find', not 'state' or 'give').

(c) [5, unseen] Following the argument in (b) we find that

$$y_{1j} \mid y_{1j} + y_{1j} = m_j \overset{\text{ind}}{\sim} B(m_j, p_j), \quad p_j = \frac{e^{\alpha_1 - \alpha_2 + \gamma(I_{1j} - I_{2j})}}{1 + e^{\alpha_1 - \alpha_2 + \gamma(I_{1j} - I_{2j})}},$$

so the intercept in the binomial logistic model estimates $\alpha_1 - \alpha_2$ and the coefficient of `lim1-lim2` (which corresponds to $I_{1j} - I_{2j}$) estimates γ.

(i) There seems to be no evidence that $\alpha_1 \neq \alpha_2$.

(ii) There is very strong evidence of a speed limit effect (the z-test gives -6.78, which is massive when compared to the standard normal distribution).

(iii) $\hat{\gamma} \doteq -0.29$, so $e^{\hat{\gamma}} \approx 0.75$, corresponding to a one-quarter reduction of the mean number of accidents on days with the limit. The corresponding approximate 95% confidence intervals for γ and e^{γ} are $-0.29 \pm 1.96 \times 0.043 = (-0.374, -0.205)$ and $(0.69, 0.81)$. Applying the limit has a very significant downward effect on the number of these accidents.

(iv The residual deviance D is around 108 on 90 degrees of freedom, which is not at all unusual: with 90 df we have $D \dot{\sim} \mathcal{N}(90, 2 \times 90)$ so the significance level of the observed D is around $1 - \Phi\{(108 - 90)/\sqrt{180}\} \doteq 0.1$.

The fit appears to be adequate, though one would like to inspect residual plots to be sure, because the residuals may be a bit more dispersed than we would expect. This is not surprising, as we would expect there to be some overdispersion relative to the Poisson model, though this will probably not affect the overall conclusions.

Solution 5

(a) [4, seen] See slides 229–231. We expect a full derivation here, with all details (no bluffing, no gaps).

(b) [4, seen/unseen] We need

$$X = \frac{\partial \eta}{\partial \beta^{\mathrm{T}}}, \quad W = \mathrm{diag}\{-\mathrm{E}\{\partial^2 \ell(\eta)/\partial \eta^2\} = \mathrm{diag}\{\kappa''(\eta_1), \ldots, \kappa''(\eta_n)\}, \quad u = y - \kappa'(\eta),$$

and the adjusted dependent variable is $z = X\beta + W^{-1}u = \eta + (y - \kappa'(\eta)\}/\kappa''(\eta)$. Probably we will get the initial value of η as $\kappa'^{-1}(y)$.

(c) [2, seen] If the step goes outside Θ, then we might need to decrease the step length. But this is an exponential family, so the log likelihood is concave unless Θ is restricted in some way.