Modern Regression: Solution 2020

26 August 2020

1.

(a) **2pt (seen)**

- Linear mixed model where β contains both fixed-effects and random-effects, the latter penalized through the matrix D_λ. This also can be interpreted as a Bayesian setup where there is a
- roughness penalty placed on the coefficients of some spline basis, such as P-splines (diagonal matrix), B-splines (banded matrix corresponding to a quadratic form of finite differences in the components of β).

(b) **5pt (seen)** We apply the argument leading to the PIWLS algorithm to the penalized log likelihood $\ell_{\mathrm{p}}(\beta)$. This leads to

- Set $i = 0$, and set initial values $\beta_i = \beta_{p\times 1}$.
- Repeat until converged
  - compute

    $$X_{n\times p} = \frac{\partial \eta}{\partial \beta^{\mathrm{T}}}(\beta_i), \quad W_{n\times n} = \mathrm{diag}(w_1, \dots, w_n), \quad u_{n\times 1} = \frac{\partial \ell}{\partial \eta}, \quad z_{p\times 1} = X\beta_i + W^{-1}u,$$

    where $w_j(\eta_j) = \mathrm{E}\{-\partial^2 \ell_j / \partial \eta_j^2\}$ (if available) and otherwise $w_j(\eta_j) = -\partial^2 \ell_j / \partial \eta_j^2$;
  - compute new ridge regression estimate of β as

    $$\beta_{i+1} = (X^{\mathrm{T}}WX + D_\lambda)^{-1}X^{\mathrm{T}}Wz;$$
  - check $|\ell_p(\hat\beta_{j+1}) - \ell_p(\hat\beta_j)| < \epsilon$ or $|\hat\beta_{j+1}) - \hat\beta_j)| < \epsilon$ for some suitably small ϵ to see if the steps have converged;
  - if not converged, set $i = i + 1$ and iterate again.

The components of the PIWLS are respectively

$$\eta_j, \quad l_j = l(y_j, \eta_j) = \log f(y_j; \beta), \quad u_j(\eta_j) = \frac{\partial l_j}{\partial \eta_j}, \quad w_{jj} = -\frac{\partial^2 l_j}{\partial \eta_j^2}$$

(c) **3pt (mostly seen)** In this situation $\eta_j = x_j^{\mathrm{T}}\beta$, and $W^{-1}u = (y - X\beta)$, so the iterative weighted least squares converges in a single step to the ridge regression estimate $(X^{\mathrm{T}}X + D_\lambda)^{-1}X^{\mathrm{T}}y$, whatever the choice of initial values. This is unrealistic because in applications λ measures the degree of penalisation of β and must be estimated, for example using generalized cross-valuation. This will involve more iterations to find the optimal λ.

2.

(a) **4pt (seen)** Describe the quantities on the axes of each of the four panels: residuals, deviance residuals, Cook statistic, leverage, normal order statistics, case (observation number). How are they used to assess the quality of the fit?

(b) **3pt (unseen)**
- Model 1 clearly fits the data better: the ordered residuals are close to standard normal, the residuals lie more or less in the random ± 2, and there seems to be one observation with a rather large Cook statistic, due to a high leverage value. For model 2 the quantile plot is not so good, suggesting overdispersion, presumably because the variable x_2 has some explanatory power, and leaving it out leads to somewhat poorer residuals.
- Points with a large Cook's distance are considered to merit closer examination in the analysis. In model one there is one point that has a large Cook statistic that corresponds to a large value of $h_{jj}/(1-h_{jj}) = 1.4 > 1$, and a large standardized residual $r_j < -2$, it is likely an influential point. Whether it should be dropped or not is moot, but at least one should check whether it seems to be a correct value.
- The second model has more points with $|r_j| > 2$, these points appear in the plot of C_j against $h_{jj}/(1-h_{jj})$, where we see that high value of C_j are not a result of high leverage $h_{jj}/(1-h_{jj}) < 0.4$, unusual observations.

(c) **3pt** (unseen)
Plotting the residuals from Model 2 against the values of x_2 is likely to show a trend, if there is a need for x_2 in the model (as seems clear from the plots for Model 1).

3.

(a) **3pt (seen)** $\Pr(Z_j = 1) = 1 - F\{-x_j^T(\gamma/\sigma)\}$, therefore $\beta = \gamma/\sigma$ is estimable, but γ and σ are unidentifiable. Knowing Z can only tell us about the probability that $Y > 0$, but it cannot tell us how spread out the distribution of Y is, i.e., it cannot tell us σ.

(b) **5pt (mostly unseen)**
The density of Z_j can be written in terms of $\pi_j = \Pr(Z_j = 1)$ as $\pi_j^{z_j}(1-\pi_j)^{1-z_j}$ and the Zs are independent, so the likelihood is

$$\prod_{j=1}^{n} \pi_j^{z_j}(1-\pi_j)^{1-z_j}.$$

Now $\mathrm{E}(Z_j) = \mu_j = \pi_j(-\eta_j) = 1 - F(-\eta_j)$, where $\eta_j = x_j\beta$ is the linear predictor. The response distribution for Z is binary and the link function is given by $\eta = g(\mu) = -F^{-1}(1-\mu)$.

In both cases (i) and (ii) the response distribution is binary (obviously).

In (i), F is the standard normal distribution Φ, so the link function is $g(\mu) = -\Phi^{-1}(1-\mu) = \Phi^{-1}(\mu)$, which is the probit link function.

In (ii), $F(\eta) = \exp\{-\exp(-\eta)\}$, so $g(\mu) = -F^{-1}(1-\mu) = \log\{-\log(1-\mu)\}$, which is the complementary log-log link function.

(c) **3pt (partly unseen)**
If we have n independent individuals whose responses $I_1, \dots, I_K$ fall into the set $\{1, \dots, K\}$, corresponding to K ordered categories, and that

$$\gamma_l = \mathrm{P}(I_j \le l) = \pi_1 + \cdots + \pi_l, \quad l = 1, \dots, K, \quad \gamma_K = 1,$$

this is an ordinal odds model, useful when there are several ordered categories (i.e., curry is mild, medium spicy, volcanic) and each individual (curry) is classified into one of them. There may be other variables x (e.g., the cook, the restaurant, tandoori, tikka, balti, ...) that can influence the category. The likelihood is is

$$\prod_{j=1}^{n}\prod_{k=1}^{K}\Pr(Y_j \in \mathcal{I}_k) = \prod_{j=1}^{n}\prod_{k=1}^{K}\pi_k^{I(Y_j\in\mathcal{I}_k)}$$

where

$$\pi_l = \mathrm{P}\left(\zeta_{l-1} < x_j^{\mathrm{T}}\beta + \sigma\varepsilon \le \zeta_l\right) = F\left(\frac{\zeta_l - x_j^{\mathrm{T}}\beta}{\sigma}\right) - F\left(\frac{\zeta_{l-1} - x_j^{\mathrm{T}}\beta}{\sigma}\right), \quad l = 1,\dots,K.$$

Here we see that if we map $\zeta_1,\dots,\zeta_{K-1} \mapsto \zeta_1 + a,\dots,\zeta_{K-1} + a$ and $\beta_0 \mapsto \beta_0 - a$, then the model is unchanged for any a, so we must set $\beta_0 = 0$ or fix one of the ζs to get an estimable model. In this case σ and γ_1 can both be estimated, as having $K > 2$ categories will provide information about the spread of ϵ.

4.

(a) **3pt (partly seen)** μ_s is the 'true' specimen mean, treated as fixed. α_{sl} is the mean amount by which the measurement of the specimen made by laboratory l would differ from μ_s, treated as random (laboratories are random). β_{slb} is the mean amount by which measurements on the bth batch of specimen s made by laboratory l would differ from the mean amount for that specimen and laboratory, $\mu_s + \alpha_{sl}$, treated as random (batches are random). ε_{slbr} is an error term, treated as random. Conventionally all the random terms are treated as independent normal variables with means zero and variances σ_α^2, σ_β^2 and σ^2 (though outliers are common in such models, and difficult to deal with).

(b) **5pt (partly seen)**
A= 'between batches within laboratories', B= $\sigma^2 + R\sigma_B^2$, C $=LB(R-1)$, and D=σ^2.

The components of variance are estimated by equating the numerical values of the mean squares (penultimate column) to their expectations (final column). These are method of moments estimates; the complication is that a negative estimate should be replaced by zero.

(c) **3pt (unseen)** Solve the equations, to obtain the estimates

$$\hat{\sigma}^2 = 0.0063, \quad \hat{\sigma}_B^2 =, \quad \hat{\sigma}_L^2 = .$$

The estimated variance for a single new observation on this specimen in a random lab and batch is then $\hat{\sigma}_L^2 + \hat{\sigma}_B^2 + \hat{\sigma}^2 =$.

5.

(a) **4pt (unseen)** $x_t = \log(\sum w_s y_{t-s})$ is a known offset and $\beta_t = \log R_t$, and we have written the joint density of $Y_{t_0}, \ldots, Y_n$ in the (prediction decomposition) form

$$\prod_{t=t_0}^{n-1} \Pr(Y_{t+1} = y_{t+1} \mid Y_t = y_t, \ldots, y_1 = 1).$$

Under the given distributional assumptions, this will be a product of conditional Poisson densities with log rates $\log \mu_{t+1} = x_t + \beta_t$, so the log likelihood will be

$$\sum_{t=t_0}^{n-1} \left(y_{t+1} \log \mu_{t+1} - \mu_{t+1} - \log y_{t+1}! \right),$$

which reduces to the expression given in the question on setting $\log \mu_{t+1} = x_t + \beta_t$ and noting that the constants $\log y_t!$ can be dropped.

For the MLE we just differentiate and obtain $\hat{\beta}_t = \log y_{t+1} - x_t$, which is obviously a terrible estimator, because it is based on just one day of data. Also the weights w_s have to be treated as known.

(b) **6pt (partly seen)** - Explain what this means? GAMs, Poisson response distribution, equivalent degrees of freedom (all bookwork).

The obvious idea is to use a penalty to ensure that the β_t vary smoothly, for example using a spline basis for them, with a suitable penalisation. Then rather than having a parameter for every day of data, we have many fewer equivalent parameters (just 8.9 for around 140 days of data, it appears from the plot). Then if the model is well-specified, the estimation of $R_t = \exp(\beta_t)$ will be MUCH better, and this is suggested by the quite narrow confidence bands in the figure.

Figure 1 looks reasonably plausible from around day 100, but before then it looks like the estimates of R_t are much too big. It's not really clear why this is. It's difficult to read the axis on the upper panel (log vertical axis might have been better), but the estimate of R_t seems to increase from around mid-June, while the numbers of cases increase from around the end of June, which seems plausible. The approach could be improved if one could estimate the weights w_s (which are here assumed to be constant over time), and it could also be used to estimate the effect of different restrictions (e.g., including indicators for total lockdown, ...).