STATISTICAL LEARNING IN PRACTICE Part III / Lent 2020
Example Sheet 1 (of 4) By courtesy of Dr T. Wang

You have the option to submit your answers to questions 2 and 3 to be marked. If you wish your answers to be marked, please leave them in my pigeon hole in the central core of the CMS preferably by 2pm on 11^{th} February, and no later than 11am on 12^{th} February.

1. Suppose $Y = X\beta + \epsilon$, where $X \in \mathbb{R}^{n\times p}$ has full rank $p \leq n$ and $\epsilon \sim N_n(0, \sigma^2 I_n)$.

   (i) Show that the maximum likelihood estimators for β and σ^2 are $\hat{\beta} = (X^\top X)^{-1}X^\top Y$ and $\hat{\sigma}^2 = n^{-1}\|Y - X\hat{\beta}\|_2^2$ respectively.

   (ii) Show that the vector of residuals satisfies $R = (I - P)Y$, where $P = X(X^\top X)^{-1}X^\top$.

   Suppose from now on that $p \leq n - 1$.

   (iii) Let $x_i^\top$ be the ith row of X and $\hat{\beta}^{(i)}$ the maximum likelihood estimator of β using all observations but the ith one. Show that

   $$\hat{\beta} - \hat{\beta}^{(i)} = \frac{(X^\top X)^{-1}x_i R_i}{1 - P_{ii}} \quad \text{and} \quad (\hat{\beta} - \hat{\beta}^{(i)})^\top (X^\top X)(\hat{\beta} - \hat{\beta}^{(i)}) = \frac{R_i^2 P_{ii}}{(1 - P_{ii})^2}.$$

   [*Hint: use* $(A + uv^\top)^{-1} = A^{-1} - (1 + v^\top A^{-1}u)^{-1}A^{-1}uv^\top A^{-1}$.]

   (iv) Explain how the leave-one-out cross-validation error $\mathrm{err}_{\mathrm{CV}}$ can be computed for this linear model. Show that the leave-one-out cross-validation error can be represented as a weighted mean squared residuals in this case:

   $$\mathrm{err}_{\mathrm{CV}} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i^2}{(1 - P_{ii})^2}.$$

Solution

(i) (Omitted)

(ii) (Omitted)

(iii) We have

$$\begin{aligned}
\hat{\beta} - \hat{\beta}^{(i)} &= (X^\top X)^{-1}(X^\top Y) - (X^\top X - x_i x_i^\top)^{-1}(X^\top Y - x_i Y_i) \\
&= (X^\top X)^{-1}X^\top Y - \left\{(X^\top X)^{-1} + \frac{(X^\top X)^{-1}x_i x_i^\top (X^\top X)^{-1}}{1 - x_i^\top (X^\top X)^{-1}x_i}\right\}(X^\top Y - x_i Y_i) \\
&= (X^\top X)^{-1}x_i Y_i - \frac{(X^\top X)^{-1}x_i x_i^\top (X^\top X)^{-1}(X^\top Y - x_i Y_i)}{1 - x_i^\top (X^\top X)^{-1}x_i} \\
&= \frac{(X^\top X)^{-1}x_i (Y_i - x_i^\top (X^\top X)^{-1}X^\top Y)}{1 - x_i^\top (X^\top X)^{-1}x_i} = \frac{(X^\top X)^{-1}x_i R_i}{1 - P_{ii}},
\end{aligned}$$

as desired. Consequently,

$$(\hat{\beta} - \hat{\beta}^{(i)})^\top (X^\top X)(\hat{\beta} - \hat{\beta}^{(i)}) = \frac{R_i^2 x_i^\top (X^\top X)^{-1} x_i}{(1 - P_{ii})^2} = \frac{R_i^2 P_{ii}}{(1 - P_{ii})^2}.$$

(iv) The leave-one-out cross-validation works by fitting the linear model on all data except the ith data point to obtain a coefficient estimate $\hat{\beta}^{(i)}$, then evaluate the fit on the remaining data point to obtain the cross-validation error on the ith observation $(Y_i - x_i^\top \hat{\beta}^{(i)})^2$. This is repeated for $i = 1, \ldots, n$ and the overall cross-validation error is the average of cross-validation errors for each observation. Thus,

$$\begin{aligned} \text{err}_{\text{CV}} &= \frac{1}{n} \sum_{i=1}^{n} (Y_i - x_i^\top \hat{\beta}^{(i)})^2 = \frac{1}{n} \sum_{i=1}^{n} \big\{ R_i + x_i^\top (\hat{\beta} - \hat{\beta}^{(i)}) \big\}^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \left\{ R_i + \frac{x_i^\top (X^\top X)^{-1} x_i R_i}{1 - P_{ii}} \right\}^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i^2}{(1 - P_{ii})^2} \end{aligned}$$

2. (i) Show that $\{\text{Poi}(\lambda) : \lambda \in (0, \infty)\}$ is an exponential dispersion family.

Suppose we observe data $Y_i \sim^{ind.} \text{Poi}(e^{x_i^\top \beta})$ for $i = 1, \ldots, n$, with $X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n \times p}$ of full rank $p \leq n$.

(ii) Show that the observed Fisher information matrix (i.e. negative Hessian of the log-likelihood) is positive definite at every β. Write down explicit expressions for one Newton–Raphson and one Fisher scoring iteration starting from an initial point β. Would they have been equal if the link were not canonical? Justify your answer.

(iii) Suppose a new observation is made with covariates $x_* \in \mathbb{R}^p$. Derive a (small-dispersion) asymptotic $1 - \alpha$ confidence interval for the associated mean response under the Poisson model.

(iv) What are the residual deviance and deviance residuals for the maximum likelihood fitted model? Show that when the fitted values are close to the observed values, the deviance residuals can be approximated by the Pearson's residuals $r_i = \hat{\lambda}^{-1/2}(Y_i - \hat{\lambda}_i)$, where $\hat{\lambda}_i = e^{x_i^\top \hat{\beta}}$.

Solution

(i) With respect to the counting measure on non-negative integers, $\text{Poi}(\lambda)$ has density

$$f(y; \lambda) = \frac{e^{-\lambda} \lambda^y}{y!} = \frac{1}{y!} e^{y \log \lambda - \lambda}.$$

Thus, $\{\text{Poi}(\lambda) : \lambda \in (0, \infty)\}$ is an exponential dispersion family with natural parameter $\log \lambda$ and dispersion parameter 1.

(ii) Writing $Y = (Y_1, \ldots, Y_n)^\top$, the log-likelihood is

$$\ell(\beta; Y) = c(Y) + \sum_{i=1}^{n} \big(Y_i x_i^\top \beta - e^{x_i^\top \beta} \big).$$

Differentiate with respect to β to obtain

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i \big(Y_i - e^{x_i^\top \beta}\big), \qquad H(\beta) := \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^\top} = -\sum_{i=1}^{n} x_i x_i^\top e^{x_i^\top \beta}.$$

To verify that the negative Hessian $-H(\beta)$ is positive definite, we compute that

$$-u^\top H(\beta) u = \|X\tilde{u}\|_2^2,$$

where $\tilde{u}_i = e^{x_i^\top \beta/2} u_i$. If $u \neq 0$, we have $\tilde{u} \neq 0$ and consequently $\|X\tilde{u}\|_2 \neq 0$ since X has full rank. Therefore, $-H$ is positive definite.
A Newton–Raphson update from β is of the form

$$\tilde{\beta} \leftarrow \beta + \left(\sum_{i=1}^{n} x_i x_i^\top e^{x_i^\top \beta}\right)^{-1} \sum_{i=1}^{n} x_i (Y_i - e^{x_i^\top \beta}).$$

For Fisher scoring, rather than using $-H(\beta)$ we use $\mathbb{E}[-H(\beta)]$. Since the former does not depend on the data, they are both equal and so are their updates from β.
If the link is not the canonical one, $\mathbb{E}[H(\beta)]$ may not equal $H(\beta)$. To argue this, one can do the (tedious) calculations for a general link or simply find a link where they are not equal: e.g. the identity link (which is not a sensible choice for this model, but possible as it satisfies the link assumptions of GLMs). For this link we have

$$\ell(\beta; Y) = c(Y) + \sum_{i=1}^{n} \big(Y_i \log(x_i^\top \beta) - x_i^\top \beta\big),$$

so

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i \left(\frac{Y_i}{x_i^\top \beta} - 1\right), \qquad H(\beta) := \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^\top} = -\sum_{i=1}^{n} x_i x_i^\top \frac{Y_i}{\big(x_i^\top \beta\big)^2}.$$

Therefore,

$$\mathbb{E}[-H(\beta)] = \sum_{i=1}^{n} x_i x_i^\top \frac{1}{x_i^\top \beta} \neq -H(\beta).$$

(iii) The associated response has distribution $Y_* \sim \text{Poi}(\exp\{x_*^\top \beta\})$. By Small Dispersion Asymptotics, $\hat{\beta} - \beta \approx^d N(0, (X^\top W X)^{-1})$ (the larger then means, the better the approximation), where $W = \text{diag}(e^{x_i^\top \beta} : i = 1, \ldots, n)$. Therefore,

$$x_*^\top (\hat{\beta} - \beta) \approx^d N(0, x_*^\top (X^\top W X)^{-1} x_*).$$

So we can construct an asymptotic $(1 - \alpha)$ confidence interval for $\log \mathbb{E} Y_* = x_*^\top \beta$ with endpoints

$$x_*^\top \hat{\beta} \pm z_{\alpha/2} \sqrt{x_*^\top (X^\top W X)^{-1} x_*},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. Exponentiating the above interval gives an asymptotic $1 - \alpha$ confidence interval of the mean response.

(iv) The residual deviance is

$$D(Y;\hat{\lambda}) = 2\sum_{i=1}^{n}\{\log f(Y_i;Y_i) - \log f(Y_i;\hat{\lambda}_i)\} = 2\sum_{i=1}^{n}\left\{Y_i \log\frac{Y_i}{\hat{\lambda}_i} - (Y_i - \hat{\lambda}_i)\right\}$$

The deviance residuals are

$$d_i = \operatorname{sgn}(Y_i - \hat{\lambda}_i)\sqrt{2\left\{Y_i \log\frac{Y_i}{\hat{\lambda}_i} - (Y_i - \hat{\lambda}_i)\right\}}.$$

Writing $\delta_i := Y_i - \hat{\lambda}_i$, we have by Taylor expanding the logarithm around 1,

$$Y_i \log\frac{Y_i}{\hat{\lambda}_i} - (Y_i - \hat{\lambda}_i) = (\hat{\lambda}_i + \delta_i)\log(1 + \delta_i/\hat{\lambda}_i) - \delta_i = \frac{\delta_i^2}{2\hat{\lambda}_i} + O(\delta_i^3).$$

Thus, for small δ_i, we can approximate the deviance residuals by

$$\operatorname{sgn}(Y_i - \hat{\lambda}_i)\sqrt{\frac{(Y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i}} = r_i,$$

as desired.

3. This is Question 6 of the 2017–2018 past paper. You may find it in the following link:
https://www.maths.cam.ac.uk/postgrad/part-iii/files/pastpapers/2018/paper_218.pdf.

Solution

(a) First plot: the x-coordinates are the fitted values $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2}$, $i = 1, \ldots, n$, where $\hat{\beta}_0 = 11.2$, $\hat{\beta}_1 = 1.87$ and $\hat{\beta}_2 = -0.721$, and the y-coordinates are the residuals $\hat{\epsilon}_i = \hat{Y}_i - Y_i$, $i = 1, \ldots, n$.

Second plot: the x-coordinates are quantiles $\Phi^{-1}(\frac{i}{n+1})$ of the standard normal distribution, and the y-coordinates are the (increasingly) ordered standarised residuals $\hat{\epsilon}^*_{(i)}$, where $\hat{\epsilon}^*_i = \frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2(1-P_{i,i})}}$ for

$$P = X(X^\top X)^{-1}X^\top \qquad \text{and} \qquad X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} \end{pmatrix}.$$

In the first plot we may appreciate a small decrease in the absolute values of the residuals, although it is quite light and could be by chance. The Q-Q plot seems fine, so neither diagnostic plot clearly suggests any violation of modelling assumptions.

(b) This often happens when the covariates are highly correlated. Indeed, we see this in the correlation matrix in the output: the correlation between `x1` and `x2` is 0.981, which

means the design matrix X has large condition number and so the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ have large standard errors, making both of them insignificant.

(c) We consider models of the form $Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$, $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ for some $\beta_0, \beta_1, \beta_2 \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$. Then, the `anova` command performs a likelihood-ratio test between the null hypothesis that $\beta_1 = 0$ against the alternative that $\beta_1 \neq 0$.

Let $f(y; \mu, \sigma^2)$ be the pdf of a univariate normal distribution with mean μ and variance σ^2. Define $\mu_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}$ and let RSS_0 and RSS_1 are respectively the residual sum of squares of the maximum likelihood fit under the null model and under the alternative model respectively. Then, the test statistic is

$$T = 2 \log \frac{\sup_{\beta_0, \beta_1, \beta_2 \in \mathbb{R}, \sigma^2 > 0} \prod_{i=1}^n f(Y_i; \mu_i, \sigma^2)}{\sup_{\beta_0, \beta_2 \in \mathbb{R}, \sigma^2 > 0, \beta_1 = 0} \prod_{i=1}^n f(Y_i; \mu_i, \sigma^2)} = \frac{(\text{RSS}_0 - \text{RSS}_1)}{\text{RSS}_1/(n-3)},$$

Under the null hypothesis, $T \sim F_{1,n-3}$ and the p-value is computed as $1 - F(T)$, where F is the distribution function of the $F_{1,n-3}$ distribution. Recall that an entry in the third column of any of the summaries represents the t-statistic for testing if the corresponding coefficient is zero. Then, the t-statistic for `x1` in the third model tests whether $\beta_1 = 0$ versus $\beta_1 = 1$ and, since it follows a t_{n-3} distribution, we have that $T = 0.800^2 = 0.640$ and the p-value is 0.43.

(d) The optimisation problem is

$$\max_{\beta_0, \beta_1, \beta_2 \in \mathbb{R}} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_{i,1} - \beta_2 x_{i,2})^2 + \lambda(\beta_1^2 + \beta_2^2).$$

The left asymptotes are horizontal lines at 1.87 and -0.721 (coefficient estimates in `model3`) and right asymptotes are both horizontal line at 0.

(e) Let $\beta = (\beta_1, \beta_2)^\top$ and $\hat{\beta}^{\text{r}}_\lambda = (\hat{\beta}^{\text{r}}_{1,\lambda}, \hat{\beta}^{\text{r}}_{1,\lambda})^\top$. Let X_{-1} be X without its first column of all ones. Then,

$$\hat{\beta}^{\text{r}}_\lambda = (X_{-1}^\top X_{-1} + \lambda I)^{-1} X_{-1}^\top \left(Y - \frac{1}{n} \sum_{i=1}^n Y_i \right),$$

and, since $\text{var}(Y) = \sigma^2 I$,

$$\text{var}(\hat{\beta}^{\text{r}}_\lambda) = \sigma^2 (X_{-1}^\top X_{-1} + \lambda I)^{-1} X_{-1}^\top X_{-1} (X_{-1}^\top X_{-1} + \lambda I)^{-1} =: \Sigma.$$

It follows that $\text{var}(\hat{\beta}^{\text{r}}_{1,\lambda}) = \Sigma_{1,1}$ and $\text{var}(\hat{\beta}^{\text{r}}_{2,\lambda}) = \Sigma_{2,2}$.

These variances are not useful in constructing confidence intervals because the ridge regression estimators are biased estimators, so will be shifted towards the origin away from the true value (and even if could centre them around true value, their variances would be smaller than needed for the correct coverage as ridge trades it off by bias).

(f) We can select among the three linear models using AIC (or BIC, cross-validation). We can select λ in the ridge regression model via cross-validation.

4. Consider the following linear regression problem without an intercept term

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \quad \text{for } i = 1, \ldots, n.$$

Suppose estimates of the regression parameters (β_1, β_2) of this model are obtained through the minimization of the sum-of-squares augmented with a ridge-type regularisation

$$\sum_{i=1}^{n} (Y_i - \beta_1 X_{i1} - \beta_2 X_{i2})^2 + \lambda(\beta_1^2 + \beta_2^2 + 2\nu\beta_1\beta_2),$$

with tuning parameters $\lambda \in [0, \infty)$ and $\nu \in [-1, 1]$.

(i) Write down the above optimisation problem in an equivalent constrained form. Sketch for both $\nu = 0$ and $\nu = 0.9$ the shape of the parameter constraint induced by the penalty above and describe in words the qualitative difference between both shapes.

(ii) When $\nu = -1$ and $\lambda \to \infty$, the estimates of β_1 and β_2 (resulting from minimisation of the penalized loss function above) converge towards each other: $\lim_{\lambda\to\infty} \hat{\beta}_1(\lambda, -1) = \lim_{\lambda\to\infty} \hat{\beta}_2(\lambda, -1)$. Motivated by this observation a data scientists incorporates the equality constraint $\beta_1 = \beta = \beta_2$ explicitly into the model, and she estimates the 'joint regression parameter' β through the minimisation (with respect to β) of:

$$\sum_{i=1}^{n} (Y_i - \beta X_{i,1} - \beta X_{i,2})^2 + \rho\beta^2,$$

with a tuning parameter $\rho \in [0, \infty)$. The data scientist is surprised to find that resulting estimate $\hat{\beta}(\rho)$ does not have the same limiting (in the penalty parameter) behaviour as the $\hat{\beta}_1(\lambda, -1)$, i.e. $\lim_{\rho\to\infty} \hat{\beta}(\rho) \neq \lim_{\lambda\to\infty} \hat{\beta}_1(\lambda, -1)$. Explain the misconception of the data scientist.

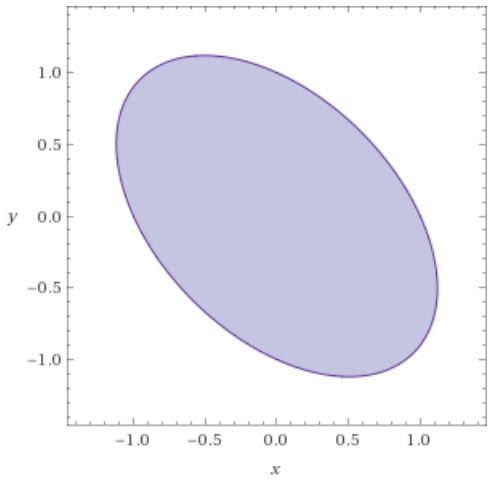
(iii) Assume that (a) $n \gg 2$, (b) the unpenalised least squares estimates $(\hat{\beta}_1(0, 0), \hat{\beta}_2(0, 0))$ are equal to $(-2, 2)$, and (c) that the two covariates $X_1 = (X_{11}, \ldots, X_{n1})^\top$ and $X_2 = (X_{12}, \ldots, X_{n2})^\top$ are zero-centred, have unit variance, and have correlation ρ with $|\rho| < 1$. Consider $(\hat{\beta}_1(\lambda, \nu), \hat{\beta}_2(\lambda, \nu))$ for both $\nu = -0.9$ and $\nu = 0.9$. For which value of ν do you expect the sum of the absolute value of the estimates to be larger? Does your answer change depending on whether X_1 and X_2 are negatively or positively correlated? Explain.

Solution

(i) The constrained form is

$$\min_{\beta_1, \beta_2} \sum_{i=1}^{n} (Y_i - \beta_1 X_{i1} - \beta_2 X_{i2})^2 \qquad \text{subject to} \quad \beta_1^2 + \beta_2^2 + 2\nu\beta_1\beta_2 \leq C.$$

When $\nu = 0$, the constraint region is a disk of radius $\sqrt{C}$ centred at the origin. When $\nu = 0.9$ the constraint region is an ellipse with major axis along the line $x_1 + x_2 = 0$.



(ii) When $\nu = -1$, and $\lambda \to \infty$, the original optimisation problem in the constrained form has $C = 0$, i.e. $\beta_1 = \beta_2$. Hence, it amounts to optimising

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - \beta X_{i1} - \beta X_{i2})^2$$

However, the optimisation problem solved by the data scientist has an additional constraint on the size of β. As $\rho \to \infty$, the corresponding constraint form for $\sum_{i=1}^{n}(Y_i - \beta X_{i1} - X_{i2})^2 + \rho\beta^2$ is minimising $\sum_{i=1}^{n}(Y_i - \beta X_{i1} - X_{i2})^2$ subject to $\beta^2 = 0$.

To summarise, $\hat{\beta}(\lambda, -1)$ is the least square coefficient for regressing $Y_1, \ldots, Y_n$ against $X_{11} + X_{12}, \ldots, X_{n1} + X_{n2}$ with no intercept term, whereas $\hat{\beta}(\rho)$ is equal to zero.

(iii) This can be argued geometrically or algebraically. We do so in this order.

Note that the boundary of the constraint in part (i) is an ellipse centred at 0 with major axis along $x_1 + x_2 = 0$ if $\nu > 0$ and along $x_1 - x_2 = 0$ if $\nu < 0$. In addition, writing $\beta = (\beta_1, \beta_2)^\top, \hat{\beta} = (\hat{\beta}_1(0,0), \hat{\beta}_2(0,0))^\top$, $X = (X_1 \ X_2)$ and $\rho = X_1^\top X_2$, so that

$$X^\top X = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

the residual sum of squares can be written as

$$\begin{aligned} \|Y - X\beta\|_2^2 &= (Y - X\beta)^\top (Y - X\beta) \\ &= Y^\top Y - \hat{\beta}^\top \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \hat{\beta} + \left(\beta - \hat{\beta}\right)^\top \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \left(\beta - \hat{\beta}\right) \\ &= \|Y - X\hat{\beta}\|_2^2 + \tilde{\beta}^\top \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \tilde{\beta}, \end{aligned}$$

where $\tilde{\beta} = \beta - \hat{\beta}$. The first term corresponds to the minimum value of the residual sum of squares, so is fixed, and the second represents the increase in the residual

sum of squares for a perturbation $\tilde{\beta}$ of $\hat{\beta}$. By the form of the second term, level sets of the residual sum of squares are ellipses (centred at $\hat{\beta}$) with major axis along $x_1 + x_2 = 0$ if $\rho > 0$ and along $x_1 - x_2 = 0$ if $\rho < 0$. Since $\hat{\beta} = (-2, 2)^\top$ falls on the line $x_1 + x_2 = 0$, for a given small enough C in the constraint of part (i), the first level set (starting from the OLS) that intersects the ellipse of the constraint will always do so on $x_1 + x_2 = 0$ and, hence, the intersection will occur closer to the OLS for $\nu > 0$ and will consequently return larger abosolute values of the entries of $\hat{\beta}(\lambda, \nu)$. (Draw a picture with every step to convince yourself.)

We can also check this directly. Writing $R = \begin{pmatrix} 1 & \nu \\ \nu & 1 \end{pmatrix}$, the optimisation problem is minimising the loss function

$$L(\beta) = \|Y - X\beta\|_2^2 + \lambda \beta^\top R \beta.$$

Differentiate with respect to β, we obtain that the optimiser is

$$\begin{aligned}
\hat{\beta}(\lambda, \nu) &= (X^\top X + \lambda R)^{-1} X^\top Y = (X^\top X + \lambda R)^{-1} (X^\top X) \hat{\beta}_0 \\
&= \begin{pmatrix} 1 + \lambda & \rho + \lambda\nu \\ \rho + \lambda\nu & 1 + \lambda \end{pmatrix}^{-1} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} -2 \\ 2 \end{pmatrix} \\
&= \frac{2(1-\rho)(1 + \lambda + \rho + \lambda\nu)}{(1+\lambda)^2 - (\lambda\nu + \rho)^2} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \\
&= 2 \frac{1 - \rho}{1 - \rho + \lambda(1 - \nu)} \begin{pmatrix} -1 \\ 1 \end{pmatrix}.
\end{aligned}$$

Therefore, for any ρ (note $\rho < 1$ by assumption), the absolute value of the coefficients of $\hat{\beta}(\lambda, \nu)$ increases with ν, and so $\nu = 0.9$ gives a larger sum of absolute values of the estimates than $\nu = -0.9$ irrespectively of the sign of ρ.

5. Recall that the negative binomial distribution $\mathrm{NB}(r, p)$ models the total number of successes until $r \in \mathbb{N}$ failures have occurred in a sequence of i.i.d. Bernoulli trials each with a success probability $p \in [0, 1]$.

   (i) Write down the probability mass function of $Y \sim \mathrm{NB}(r, p)$ and show that $\{\mathrm{NB}(r, p) : r \in \mathbb{N}, p \in (0, 1)\}$ is an exponential dispersion family if and only if r is known.

   (ii) Suppose $\nu \sim \mathrm{Gamma}(\theta, \theta)$ for some $\theta \in \mathbb{N}$ and $Y \mid \nu \sim \mathrm{Poi}(\lambda\nu)$ for some $\lambda > 0$. Show that $Y \sim \mathrm{NB}(\theta, \frac{\lambda}{\lambda + \theta})$. What are the resulting parameters in the alternative parametrisation introduced in class? Make sure you are convinced about the validity of the generalisation of this reparametrisation to $\theta > 0$.

Solution

(i) The pmf is $f(y; r, p) = \binom{r+y-1}{y} p^y (1-p)^r$, which we can rewrite as

$$f(y; r, p) = \binom{r + y - 1}{y} \exp\left(y \log p + r \log(1 - p)\right).$$

If we know r then it is clear that this comes from an exponential dispersion family with natural parameter $\theta = \log p$ (and $K(\theta) = \log(1 - e^{\theta})$). If r is unknown, then we cannot disentangle y and r in the binomial coefficient in such a way that r is absorbed by the natural or dispersion parameters, and $\{\text{NB}(r, p) : r \in \mathbb{N}, p \in (0, 1)\}$ is not an exponential dispersion

(ii) We compute the pmf of Y as

$$\mathbb{P}(Y = y) = \int_{\nu=0}^{\infty} e^{-\lambda\nu} \frac{(\lambda\nu)^y}{y!} \frac{1}{\Gamma(\theta)} \theta^{\theta} \nu^{\theta-1} e^{-\theta\nu} \, d\nu = \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)} \frac{\theta^{\theta}\lambda^y}{(\lambda+\theta)^{y+\theta}},$$

which is precisely the pmf of the $\text{NB}(\theta, \frac{\lambda}{\lambda+\theta})$. In the second parametrisation given in class, we have $\text{NB}(\mu, \tau)$, where $\mu = \lambda$ and $\tau = 1/\theta$.

6. To understand the relationship between the number of typographical errors (dependent variables Y_i, $i = 1, \ldots, n$) and lengths of manuscripts in words (independent variables x_i, $i = 1, \ldots, n$), a researcher fitted both a Poisson model ω_1 and a negative binomial model ω_2 to the data.

(i) Write down the two models algebraically.

(ii) If the log-likelihoods for ω_1 and ω_2 are -193.4 and -192.2 respectively. What is the result of the likelihood ratio test between the null hypothesis that ω_1 is correct and the alternative hypothesis that model ω_2 is correct (at 5% level)? Which model will be selected based on AIC?

Solution

(i) Under ω_1, we assume $Y_i \overset{\text{ind}}{\sim} \text{Poi}(\lambda_i)$, where $\log \lambda_i = \beta_0 + \beta_1 x_i$ for some $\beta = (\beta_0, \beta_1)^{\top} \in \mathbb{R}^2$. Under ω_2, we assume that there exist hidden variables $\nu_i \overset{\text{iid}}{\sim} \text{Gamma}(\theta, \theta), \theta > 0$, and, conditional on $(\nu_1, \ldots, \nu_n)$, we have $Y_i \overset{\text{ind}}{\sim} \nu_i \sim \text{Poi}(\lambda_i \nu_i)$, where $\log \lambda_i = \beta_0 + \beta_1 x_i$ for some $\beta = (\beta_0, \beta_1)^{\top} \in \mathbb{R}^2$. You may also write that, under ω_2, $Y_i \overset{\text{ind}}{\sim} \text{NB}(\lambda_i, \tau)$, where $\log \lambda_i = \beta_0 + \beta_1 x_i$ for some $\beta = (\beta_0, \beta_1)^{\top} \in \mathbb{R}^2$ and $\tau > 0$.

(ii) The likelihood ratio test statistic is $2(193.4 - 192.2) = 2.4$. Since ω_1 is on the boundary of the parameter space in ω_2 ($\theta = \infty$ or $\tau = 0$), the likelihood ratio test under the null hypothesis is distributed according to χ^2_1 with probability 0.5 and is equal to 0 with probability 0.5. Hence, we should compare the test statistic with $\chi^2_1(0.1) = 2.71$. There is not sufficient evidence to reject the null hypothesis. The AIC for the two models are $2(193.4) + 4 = 390.8$ and $2(192.2) + 6 = 390.4$. Hence ω_2 will be chosen, so which we choose may change depending on the criterion and, thus, we should always do so with a certain initial goal in mind.

7. (*Exercise with* `R`) The dataset `ships` in the library `MASS` contains the number of incidents for a set of ships. Investigate the relationship between the number of incidents (response variable `incidents`) and the months of service (variable: `service`) and the type of the

ships (variable: `type`). Ignore the other columns of the dataset. Explore the data graphically (it may be necessary to apply transformations to the data) and fit the appropriate model to predict the number of incidents.

Remark: this type of question is not something that can happen in the exam, which is pen and paper only. However, it is a good exercise to become more familiar with `R` and to tackle a small data analysis example.