Modern Regression: Solution 2019

25 June 2019

1. Independent random variables $Y_1, \ldots, Y_n$ have probability density functions

$$f(y_j;\beta) = \frac{1}{\pi\left\{1+\left(y_j - x_j^{\mathrm{T}}\beta\right)^2\right\}}, \quad y_j \in \mathbb{R}, \quad j = 1, \ldots, n,$$

where β is a $p \times 1$ vector of unknown real-valued parameters and $x_1, \ldots, x_n$ are $p \times 1$ vectors of explanatory variables.

(a) **Seen 5pt**
Steps of the algorithm: slide 30 of the lecture notes
Take initial values of $\hat{\beta}$. Repeat

- compute X, W, u, z
- compute new $\hat{\beta}$
- check $l(\hat{\beta}_{j+1}) - l(\hat{\beta}_j)$

Components of the IWLS

$$\begin{aligned}
&\eta_j = x_j^{\mathrm{T}}\beta, \ \frac{\partial \eta_j}{\partial \beta_r} = x_{jr}, \\
&l_j = l(y_j, \eta_j) = \log f(y_j;\beta) = -\log\left[\pi\left\{1+(y_j-\eta_j)^2\right\}\right] \ , \\
&u_j(\eta_j) = \frac{\partial l_j}{\partial \eta_j} = -2\frac{y_j - \eta_j}{1+(y_j-\eta_j)^2} \ , \\
&w_{jj} = -\frac{\partial^2 l_j}{\partial \eta_j^2} = 2\frac{1-(y_j-\eta_j)^2}{\{1+(y_j-\eta_j)^2\}^2}.
\end{aligned}$$

(b) **Seen 1pt**
If $y_j - x_j^{\mathrm{T}}\beta$ was replaced by $(y_j - x_j^{\mathrm{T}}\beta)/\sigma$, then we need to estimate an unknown scale parameter ϕ, one approach is to use the profile log-likelihood $l_p(\phi)$.

(c) **Unseen 4pt**
$a(u) = \mathrm{d}^2 \log(1+u^2)/\mathrm{d}u^2 = 2\dfrac{1-u^2}{(1+u^2)^2}$.

$$\Pr\{a(U) > 0\} = \Pr(U^2 < 1) = \int_{-1}^{1} \frac{1}{\pi(1+u^2)} du = \frac{2}{\pi}\int_0^1 \frac{1}{\pi(1+u^2)} du = \frac{2}{\pi}\tan^{-1}(1) = \frac{1}{2}.$$

Unseen 2pt
This implies that the diagonal elements of the weight matrix W can be negative which causes positive definite problem. We can fix this by taking the expectation or ridge regression. For λ sufficiently large, $X^T W X + \lambda I$ will be diagonally dominant and therefore PD.

2. (a) **Seen 2pt**

- The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a **link** function and by allowing the magnitude of the **variance** of each measurement to be a function of its predicted value.
- Give two examples of generalized linear models.

(b) **Unseen 4pt**
The density function is

$$f(y;\mu,\nu) = \frac{y^{\nu/2-1}}{2^{\nu/2}\sigma^{\nu}\Gamma(\nu/2)}\exp\left(-\frac{y}{2\sigma^2}\right),$$
$$= \exp\left\{-\frac{y}{2\sigma^2}+\frac{\nu}{2}\log\left(\frac{1}{\sigma^2}\right)+(\frac{\nu}{2}-1)\log y-\log\Gamma(\frac{\nu}{2})\right\}.$$

then we deduce

$$\theta = -\frac{1}{\nu\sigma^2}, \quad b(\theta) = -\log(-\nu\theta), \quad \phi = \frac{2}{\nu}, \quad c(y,\phi) = (\frac{1}{\phi}-1)\log y - \log\Gamma(\frac{1}{\phi}).$$

For the mean, we have

$$\mathrm{E}(Y) = b'(\theta) = -\frac{1}{\theta} = \nu\sigma^2,$$

Similarly, the variance function is

$$V(\theta) = b''(\theta) = \frac{1}{\theta^2} = \nu^2(\sigma^2)^2.$$

Finally, this leads to

$$\mathrm{Var}(X) = \phi V(\theta) = 2\nu(\sigma^2)^2.$$

(c) **Unseen 4pt**
Let, $Y_i = \sum_{j=1}^{r_i}(Z_{ij}-\bar{Z}_i)^2$, where $\bar{Z}_i = r_i^{-1}\sum_{j=1}^{r_i} Z_{ij}$. Define $SS_B = \sum_{i=1}^{m} r_i(\bar{Z}_{i\cdot} - \bar{Z}_{ii})^2$ and $SS_W = \sum_{i=1}^{m} Y_i$, where SS_B is the sum of squares due to variation between times and SS_W is the sum of squares due to error within times. One can use the $\sigma_i^2 = \sigma_0^2 + t_i\sigma_i^2$ to assess the dependence of variance on temperature.

- we can use reciprocal link function.
- Drawback: negative values

3. (a) **Seen 2pt**
Binomial model with $R_i = \begin{cases} 1, \text{presence of calcium oxalate crystals in the ith sample,} \\ 0, \text{otherwise.} \end{cases}$
and

$$\Pr(R_i = 1) = \pi_i = 1 - \Pr(R_i = 0) = \frac{\exp(x_i^T\beta)}{1+\exp(x_i^T\beta)}, \quad j = 1\ldots n.$$

(b) **Seen/unseen 2pt**
The ANOVA table shows how the addition of independent variables reduces the Residual deviance (Resid. Dev). It is used to determine the significance of the independent variables. According to the table, the deviance reduction due to `ph`, `urea` and `cond` are less than 1% of the total deviance. One can remove these variables and check the goodness of the fit, use a stepwise selection approach or change the order of inclusion of these dependent variables.

(c) **Seen/unseen 2pt**
Including `grav`, reduces the residual deviance by 14.93 for 1 degree of freedom, showing that `grav` is significant. According to the estimated regression coefficients, the corresponding p-value of `grav` is not significantly small which suggests that we have a collinearity problem.

(d) **Seen/unseen 1pt** 95% confidence interval for `calc` is $[0.299, 1.268]$.

(e) **Seen/unseen 1pt** The fitted values increase with a decrease in `ph` .

(f) **Seen/unseen 2pt**

- The only terms significant at 5% level are `calc` and `urea`.
- For our example, we have a null deviance of 105 on 76 degrees of freedom. Including 6 independent variables decreased the deviance to 57.56 on 70 degrees of freedom. The Residual Deviance has reduced by 47.61 with a loss of 6 degrees of freedom. A significant reduction (i.e., the alternative model is better than the null model (intercept)).
- The chi square distribution is an asymptotic approximation and for small counts it is unreliable, so we can not say anything about this fit.

```
> anova(urine.glm)
Analysis of Deviance Table

Model: binomial, link: logit

Terms added sequentially (first to last)

                     Df Deviance Resid. Df Resid. Dev
NULL                                    76    105.168
grav                  1  14.9327        75     90.235
ph                    1   0.0723        74     90.163
osmo                  1   9.5573        73     80.606
cond                  1   0.0106        72     80.595
urea                  1   1.3343        71     79.261
calc                  1  21.7007        70     57.560

> summary(urine.glm)
Call:
glm(formula = r ~ grav + ph + osmo + cond + urea + calc, family = binomial, data = urine)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6215  -0.5967  -0.2849   0.3176   2.7445

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.60609    3.79582   0.160  0.87314
grav          3.55944    2.22110   1.603  0.10903
ph           -0.49570    0.56976  -0.870  0.38429
osmo          0.01681    0.01782   0.944  0.34536
cond         -0.43282    0.25123  -1.723  0.08493 .
urea         -0.03201    0.01612  -1.986  0.04703 *
calc          0.78369    0.24216   3.236  0.00121 **
---
Signif. codes:  0 ‘***‘ 0.001 ‘**‘ 0.01 ‘*‘ 0.05 ‘.‘ 0.1 ‘ ‘ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 105.17  on 76  degrees of freedom
Residual deviance:  57.56  on 70  degrees of freedom
AIC: 71.56
```

4. (a) **Seen 2pt**

We minimize $\mathrm{E}_{b,y}\left[\left\{\tilde{b}(y)-b\right\}^{\mathrm{T}}\left\{\tilde{b}(y)-b\right\}\right]$

$$\begin{aligned}\frac{\partial \mathrm{E}_{b,y}\left[\left\{\tilde{b}(y)-b\right\}^{\mathrm{T}}\left\{\tilde{b}(y)-b\right\}\right]}{\partial \widetilde{b}} &= \mathrm{E}_{b,y}\left[\frac{\partial\left\{\tilde{b}(y)-b\right\}^{\mathrm{T}}\left\{\tilde{b}(y)-b\right\}}{\partial \widetilde{b}}\right] \\ &= \mathrm{E}_{b,y}\left\{\tilde{b}(y)-b\right\} \\ &= \tilde{b}(y)-\mathrm{E}(b \mid y).\end{aligned}$$

Thus, we can conclude that $\tilde{b}(y) = \mathrm{E}(b \mid y)$.

(b) **Seen 5pt**

$$\begin{pmatrix} y \\ b \end{pmatrix} \sim \mathcal{N}_{n+q}\left\{\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega + Z\Omega_b Z^{\mathrm{T}} & Z\Omega_b \\ \Omega_b Z^{\mathrm{T}} & \Omega_b \end{pmatrix}\right\},$$

Seen
The conditional mean and variance of b given y are respectively

$$\Omega_b Z^{\mathrm{T}} \left(\Omega + Z\Omega_b Z^{\mathrm{T}}\right)^{-1} (y - X\beta), \quad \Omega_b - \Omega_b Z^{\mathrm{T}} \left(\Omega + Z\Omega_b Z^{\mathrm{T}}\right)^{-1} Z\Omega_b.$$

The Woodbury formula applied to $\left(\Omega + Z\Omega_b Z^{\mathrm{T}}\right)^{-1}$ yields

$$\mathrm{E}(b \mid y) = \Omega_b Z^{\mathrm{T}} \left\{\Omega^{-1} - \Omega^{-1} Z \left(\Omega_b^{-1} + Z^{\mathrm{T}}\Omega^{-1} Z\right)^{-1} Z^{\mathrm{T}}\Omega^{-1}\right\} (y - X\beta)$$

and applied to the variance formula gives

$$\mathrm{var}(b \mid y) = \left(Z^{\mathrm{T}}\Omega^{-1} Z + \Omega_b^{-1}\right)^{-1}$$

as required. The conditional mean equals

$$\Omega_b \left\{I_q - Z^{\mathrm{T}}\Omega^{-1} Z \left(\Omega_b^{-1} + Z^{\mathrm{T}}\Omega^{-1} Z\right)^{-1}\right\} Z^{\mathrm{T}}\Omega^{-1}(y - X\beta)$$

and the term in braces can be written as $I - B(A + B)^{-1} = A(A + B)^{-1}$, with $A = \Omega_b^{-1}$ and $B = Z^{\mathrm{T}}\Omega^{-1}Z$, which yields the stated formulae,

$$\begin{aligned} \mathrm{E}(b \mid y) &= \left(Z^{\mathrm{T}}\Omega^{-1} Z + \Omega_b^{-1}\right)^{-1} Z^{\mathrm{T}}\Omega^{-1} (y - X\beta), \\ \mathrm{var}(b \mid y) &= \left(Z^{\mathrm{T}}\Omega^{-1} Z + \Omega_b^{-1}\right)^{-1}. \end{aligned}$$

Seen 1pt
To obtain the BLUP and its variance, we replace the parameters β and σ by their estimates, giving

$$\tilde{b} = \left(Z^{\mathrm{T}}\widehat{\Omega}^{-1} Z + \widehat{\Omega_b}^{-1}\right)^{-1} Z^{\mathrm{T}}\widehat{\Omega}^{-1} \left(y - X\widehat{\beta}\right),$$

and

$$\mathrm{var}(\tilde{b}) = \mathrm{var}(b \mid y) = \left(Z^{\mathrm{T}}\widehat{\Omega}^{-1} Z + \widehat{\Omega}_b^{-1}\right)^{-1}.$$

(c) **Unseen 2pt**
If $\sigma_b^2/\sigma^2 \ll 1$ then, $\mathrm{var}(\tilde{b}) \to 0$, there is no variation between groups, and hence $b_i = 0$ with probability one.

5. (a) **Seen 1pt**
When the shape of the functional relationships between the response (dependent) and the explanatory (independent) variables are not predetermined but can be adjusted to capture unusual or unexpected features of the data.
Seen 3pt
This asks about the choice of the **degree of smoothing**, not about methods of smoothing. The methods for choosing the degree of smoothing described in the course are: minimising a cross-validation sum of squares, a generalised cross-validation sum of squares, or maximising a marginal likelihood for the smoothing parameter (the first two tend to give curves that are too variable).

(b) **Unseen 3pt**
We have

$$\begin{aligned} \sum_{i=1}^{n} \{y_i^* - \mu(x_i)\}^2 + \lambda \int \mu''(x)^2 \,\mathrm{d}x &\geq \sum_{i \neq j}^{n} \{y_i^* - \mu(x_i)\}^2 + \lambda \int \mu''(x)^2 \,\mathrm{d}x \\ &\geq \sum_{i \neq j}^{n} \{y_i^* - \hat{\mu}^{-j}(x_i)\}^2 + \lambda \int \{\hat{\mu}^{-j}\}''(x)^2 \,\mathrm{d}x \\ &= \sum_{i=1}^{n} \{y_i^* - \hat{\mu}^{-j}(x_i)\}^2 + \lambda \int \{\hat{\mu}^{-j}\}''(x)^2 \,\mathrm{d}x, \end{aligned}$$

the first line by dropping a positive quantity, the second by definition of $\hat{\mu}^{-j}$, and the third by definition of y_j^*. Thus, we can deduce that $\hat{\mu}^{-j} = Sy^*$.

(c) **Unseen 3pt**

$$\begin{aligned} y_j - \hat{\mu}^{-j}(x_j) &= y_j - \sum_{i=1}^{n} S_{ji} y_i^* \\ &= y_j - \left(\sum_{i \neq j}^{n} S_{ji} y_i + S_{jj} \hat{\mu}^{-j}(x_j) \right) \\ &= y_j - \left\{ \hat{\mu}(x_j) - S_{jj} y_j + S_{jj} \mu^{-j}(x_j) \right\}, \end{aligned}$$

which implies that

$$(1 - S_{jj}) \{ y_j - \hat{\mu}(x_j) \} = y_j - \hat{\mu}(x_j),$$

thus giving

$$y_j - \hat{\mu}^{-j}(x_j) = \frac{y_j - \hat{\mu}(x_j)}{1 - S_{jj}}.$$

The optimal λ is chosen to minimize the cross-validation sum of squares

$$CV(\lambda) = \sum_{j=1}^{n} \{ y_j - \hat{\mu}^{-j}(x_j) \}^2 = \sum_{j=1}^{n} \left\{ \frac{y_j - \hat{\mu}(x_j)}{1 - S_{jj}} \right\}^2,$$

which can be computed from a single fit of the model for a given λ, rather than from the ostensible $n + 1$ fits needed.