Exercise 6: Linear Models and Generalizations

2018 Spring

1 Background Knowledge

1.1 Maximum Likelihood Estimator

1.2 Exponential Families

A one-parameter exponential-family distribution has the probability density function (pdf) of the following form:

$$f(y;\theta) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(\phi, y)\right)$$

where

- θ is the **canonical parameter**.
- ϕ is the (optional) **dispersion parameter**.
- The expected value of Y: $E(Y) = \mu = b'(\theta)$
- The variance of μ is: $V(\mu) = b''(\theta)$
- The variance of Y is: $Var(Y) = V(\mu)a(\phi)$
- The **link function** $g(\mu) = \eta = x^T\beta$
- **Canonical link function** is obtained through $\eta = \theta$.

Thus the log-likelihood function of the data is:

$$l(\theta; y) = \frac{y\theta - b(\theta)}{a(\phi)}$$

1.3 Generalized Linear Models (GLMs) Fitting

With this we can obtain the **Fisher's score vector u** $= (u_j)$ with

$$u_j = \frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta}\frac{\partial \theta}{\partial \mu}\frac{\partial \mu}{\partial \eta}\frac{\partial \eta}{\partial \beta_j}$$

Since

$$\begin{array}{rcl} \frac{\partial l}{\partial \theta} & = & \frac{y - b'(\theta)}{a(\phi)} \\ \frac{\partial \theta}{\partial \mu} & = & \frac{1}{b''(\theta)} = \frac{1}{V(\mu)} = \frac{a(\phi)}{Var(y)} \\ \frac{\partial \eta}{\partial \beta_j} & = & x_{ij} \end{array}$$

Therefore

$$\frac{\partial l}{\partial \beta_j} = \frac{y - \mu}{Var(y)}\left(\frac{\partial \mu}{\partial \eta}\right)x_{ij}$$

When we use the canonical link function

$$\frac{\partial \mu}{\partial \eta} = \frac{\partial \mu}{\partial \theta} = b''(\theta)$$

therefore

$$\frac{\partial l}{\partial \beta_j} = \frac{y-\mu}{Var(y)} b''(\theta) x_{ij} = \frac{y-\mu}{a(\phi)} x_{ij}$$

And **Fisher's information matrix** can be obtained by:

$$\begin{aligned} -E\left(\tfrac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right) &= E\left[\tfrac{\partial l}{\partial \beta_j}\tfrac{\partial l}{\partial \beta_k}\right] \\ &= E\left(\tfrac{y-\mu}{Var(y)}\right)^2 \left(\tfrac{\partial \mu}{\partial \eta}\right)^2 x_{ij}x_{ik} \\ &= \tfrac{1}{Var(y)}\left(\tfrac{\partial \mu}{\partial \eta}\right)^2 x_{ij}x_{ik} \end{aligned}$$

For general link function, the score function for only **1-observation** becomes

$$\frac{\partial l}{\partial \beta_j} = \frac{y-\mu}{Vary}\left(\frac{\partial \mu}{\partial \eta}\right) x_{ij}$$

And the score function for n-observation becomes:

$$\frac{\partial l}{\partial \beta} = X^T A(y-\mu)$$

Similarly the Fisher's information matrix can also be simplified as:

$$-E\left(\frac{\partial^2 l}{\partial \beta_j \beta_k}\right) = \frac{1}{Var(y)}\left(\frac{\partial \mu}{\partial \eta}\right)^2 x_{ij}x_{ik}$$

and also the matrix form:

$$-E\left(\frac{\partial^2 l}{\partial \beta \partial \beta^T}\right) = X^T W X$$

where

$$W = \operatorname{diag}(w_1, \ldots, w_n)$$

and

$$w_i = \frac{1}{Var(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = [b''(\theta_i)]^{-1}\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-2}$$

1.4 Iteratively Reweighted Least Squares (IRWLS)

Fisher's scoring algorithm can iteratively compute the coefficient by:

$$\beta^{(t+1)} = \beta^{(t)} + (X^T W X)^{-1} X^T A(y-\mu)$$

The score equation can be solved using the numerical method (Newton-Raphson), iteratively reweighted least squares (IRWLS):

$$\beta^{(t+1)} = \left(X^T W X\right)^{-1}\left[X^T W X \beta^{(t)} + X^T A(y-\mu)\right]$$

Since $X\beta = \eta$, we have

$$A = W\left(\frac{\partial \eta}{\partial \mu}\right)$$

Replace it into the equation:

$$\beta^{(t+1)} = \left(X^T W X\right)^{-1} X^T W z$$

which is similar to the closed-form of least squares fitting, where

$$z = \eta + \left(\frac{\partial \eta}{\partial \mu}\right)(y - \mu)$$

is called the **adjusted dependent variable**.

Example 1 (Logistic Regression) *Let $y_1, \dots, y_n$ where $y_i \sim Bin(n_i, p_i)$*

$$\begin{aligned} f(y;p) = \tbinom{n}{y} p^y (1-p)^{n-y} &= \exp\left(y \log p + (n-y)\log(1-p) + \log\tbinom{n}{y}\right) \\ &= \exp\left(y \log\tfrac{p}{1-p} + n\log(1-p) + \log\tbinom{n}{y}\right) \end{aligned}$$

- $\theta = \log\frac{p}{1-p} \Rightarrow p = \frac{e^\theta}{1+e^\theta}$;
- $b(\theta) = n\log\left(\frac{1}{1+e^\theta}\right)$
- $\mu = b'(\theta) = n\frac{e^\theta}{1+e^\theta} = np$
- $Var(\mu) = np(1-p)$

For the canonical link function:

$$\eta = \theta = \log\frac{p}{1-p} = \log\frac{\mu}{n-\mu}$$

2 Exercises

1. (20 points) For a set of independent observations $(Y_1, \dots, Y_N), N = 2n$; $Y_i \sim Bin(n_i, p_i), i = 1, \dots, N$ (n_i known); We can fit a GLM:

$$\log\frac{p_i}{1-p_i} = x_i^T\beta$$

where the design matrix has the form

$$X = \begin{bmatrix} a_1 & 0 \\ \vdots & \vdots \\ a_n & 0 \\ 0 & c_1 \\ \vdots & \vdots \\ 0 & c_n \end{bmatrix}$$

and the parameter $\beta = (\beta_1, \beta_2)^T$.

(a) (10 points) Find the score vector u in terms of Y_i, n_i, p_i, a_i, c_i and show that $\hat{\beta}_1$ are independent of $Y_{n+1}, \ldots, Y_N$ while $\hat{\beta}_2$ are independent of $Y_1, \ldots, Y_n$.

(b) (10 points) Find the Fisher's information matrix and show that the method of scoring for $\hat{\beta}$ has the form:

$$\begin{aligned}\hat{\beta_1}^{(k)} &= \hat{\beta}_1^{(k-1)} + \frac{\sum_{i=1}^{n}(Y_i - n_i p_i^{(k-1)})a_i}{\sum_{i=1}^{n} n_i p_i^{(k-1)}(1-p_i^{(k-1)})a_i^2} \\ \hat{\beta_2}^{(k)} &= \hat{\beta}_2^{(k-1)} + \frac{\sum_{i=n+1}^{N}(Y_i - n_i p_i^{(k-1)})c_{i-n}}{\sum_{i=n+1}^{N} n_i p_i^{(k-1)}(1-p_i^{(k-1)})c_{i-n}^2}\end{aligned}$$

2. (30 points) We have n independent observations $Y_1, \ldots, Y_n$ following normal distributions with the same variance σ^2:

$$EY_1 = \mu_1 = \beta_1 + \beta_2$$
$$EY_2 = EY_3 = \cdots = EY_n = \mu = \beta_1$$

where $\beta = (\beta_1, \beta_2)^T$ are parameters of interest.

(a) (10 points) What is the design matrix X in this model?

(b) (10 points) Show that only the first observation is highly influential, using the simple rule that $h_{ii} > 2p/n$ where $H = [h_{ij}]$ is the hat matrix $H = X(X^T X)^{-1} X^T$, and p is the number of paramters, here we have $p = 2$.

(c) (10 points) Find the maximum likelihood estimator of β.

3. (20 points) Table 3 from [1] displays results from a case-control study of oropharyngeal cancer patients. The investigators were looking for associations between HPV and oropharyngeal cancer. Use the table to answer the following questions.

(a) (5 points) Draw a conclusion on the association between the seropositive HPV-16 L1 serologic status and oropharyngeal cancer.

(b) (5 points) Can you calculate the unadjusted risk ratio for the risk of oropharyngeal cancer in patients who were positive for oral HPV-16 infection, using only the information given in this table?

(c) (5 points) What statistical method was used to calculate the "adjusted odds ratios" given in the table? The unadjusted odds ratio for HPV-16 L1 seropositivity is 17.6 but the adjusted odds ratio is 32.2. How do you explain this difference?

(d) (5 points) Which statistical test was used to compare the oral HPV infection prevalence in cases versus controls? Write down the correct statistic for comparing.

4. (20 points) Table 2 from [2] displays the beta coefficients from linear regression analysis. Refer to this table to answer the following questions.

(a) (5 points) What is your interpretation of the Beta coefficient relating BMI (kg/m^2) to vitamin D levels [Beta coefficient (and 95% CI) = -0.811 (-1.081, -0.541)]?

(b) (5 points) Specify the models used to generate the Beta coefficient relating BMI (kg/m^2) to vitamin D levels?

(c) (5 points) Which factor is associated with the greatest decrease in DSST score?

(d) (5 points) What is another test that the authors could use to determine whether people in the different BDI categories (normal, mild to borderline, moderate to extreme) have significantly different mean DSST scores?

5. (10 points) A multiple regression model is fitted on a data set:

$$Y_i = x_i^T \beta + \epsilon_i, \epsilon_i \overset{\text{i.i.d}}{\sim} \mathbf{N}(0, \sigma^2)$$

where $x_i = (1, x_{i1}, \ldots, x_{i4})^T, \beta = (\beta_0, \beta_1, \ldots, \beta_4)^T$ and we can obtain the result through R:

```

Coefficients:
                Estimate    Std. Error   t value   Pr(>|t|)
(Intercept)          ???        0.1960     8.438   3.57e-13
x1                5.3036        2.5316       ???   0.038834
x2                4.0336        2.4796     1.627   0.107111
x3               -9.3153        2.4657    -3.778   0.000276
x4                0.5884        2.2852     0.257   0.797373

Residual standard error: 1.892 on 95 degrees of freedom
Multiple R-squared: 0.1948, Adjusted R-squared: ???
F-statistic: 5.745 on 4 and 95 DF,  p-value: 0.0003483
```

(a) (2 points) What is the value of the t-statistics of $\hat{\beta}_1$?

(b) (2 points) How many observations are in the data set?

(c) (2 points) Has the null hypothesis $H_0 : \beta_3 = 0$ to be rejected on $\alpha = 0.05$?

(d) (2 points) What is the estimate of the intercept $\hat{\beta}_0$?

(e) (2 points) What is the estimate of $Var(\epsilon_i)$?

(f) (2 points) What is the 95% confidence interval for β_3?

References

[1] Gypsyamber D'souza, Aimee R Kreimer, Raphael Viscidi, Michael Pawlita, Carole Fakhry, Wayne M Koch, William H Westra, and Maura L Gillison. Case–control study of human papillomavirus and oropharyngeal cancer. *New England Journal of Medicine*, 356(19):1944–1956, 2007.

[2] David M Lee, Abdelouahid Tajar, Aslan Ulubaev, Neil Pendleton, Terence W O' neill, Daryl B O' connor, Gyorgy Bartfai, Steven Boonen, Roger Bouillon, Felipe F Casanueva, et al. Association between 25-hydroxyvitamin d levels and cognitive performance in middle-aged and older european men. *Journal of Neurology, Neurosurgery & Psychiatry*, 80(7):722–729, 2009.