Exercise 6B: Genalized Linear Models and Applications
2018 Spring

1 Background Knowledge

1.1 Maximum Likelihood

For a random variable $X \sim f(x|\theta)$, we define $l(x|\theta) = \log f(x|\theta)$ as the log-likelihood function, and

$$l'(x|\theta) = \frac{\partial}{\partial\theta} \log f(x|\theta) = \frac{f'(x|\theta)}{f(x|\theta)}$$

where $f'(x|\theta)$ is the derivative of $f(x|\theta)$ w.r.t. θ. Similarly, the second derivative $f''(x|\theta)$.
Fisher's information (for θ) contained in the random variable X is defined as:

$$I(\theta) = E_\theta \left\{ [l'(X|\theta)]^2 \right\} = \int [l'(X|\theta)]^2 f(x|\theta) dx \tag{1}$$

Assume that

$$\begin{array}{rcl} \int f'(x|\theta)dx & = & \frac{\partial}{\partial\theta} \in f(x|\theta) = 0 \\ \int f''(x|\theta)dx & = & \frac{\partial^2}{\partial\theta^2} \int f(x|\theta)dx = 0 \end{array}$$

It is easy to see that the expected value of the score function:

$$E_\theta[l'(X|\theta)] = \int l'(x|\theta) f(x|\theta) dx = \int \frac{f'(x|\theta)}{f(x|\theta)} f(x|\theta) dx = \int f'(x|\theta) dx = 0$$

Therefore, the definition of Fisher information can be written as:

$$I(\theta) = \mathrm{Var}_\theta[l'(X|\theta)] \tag{2}$$

Also,

$$l''(x|\theta) = \frac{\partial}{\partial\theta}\left[\frac{f'(x|\theta)}{f(x|\theta)}\right] = \frac{f''(x|\theta)f(x|\theta) - [f'(x|\theta)]^2}{[f(x|\theta)]^2} = \frac{f''(x|\theta)}{f(x|\theta)} - [l'(x|\theta)]^2$$

Therefore,

$$E_\theta[l''(X|\theta)] = \int \left[\frac{f''(x|\theta)}{f(x|\theta)} - [l'(x|\theta)]^2\right] f(x|\theta)dx = \int f''(x|\theta)dx - E_\theta\{[l'(X|\theta)]^2\} = -I(\theta)$$

Finally, we can compute the Fisher information:

$$I(\theta) = -E[l''(X|\theta)] = -\int \left[\frac{\partial^2}{\partial\theta^2} \log f(x|\theta)\right] f(x|\theta) dx \tag{3}$$

So, we have three methods, equation 1, 2, 3 to compute $I(\theta)$.

1.1.1 Newton-Raphson and Fisher Scoring Method

If there is no closed-form solution, MLE should be solved using iterative procedures.

Expanding the score function $u(\hat{\theta})$ around a trial value θ_0 using the first-order Taylor series:

$$u(\hat{\theta}) = u(\theta_0) + \frac{\partial u(\theta_0)}{\partial \theta}(\hat{\theta} - \theta_0) + o^{(n)}(\hat{\theta} - \theta_0).$$

Equating $u(\hat{\theta}) = 0$, then we have

$$\hat{\theta} \approx \theta_0 - \left(\frac{\partial u(\theta_0)}{\partial \theta}\right)^{-1} u(\theta_0)$$

Then **Newton-Raphson (NR)** procedure to iterate:

$$\theta^{(k+1)} = \theta^{(k)} - \left(\frac{\partial^2 l(\theta)}{\partial \theta^2}\right)^{-1} \frac{\partial l(\theta)}{\partial \theta}\Big|_{\theta=\theta^{(k)}}$$

An alternative approach replace the information marix $I_o(\theta)$ by its expected value $I_e(\theta)$:

$$\theta^{(k+1)} = \theta^{(k)} - E\left(\frac{\partial^2 l(\theta)}{\partial \theta^2}\right)^{-1} \frac{\partial l(\theta)}{\partial \theta}\Big|_{\theta=\theta^{(k)}}$$

This is called the **Fisher-scoring (FS)** procedure.

1.2 Model Diagnosis of the General Linear Models

1.2.1 Checking the assumptions of the residuals

$$\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- **normality assumption** can be assessed using **Q-Q plot**;
- **homoskedasticity assumption** can be assessed using **residuals vs. predicted responses scatterplot**;
- **independence assumption** can be assessed using **residuals plot**.

1.2.2 Checking the influential data

Influential data can be **outliers (离群数据)**, **leverage (杠杆数据)**.

- Outliers: Unusual dependent variables from the predicted model indicated by **large residuals**.
  - Studentized residuals r_i for measuring "outlierness".
  - $r_i = \frac{(y_i - \hat{y}_i)}{\text{SE}(r_i)}$
  - $\text{SE}(r_i) = \hat{\sigma}\sqrt{1 - h_i}$
  - $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}$
- Leverage: Unusual independent variables from the other observations indicated by **extreme predictor variable**.
  - h_i for measuring "unusualness" of x's

- $h_i = \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2} + \frac{1}{n}$

- Influence: Influence can be thought of as the product of leverage and outlierness. Removing the observation substantially changes the estimate of coefficients.
  - Cook's Distance $D_i = \sum_{j=1}^{n} \frac{(\hat{y}_{j(i)} - \hat{y}_j)^2}{psi\hat{g}ma^2}$
  - $\hat{y}_{j(i)}$ is the estimated y_j, based on the reduced data set without observation i.
  - p is the number of regression coefficents.
  - $\hat{\sigma}^2$ is the estimated variance from the fit based on all observations.

1.2.3 Testing for Collinearity

Multicollinearity means that two or more predictors are highly correlated with each other. Although this does not bias the estimates of the dependent variable, but it does affect the inference about the significance of the collinear variables.
We can use **variance inflation factors (VIF)** to help detect multicollinearity:

$$\text{VIF}_k = \frac{1}{1 - R_k^2}$$

where R_k^2 is the R^2-value obtained by regressing the k-th predictor on the remaining predictors. Any VIF value over 10 is worrisome.

1.2.4 Model Selection Strategies

In GLM, the **deviance** for model ω is defined as:

$$D(\omega) = 2a(\phi)[l_\Omega(\tilde{\theta}) - l_\omega(\hat{\theta})] = 2\sum_{i=1}^{n}[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)].$$

and

$$\frac{D(\omega_1) - D(\omega_2)}{\phi} \overset{n\to\infty}{\sim} \chi^2_{p_2 - p_1}$$

The scale parameter ϕ is either known or estimated using the larger model ω_2.

- (Ajusted) R-square (R^2)
- Mallow's C_p metric
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)
- Cross-validation

1.3 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:

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

1.4 Generalized Linear Models (GLMs) Fitting

With this we can obtain the **Fisher's score vector** $\mathbf{u} = (u_j)$ 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

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

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:

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

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

$$W = \text{diag}(w_1, \dots, w_n)$$

and

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

1.5 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 $y_i \sim Bin(n_i, p_i)$*

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

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

For the canonical link function:

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

2 Exercises

1. (30 points) Tell True or False for the following statements. And tell the reason.

   (1) ___ It can happen that all the individual t-tests for each coefficient in a regression model do not reject the null hypothesis, although the global F-test is significant.

   (2) ___ Let $[-0.01, 1.05]$ be the 95% confidence interval for the coefficient β_1 in the model $Y = \beta_0 + \beta_1 x + \epsilon$. Then a t-test will reject the null hypothesis $H_0 : \beta_1 = 0$ with $\alpha = 0.01$.

   (3) ___ The coefficient of determination R^2 cannot be used to compare the goodness of fit for a single model, on two separate dataset.

   (4) ___ A complicated models with many parameters are better for prediction than a simple model with only a limited number of parameters.

   (5) ___ The three fomula specify a same model: `z ~ x + y + x:y`, `z ~ x*y`, and `z ~ (x+y)^2`.

   (6) ___ The canonical link function for a Poisson distribution is $\log(\cdot)$ function.

   (7) ___ Compared with the R^2, AIC can balance between the goodness-of-fit and the model complexity.

   (8) ___ ℓ_2-regularzation term can be used to generate a sparse model.

   (9) ___ Penalized quasi-likelihood maximization can be used to fit a generalized linear mixed model.

   (10) ___ A larger BIC metric for a common data set indicates a better model for the data.

   (11) ___ Both GEE model and random effects model can conduct individual-level predictitions.

   (12) ___ One of the assumption for a general linear model is that the response variable should follow a normal distribution.

   (13) ___ Score function is the first derivative of the likelihood function, while Fisher's information matrix is the second moment of the score function.

   (14) ___ The assumption of linearity can be checked using the residual plots in either linear or generalized linear models.

(15) ___ Wald test is equivalent to t-test in general linear regression model.

2. (20 points) Let l be the log-likelihood function w.r.t. parameter θ, show that $E\left[\frac{\partial^2 l}{\partial\theta_i\partial\theta_j} + \frac{\partial l}{\partial\theta_i}\frac{\partial l}{\partial\theta_j}\right] = 0$. (Hint: Use the definition of Fisher information matrix.)

3. (20 points) Here is a 5×4 table giving the number of girls with 4 different rating for disturbed dreams in 5 different age groups. The higher the rating the more the girls suffer from disturbed dreams.

	Rating			
Age	4	3	2	1
5-7	7	3	4	7
8-9	13	11	15	10
10-11	7	11	9	23
12-13	10	12	9	28
14-15	3	4	5	32

One way of modeling the data is to let the number in the (i, j)-th cell be Y_{ij} and assume that $Y_{ij} \sim \text{Poisson}(\mu_{ij})$.

We seek for a pattern in the μ_{ij} relating to the factors `age` and `rating` using `log` link.

The saturated model Ω is:

$$\log \mu_{ij} = \mu_0 + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

where α_i is the age group effect, while β_j is the dream rating effect, $(\alpha\beta)_{ij}$ is the interacting effects. We set $\alpha_1 = \beta_1 = 0$ and $(\alpha\beta)_{1j} = 0 = (\alpha\beta)_{i1}$.

(1) Compute the deviance for the null model.

(2) Compute the deviance for the additive model without the interaction terms and also the AIC.

(3) Test if $\beta_1 = \beta_2 = \beta_3$?

(4) Compute the Pearson (or standardized) residuals: $R_{pi} = \frac{Y_i - \mu_i}{\sqrt{\text{Var}(Y_i)}}$ where $\text{Var}(Y_i) = a(\phi)b''(\theta_i)$ and try to interpret the residuals plot.

4. (20 points) The Gamma distribution has a density function $f(y) = \lambda^\alpha/\Gamma(\alpha) y^{\alpha-1}\exp(-\lambda y)$. With the reparameterization $\mu = \alpha/\lambda$, it can also be written as $f(y) = \frac{(\alpha/\mu)^\alpha}{\Gamma(\alpha)} y^{\alpha-1}\exp(-\alpha y/\mu)$.

(1) Suppose that the Gamma distribution belongs to the exponential family. Identify θ, $b(\theta)$.

(2) Suppose that the mean of y, μ is related to covariate vector x via $g(\mu) = \mu^{-1} = x^T\beta$, where β is the unknown regression coefficient vector. Assume n independent and identically distributed observations. Describe the IWLS algorithm for estimating β. Clearly identify the weight function.

5. (10 points) Write down the deviance and scaled deviance for

(1) Poisson distribution

(2) Binomial distribution

(3) Gamma distribution

(4) Inverse Gaussian distribution

(5) Negative binomial distribution.