BIOS 6110
Applied Categorical Data Analysis

Instructor: Yuan Huang, Ph.D.

Department of Biostatistics

Fall 2017

Part III

Generalized Linear Models

Objectives

Key concepts for GLM:

Three components of GLM

- Random Component, Systematic Component, Link Function

Fitting GLM

- Newton-Raphson Algorithm, Fisher Scoring, IRWIS

GLM Diagnostics

- Residuals: Person Residuals, Standardized/Adjusted Residuals, Deviance Residuals
- Hypothesis testing for coefficients: Wald Test, Score Test, Likelihood Ratio Test
- Model discrepancy/Goodness of fit: Deviance, Pearson X^2.

Reading: Agresti (2002), Section 3.1, 3.4 - 3.5

Review: Simple Linear Regression Model

- Data: $(x_1, y_1), \ldots, (x_n, y_n)$ for n subjects
- Objective: model the expected value of a continuous variable, Y , as a linear function of the continuous predictor, X. Denote the mean of Y_i as μ_i, then $\mu_i = E(Y_i) = \beta_0 + \beta_1 x_i$.
- Model structure: $Y_i = \mu_i + \epsilon_i$
- Model assumption: $Y_1, \ldots, Y_n$ are independent and normally distributed. $\epsilon_i \sim N(0, \sigma^2)$.
- Model fitting: Ordinary Least Square (OLS), find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ is minimized. Then $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
- Model fit: R^2, residual analysis, F-statistics
- Model selection: stepwise, AIC, BIC, ...

Question: how to handle the case with binary or count data as the response, Y.

Generalized Linear Models: Three Components

Data: $(x_i, y_i), i = 1, \ldots, n$ for n independent subjects, where $x_i = (x_{i1}, \ldots, x_{ip})$ is a vector that x_{ij} is the j-th variable for the i-th subject.

Generalized Linear Models (GLM) refers to a larger class of models where the response variable y_i is assumed to follow an exponential family distribution with mean μ_i that is some function of $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$.

Three Components:

- Random components: specify a probability distribution from the exponential family for Y. Denote the mean of Y_i as μ_i.
- Systematic components: specify a linear combination of covariates $\eta_i = x_i^t \beta$, where $x_i^t \beta = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$.
- Link function $g(\cdot)$: specify how the mean of the specified probability distribution (random component) relates to the linear predictor (systematic component). That is, $g(\mu_i) = \eta_i$.

Exponential Family

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

- θ: Canonical/Natural parameter. If $g(\mu) = \theta$, then the link is called canonical link.
- ϕ: Dispersion/Scale parameter, which is fixed and constant over observations. Because ϕ is known, we can also write $f(y;\theta)$. If ϕ is unknown, then we will have exponential-dispersion family, or two-parameter exponential family.
- $a(\phi)$ is commonly of the form $a(\phi) = \phi/w$, where w is a known prior weight that may vary from observation to observation

Properties: if $y \sim f(y;\theta,\phi)$, then

- Mean of y is $E(y) = b'(\theta)$
- Variance of y is $Var(y) = b''(\theta)a(\phi)$. We also denote the $b''(\theta)$ as $V(\theta)$, the variance function.

Link Function $g(\cdot)$

For one random component, there are possibly multiple choices for $g(\cdot)$.

[Example] For the binary response, we can specify binomial distribution as the random component and the mean is $\pi(x)$. For the link function, we can have

- Identity link: $\pi(x) = \alpha + \beta x$. Here the β corresponds to the change in the probability per unit change in x. The problem is that this link may lead to the $\pi(x)$ outside the $(0, 1)$ range.
- Logit Link: $\log(\pi(x)/(1 - \pi(x))) = \alpha + \beta x$. This is the canonical link for binomial random component.
- Probit Link: $\Phi^{-1}(\pi(x)) = \alpha + \beta x$, where Φ^{-1} is the inverse of the cumulative probability function of normal random variable.
- Complementary log-log link: $\log(-\log(1 - \pi(x))) = \alpha + \beta x$.

Except the identity link, the rest links project the $\pi(x)$ from (0,1) to the real line $(-\infty, \infty)$. It is often convenient to use a canonical link. But convenience does not imply that the data actually conform to it.

Examples of Exponential Family

To validate an exponential family distribution, the key is to be able to format the distribution function into the above form. We will use the properties of log and exp: $\exp(\log(a)) = \log(\exp(a)) = a$

$$\log(a \times b/c) = \log(a) + \log(b) - \log(c).$$

[Example] Possion Distribution

$$\begin{aligned} f(y;\theta) &= \frac{\lambda^y e^{-\lambda}}{y!} \\ &= \exp\left(\log\left(\frac{\lambda^y e^{-\lambda}}{y!}\right)\right) \\ &= \exp\left(y \log \lambda - \lambda - \log y!\right) \end{aligned}$$

Hence,

- $\theta = \log \lambda$, therefore, $\lambda = e^{\theta}$
- $b(\theta) = \lambda$, that is, $b(\theta) = e^{\theta}$
- $\mu = b'(\theta) = e^{\theta} = \lambda$
- Therefore, the canonical link is $\log(\mu)$

[Example] Binomial Distribution

For the Binomial case, the distribution of $y = s/n$ is used, where s is the number of success that $s \sim Binom(n, \pi)$.

$$\begin{aligned} f(y;\theta) &= \binom{n}{ny}\pi^{ny}(1-\pi)^{n-ny} = \binom{n}{ny}\left(\frac{\pi}{1-\pi}\right)^{ny}(1-\pi)^{n} \\ &= \exp\left(\log\left(\binom{n}{ny}\left(\frac{\pi}{1-\pi}\right)^{ny}(1-\pi)^{n}\right)\right) \\ &= \exp\left(\log\binom{n}{ny} + ny\log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi)\right) \\ &= \exp\left(\log\binom{n}{ny} + \frac{y\log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)}{1/n}\right) \end{aligned}$$

Hence,

- $\theta = \log\left(\frac{\pi}{1-\pi}\right)$, therefore, $\pi = \exp\theta/(1+\exp\theta)$
- $b(\theta) = -\log(1-\pi) = \log(1+\theta)$
- $\mu = b'(\theta) = \exp\theta/(1+\exp\theta) = \pi$
- Therefore, the canonical link is $\text{logit}(\mu) = \log(\mu/(1-\mu))$

Advantages of GLM

- We do not need to transform the response Y to have a normal distribution
- The choice of link is separate from the choice of random component thus have more flexibility in modeling
- The models are fitted via Maximum Likelihood estimation; thus optimal properties of the estimators.

[Caution]

The normal linear regression model has an additive error, so we may write $y = \beta_0 + \beta_1 x + \epsilon$, where ϵ is the random error.

However, GLM does not have this structure. For example, in the logit model, we cannot write $\log(\pi/(1-\pi)) = \beta_0 + \beta_1 x + \epsilon$. For this model, the random error is contained in the random component $y \sim n^{-1} Binom(n, \pi)$, and $g(\mu) = \eta$ is a purely functional (deterministic) relationship.

Relation to The Rest of The Chapters

Model	Ranodm Component	Link Function	Covariates Type	Chapter
Linear Regression	Normal	Identity	Mixed	Classic
Logistic Regression	Binomial	Logit	Mixed	Ch 4 and 5
Multinomial	Multinomial	Generalized	Mixed	Ch 6
Loglinear Model	Poisson	Log	Categorical	Ch7
Poisson Regression	Poisson	Log	Mixed	Ch 3.3

- The above models concern more on the independent observations.
- For the correlated (repeated, matched, etc) observations, we will study them on Ch 8, 9, and 10.

Fit GLM

GLM is estimated by MLE.

From the random component $\rightarrow$ liklihood function $l = \prod f(y_i; \theta) \rightarrow$ log-likelihood function $L = \sum \log f(y_i; \theta)$. The rest of the work is to maximize L. Equivalently, to solve the Score equations

$$U_j(\beta) = \frac{\partial}{\partial \beta_j} L, \quad j = 1, 2, \ldots, p.$$

Because L is defined via θ, chain rule will be applied to get the $U_j(\beta)$.

$$U_j(\beta) = \sum_i^n \frac{\partial \log L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\beta_j}$$

The $U_j(\beta)$ are called scores and can be written in the vector form.

$$\begin{pmatrix} U_1(\beta) \\ U_2(\beta) \\ \ldots \\ U_p(\beta) \end{pmatrix}$$

Algorithms

Newton-Raphson algorithm. is an iterative algorithm. As an illustration, to solve $t(x) = 0$. At the m-th step, denote $t'(x^{(m-1)})$ as the derivative of function t evaluated at $x^{(m-1)}$. From $t'(x^{(m-1)}) = \frac{t(x^{(m)}) - t(x^{(m-1)})}{x^{(m)} - x^{(m-1)}}$ and use $t(x^{(m)}) = 0$, we can obtain the updated $x^{(m)} = x^{(m-1)} - \frac{t(x^{(m-1)})}{t(x^{(m)})}$.

In our case, we have p-dimensional vector to estimate, we will use the matrix form of inversion. Note that we will need first-order derivative of score functions, which is second derivative of L.

- Newton-Raphson: uses the observed derivative of scores $\frac{\partial^2 L}{\partial \beta_i \beta_j}$
- Fisher scoring: uses the expected derivative of scores $E(\frac{\partial^2 L}{\partial \beta_i \beta_j})$
- Fisher scoring algorithm can be written as an iterative re-weighted least squares (IRWLS).

The expected and observed second derivative of scores are the same under canonical links. The IRWLS form suggests diagnostic techniques to check the appropriateness of the model.

Measuring Goodness of Fit

Null model: one parameter, representing a common mean for all y_is. That is, for all subjects i, $\hat{\mu}_i = \sum_{i=1}^{n} y_i/n$

Full/Saturated model: n parameters, no errors, the mean matches the data exactly. That is, $\hat{\mu}_i = y_i$.

It is often convenient to express the log-likelihood function in terms of mean value parameter μ rather than the canonical parameter θ. So the log-likelihood functions for the full models are $L(y, \phi; y)$.

Most often, we have a model M between these two extreme. We may write it as $L(\hat{\mu}, \phi; y)$.

- Deviance $= 2\left[L(y, \phi; y) - L(\hat{\mu}, \phi; y)\right]$
  This is also the test statistic for the hypothesis testing that all parameters that are in the full model but not in the model M are 0s. It is a χ^2 with df $= (n-$ number of parameters in model $M)$.
- Pearson's $\chi^2 = \sum_{i=1}^{N} \dfrac{(y_i - \hat{\mu}_i)^2}{\hat{Var}(y_i)}$

Residuals

Pearson residuals:

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{Var}(y_i)}}$$

Standardized/Adjusted residuals:

$$r_i^S = \frac{y_i - \hat{\mu}_i}{\text{SE of } (y_i - \hat{\mu}_i)}$$

Deviance residuals:

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i},$$

where d_i is the contributions of the ith unit in the Deviance. That is, $d_i = 2\left[L(y_i, \phi; y) - L(\hat{\mu}_i, \phi; y)\right]$ and Deviance $= \sum_{i=1}^{N} d_i$.

Inference for β

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \begin{matrix} q \times 1 \\ (p-q) \times 1 \end{matrix}$$

If we would like to test the second part of β, that is, $H_0 : \beta_2 = \beta_2^0$

- Likelihood Ratio Test $= 2\left(L(\hat{\beta}_1, \hat{\beta}_2) - L(\hat{\beta}_1^0, \beta_2^0)\right)$, where $\hat{\beta}_1$ and $\hat{\beta}_2$ are obtained from MLE under the alternatives, and the $\hat{\beta}_1^0$ is the MLE under the null that $\beta_2 = \beta_2^0$.
- Wald Test $= \left(\hat{\beta}_2 - \beta_2^0\right)' \left(Cov(\hat{\beta}_2)\right)^{-1} \left(\hat{\beta}_2 - \beta_2^0\right)$
- Score Test $= U_2(\beta_2^0)' \left\{Cov(U_2(\beta_2^0))\right\}^{-1} U_2(\beta_2^0)$, where $U_2(\beta_2)$ correspond to the score function for the part of β_2. That is,

$$U(\beta) = \begin{pmatrix} U_1(\beta_1) \\ U_2(\beta_2) \end{pmatrix} \begin{matrix} q \times 1 \\ (p-q) \times 1 \end{matrix}$$

All three tests are χ^2 with df $= (p-q)$ under large samples. For small samples, the likelihood ratio test is more reliable.

Model Diagnostics

Plot the residuals on the vertical axis versus the linear predictor η on the horizontal axis.

- We hope to see something like a "horizontal band" with mean ≈ 0 and constant variance as we move from left to right.
- Curvature in the plot may be due to a wrong link function or the omission of a nonlinear (e.g. quadratic) term for an important covariate.
- Non-constancy of range suggests that the variance function may be incorrect.

For binary responses, this plot is not very informative; all the points will lie on two curves, one for $y = 0$ and the other for $y = 1$. However, the plot may still help us to find outliers.

Plot residuals versus individual covariates.

- Again, we hope to see something like a "horizontal band".
- Curvature in this plot suggests that the x-variable in question ought to enter into the model in a nonlinear fashion. For example, we might add a quadratic term x^2 or consider other transformation $\sqrt{x}$.

Plot the absolute residuals with fitted values μ, to check the appropriateness of variance function.

- If there is no trend, the variance function is probably okay.
- An increasing trend (positive slope) suggests that the variance function is increasing too slowly with the mean; for example, $V(\mu) = \mu$ might have to be replaced with $V(\mu) = \mu^2$.

Checking the link function.

The simplest way to check the link function is to plot the final value of the working variate z against the linear predictor η. The plot should resemble a straight line. Curvature in this plot suggests that the link function is not appropriate.

For binary data, this plot is uninformative.

Hinkley (1985) suggests the following test: After fitting the GLM, calculate η^2 and then try adding it to the model as a covariate. If this covariate is significant, then there is a problem. Significance in this test could be caused by a wrong link function, a wrong scale for one or more predictors, or both.