Course Format.

一节 Lecture + 一节 Lab.

=

GLM. — Revisit to GLM. — from math./stat prospective.

1. Why GLM.

Linear Regression

Assume. $Y_1, \cdots Y_n$ ind., satisfy $E(Y_i) = x_i^T\beta$.

x_i — observed predictors / covariates.

β — regression coefficients.

→ If we further assume

$Y_i \overset{ind}{\sim} N(x_i^T\beta, \sigma^2)$. $i = 1, \cdots n$ $\sigma^2 > 0$. unknown.

ordinary linear regression.

exact inference about β. based on t, & F tests.

→. If we drop normality assumption.

$Y_i \overset{ind}{\sim} (x_i^T\beta, \sigma^2)$ $i = 1, \cdots n$.

when n is large.

standard tests. CI. are approximately correct. robust. to voilation

*. In many situations even these weakened assumption are untenable.

- $\mu(x) := E(Y|x)$ is not a linear function of x
- $Var(Y_i)$ is not constant in i

Possible solution

- weighted least squares. if $Var(Y_i) = a_i^2\sigma^2$. $a_1, \cdots a_n$ are known constants
- transformation of $\vec{x}$ to correct nonlinerity. Focus. [Figure 8.3. CASI].
- transformation of Y if either $\mu(x) = E(Y|x)$ is nonlinear.

  or $\sigma^2(x) := Var(Y|x)$ is not constant in $\vec{x}$

2. Defining GLM.

Example.

- Gaussian — identity.
- binary. — logit.

Systematic Component. — relates predictors. $\vec{x} = (x_1, \cdots, x_p)$. to the mean response $\mu = E(Y)$.

- the linear predictor : $\eta = \vec{x}^T \vec{\beta}$
- the link function : $g(\mu) = \eta$, where g is a smooth. monotonic function.

Random (Stochastic) Component — specifies distributional form of the responses. It assume

- $Y_1, \cdots, Y_n$ ind.
- Y_i has density $f(\cdot\,; \theta_i, \phi_i)$, $\phi_i = \phi / a_i$, $\phi > 0$. where $a_1, \cdots, a_n$ are known.

  and f has the form

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

$\Rightarrow$ dispersion parameter.

For fixed ϕ, $f(\cdot\,; \theta, \phi)$ defines a one-parameter exponential family.

↑ canonical parameter.

HW ?

Let $l(\theta, \phi) = \log f(y; \theta, \phi) = \frac{1}{\phi}[y\theta - b(\theta)] + c(y; \phi)$.

Then $\frac{\partial l}{\partial \theta} = \frac{1}{\phi}[y - b'(\theta)]$.

$\frac{\partial^2 l}{\partial \theta^2} = -\frac{1}{\phi} b''(\theta)$.

From the first two Bartlett identities

$$E_{\theta,\phi}\left(\frac{\partial l}{\partial \theta}\right) = 0, \quad E_{\theta,\phi}\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right] = E_{\theta,\phi}\left(-\frac{\partial^2 l}{\partial \theta^2}\right).$$

$\Rightarrow$ $\underset{\mu.}{\underbrace{E_{\theta,\phi}(Y)}} = b'(\theta)$, $\mathrm{Var}_{\theta,\phi}(Y) = \phi\, b''(\theta)$. $\xrightarrow[\phi > 0]{\mathrm{Var}(Y) > 0}$ $b''(\theta) > 0$. $\Downarrow$ $b'(\theta) \uparrow$.

$\Rightarrow$ $\theta = (b')^{-1}(\mu)$.

$\mu = g^{-1}(\eta) = g^{-1}(\vec{x}^T \vec{\beta})$.

$\Rightarrow$ $\theta = (b')^{-1}(g^{-1}(\vec{x}^T \vec{\beta}))$.

- The function $(b')^{-1}(\mu)$ — canonical link.

  If we choose $g = (b')^{-1}$

  $\theta = \vec{x}^T \vec{\beta} = \eta$. — simplifies calculations.

  $g'(\mu) = \frac{1}{b''((b')^{-1}(\mu))} = \frac{1}{V(\mu)}$

$\mathrm{Var}(Y) = \phi\, b''((b')^{-1}(\mu)) = \phi\, V(\mu)$.

$b''((b')^{-1}(\mu)) \;\text{!!}\; V(\mu)$ — variance function.

Give an Example. poisson (μ). [Example 1.7]

HW on. Bionomial. $\Rightarrow$ logistic.

3 The GLM Likelihood.

Recall that in a GLM, $Y_1, \dots Y_n$ are independent with.

$$Y_i \sim f(\cdot\,; \theta_i, \phi_i), \quad \phi_i = \phi / a_i. \qquad \text{where.}$$

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

Further. θ_i is a function of μ_i. : $\mu_i = b'(\theta_i) \Rightarrow \theta_i = (b')^{-1}(\mu_i)$;

μ_i is a function of η_i : $g(\mu_i) = \eta_i \Rightarrow \mu_i = g^{-1}(\eta_i)$;

η_i is a function of β : $\eta_i = \vec{x}_i^T \vec{\beta}$

$\Rightarrow$. each θ_i is a function of the unknow $\vec{\beta}$, while ϕ (known or unknown) is free of $\vec{\beta}$.

↳. Need to estimate $\vec{\beta}$ and possibly ϕ.

- Likelihood Equations for the Regression Coefficients

The contribution of the i-th observation to the log-likelihood is.

$$l_i(\vec{\beta}, \theta) = \log f(y_i; \theta_i, \phi/a_i) = \frac{y_i\theta_i - b(\theta_i)}{\phi / a_i} + c(y_i; \phi/a_i)$$

$\Rightarrow l = \sum_{i=1}^{n} l_i$, ϕ ~~is~~ does not depend on $\vec{\beta}$.

By the chain rule, $$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \beta_j} + \underbrace{\frac{\partial l_i}{\partial \phi} \cdot \frac{\partial \phi}{\partial \beta_j}}_{=0} = \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j}$$

Let $V(\mu) = b''((b')^{-1}(\mu))$ be the variance function. $\Rightarrow \mathrm{Var}(Y) = \phi V(\mu)$.

Then $$\frac{\partial l_i}{\partial \theta_i} = \frac{1}{\phi} a_i (y_i - \mu_i). \qquad \rightarrow \mu_i = b'(\theta_i).$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\partial \mu_i / \partial \theta_i} = \frac{1}{b''(\theta_i)} = \frac{1}{b''((b')^{-1}(\mu_i))} = \frac{1}{V(\mu_i)}$$

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{\partial \eta_i / \partial \mu_i} = \frac{1}{g'(\mu_i)}$$

$$\eta_i = \vec{x}_i^T \vec{\beta} = \sum_{j=1}^{p} x_{ij}\beta_j \qquad \frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$$

$$\Rightarrow \frac{\partial l_i}{\partial \beta_j} = \frac{1}{\phi} \frac{a_i(y_i - \mu_i)}{V(\mu_i) g'(\mu_i)} \cdot x_{ij}$$

$$\Rightarrow \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{a_i(y_i - \mu_i)}{V(\mu_i) g'(\mu_i)} x_{ij}$$

MLE of $\vec\beta$. $\frac{\partial l}{\partial \beta_j} = 0$. Likelihood equation for $\vec\beta$

$$\Rightarrow \sum_{i=1}^n \frac{a_i(y_i - \mu_i)}{V(\mu_i)\, g'(\mu_i)} \cdot x_{ij} = 0 \quad j = 1, \cdots, p. \quad \text{for } \vec\beta = (\beta_1, \cdots, \beta_p)^T.$$

- Fisher Information.

[Definition]. - quantifies how much information an observed r.v. X carries about an unknown parameter θ of a statistical model.

Formally, for a likelihood $L(\theta; X)$

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta} \log L(\theta; X)\right)^2\right].$$

Assume ϕ is known. the observed Fisher Information is $I(\hat\beta)$. $\hat\beta$ is the MLE of $\vec\beta$.

$$I(\beta) = -\frac{\partial^2 L}{\partial\beta\,\partial\beta^T} = \left(-\frac{\partial^2 L}{\partial\beta_j\,\partial\beta_k}\right)_{1\le i,j\le p} \quad \text{- negative - Hessian of the log-likelihood.}$$

Note $\frac{\partial^2 L_i}{\partial\beta_j\,\partial\beta_k} = \frac{\partial}{\partial\beta_k}\left(\frac{1}{\phi} \frac{a_i(y_i-\mu_i)}{V(\mu_i)g'(\mu_i)} x_{ij}\right) = \frac{\partial}{\partial\mu_i}\left(\frac{1}{\phi}\frac{a_i(y_i-\mu_i)}{V(\mu_i)g'(\mu_i)} \cdot x_{ij}\right)\cdot\frac{\partial\mu_i}{\partial\beta_k}$

$$\frac{\partial}{\partial\mu_i}\left(\frac{1}{\phi}\frac{a_i(y_i-\mu_i)}{V(\mu_i)\,g'(\mu_i)} x_{ij}\right) = -\frac{1}{\phi}\frac{a_i}{V(\mu_i)\cdot g'(\mu_i)} x_{ij} + \frac{1}{\phi} a_i(y_i-\mu_i)\cdot\frac{\partial}{\partial\mu_i}\left(\frac{1}{V(\mu_i)\,g'(\mu_i)}\right).$$

$$\frac{\partial\mu_i}{\partial\beta_k} = \frac{\partial\mu_i}{\partial\eta_i}\cdot\frac{\partial\eta_i}{\partial\beta_k} = \frac{1}{g'(\mu_i)} x_{ik}.$$

$$\Rightarrow I(\beta)_{jk} = -\sum_{i=1}^n \frac{\partial^2 l_i}{\partial\beta_j\,\partial\beta_k} = \frac{1}{\phi}\sum_{i=1}^n\left\{\frac{a_i}{V(\mu_i)\,g'(\mu_i)^2} - \frac{a_i(y_i-\mu_i)}{g'(\mu_i)}\frac{\partial}{\partial\mu_i}\left(\frac{1}{V(\mu_i)g'(\mu_i)}\right)\right\} x_{ij}\, x_{ik}.$$

depend on $\vec\beta$ through μ_i

$\Rightarrow$ $I(\vec\beta) = E_{\vec\beta}[\,I(\vec\beta)\,]$. since only y_i are random. & $E[(Y_i - \mu_i)] = 0$.

$$I(\beta)_{j,k} = \frac{1}{\phi}\sum_{i=1}^n \frac{a_i}{V(\mu_i)\,g'(\mu_i)^2} x_{ij}\, x_{ik}.$$

Let $W = W(\vec\beta) = \operatorname{diag}\left\{\frac{a_i}{V(\mu_i)\,g'(\mu_i)^2} : i = 1, \cdots, n\right\}$.

$$\Rightarrow I(\vec\beta) = \frac{1}{\phi} X^T W X.$$

? HW. canonical case.

Large sample theory. MLE of $\vec\beta$ $\hat{\vec\beta} \sim AN(\beta, \phi(X^TWX)^{-1})$. as. $n\to\infty$.

4. Computation of Estimators.

- Newton's method.
- Iteratively reweighted least squares.

5. Deviance - a measure of fit.

analogous the ~~sum~~ residual sum of squares in Linear regression.

Let $l(\mu, \phi; y) = \sum_{i=1}^{n} \log f(y_i; \theta_i, \phi/a_i) = \frac{1}{\phi} \sum_{i=1}^{n} a_i [y_i \theta_i - b(\theta_i)] + \sum_{i=1}^{n} c(y_i; \phi/a_i)$

where $\theta_i = (b')^{-1}(\mu_i)$.

For GLM with $\eta_i = x_i^T \beta$ & $g(\mu_i) = \eta_i$, $i = 1, \cdots n$

let $\hat{\beta}$ be the MLE, and let

$\hat{\eta}_i = x_i^T \hat{\beta}$, $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$. $\hat{\theta}_i = (b')^{-1}(\hat{\mu}_i)$. and $\tilde{\theta}_i = (b')^{-1}(y_i)$.

Def. With the notation above, the deviance for the fitted GLM model is

$$D(\vec{y}; \hat{\mu}) = 2[l(\vec{y}, \phi; \vec{y}) - l(\hat{\mu}, \phi; \vec{y})] \cdot \phi$$

$$= \sum_{i=1}^{n} 2 a_i \{ y_i(\tilde{\theta}_i - \hat{\theta}_i) - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)] \}.$$

$\rightarrow$ Does not depend on ϕ.

$\downarrow$

It is twice the "Kullback-Leibler" distance"

and the scaled deviance is

$$D^*(\vec{y}; \hat{\mu}) = \frac{1}{\phi} D(\vec{y}; \hat{\mu}).$$

* GLM. Maximum likelihood fitting is "least total deviance" in the same way that Linear regression is least sum of squares.

i.e. the MLE $\hat{\beta}$ is the choice of β that minimizes the total deviance.

- Analysis of Deviance.

Suppose $M_0 \subset M_1$. $\hat{\beta}_0$, $\hat{\beta}_1$ corresponding MLE with $\hat{\mu}_0$, $\hat{\mu}_1$. Assume ϕ is known.

the likelihood ratio test statistics for H_0: M_0 is the true model

vs H_1: M_1 is the true model.

$$T_{LR} =: -2[l(\hat{\mu}_0, \phi; \vec{y}) - l(\hat{\mu}_1, \phi; \vec{y})] = \frac{1}{\phi}[D(\vec{y}; \hat{\mu}_0) - D(\vec{y}; \hat{\mu}_1)] = D^*(y, \hat{\mu}_0) - D^*(y, \hat{\mu}_1)$$

$\xrightarrow[H_0]{d} \chi^2_r$ as $n \to \infty$.

r: difference in the # of para. between M_0 & M_1.

We reject M_0 if $T_{LR} > \chi^2_{r,\alpha}$.