GLM summary

$\boxed{\text{Recall:}}$

Linear Regression Model

$\rightarrow \boxed{Y \sim N(X\beta, \Sigma)}$ where $\Sigma = \begin{pmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{pmatrix}$, $X = \begin{pmatrix} 1 & x_1^T \\ \vdots & \vdots \\ 1 & x_n^T \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$

$\downarrow$

$\boxed{\text{response}}$

$\boxed{Y_i = \beta_0 + x_i^T \beta + \varepsilon_i}$

$\boxed{E[Y_i] := \mu_i = \beta_0 + x_i^T \beta}$

$$\Rightarrow \log f(y; \beta) = \sum_{n=1}^{N} \log \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - x_i^T\beta)^2}{2\sigma^2}\right)\right)$$

$$= \sum_{n=1}^{N} \left[\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(y_i - x_i^T\beta)^2}{2\sigma^2} \right)$$

$$\Rightarrow \begin{cases} \hat{\beta} = (X^TX)^{-1}X^TY \\ \hat{\sigma}^2 = \dfrac{\sum_{n=1}^{N}(y_i - \hat{y}_i)^2}{N} \end{cases}$$

$$\left[\begin{array}{l} E[\hat{\beta}] = (X^TX)^{-1}X^T E[Y] = (X^TX)^{-1}X^TX\beta = \beta \\ Var[\hat{\beta}] = (X^TX)^{-1}X^T Var[Y] X (X^TX)^{-1} \\ \qquad = \sigma^2 (X^TX)^{-1} \end{array} \right.$$

$$:= \frac{RSS}{N}$$

$$E[\widehat{\sigma^2_{MLE}}] = \frac{N-p-1}{N}\sigma^2 \Longrightarrow \text{Biased estimator}$$

Modification

$$\boxed{\widehat{\sigma^2}_{unbias} = \frac{RSS}{N-p-1}}$$

$\boxed{\text{Decomposition of } \sum_{i=1}^{N}(y_i - \bar{y})^2}$

$\boxed{\text{only holds}}$ for LR without Regularization Term

Property:

$$\begin{cases} \boxed{\text{rank}(H) = p+1} \\ H = X(X^TX)^{-1}X^T \in \mathbb{R}^{n \times n} \end{cases}$$

$$\boxed{HX = X \Rightarrow H \cdot \mathbb{1}_n = \mathbb{1}_n}$$

$$\sum_{i=1}^{N}(y_i - \bar{y})^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 + \cancel{\sum_{i=1}^{N} 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}$$

$$\hat{e}^T(H - \tfrac{1}{n}\mathbb{I})y$$

$$= y^T(I - H)(H - \tfrac{1}{n}\mathbb{I})y$$

$$\Rightarrow \boxed{TSS = RSS + \text{Regression sum of squares error}}$$

① $\boxed{R^2 = 1 - \frac{RSS}{TSS}}$

$\in [0,1]$

② when model is larger (more features ≈ 0),

R^2 must be bigger !!!

$\Downarrow$

$\boxed{\text{do feature selection}}$

$\boxed{\text{GLM}}$ $\rightarrow$ when response is no longer Gaussian distributed given the params

1) $\boxed{\eta_i = x_i^T \beta}$ $\rightarrow$ linear predictor [systematic component]

2) $\boxed{g(\mu_i) = \eta_i}$ $\rightarrow$ link function

3) $\boxed{Y_i \sim \text{exp. family}}$ $\quad f(Y_i \mid \theta_i, \phi) = \exp\left\{ \frac{\theta_i Y_i - b(\theta_i)}{a_i(\phi)} + c(\phi, Y_i) \right\}$

random component

(distributed of response)

$\Downarrow$

$$\begin{cases} E[Y_i] = b'(\theta_i) = \mu_i \\ Var[Y_i] = a_i(\phi)\, b''(\theta_i) \end{cases}$$

$$\theta_i \xleftarrow{(b')^{-1}(\mu_i)} \mu_i \xleftarrow{g^{-1}(\eta_i)} \eta_i \xleftarrow{x_i^T\beta} \beta$$

$\Rightarrow$ if canonical link function is used,

then $g(\mu_i) = \theta_i \Rightarrow \boxed{g = (b')^{-1}}$

$\Downarrow$

$\boxed{\theta_i = \eta_i = X_i^T \beta}$

$\Rightarrow f(Y_i; \theta_i, \phi) = \exp\left\{ \frac{\beta^T X_i Y_i - b(\beta^T X_i)}{a_i(\phi)} + c(Y_i, \phi) \right\}$

$\boxed{\text{Conclusion}}$: (for canonical link)

① $X^T Y$ is a Sufficient Statistic for β

② $J(\beta) = -\frac{\partial^2 \log f}{\partial \beta^2}$ & $I(\beta) = E[J(\beta)]$

observed info matrix $\quad$ Fisher info matrix we have $J(\beta) \equiv I(\beta)$

③ $\boxed{\text{Newton-Raphson Method}}$

$\Updownarrow$

$\boxed{\text{Fisher-Scoring Algo}}$

$\boxed{\text{Notation}} \longrightarrow \underline{S(\beta) = \frac{\partial \log f}{\partial \beta}}$

$$\boxed{\begin{aligned} f(\beta) &:= \sum_{i=1}^{n} f_i(\beta) \\ &= \sum_{i=1}^{n} \log\{ f(y_i \mid x_i, \beta)\} \end{aligned}}$$

$\longrightarrow$ $\boxed{\text{Newton-Raphson Method}}$

$$\beta^{k+1} = \beta^k + [J(\beta^k)]^{-1} \cdot S(\beta^k)$$

$\longrightarrow$ $\boxed{\text{Fisher-Scoring Method}}$

$$\beta^{k+1} = \beta^k + [I(\beta^k)]^{-1} S(\beta^k)$$

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n} S_i(\beta) \\ &= \sum_{i=1}^{n} \frac{\partial f_i}{\partial \beta} \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \log\{f(y_i \mid x_i, \beta)\} \end{aligned}$$

$E_\beta[S_i(\beta)] = 0$

when have β^k

$\downarrow$

$\eta^k = x\beta^k$

$\downarrow$

$g(\mu^k) = \eta^k$

$\downarrow$

$b'(\theta^k) = \mu^k$

$\boxed{S(\beta)}$

$$\begin{aligned} \frac{\partial \log f(\beta)}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{a_i(\phi)} \cdot \frac{\partial \theta_i}{\partial \beta_j} \\ &= \sum_i \frac{y_i - \mu_i}{a_i(\phi)} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} \\ &= \sum_i \frac{y_i - \mu_i}{a_i(\phi)} \cdot \frac{1}{b''(\theta_i)} \cdot \frac{1}{g'(\mu_i)} \cdot x_{ij} \\ &= \frac{1}{\phi} \sum_i (y_i - \mu_i) \cdot w_i \widetilde{w}_i(\mu_i)\, x_{ij} \end{aligned}$$

$$\boxed{\begin{aligned} \widetilde{w}_i(\mu_i) &= \frac{1}{b''(\theta_i) g'(\mu_i)} \\ &= \frac{1}{V(\mu_i) g'(\mu_i)} \end{aligned}}$$

$$\begin{aligned} \Rightarrow S(\beta) &= (x_1, \ldots\ldots, x_n)\, \widetilde{W}\, (Y - \mu) \\ &= X^T \widetilde{W} (Y - \mu) \end{aligned}$$

---

①

$$\frac{\partial^2 \log f}{\partial \beta_j \partial \beta_k} = \frac{1}{\phi} \sum_i w_i x_{ij} \left[(y_i - \mu_i) \frac{\partial \widetilde{w}_i}{\partial \beta_k} - \frac{\partial \mu_i}{\partial \beta_k} \cdot \widetilde{w}_i(\mu_i) \right]$$

$$\Rightarrow E\left[\frac{\partial^2 \log f}{\partial \beta_j \partial \beta_k}\right] = \frac{1}{\phi} \sum_i - w_i \widetilde{w}_i \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_k} x_{ij}$$

$$= \frac{1}{\phi} \sum_i - w_i \tilde{w}_i(\mu_i) \frac{1}{g'(\mu_i)} \cdot x_{ik} x_{ij}$$

$$= \sum - \hat{w}_i x_{ik} x_{ij}$$

$$\Rightarrow I(\beta) = \mathbb{E}[J(\beta)] = [X_1, \ldots, X_n] \cdot \hat{W} \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix}$$

$$= X^T \hat{W} X$$

② When choose canonical link function, then $\tilde{w}_i(\mu_i) = \dfrac{1}{b''(\theta_i) \cdot g'(\mu_i)}$

$$\boxed{g(\mu_i) = \theta_i = \eta_i}$$

$$= \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i}$$

$$= \frac{\partial \theta_i}{\partial \eta_i} \equiv 1$$

observed info matrix

$$\Rightarrow \frac{\partial^2 \log f}{\partial \beta_j \partial \beta_k} = \frac{1}{\phi} \sum_i - \frac{\partial \mu_i}{\partial \beta_k} \cdot \tilde{w}_i w_i x_{ij}$$

coincidence

$$= \mathbb{E}\left[\frac{\partial^2 \log f}{\partial \beta_j \partial \beta_k}\right]$$

Fisher info matrix

when picking canonical link function

$\Rightarrow$ Newton-Raphson Algo $\Longleftrightarrow$ Fisher-scoring algo

---

Details of Fisher-Scoring Algo

where $\tilde{W} = \text{diag}\left(\dfrac{w_i}{\phi b''(\theta_i) g'(\mu_i)}\right)$, $\hat{W} = \text{diag}\left(\dfrac{w_i}{\phi b''(\theta_i) g'(\mu_i)^2}\right)$

$$\beta^{k+1} = \beta^k + [I(\beta^k)]^{-1} S(\beta^k)$$

$$= \beta^k + (X^T \hat{W} X)^{-1} X^T \tilde{W} (y - \mu^k)$$

$\eta^k = X\beta^k$

$$= (X^T \hat{W} X)^{-1} (X^T \hat{W} X \beta^k + X^T \tilde{W} (y - \mu^k))$$

$$= (X^T \hat{W} X)^{-1} X^T \hat{W} (\eta^k + \hat{W}^{-1} \tilde{W} (y - \mu^k))$$

$:= \boxed{(X^T \hat{W} X)^{-1} X^T \hat{W} z^k}$

where $z^k = \eta^k + \hat{W}^{-1} \tilde{W} (y - \mu)$

linear predictor of (i-th) sample based on k-iter β

$\boxed{\eta_i^k = x_i^T \beta^k}$

can be viewed as:
$\mathcal{D} = \{(x_i, z_i^k)\}_{i=1}^n$
& $Z^k \sim N(X\beta, \hat{W})$

$z_i^k = \eta_i^k + g'(\mu_i^k)(y_i - \mu_i^k)$

$\approx \boxed{g(y_i)}$

$\boxed{g(\mu_i^k) = \eta_i^k = x_i^T \beta^k}$

$\boxed{\hat{\beta}_{MLE} \rightarrow \beta^{k+1}}$

$g(y_i)$ 在不同 μ_i^k 处的 Taylor approximation $\rightarrow \boxed{z_i^k}$

adjusted dependent variable

Fisher-scoring Algo (Iterative WLS)

① initial β^0, then μ^0 comes from $g(\mu^0) = X^T \beta^0$

② given β^k & μ^k, do the following calculation:

a) $z_i^k = \eta_i^k + g'(\mu_i^k) \cdot (y_i - \mu_i^k)$

where $\eta_i^k = x_i^T \beta^k$

b) $(\hat{W}^k)_i = \dfrac{w_i}{\phi V(\mu_i^k) g'(\mu_i^k)^2}$

c) $\beta^{k+1} = \hat{\beta}_{MLE}$ for $z^k \sim N(X\beta, W^k)$

$= (X^T \hat{W}^k X)^{-1} X^T \hat{W}^k z^k$

weighted Least Square Solution

adjusted dependent variable!

Remark: for each iteration, (k-th)

z^k (response) is updated

W^k (variance) is updated

(IWLS) $\Longrightarrow$ iterative reweighted LS

Calculation is based on β^k (and $\mu^k = g^{-1}(X\beta^k)$)