STOR 556 3/26/19

New HW due 04/02: Ch. 8, Q4, 6 (Assignment 7)

GLM Diagnostics

1. Residuals
2. Leverage & Influence
3. Model Diagnostics
4. Outliers

1. Residuals

(a) Pearson Residuals

Definition of a GLM: Mean function μ

$$\eta_i = g(\mu_i) = \sum_j x_{ij} \beta_j$$

Variance component $V(\mu)$, Dispersion ϕ

e.g. $V(\mu) = \text{const}$ Normal GLM

$V(\mu) = \mu$ Poisson

$V(\mu) = \mu(1-\mu)$ Binomial

$V(\mu) = \mu^2$ Gamma

$(V(\mu) = \mu^3$ Inverse Gaussian

$$r_{P,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$ Pearson Residual

$V(\mu) = b''(\theta)$ <u>Note</u> $\sum r_{P,i}^2 = X^2$

$\Downarrow$

Pearson χ^2 statistic

(b) <u>Deviance Residuals</u>

Recall: $D = \frac{2}{\phi} \sum_i w_i \left\{ y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \right.$

$\hat{\theta}_i$: MLE under assumed model

$\tilde{\theta}_i$: — — — — full —

$\tilde{\theta}_i = g(\tilde{\mu}_i) = g(y_i)$

$$D = \sum_i d_i$$

$$r_{D,i} = \text{sign}(y_i - \hat{\mu}_i) \cdot \sqrt{d_i}$$

Example: Poisson $b(\theta) = e^{\theta} = \mu$

$$r_{D,i} = \text{sign}(y_i - \hat{\mu}_i) \cdot \sqrt{2\left\{ y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \right\}}$$

Estimate dispersion parameter (when needed)

Usual: $\hat{\phi} = \dfrac{X^2}{n-p}$ X^2: Pearson χ^2

n: # obsns

p: # parameters

(c) Response residuals

$r_{R,i} = y_i - \hat{\mu}_i$

(d) "Working Residuals"

Side-product of GLM optimization

$\underbrace{lm(z \sim . , weights = w)}$

residuals of this

In R:

residuals (model, type = "deviance")

"pearson"

"response"

"working"

2. Leverage and Influence

Recall: for a linear model

$$H = X(X^T X)^{-1} X^T$$

h_i: diagonal entries of H

Doesn't depend on the y_i's

Influence stats (e.g. Cook's D)

do depend on y_i's

With GLMs: weights w_i

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

$$H = W^{\frac{1}{2}} X (X^T W X)^{-1} X^T W^{\frac{1}{2}}$$

h_i: diagonal entries of H

In R: influence(model)$hat

Studentized Residuals

$$r_{SD} = \frac{r_D}{\sqrt{\hat{\phi}(1-h_i)}}$$

$\hat{\phi}$ is needed for ~~unscaled~~ unscaled residuals

"Jackknife" residuals — based on deleting observations and refitting model

"Williams approximation"

$$\mathrm{sign}(y-\hat{\mu})\left\{(1-h_i)\, r_{SD}^2 + h_i\, r_{SP}^2\right\}$$

$$r_{SP} = \frac{r_P}{\sqrt{1-h_i}}$$

In R: rstudent(model)

Other kinds of diagnostics

influence(model)$coef — influence on individual parameter estimates (DFBETAS)

or combine everything in Cook's D

$$D_i = \frac{(\hat{\beta}_{(i)}-\hat{\beta})^T (X^T W X)(\hat{\beta}_{(i)}-\hat{\beta})}{p\,\hat{\phi}}$$

3. Model Diagnostics

Galapagos

(a) three types of residual plot

(b) Transformation of X variables

specifically

replace "Area" by "log(Area)"

"Linearized Response"

$$z = \eta + (y-\mu)\frac{d\eta}{d\mu}$$

$$\bar{y}_i \to \mu_i \to g(\mu_i) = \eta_i$$

$\eta = g(\mu)$

Think about

$$g(y_i) - g(\mu_i) \approx g'(\mu_i)\cdot(y_i - \mu_i)$$

$$\downarrow \qquad = \frac{d\eta}{d\mu}\cdot(y_i-\mu_i)$$

z_i

$$z_i = \eta_i + (y_i - \mu_i)\left.\frac{d\eta}{d\mu}\right|_{\mu=\mu_i}$$