Logistic regression assumptions and diagnostics

- HW 1 due tomorrow (Tuesday) at noon
- HW 2 and Challenge 3 released on course website

practice w/ Fisher scoring + quantile residuals

neural networks

Recap: IRLS for logistic regression

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & & \ddots & \\ \vdots & & & \\ 1 & & & X_{nk} \end{pmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$W = \operatorname{diag}(p_1(1-p_1), \ldots, p_n(1-p_n)), \quad p = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix} \qquad p_i = \frac{e^{\beta^T X_i}}{1 + e^{\beta^T X_i}}$$

$$\beta^{(r+1)} = (X^T W^{(r)} X)^{-1} X^T W^{(r)} \underbrace{\left(X\beta^{(r)} + (W^{(r)})^{-1}(Y - p^{(r)})\right)}_{Z^{(r)}}$$

$Z^{(r)} \leftarrow$ working responses

$$= (X^T W^{(r)} X)^{-1} X^T W^{(r)} Z^{(r)}$$

weighted least squares

$$Y = X\beta + \varepsilon$$

$$\Rightarrow \underbrace{W^{\frac{1}{2}}Y}_{Y_w} = \underbrace{W^{\frac{1}{2}}X}_{X_w}\beta + \underbrace{W^{\frac{1}{2}}\varepsilon}_{\varepsilon_w}$$

$$\Rightarrow Y_w = X_w\beta + \varepsilon_w$$

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

$$\Rightarrow \hat{\beta} = (X_w^T X_w)^{-1} X_w^T Y_w$$

$$= (X^T W X)^{-1} X^T W Y$$

$$(X_w^T X_w)^{-1} X_w^T Y_w$$

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

$$= (X^T W X)^{-1} X^T W Y$$

$$\varepsilon \sim N(0, W^{-1}) \qquad W^{\frac{1}{2}} W^{\frac{1}{2}} = W$$

$$W \in \mathbb{R}^{n\times n} \quad (W = \mathrm{diag}(w_1, \dots, w_n))$$

$$\Rightarrow W^{\frac{1}{2}} = \mathrm{diag}(\sqrt{w_1}, \dots, \sqrt{w_n})$$

$$\varepsilon_w \sim N(0, I)$$

$$\mathrm{Var}(W^{\frac{1}{2}}\varepsilon) = W^{\frac{1}{2}} \underbrace{\mathrm{Var}(\varepsilon)}_{W^{-1}} W^{\frac{1}{2}}$$

$$= I$$

$$\hat{Y}_w = X_w \hat{\beta}$$

$$= X_w (X_w^T X_w)^{-1} X_w^T Y_w$$

$$= \underbrace{W^{\frac{1}{2}} X (X^T W X)^{-1} X^T W^{\frac{1}{2}}}_{\text{hat matrix } H} Y_w$$

<u>Intuition:</u> $\beta^{(r+1)}$ is the estimate from weighted least squares regression of $Z^{(r)}$ on X w/ weights $W^{(r)}$

$$Z^{(r)} = X\beta^{(r)} + (W^{(r)})^{-1}(Y - p^{(r)})$$

<u>If</u> $\beta^{(r)} = \beta$ (fixed), $\text{Var}(Z^{(r)}) =$

$$\text{Var}\left((W^{(r)})^{-1}(Y - p^{(r)})\right)$$

$$= (W^{(r)})^{-1} \underbrace{\text{Var}(Y)}_{W} (W^{(r)})^{-1}$$

$\text{var}(Y_i) = p_i(1-p_i)$

$$= W^{-1} W W^{-1}$$

$$= W^{-1}$$

Leverage and Cook's Distance in logistic regression

Let $\hat{\beta}$, $\hat{W}$ be the estimates of β, W at convergence

Hat matrix: $\hat{W}^{\frac{1}{2}} X (X^T \hat{W} X)^{-1} X^T \hat{W}^{\frac{1}{2}}$

It can be shown that:

$$Var(Y_i - \hat{p}_i) \approx \hat{p}_i(1-\hat{p}_i)(1-h_i)$$

leverage of i^{th} observation

(i^{th} diagonal element of hat matrix H)

Cook's distance:

$$D_i = \frac{(Y_i - \hat{p}_i)^2 \; h_i}{k \; \hat{p}_i(1-\hat{p}_i) \; (1-h_i)^2}$$

k = # of explanatory variables

(worried if $D_i > 0.5$ or 1)

as $h_i \uparrow$, $D_i \uparrow$

as $h_i \downarrow$, $D_i \downarrow$

as $(Y_i - \hat{p}_i)^2 \uparrow$, $D_i \uparrow$

$$D_i = \left(\frac{Y_i - \hat{p}_i}{SD(Y_i - \hat{p}_i)} \right)^2 \cdot \frac{h_i}{(1 - h_i) K}$$

Class activity

https://sta712-f22.github.io/class_activities/ca_lecture_7.html

- Generate data with a potentially influential point
- Explore leverage and Cook's distance

Variance inflation factors for logistic regression

Addressing model issues

How should we handle each of the following issues in a fitted model?

- Violations of the shape assumption
- An influential point with high Cook's distance
- High multicollinearity in the explanatory variables

Discuss with your neighbor for 3--5 minutes, then we will discuss as a group.