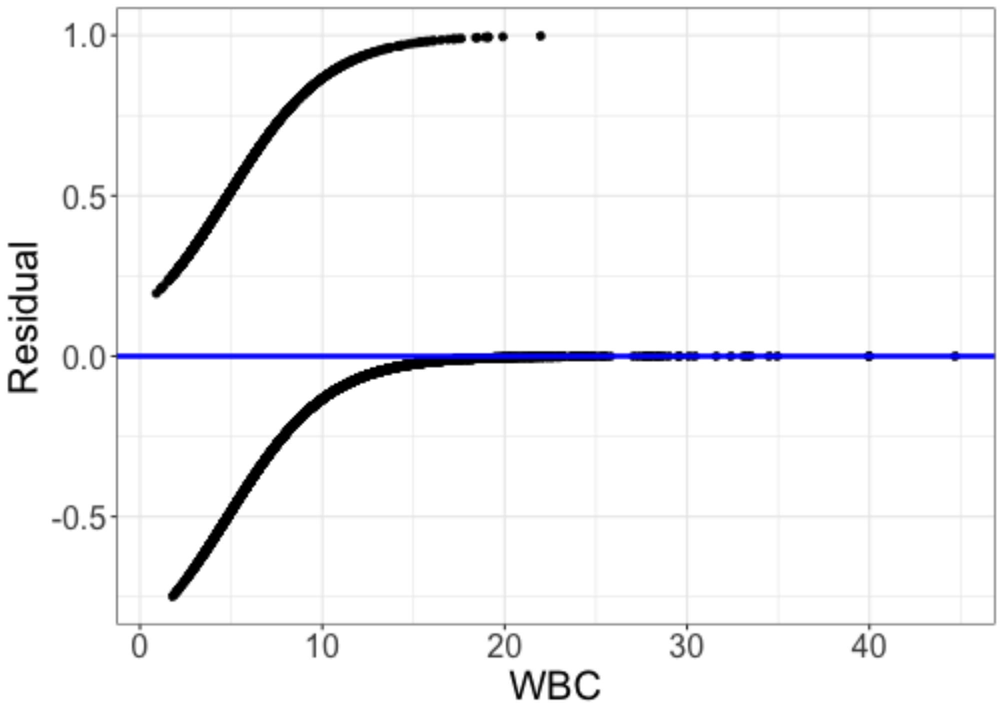
Logistic regression assumptions and diagnostics

(residual plots $Y - \hat{p}$)

Don't use usual residuals for logistic regression

Fitted model: $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = 1.737 - 0.361\, WBC_i$

Residuals $Y_i - \hat{p}_i$:



(checking shape assumption)

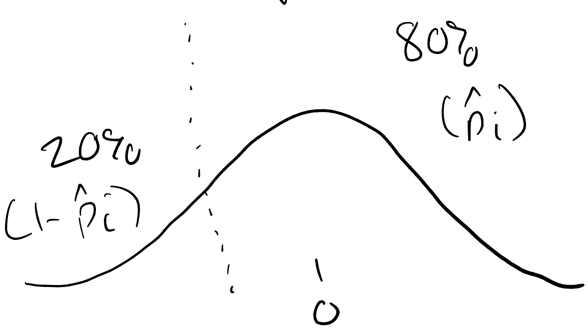
(randomized)

Quantile residuals for logistic regression

Motivation: Suppose $\hat{p}_i = 0.8$. I want to create residual r_Q that behaves like linear regression residuals: want

- If $\hat{p}_i \approx p_i$ (good estimate) then $E[r_Q | X_i] \approx 0$
- If $\hat{p}_i > p_i$ (overestimate), then $E[r_Q | X_i] < 0$
- If $\hat{p}_i < p_i$ (underestimate), then $E[r_Q | X_i] > 0$
- want $r_Q \approx$ Normal (if $\hat{p}_i \approx p_i$)

Idea: $\hat{p}_i = 0.8$, Divide $N(0,1)$ into 2 regions:



- If $Y_i = 1$, sample r_Q from right side
- If $Y_i = 0$, sample r_Q from left side

If $\hat{p}_i \approx p_i$, then I'm sampling from a $N(0,1)$ on average

Pseudo-code:

for each $i = 1, \dots, n$:

Calculate $\hat{p}_i$

If $Y_i = 1$:

Sample once from the upper $\hat{p}_i$ area of $N(0,1)$

if $Y_i = 0$

Sample once from the lower $1-\hat{p}_i$ area of $N(0,1)$

use r_Q to make residual plots

quantile residuals

If $\hat{p}_i \approx p_i$, then $r_{Q_i} \sim N(0,1)$ (over many datasets)

If all $\hat{p}_i \approx p_i \ \forall i$, then marginally $r_Q \sim N(0,1)$

Class activity, Part I

https://sta711-s23.github.io/class_activities/ca_lecture_10.html

Leverage and Cook's distance

Linear regression: $\hat{Y} = X\hat{\beta}$

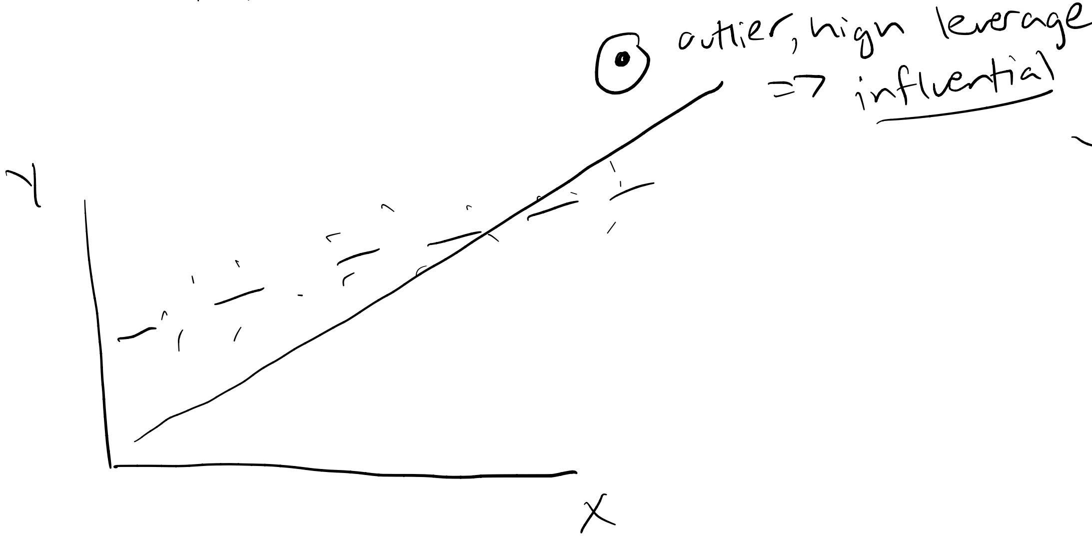
$$= \underbrace{X(X^TX)^{-1}X^T}_{\text{"hat matrix" } H} Y$$

(not Hessian! I will use $H(\beta)$ to denote Hessian)

$$\mathrm{Var}(Y - \hat{Y}) = \sigma^2(I - H) \Rightarrow \mathrm{Var}(Y_i - \hat{Y}_i) = \sigma^2(1 - h_i)$$

leverage = potential of observation to influence fit

$h_i = [H]_{ii}$ leverage of the i^{th} observation



Cook's distance (linear regression):

$$D_i = \frac{(Y_i - \hat{Y}_i)^2}{\underbrace{(k+1)}_{\text{\# of }\beta\text{s in model}}\hat{\sigma}^2} \cdot \underbrace{\frac{h_i}{(1-h_i)^2}}_{\text{high leverage?}}$$

outlier?

concerned that a point is influential when $D_i >$ threshold (e.g. 0.5 or 1)

Logistic regression:

$(W = \text{diag}(p_i(1-p_i)))$

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

h_i = leverage

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

concerned when $D_i > 0.5$ or 1

Class activity, Part II

https://sta711-s23.github.io/class_activities/ca_lecture_10.html