Logistic regression assumptions and diagnostics

Multicollinearity

Definition: Multicollinearity occurs when one explanatory variable can be approximated by a linear combination of other explanatory variables

E.g. $Y_i \sim \text{Bernoulli}(p_i)$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$$

worst case: $X_{i1} = \alpha_2 X_{i2} + \alpha_3 X_{i3}$

$$\Rightarrow \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + (\beta_1\alpha_2 + \beta_2)X_{i2} + (\beta_1\alpha_3 + \beta_3)X_{i3}$$

⇒ Too many unknowns ⇒ can't estimate βs

higher multicollinearity → more trouble with estimation

Class activity

https://sta712-f22.github.io/class_activities/ca_lecture_8.html

> - Simulate correlated data
> - Assess the impact on estimated coefficients

The impact of multicollinearity

Problems

- inflates variability of $\hat{\beta}$s

  => problems in inference

- difficult to interpret $\hat{\beta}$

Need: a method for diagnosing multicollinearity

option 1 : pairs plot of X

+ correlation matrix for X

but, only looks @ pairwise relationship

option 2 : Variance inflation factor

involve coefficients of determination R^2

Variance inflation factors

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta^T X_i \qquad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

$$VIF_j = \frac{Var(\hat{\beta}_j) \text{ using all our explanatory variables}}{Var(\hat{\beta}_j) \text{ using only } \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}}$$

Turns out (HW 3!) that

$$VIF_j = \frac{1}{1-R_j^2} \qquad R_j^2 = R^2 \text{ for regression of } \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix} \text{ on all other explanatory variables}$$

Thresholds: usually concerned if VIF > threshold (5 or 10)

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.

Assumption	Diagnostics	Fixing violations
Shape	· quantile residuals · empirical logit plot	· transform · more flexible models (GAMs, forests, NNs, etc.)
No outliers	· cook's distance · other measures: DFFITS, DFBETAS, etc.	· report results w/ and w/out outliers · transform skewed predictors
No issues w/ multicollinearity	· VIFs · correlation matrix	· remove some columns · combine variables · Ignore! (if we care about prediction)

Asymptotic distribution of the MLE

Multicollinearity can cause problems in the variance of the estimated coefficients $\widehat{\beta}$. But what is $Var(\widehat{\beta})$?