Logistic regression assumptions and diagnostics

- HW 4 released on course website, due next Friday
- Exam 1 released next Friday (Feb. 10)
  - take-home
  - closed notes
  - covers HW 1-3

Multicollinearity

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

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

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

Worst-case scenario: $X_{i1} = \alpha_2 X_{i2} + \alpha_3 X_{i3} \quad \forall i$

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

$\Rightarrow$ too many unknowns, can't estimate βs

higher multicollinearity $\Rightarrow$ more trable with estimation (e.g., higher variability in estimates)

Class activity

https://sta711-s23.github.io/class_activities/ca_lecture_11.html

- Simulate correlated data
- Assess the impact on estimated coefficients

The impact of multicollinearity

- Standard error of $\hat{\beta}$s increases, i.e. variability of $\hat{\beta}$ ↑

  If perfect collinearity between some explanatory variables, we can't estimate standard errors

  Sneak peek: $Var(\hat{\beta}) \approx \hat{\mathcal{I}}^{-1}(\beta) = (X^T W X)^{-1}$

- Multicollinearity also makes it hard to interpret $\hat{\beta}$s

Diagnosing multicollinearity

- Scatterplot matrix of quantitative explanatory variables
- correlation matrix "

- Variance inflation factors

Variance inflation factors

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

$$VIF_j = \frac{\text{Var}(\hat{\beta}_j) \text{ using all explanatory variables in model}}{\text{Var}(\hat{\beta}_j) \text{ using only } X_{ij}\text{s}} = \begin{bmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{nj} \end{bmatrix}$$

$$VIF_j = \frac{1}{1-R_j^2} \qquad R_j^2 = R^2 \text{ for regression of } X_{ij}\text{s on other explanatory variables}$$

(true for both linear & logistic)

Thresholds: usually concerned if VIF > threshold (e.g. 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	• empirical logit plot • quantile residual plot	• transformations! • different (more flexible) model (GAMs, forest, NNs)
No outliers	• Cook's distance • other related measures (DFFITS, DFBETAS, etc.)	• remove errors in data (e.g. negative SAT score) • Fit with & without, see if our conclusions change • try transformations for skewed explanatory variables
No issues w/ multicollinearity	• VIFs • Scatterplot / correlation matrix	• remove some variables • use a different model • Add penalty terms • combine variables • Ignore!