5.2. Nominal/Ordinal Logistic Regression

Dr. Bean - Stat 5100

1 What to do when your categorical data isn't binary?

Recall binary response logistic regression:

- Model framework:

$$Y \in \{0,1\} \qquad \pi = P(Y=1)$$

$$L = \log\frac{\pi}{1-\pi} \quad = \quad \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{i,p-1}$$

- logit link function:

$$L_i \quad = \quad \log\frac{P(Y=1|profile_i)}{P(Y=0|profile_i)}$$

Two kinds of multi-class categorical data

- Nominal: no apparent ordering of classes (ex: race, major, color)
- Ordinal: ordering of data makes sense (product rating, pain scale, etc.)

2 Nominal Logistic Regression

- pick one level as reference (say $Y = r$)
- generalized logit (glogit) link function:

$$L_{k|i} \quad = \quad \log\frac{P(Y=k|profile_i)}{P(Y=r|profile_i)}$$

- coefficient $\beta_{j,k}$ for the (marginal) effect of predictor X_j for $Y = k$ vs. $Y = r$:

$$L_{k|i} \quad = \quad \beta_{0,k} + \beta_{1,k} X_{i,1} + \ldots + \beta_{p-1,k} X_{i,p-1}$$

- odds ratio interpretation involves the base class r
  - 5.2.1 Example: Coefficient associated with A3 and Importance of 3:
    "Holding all other predictors constant, the odds that a 40+ year old rates AC and power steering as 'very important' *versus not important* are $100(e^{2.9165}-1) = 1747\%$ greater than for an 18-23 year old."

Other Comparisons

To compute the log odds ratio for two *non-base* classes simply compute:

$$L_{k_1|k_2} = L_{k_1|i} - L_{k_2|i}$$

The estimated probability of each (non-base) class can be computed as

$$\hat{\pi}_k = \frac{e^{L_{k|i}}}{1 + \sum_{j=1}^{J-1} e^{L_{j|i}}}$$

(Note that $\hat{\pi}_i$ will be fully determined from the estimated probabilities of the other classes.)

3 Ordinal Logistic Regression

- $Y \in \{1, 2, \ldots, r\}$ and $1 < 2 < \ldots < r$
- accumulate probability over lower levels:

$$p_k^c = P(Y \leq k)$$

- logit function accounts for this accumulation ("proportional odds" model):

$$\begin{aligned} L_{k|i} &= \log \frac{p_k^c}{1 - p_k^c} \\ &= \log \frac{P(Y \leq k|profile_i)}{P(Y > k|profile_i)} \end{aligned}$$

- coefficient $\beta_{j,k}$ for the (marginal) effect of predictor X_j for $Y \leq k$ vs. $Y > k$:

$$L_{k|i} = \beta_{0,k} + \beta_{1,k} X_{1,i} + \ldots + \beta_{p-1,k} X_{i,p-1}$$

- odds ratio interpretation involves direction of k:
  - "Holding all other predictors constant, the odds that a 40+ year old rates AC and power steering as either important or very important are $100(e^{2.2322} - 1) = 832\%$ greater than for an 18-23 year old."
- In ordinal logistic regression, coefficient interpretation relies on direction in Y (higher or lower) because we assume the coefficient is the same for all levels of Y:
  - Let $\beta_{j,k}$ be coeff. for predictor X_j in model for $L_{k|i}$

$$L_{k|i} = \beta_{0,k} + \beta_{1,k} X_{1,i} + \ldots + \beta_{p-1,k} X_{i,p-1}$$

$$\begin{aligned} H_0 &: \beta_{j,1} = \beta_{j,2} = \ldots = \beta_{j,r} \\ H_0 &: L_{k|i} = \beta_{0,k} + \beta_1 X_{1,i} + \ldots + \beta_{p-1} X_{i,p-1} \end{aligned}$$