Nominal Logistic Regression

- Random variable **Y** that corresponds to K categories (blue, red, green).
- $\pi_1, \pi_2, \ldots, \pi_K$ probabilities for each category (population).
- n independent observations of **Y**.
- We count $Y_i, i = 1 \ldots, K$ number of times we observe category i.
- Category 1 is a *reference* category.

$$\begin{aligned} log\left(\frac{\pi_j}{\pi_1}\right) &= \beta_{0j} + \beta_{1j}X_{1j} + \ldots + \beta_{pj}X_{pj}; \\ &= \mathbf{x}_j^T\beta_j; j = 2, \ldots, K \end{aligned}$$

- $\mathbf{x}_j$ are the predictors or covariates.
- β_j are the coefficients for category j.
- For an estimate of coefficients $\hat{\beta}_j$ (or $\mathbf{b}_j$),

$$\hat{\pi}_j = \frac{exp(\mathbf{x}_j^T \beta_j)}{1 + \sum_{j=2}^{K} exp(\mathbf{x}_j^T \beta_j)}$$

  provides an estimate of the probability for category j.
- The $j = 1$ case

$$\hat{\pi}_1 = \frac{1}{1 + \sum_{j=2}^{K} exp(\mathbf{x}_j^T \beta_j)}$$

- **Residuals:** Defined in a similar way as in *logistic* regression.

$$r_i = \frac{o_i - e_i}{\sqrt{e_i}}, i = 1, 2, \ldots, K$$

where o_i is the observed value and e_i is the expected frequency.
- e_i computed as $n\hat{\pi}_i$
- Compared with $o_i = y_i$.
- Similar to ideas in a one-way *contingency tables*
- **Chi-square statistic:** applies if each category has a unique set of covariates

$$X^2 = \sum_{i=1}^{K} r_i^2$$

UNM

Odds ratio

- Suppose we have a covariate X, where $X = 0$ factor is absent and $X = 1$ factor is present.
- Since
$$log\left(\frac{\pi_j}{\pi_1}\right) = \beta_{0j} + \beta_{1j}X; j = 2, \ldots, K$$
- If π_{jp} (π_{ja}) is the response probability associated to factor present (absent).
- For $X = 0$,
$$log\left(\frac{\pi_j}{\pi_1}\right) = \beta_{0j}$$
- and for $X = 1$,
$$log\left(\frac{\pi_j}{\pi_1}\right) = \beta_{0j} + \beta_{1j}$$

- **Odds ratio** for exposure for response *j*

$$
\begin{aligned}
OR_j &= \frac{\pi_{jp}/\pi_{1p}}{\pi_{ja}/\pi_{1a}} \\
&= \frac{\pi_{jp}/\pi_{ja}}{\pi_{1p}/\pi_{1a}} = exp(\beta_{1j})
\end{aligned}
$$

  or $log(OR_j) = \beta_{1j}$.
- Odds ratio is relative to category 1.
- Somewhat a measure of factor effect.
- Wouldn't it be easier to compare π_{jp} with π_{ja}?

$$RR_j = \frac{\pi_{jp}}{\pi_{ja}}$$

Ordinal Logistic regression

- Random variable Z difficult to measure (latent variable).
- Z could be "income", perhaps we know *cutoff* points

$$-\infty < c_1 < c_2 < c_3 \ldots < c_{j-1} < \infty$$

- Category j is established if

$$c_{j-1} < Z < c_j.$$

- Therefore

$$\pi_j = Pr(c_{j-1} < Z < c_j)$$

  is the probability associated to category j.
- Notice that larger Z values, the more "preferable" the category.
- Defines an ordinal scale on j.

Cummulative Logit model

- Cummulative odds for the $j-th$ category.

$$\frac{P(Z \leq c_j)}{P(Z > c_j)} = \frac{\pi_i + \pi_2 + \ldots + \pi_j}{\pi_{j+1} + \pi_{j+2} + \ldots + \pi_J}$$

- **Proportional odds model:**

$$log\left(\frac{\pi_i + \pi_2 + \ldots + \pi_j}{\pi_{j+1} + \pi_{j+2} + \ldots + \pi_J}\right) = \beta_{0j} + \beta_1 X_1 + \ldots + \beta_p X_p$$

- only intercept β_{0j} depends on j.
- $\beta_1, \beta_2, \ldots, \beta_p$ are constant across j.

- Alternatively, consider ratios

$$\frac{\pi_1}{\pi_2}, \frac{\pi_2}{\pi_3}, \ldots, \frac{\pi_{j-1}}{\pi_j}, \ldots$$

- **Adjacent category logic model**

$$log\left(\frac{\pi_j}{\pi_{j+1}}\right) = \beta_{0j} + \beta_1 X_1 + \ldots + \beta_p X_p$$

or

$$log\left(\frac{\pi_j}{\pi_{j+1} + \pi_{j+2} \ldots + \pi_K}\right) = \beta_{0j} + \beta_1 X_1 + \ldots + \beta_p X_p$$

- Odds of being in category j ($c_{j-1} \leq Z \leq c_j$), conditional on $Z > c_{j-1}$.
- Other link functions can be considered.