BIOSTAT 651
Notes #12:
Multinomial Data

- Lecture Topics:
  - Models for multinomial data
  - Interpretation of parameters
  - Examples
- Text (Dobson & Barnett, 3rd Ed.): Chapter 8

Nominal, Ordinal Scales

- *Nominal* data:
  - state of residence after graduation:

    MI, OH, WI, IL, NC
  - political affiliation:

    Republican, Independent, Democrat
  - T.V. channel preferences:

    Fox, CNN, MSNBC, do-not-watch-TV
- *Ordinal* data:
  - pain:

    slight, moderate, severe
  - grade:

    A+, A, A-, B+, B, B-

Multinomial Data

- Often in practice, the response is multinomial

  ($>$ 2 categories)

- Simplification: reduce to 2 categories; employ logistic regression
  - may be met with skepticism by investigators and/or reviewers
  - how such grouping affected the results will be an issue
  - may result in a considerably less informative analysis

Multinomial : Exponential Family

- Suppose $\mathbf{Y} \sim \text{Multinomial}(n, \boldsymbol{\pi})$ with $J+1$ categories.

  $\mathbf{Y} = (Y_0, Y_1, \ldots, Y_J)'$

  $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_J)'$

  Y_j: number of subjects with $Y = j$

  π_j: probability of $Y = j$

- Moments:

  $E[Y_j] =$

  $V(Y_j) =$

  $\text{cov}(Y_j, Y_k) =$

- Probability function:

$$p(\mathbf{Y}; \boldsymbol{\pi}) \quad = \quad \begin{pmatrix} n \\ Y_0 \; Y_1 \; \ldots Y_J \end{pmatrix} \prod_{j=0}^{J} \pi_j^{Y_j}$$

Multinomial : Exponential Family (continued)

- Write as an exponential family:

$$p(\mathbf{Y};\boldsymbol{\pi}) \quad = \quad \exp\left\{\sum_{j=0}^{J} Y_j \log \pi_j + \log C(n;\mathbf{Y})\right\}$$

- Consider the constraints:
  - $\sum_{j=0}^{J} Y_j = n$
  - $\sum_{j=0}^{J} \pi_j = 1$

Multinomial : Exponential Family (cont'd)

- Incorporating these constraints, we re-write

$p(\mathbf{Y};\boldsymbol{\pi}) =$

$$
\begin{aligned}
& \exp\left\{ Y_0 \log \pi_0 + \sum_{j=1}^{J} Y_j \log \pi_j + \log C(n;\mathbf{Y}) \right\} \\
= \quad & \exp\left\{ (n - \sum_{j=1}^{J} Y_j) \log \pi_0 + \sum_{j=1}^{J} Y_j \log \pi_j \right. \\
& \quad \left. + \log C(n;\mathbf{Y}) \right\} \\
= \quad & \exp\left\{ \sum_{j=1}^{J} Y_j \log\left(\frac{\pi_j}{\pi_0}\right) + n \log \pi_0 + \log C(n;\mathbf{Y}) \right\}
\end{aligned}
$$

- This is an exponential family with:

$t(\mathbf{Y}) =$ $\qquad\qquad$ $\boldsymbol{\theta} =$

Deriving Multinomial Parameters

- Examining the natural parameter:

$$\begin{aligned} \frac{\pi_1}{\pi_0} &= e^{\theta_1} \\ &\vdots \\ &\vdots \\ \frac{\pi_J}{\pi_0} &= e^{\theta_J} \end{aligned}$$

- Therefore, we obtain

$$\sum_{j=1}^{J} e^{\theta_j} = \frac{1}{\pi_0}\sum_{j=1}^{J}\pi_j = \frac{1}{\pi_0}(1-\pi_0)$$

such that

$\pi_0 =$ $\pi_j =$

Properties of Multinomial

- We then have

$$\begin{aligned} b(\boldsymbol{\theta}) &= -n \log \pi_0 \\ &= n \log \left(1 + \sum_{j=1}^{J} e^{\theta_j} \right) \end{aligned}$$

- The canonical link: since $\theta_j = \log(\pi_j / \pi_0)$

$$\log \frac{\pi_{ij}}{\pi_{i0}} = \mathbf{x}_i^T \boldsymbol{\beta}_j$$

for $j = 1, \ldots, J$

Multinomial: Regression Framework

- We now consider regression modeling of multinomial outcomes
- General set-up:

  Y_i = response, subject i

  set $Y_{ij} = I(Y_i = j) \qquad\qquad j = 0, 1, \ldots, J$

  such that $Y_i = \sum_{j=0}^{J} j \times Y_{ij}$

  covariate: $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{iq})'$

- Outcome probabilities:

$$\pi_{ij} \quad = \quad \pi_j(\mathbf{x}_i) = P(Y_i = j|\mathbf{x}_i)$$

Generalized Logits Model

- Generalized logits model:

$$\log\left(\frac{\pi_{ij}}{\pi_{i0}}\right) \quad = \quad \mathbf{x}_i^T\boldsymbol{\beta}_j$$

  - one category defined as the *reference*
  - distinct intercepts and regression coefficients

- Response probabilities:

$$P(Y_i = j|\mathbf{x}_i) = \pi_{ij} = \frac{\exp(\mathbf{x}_i^T\boldsymbol{\beta}_j)}{1+\sum_{j=1}^{J}\exp(\mathbf{x}_i^T\boldsymbol{\beta}_j)}$$

$$P(Y_i = 0|\mathbf{x}_i) = \pi_{i0} = \frac{1}{1+\sum_{j=1}^{J}\exp(\mathbf{x}_i^T\boldsymbol{\beta}_j)}$$

- Direct generalization of logistic regression to > 2 response categories

MLE: Generalized Logits Model

- Set $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \ldots, \boldsymbol{\beta}'_J)$

  dimension $= (q+1) \times J$

- Likelihood is given by:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{j=0}^{J} \pi_{ij}^{Y_{ij}}$$

- Subbing in $\boldsymbol{\beta}$, then taking log:

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left\{ -Y_{i0} \log \left(1 + \sum_{\ell=1}^{J} e^{\mathbf{x}_i^T \boldsymbol{\beta}_\ell} \right) \right. \\ &\qquad \left. + \sum_{j=1}^{J} Y_{ij} \log \left(\frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}_j}}{1 + \sum_{\ell=1}^{J} e^{\mathbf{x}_i^T \boldsymbol{\beta}_\ell}} \right) \right\} \end{aligned}$$

- Recalling that $Y_{i0} = 1 - \sum_{j=1}^{J} Y_{ij}$

  we then obtain:

$$
\begin{aligned}
\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left\{ -\left(1 - \sum_{j=1}^{J} Y_{ij}\right) \log\left(1 + \sum_{\ell=1}^{J} e^{\mathbf{x}_i^T \boldsymbol{\beta}_\ell}\right) \right. \\
&\qquad \left. + \sum_{j=1}^{J} Y_{ij} \mathbf{x}_i^T \boldsymbol{\beta}_j - \sum_{j=1}^{J} Y_{ij} \log\left(1 + \sum_{\ell=1}^{J} e^{\mathbf{x}_i^T \boldsymbol{\beta}_\ell}\right) \right\} \\
&= \sum_{i=1}^{n} \left\{ -\log\left(1 + \sum_{\ell=1}^{J} e^{\mathbf{x}_i^T \boldsymbol{\beta}_\ell}\right) + \sum_{j=1}^{J} Y_{ij} \mathbf{x}_i^T \boldsymbol{\beta}_j \right\}
\end{aligned}
$$

- The score function is then given by:

$$
U(\boldsymbol{\beta}) = \begin{pmatrix} U_1(\boldsymbol{\beta}) \\ \vdots \\ U_J(\boldsymbol{\beta}) \end{pmatrix}
$$

  with jth sub-vector:

$$
U_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_{ij} - \pi_{ij}) \mathbf{x}_i
$$

MLE: Generalized Logits Model (continued)

- Information matrix:

  dimension of $J(\boldsymbol{\beta})$: $\{J(q+1)\} \times \{J(q+1)\}$

$$J(\boldsymbol{\beta}) = \begin{pmatrix} J_{11}(\boldsymbol{\beta}) & \cdots & J_{1J}(\boldsymbol{\beta}) \\ & J_{22}(\boldsymbol{\beta}) & \vdots \\ & \cdots & J_{JJ}(\boldsymbol{\beta}) \end{pmatrix}$$

  (j,j)th block: $\sum_{i=1}^{n} \pi_{ij}(1-\pi_{ij})\mathbf{x}_i^T\mathbf{x}_i$

  (j,k)th block: $-\sum_{i=1}^{n} \pi_{ij}\pi_{ik}\mathbf{x}_i^T\mathbf{x}_i$

- Note that $U_j(\boldsymbol{\beta})$ and $J_{jj}(\boldsymbol{\beta})$ are analogous to a logistic regression which considers categories j and 0 only

Gen Logit Model: Example

- Example: We reconsider the study of childhood asthma. The objective was to determine the role of gender in pre-school asthma incidence. Children enrolled in the study (n=100) were followed prospectively in order to determine whether they were hospitalized for asthma. Suppose now that some children were kept overnight after being admitted ($Y_i = 2$), while others were tested then discharged the same day ($Y_i = 1$). Of course, many children did not suffer asthma ($Y_i = 0$).

Gen Logit Model: Example (continued)

The observed data are summarized by the following table:

	Y_i=0	Y_i=1	Y_i=2	total
F_i=0	24	14	22	60
F_i=1	21	10	9	40

- The model is given by:

$$\log\left\{\frac{\pi_{ij}}{\pi_{i0}}\right\} = \beta_{0j} + \beta_{1j} F_i$$

  for $j = 1, 2$

  - need to estimate 2 intercepts, and 2 differences

- Note: this model is *saturated*

Gen Logit Example: Interpretation)

- Interpretation of parameters:
  - e.g., set $j = 1$; intercept:

$$\beta_{01} \quad = \quad \log\left\{\frac{P(Y_i = 1|F_i = 0)}{P(Y_i = 0|F_i = 0)}\right\}$$

  - Q: Does this lead to an odds?

- set $j = 1$; difference:

$$
\begin{aligned}
\beta_{01} + \beta_{11} &= \log\left\{\frac{P(Y_i = 1|F_i = 1)}{P(Y_i = 0|F_i = 1)}\right\} \\
\beta_{11} &= \log\left\{\frac{P(Y_i = 1|F_i = 1)}{P(Y_i = 0|F_i = 1)}\right\} \\
&\quad - \log\left\{\frac{P(Y_i = 1|F_i = 0)}{P(Y_i = 0|F_i = 0)}\right\} \\
\exp\{\beta_{11}\} &= \frac{P(Y_i = 1|F_i = 1)}{P(Y_i = 0|F_i = 1)} \bigg/ \frac{P(Y_i = 1|F_i = 0)}{P(Y_i = 0|F_i = 0)}
\end{aligned}
$$

- Q: Is this an odds ratio?

Gen Logit Model: Example (cont'd)

- Estimated intercepts:

$$\widehat{\beta}_{01} = \log\left(\frac{14/60}{24/60}\right) = -0.539$$

$$\widehat{\beta}_{02} = \log\left(\frac{22}{24}\right) = -0.087$$

- Estimated differences:

$$\widehat{\beta}_{01} + \widehat{\beta}_{11} = \log\left(\frac{10/40}{21/40}\right) = -0.742$$

$$\widehat{\beta}_{02} + \widehat{\beta}_{12} = \log\left(\frac{9/40}{21/40}\right) = -0.847$$

- Estimated gender effects:

$$\begin{aligned} \exp\{\widehat{\beta}_{11}\} &= 0.81 \\ \exp\{\widehat{\beta}_{12}\} &= 0.48 \end{aligned}$$

Ordinal data: Cumulative Logit Model

- Generalized logit model is applicable when there are > 2 response categories
  - applicable whether or not the response categories are ordered
  - does not exploit the ordering of categories
- Several other modeling options exist for ordinal responses
- Cumulative logit model:

$$\log\left\{\frac{P(Y_i \leq j)}{P(Y_i > j)}\right\} = \log\left\{\frac{\pi_0 + \pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J}\right\} = \gamma_{0j} + \mathbf{x}_i^T \boldsymbol{\gamma}_j$$

for $j = 0, 1, \ldots, (J-1)$

  - distinct intercepts and covariate parameters

Ordinal data: Proportional Odds Model

- Cumulative logit model: consider the following setting:

- $(J+1)$ response levels

- $\mathbf{x}_i$: $q \times 1$ covariate
    - number of parameters: $(q+1) \times J$

- Proportional odds model:
    - Assumes $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0 = \boldsymbol{\gamma}_1 = \cdots = \boldsymbol{\gamma}_{J-1}$
    - fewer parameters to explain to investigator

$$\log\left\{\frac{P(Y_i \leq j)}{P(Y_i > j)}\right\} \quad = \quad \gamma_{0j} + \mathbf{x}_i^T \boldsymbol{\gamma}$$

- Trade-off: PO model entails additional assumptions namely, *proportionality*

Ordinal data: Proportional Odds Model

- Proportionality assumption
  - Odds:

$$odds(Y_i \leq j|\mathbf{x}_1) = \frac{P(Y_i \leq j|\mathbf{x}_1)}{P(Y_i > j|\mathbf{x}_1)} = \exp(\gamma_{0j}) \exp(\mathbf{x}_1^T \boldsymbol{\gamma})$$

  - Odds ratio (independent of j):

$$\frac{odds(Y_i \leq j|\mathbf{x}_1)}{odds(Y_i \leq j|\mathbf{x}_2)} = \exp((\mathbf{x}_1 - \mathbf{x}_2)^T \boldsymbol{\gamma})$$

- Number of parameters: $J + q$

Proportional Odds Model: Example

- Example: Consider a study of incomes of students after completing their degree. Information on the student (age, gender, GPA) and program of study (region, level of program) is captured by a covariate, $\mathbf{x}_i$. The response variate has three levels: low ($Y_i = 0$), medium ($Y_i = 1$) and high ($Y_i = 2$).

Prop Odds Model: Example (continued)

- Suppose the cut-point is $j = 0$:

$$\log\left\{\frac{P(Y_i \leq 0)}{P(Y_i > 0)}\right\} = \gamma_{00} + \mathbf{x}_i^T \boldsymbol{\gamma}_{10} \qquad (1)$$

  which separates 'low' from 'medium, high'

- Another choice could be $j = 1$:

$$\log\left\{\frac{P(Y_i \leq 1)}{P(Y_i > 1)}\right\} = \gamma_{01} + \mathbf{x}_i^T \boldsymbol{\gamma}_{11} \qquad (2)$$

  which separates 'high' from 'medium, low'

- A proportional odds model would assume that (1) and (2) hold, but with the constraint that $\boldsymbol{\gamma}_{10} = \boldsymbol{\gamma}_{11}$

Prop Odds Model: Example (continued)

- Consider a reduced model, for which $\mathbf{x}_i^T = [1, A_i, G_i]$, where A_i equals age (years) and G_i is an indicator for graduate degree (ref = undergraduate).
  - proportional odds model:

$$\begin{aligned} \log\left\{\frac{P(Y_i \le 0)}{P(Y_i > 0)}\right\} &= \gamma_{00} + \gamma_A(A_i - 20) + \gamma_G G_i \\ \log\left\{\frac{P(Y_i \le 1)}{P(Y_i > 1)}\right\} &= \gamma_{01} + \gamma_A(A_i - 20) + \gamma_G G_i \end{aligned}$$

- Interpreting the parameters (e.g., $j = 0$):

  $e^{\gamma_{00}}$:

  e^{γ_G}:

Prop Odds Model: Example (cont'd)

- Based on proportional odds model:
  - can estimate odds through

$$\frac{P(Y_i \leq j)}{P(Y_i > j)} = e^{\gamma_{0j} + \mathbf{x}_i^T \gamma_1}$$

  - probabilities:

$$P(Y_i \leq j) = \frac{e^{\gamma_{0j} + \mathbf{x}_i^T \gamma_1}}{1 + e^{\gamma_{0j} + \mathbf{x}_i^T \gamma_1}}$$

  - cell probabilities:
    $P(Y_i = j) = P(Y_i \leq j) - P(Y_i \leq j - 1)$

Test for Proportionality

- Compare the models with and without the proportionality assumption
  - H0: $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0 = \boldsymbol{\gamma}_1 = \cdots = \boldsymbol{\gamma}_{J-1}$
  - Full model: without the proportionality assumption $(j = 0, \ldots, J-1)$

$$\log\left\{\frac{P(Y_i \leq j)}{P(Y_i > j)}\right\} = \gamma_{0j} + \mathbf{x}_i^T \boldsymbol{\gamma}_{1j}$$

  - Reduced model: with the assumption

$$\log\left\{\frac{P(Y_i \leq j)}{P(Y_i > j)}\right\} = \gamma_{0j} + \mathbf{x}_i^T \boldsymbol{\gamma}$$

- SAS carries out score test.
- Reference distribution: …