Multinomial Data

- Recall: in binomial GLM, observe data

  $$(x_{i1}, x_{i2}, \ldots, x_{im}, y_i),\ i = 1, 2, \ldots, I \Leftrightarrow (\mathbf{x}_i, y_i) \Leftrightarrow (\mathbf{x}, y_{\mathbf{x}})$$

  where $y_{\mathbf{x}}$ is bounded by $n_{\mathbf{x}}$=number of total trials
- Data: now observe a response with J categories, where $J > 2$:

  $$(x_{i1}, x_{i2}, \ldots, x_{im}, y_{i1}, y_{i2}, \ldots, y_{iJ}),\ i = 1, 2, \ldots, I$$
  $$\Leftrightarrow (\mathbf{x}_i, Y_i) \Leftrightarrow (\mathbf{x}, Y_{\mathbf{x}}), \text{ where (i) } Y_i = (y_{i1}, y_{i2}, \ldots, y_{iJ}),$$

  (ii) y_{ij}= # of observations falling into category j at $\mathbf{x}_i$,

  (iii) $\sum_{j=1}^{J} y_{ij} = n_i$ (or denoted by $n_{\mathbf{x}}$), and n_i's are fixed
  - $Y_{\mathbf{x}} \sim$ multinomial($n_{\mathbf{x}}$, $p_{\mathbf{x}1}$, $p_{\mathbf{x}2}$, …, $p_{\mathbf{x}J}$), where $\sum_{j=1}^{J} p_{\mathbf{x}j} = 1$
  - we may encounter grouped or ungrouped data
    - For ungrouped data, n_i=1 for i=1, 2, …, I
    - The idea of covariate classes can be similarly applied to generate grouped data from ungrouped data (Q: what benefit?)
  - The J categories can be
    - *nominal* or *ordinal*, or
    - have a *hierarchical or nested* structure

Nominal Multinomial Responses

- Q: how to link $p_{\mathbf{x}}$=($p_{\mathbf{x}1}$, $p_{\mathbf{x}2}$, …, $p_{\mathbf{x}J}$) with $\mathbf{x}$?
  - Recall: binomial logit model.

    $$(1-p_{\mathbf{x}}, p_{\mathbf{x}}) \Rightarrow (1, p_{\mathbf{x}}/(1-p_{\mathbf{x}})) \Rightarrow p_{\mathbf{x}}/(1-p_{\mathbf{x}}) \underset{\text{exp}}{\overset{\text{log}}{\rightleftarrows}} \eta_{\mathbf{x}} = X\boldsymbol{\beta}$$
- *Multinomial logit model*:

$$\begin{array}{ccccc} (p_{\mathbf{x},1}, p_{\mathbf{x},2}, \ldots, p_{\mathbf{x},J}) \Rightarrow (1, & p_{\mathbf{x},2}/p_{\mathbf{x},1}, & \ldots, & p_{\mathbf{x},J}/p_{\mathbf{x},1}) \\ \text{log}\downarrow\uparrow\text{exp} & \text{log}\downarrow\uparrow\text{exp} & \cdots & \text{log}\downarrow\uparrow\text{exp} \\ \eta_{\mathbf{x},1}=0 & \eta_{\mathbf{x},2}=X\boldsymbol{\beta}_2 & \cdots & \eta_{\mathbf{x},J}=X\boldsymbol{\beta}_J \end{array}$$

$$\boxed{X\boldsymbol{\beta}_j = \beta_{0j} + h_1(\mathbf{x})\beta_{1j} + \cdots + h_{p-1}(\mathbf{x})\beta_{p-1,j}}$$

  - The link obeys the constraints $0 \le p_{\mathbf{x}j} \le 1$, $j=1,\ldots,J$, and $\Sigma_j p_{\mathbf{x}j}=1$
  - Category 1 declared as the *baseline* (note that we add $\eta_{\mathbf{x}1}=0$)
  - $p_{\mathbf{x}j} = \exp(\eta_{\mathbf{x}j}) \Big/ \left[1 + \sum_{j=2}^{J} \exp(\eta_{\mathbf{x}j})\right]$
  - Inferences
    - Estimation of the parameters: can use maximum likelihood
    - Other inferences: can use standard likelihood-based approach
    - Concepts and methods in GLM can be similarly applied (Q: why?)

➢ Interpretation of $\boldsymbol{\beta}_j$

- The log-odds in favor of category j over category j' is

$$\log\left(p_{\mathbf{x}j}/p_{\mathbf{x}j'}\right) = \mathbf{x}^T\left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'}\right)$$

  ($\Rightarrow$ the contrasts among the vectors $\boldsymbol{\beta}_j$ are of interest)
- The interpretation of $\boldsymbol{\beta}_j$ depends on the choice of baseline category $\Rightarrow$ the k^{th} component of $\boldsymbol{\beta}_j$ represents the change in log-odds of moving from the baseline category to the j^{th} category for a unit change in the k^{th} predictor of $\mathbf{x}$ (other predictors held constant)
- Intercepts of $\boldsymbol{\beta}_j$ represent probabilities of J categories at $\mathbf{x}$=**0**

• *Multinomial log-linear model*:

➢ Recall: the connection between Poisson and multinomial $\Rightarrow$ it is possible to fit a model for multinomial response using a Poisson

$\mathbf{x}$	z: 1	…	J	
$\mathbf{x}_1$	y_{11}	…	y_{1J}	n_1
…	…	…	…	…
$\mathbf{x}_I$	y_{I1}	…	y_{IJ}	n_I

➢ Procedure:

- Arrange the data in a two-way tables
- Declare a nominal variable that has a level for each multinomial response category, called it *response factor*, z

- Fit a log-linear model $Y_{\mathbf{x},z} \sim \mathbf{x}_c + z + z{:}X$, where

  $Y_{\mathbf{x},z}$: $(y_{11}, \ldots, y_{1J}, y_{21}, \ldots, y_{2J}, \ldots, y_{I1}, \ldots, y_{IJ})^T$

  $\mathbf{x}_c$: a nominal factor which treats each covariate class as a level (i.e., $\mathbf{x}_c$ has I levels)

  $z{:}X$: the interaction term with the same X (without the intercept term) in the multinomial logit model

  - ▫ The parameters associated with $\mathbf{x}_c$ will represent the total for the cases ($\Leftrightarrow n_1, n_2, \ldots, n_I$)
  - ▫ the parameters associated with r will represent the category totals ($\Leftrightarrow$ intercepts in multinomial logit model)
  - ▫ The interaction terms $z{:}X$ tell us how the probability of falling in the different categories changes with z ($\Leftrightarrow$ coefficients of X in the multinomial logit model)
  - ▫ This multinomial log-linear model would have same deviance as its corresponding multinomial logit model
- **Q**: what information in the log-linear model not valid?

➢ Multinomial logit can be viewed as a GLM-type model, which allows us to apply all the methodology developed for GLM

- **Q**: why can multinomial log-linear model be connected to multinomial logit model?

  Suppose that

$$n_{\mathbf{x}} \overset{\text{indep}}{\sim} \text{Poisson}(\lambda_{\mathbf{x}})$$

$$(y_{\mathbf{x}1}, \ldots, y_{\mathbf{x}J} | n_{\mathbf{x}}) \overset{\text{indep}}{\sim} \text{multinomial}(n_{\mathbf{x}}, p_{\mathbf{x}1}, \ldots, p_{\mathbf{x}J})$$

  Then, $y_{\mathbf{x}j} \overset{\text{indep}}{\sim} \text{Poisson}(\mu_{\mathbf{x}j} = \lambda_{\mathbf{x}} p_{\mathbf{x}j})$

  If we fit a log-linear model for $y_{\mathbf{x}j}$, then:

$$\log\left(\frac{\mu_{\mathbf{x}j}}{\mu_{\mathbf{x}1}}\right) = \log\left(\frac{\lambda_{\mathbf{x}} p_{\mathbf{x}j}}{\lambda_{\mathbf{x}} p_{\mathbf{x}1}}\right) = \log\left(\frac{p_{\mathbf{x}j}}{p_{\mathbf{x}1}}\right) = \eta_{\mathbf{x}j} = X\boldsymbol{\beta}_j$$

$$\begin{aligned} \Rightarrow \log(\mu_{\mathbf{x}j}) &= \log(\mu_{\mathbf{x}1}) + X\boldsymbol{\beta}_j \\ &= \log(\lambda_{\mathbf{x}} p_{\mathbf{x}1}) + \underbrace{X\boldsymbol{\beta}_j}_{\beta_{0j} + h_1(\mathbf{x})\beta_{1j} + \cdots + h_{p-1}(\mathbf{x})\beta_{p-1,j}} \end{aligned}$$

❖ **Reading**: Faraway (2006, 1st ed.), 5.1

Ordinal Multinomial Responses

- Suppose that
  - We have a multinomial response with J ordered categories, and
  - For individual **x**, observe ordinal response $(y_{\mathbf{x}1}, \ldots, y_{\mathbf{x}J})$
- For an ordered multinomial response, it is often easier to work with the cumulative probabilities

$$\gamma_{\mathbf{x}j} = p_{\mathbf{x}1} + \cdots + p_{\mathbf{x}j},\ j = 1, \ldots, J$$

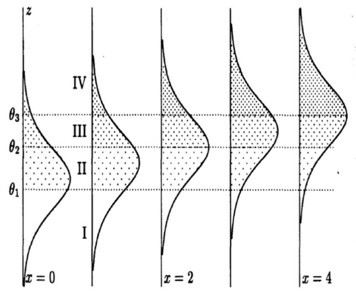
  - $p_{\mathbf{x}1} = \gamma_{\mathbf{x}1}$; $p_{\mathbf{x}j} = \gamma_{\mathbf{x},j} - \gamma_{\mathbf{x},j-1}$, $j = 2, \ldots, J-1$; $\gamma_{\mathbf{x}J} = 1$
  - **Q**: why cumulative probabilities are better?
    - the cumulative probabilities are increasing and invariant to combining adjacent categories
    - $\gamma_{\mathbf{x}J} = 1$, so we need only model $J-1$ probability
- **Q**: how to link $\boldsymbol{\gamma}_{\mathbf{x}} = (\gamma_{\mathbf{x}1}, \gamma_{\mathbf{x}2}, \ldots, \gamma_{\mathbf{x},J-1})$ with **x**?

$$\begin{array}{ccc} (\gamma_{\mathbf{x}1}, & \ldots\ , & \gamma_{\mathbf{x},J-1}) \\ g \downarrow & & g \downarrow \\ \eta_{\mathbf{x}1} = X\boldsymbol{\beta}_1 & \cdots & \eta_{\mathbf{x},J-1} = X\boldsymbol{\beta}_{J-1} \end{array}$$

$$\boxed{\begin{aligned} X\boldsymbol{\beta}_j = \ & \beta_{0j} - \\ & h_1(\mathbf{x})\beta_1 - \cdots - \\ & h_{p-1}(\mathbf{x})\beta_{p-1} \end{aligned}}$$

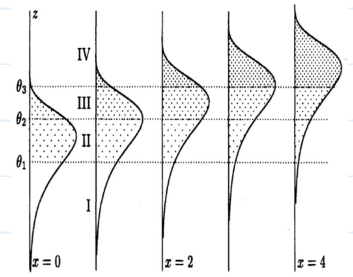
- some choices of g: (1) logit (2) probit (3) complementary log-log
  - Notice that the intercepts, β_{0j}, are *different* for the J categories
  - $\boldsymbol{\beta}=(\beta_1, \ldots, \beta_{p-1})$ *do not* depend on $j \Rightarrow$ we assume that the predictors have a uniform effect on the probabilities of response categories in a sense that we will shortly make clear
  - Negative sign before $\boldsymbol{\beta}$

- Latent variable approach for ordinal variables:
  - let $z_{\mathbf{x}}$ be some unobserved continuous variable that might be thought of as the real underlying latent response
  - we only observe a discretized version of $z_{\mathbf{x}}$ in the form of $y_{\mathbf{x}}$ where $y_{\mathbf{x}}=j$ is observed if $\beta_{0,j-1}<z_{\mathbf{x}} \leq \beta_{0,j}$



  - suppose that $z_{\mathbf{x}}-\boldsymbol{\beta}^T\mathbf{x}^*$ has a distribution F:

$$
\begin{aligned}
P(y_{\mathbf{x}} \leq j) &= P(z_{\mathbf{x}} \leq \beta_{0j}) \\
&= P(z_{\mathbf{x}} - \boldsymbol{\beta}^T\mathbf{x}^* \leq \beta_{0j} - \boldsymbol{\beta}^T\mathbf{x}^*) \\
&= F(\beta_{0j} - \boldsymbol{\beta}^T\mathbf{x}^*)
\end{aligned}
$$



- If F follows the logistic distribution, i.e., $F(z)=e^z/(1+e^z)$, then

$$
\gamma_{\mathbf{x}j} = \frac{\exp(\beta_{0j} - \boldsymbol{\beta}^T\mathbf{x}^*)}{1+\exp(\beta_{0j} - \boldsymbol{\beta}^T\mathbf{x}^*)}
$$

  - so, we would have a logit model for the cumulative probabilities $\gamma_{\mathbf{x}j}$
- Choosing Normal for latent variable (i.e, F) leads to probit model
- F=Extreme value distribution leads to complementary log-log
- Notice that if $\boldsymbol{\beta} > 0$, as $\mathbf{x}$ increases, $P(y_{\mathrm{x}}=J)$ will also increase

  $\Rightarrow$ this explain the use of the minus sign in the definition of the model because it allows for the more intuitive interpretation of the sign of $\boldsymbol{\beta}$

- *Proportional odds* model
  - Let $\gamma_{\mathbf{x}j}= \gamma_j(\mathbf{x}) = P(y_{\mathrm{x}} \leq j|\mathbf{x})$, then the proportional odds model, which use the *logit* link, is:

$$
\log\left(\frac{\gamma_j(\mathbf{x})}{1-\gamma_j(\mathbf{x})}\right) = \beta_{0j} - \boldsymbol{\beta}^T\mathbf{x}^*
$$

- It is so called because the relative odds for the event $y_{\mathbf{x}} \leq j$ comparing $\mathbf{x}=\mathbf{x}_1$ and $\mathbf{x}=\mathbf{x}_2$ are:

$$\left(\frac{\gamma_j(\mathbf{x}_1)}{1-\gamma_j(\mathbf{x}_1)}\right) \Big/ \left(\frac{\gamma_j(\mathbf{x}_2)}{1-\gamma_j(\mathbf{x}_2)}\right) = \exp\left(-\boldsymbol{\beta}^T(\mathbf{x}_1^* - \mathbf{x}_2^*)\right)$$

  - This does not depend on j
  - Of course, this assumption of proportional odds does need to be checked for a given dataset
  - It can be checked by computing the sample (observed) odds proportions with respect to $\mathbf{x}$
- Some advantages
  - The model use fewer parameters than the multinomial logit model (**Q**: the former nested in the latter?)
  - Typically, the output from the proportional odds model is easier to interpret
- Inferences follows the usual GLM likelihood-based approach
- Interpretation of the fitted coefficients: odds of moving from one category to another increase by a factor of exp($\boldsymbol{\beta}$) as $\mathbf{x}$ increases by one unit (check lab)

- *ordered probit* model
  - If the latent variable $z_{\mathbf{x}}$ has a standard normal distribution, then

$$\Phi^{-1}(\gamma_j(\mathbf{x})) = \beta_{0j} - \boldsymbol{\beta}^T\mathbf{x}^*$$

  - Coefficients estimates could be very different from proportional odds model, but predicted values usually are very similar
- *proportional hazards* model
  - Concept of hazard
    - Developed in insurance application: when issuing a life insurance policy, the insurer is interested in the prob. That the person will die during the term of the policy given that they are alive now
    - This is not the same as the unconditional probability of death
  - Suppose we use the complementary log-log link, i.e.,

$$\log(-\log(1-\gamma_j(\mathbf{x}))) = \beta_{0j} - \boldsymbol{\beta}^T\mathbf{x}^*$$

  Then, the hazard of category j is the probability of falling in category j given that your category is greater than j:

$$\text{hazard}_{\mathbf{x}}(j) = P(y_{\mathbf{x}} = j | y_{\mathbf{x}} \geq j) = P(y_{\mathbf{x}} = j)/P(y_{\mathbf{x}} \geq j)$$

$$= \frac{p_{\mathbf{x}j}}{1-\gamma_{\mathbf{x},j-1}} = \frac{\gamma_{\mathbf{x}j} - \gamma_{\mathbf{x},j-1}}{1-\gamma_{\mathbf{x},j-1}} = 1 - \frac{1-\gamma_{\mathbf{x}j}}{1-\gamma_{\mathbf{x},j-1}}$$

- These hazards are then proportional across categories as **x** varies
- The corresponding latent variable distribution is the extreme value distribution:

$$F(z) = 1 - \exp(-\exp(z))$$

- The extreme value distribution is not symmetric like the logistic and normal

• Generalization:

➢ The proportional hazards and odds models can be generalized by allowing beta to vary, i.e.,

$$\log\left(\frac{\gamma_j(\mathbf{x})}{1-\gamma_j(\mathbf{x})}\right) = \beta_{0j} - \boldsymbol{\beta}_j^T \mathbf{x}^*$$

➢ But, this loses the proportionality property

❖ **Reading**: Faraway (2006, 1st ed.), 5.3

Hierarchical or Nested Multinomial Responses

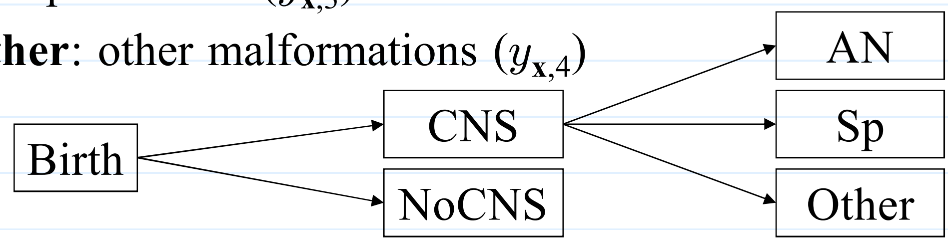
• Consider a multinomial response with the 4 categories:

➢ **NoCNS**: no central nervous system malformation ($y_{\mathbf{x},1}$)

AN: anencephalus ($y_{\mathbf{x},2}$)

Sp: spina bifida ($y_{\mathbf{x},3}$)

Other: other malformations ($y_{\mathbf{x},4}$)



➢ There exists a hierarchical structure between the 4 categories

➢ Q: what are the problems if we ignore the hierarchical structure and just treat them as 4 nominal categories in the analysis?

- In the data, most births suffer no malformation and so NoCNS dominates the other 3 categories (Q: why is this a problem?)
- Q: what happen if **x** has significant effects on the prob. of NoCNS but not on the other 3 categories, and NoCNS is chosen to be the baseline? (check lab)

- The likelihood of $(y_{\mathbf{x}1}, y_{\mathbf{x}2}, y_{\mathbf{x}3}, y_{\mathbf{x}4})$ is proportional to (note: most inference methods in GLM are based on the likelihood approach):

$$p_{\mathbf{x}1}^{y_{\mathbf{x}1}} p_{\mathbf{x}2}^{y_{\mathbf{x}2}} p_{\mathbf{x}3}^{y_{\mathbf{x}3}} p_{\mathbf{x}4}^{y_{\mathbf{x}4}}$$
$$= \left[p_{\mathbf{x}1}^{y_{\mathbf{x}1}} \left(p_{\mathbf{x}c}\right)^{y_{\mathbf{x}2}+y_{\mathbf{x}3}+y_{\mathbf{x}4}} \right] \times \left[\left(\frac{p_{\mathbf{x}2}}{p_{\mathbf{x}c}}\right)^{y_{\mathbf{x}2}} \left(\frac{p_{\mathbf{x}3}}{p_{\mathbf{x}c}}\right)^{y_{\mathbf{x}3}} \left(\frac{p_{\mathbf{x}4}}{p_{\mathbf{x}c}}\right)^{y_{\mathbf{x}4}} \right]$$

where $p_{\mathbf{x}c}=p_{\mathbf{x}2}+p_{\mathbf{x}3}+p_{\mathbf{x}4}=1-p_{\mathbf{x}1}$ (= prob. of a birth with CNS malform.)

- The 1st part is a binomial likelihood for CNS vs. NoCNS
- The 2nd part is a multinomial likelihood for the three CNS categories *conditional* on the presence of CNS
  - e.g., $p_{\mathbf{x}2}/p_{\mathbf{x}c}$ is the conditional probability of an anencephalus birth given that a malformation has occurred at $\mathbf{x}$
- We can separately
  - develop a binomial model for whether malformation occurs using data from *all objects*, and
  - develop a multinomial model for the type of malformation using data only from *subjects with CNS malformation*

❖ **Reading**: Faraway (2006, 1st ed.), 5.2