Nominal Response Multinomials (Chapter 8.1 in CDA)

Luke Toomey

October 5, 2017

Multinomial Response Data

– Y is a nominal scale response variable with J categories, where Multinomials/polytomous logistic model describes loss odds for $\binom{J}{2}$ pairs of categories

– $\pi_j(\vec{x}) = P(y = j|\vec{x}), when \sum_j \pi_j(x) = 1$

– $Y|\vec{x} \sim Multinomial(\hat{N}, \{\pi_1(x), \pi_2(x), ..., \pi_j(x)\}$

– pick baseline category, J

- $log\frac{\pi_j(x)}{\pi_J(x)} = \alpha_j + \beta_j^T(\vec{x})$ for $j = 1, ..., J-1$

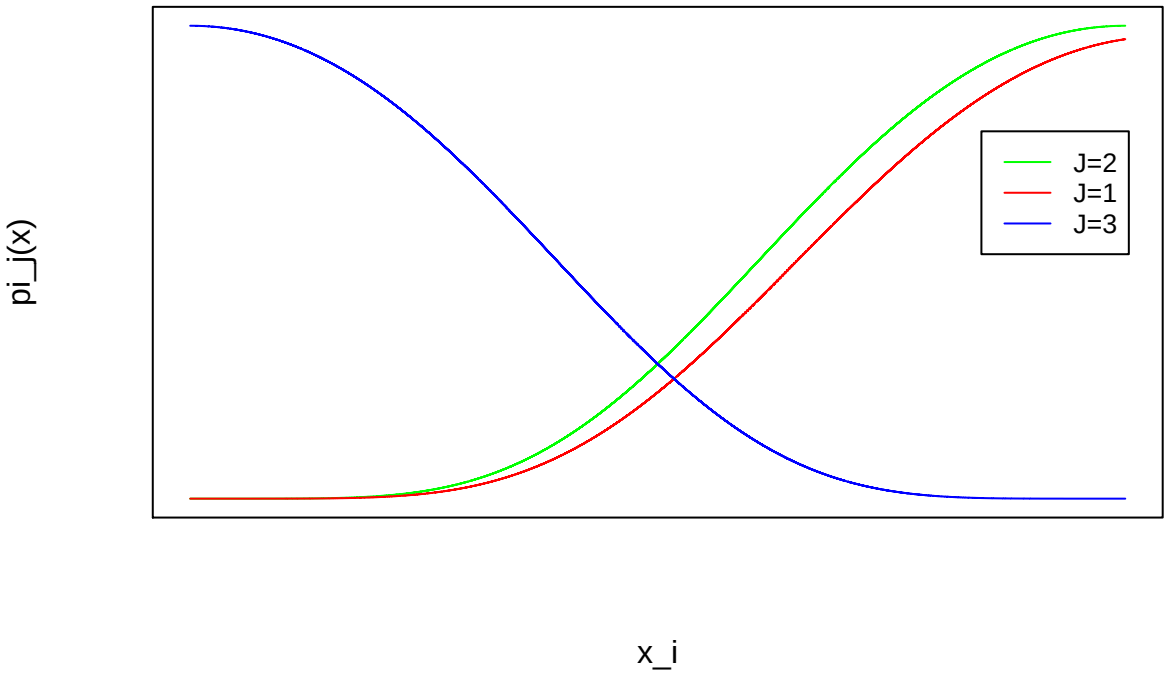
– Notes about using model:

- X^2 (Pearson G.O.F) or G^2 (likelihood G.O.F) statistics when data not sparse
- if data are sparse X^2, G^2 are valid for comparison between models with a few terms different

Response Probabilities

$$\pi_j(\vec{x}) = \frac{exp\{\alpha_j + \beta_j^T(\vec{x})\}}{1 + \sum_{h=1}^{J} exp\{\alpha_h + \beta_j^T(\vec{x})\}}$$

where $\alpha_j = \beta_j^T = 0$ and x is a vector.



Here we hold all X_j constant, and calculate the probabilities across a range of X.

How is this a multivariate GLM?

For a response vector:

$$\vec{y_i} = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ij} \end{bmatrix}, \ observations \ for \ subject \ i$$

y_{ij} are the number of occurrences in each j category with covariate profile i^{th} observation if you have a data set with only categorical X's

$$\mu_j = \begin{bmatrix} \pi_1(x_i) \\ \pi_2(x_i) \\ \vdots \\ \pi_{J-2}(x_i) \end{bmatrix}$$

with a vector $g(\mu_i) = X_i\beta^T$ *where* $g(\mu_i)$ *has dimensions* $((J-1)\times 1)$, X_i *has dimensions* $(J-1)\times(p(J-1))$, *and* β *has dimensions* $(p(J-1))\times 1$ ...

$$g_i(\mu_i) = log\frac{\mu_i}{1-\sum_{h=1}^{J-1}\mu_{ih}}$$

For some given individual assume we have:

$$X_i = \begin{bmatrix} 1 & x_i^T & 0 & 0 & \dots & 0 \\ 0 & 1 & x_i^T & 0 & \dots & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & \dots & & \ddots & \\ 0 & 0 & 0 & \dots & 1 & x_i^T \end{bmatrix} \beta^T = \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \\ \vdots \\ \alpha_{J-1} \\ \beta_{J-1} \end{bmatrix}$$

Example

–Information gathered on patients with certain symptoms through surveys given in clinical evaluations i.e age, do they have a fever, temp, rash present etc., hoping that answers to survey questions can be used to diagnose someone based on their symptoms, using a model to classify which disease you may have.

- could be used to avoid doing expensive tests on patients
- Y_{ij}= diagnosis of i^{th} individual with j^{th} disease:

$$= \begin{bmatrix} 1 & if \ person \ has \ disease \\ 0 & otherwise \end{bmatrix}$$

- This ignores possibilties of coieinfection, meaning you are sick with 2 pathogens

–The response vector would look like:

$$y_i = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

Where the sum will always be 1 and the row with the 1 indciated which diease they have

–The motive here is the fact that clinical tests are expensive and time consuming, here we can model and accurately precet or classift indivudals based on sympotons or other easily ascceibke covaeriate such as:

- Age (centered at mean), Heachace (Y/N), Rash (Y/N), Stomach Ache (Y/N), Month (categorical), etc.
- Outcomes: D = Dengue fever, Z = Zika, C = Chikingunnya, F = Flu, O = Other
- we have an aggregated data set

i	j	y_{ij}	n_1	$X_{h,i}$	$X_{r,i}$
1	D	10	50	1	0
1	Z	30	50	1	0
1	C	5	50	1	0
1	F	0	50	1	0
1	O	5	50	1	0

From this table, there were 50 people in the group with i covariate profiles, i.e. 30/50 people with covariate i have zika. Moreover, this group of people had headches but did not have rashes.

$$y_i \sim multinational(n_i, \{\pi_i(x)\})$$

Model

–Let J = "other"

$$log\frac{\pi_j(x_i)}{\pi_J(x_j)} = \alpha_i + \beta_j^A(age_i) + \beta_j^H(headache_i) + \beta_j^R(rash_i)...$$

	$Dengue$	$Zika$	Flu	$\dots$
α	0.1	2.6	$\dots$	$\dots$
$Age,\ \beta^A$	0.1	$-.2$	$\dots$	$\dots$
$Headache,\ \beta^H$	1.5	1.4	$\dots$	$\dots$
$Rash,\ \beta^R$	1.6	.1	$\dots$	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

What is α_{zika} *or* e^z ... $\alpha_{j=z}$?

$$log\frac{\pi_Z(x)}{\pi_D(x)} = \alpha_z + X\beta_z \ \ , \ log\frac{\pi_Z(x=0)}{\pi_O(x=0)} = \alpha_z$$

–e^{α_z} is the realative risk in sample of having Zika vs other infection for baseline covariate profile i.e. no symptoms and average aged individual (important to note interpretability depends on sampling scheme).

–What does $\beta_D^R = 1.6$ *or* $e^{\beta_D^R} = 4.5$ mean?

- Holding everything but rash constant, having a rash is associatd with a 4.5 times incresed risk of Dengue compared to other.

$$log\frac{(\frac{\pi_D(R=1)}{\pi_D(R=1)})}{(\frac{\pi_O(R=0)}{\pi_O(R=0)})} = \beta_D^R$$

–We know the comparision for any $log\frac{\pi_a(x)}{\pi_b(x)} = log\frac{\pi_a(x)}{\pi_J(x)} - log\frac{\pi_b(x)}{\pi_J(x)}$