LEC. #8

Nominal multinomial regression

data $(\underset{\sim}{x}_i, \underbrace{y_{i1}, \ldots, y_{iJ}}_{\text{sum to } n_i})$ $\quad i = 1, \ldots, n$

on n_i independent trials,
y_{ij} had the j-th of J
(unordered) outcomes

Binomial when $J = 2$
generalizing to $J > 2$

so

$$Pr\{Y_{i1} = y_{i1}, \ldots, Y_{iJ} = y_{iJ}\} = \left(\frac{n_i!}{y_{i1}! \cdots y_{iJ}!}\right) p_{i1}^{y_{i1}} \cdots p_{iJ}^{y_{iJ}}$$

$p_{i1} + \cdots + p_{iJ} = 1$

Ch. 5 example

Survey of $n = 944$ voters

y_i = party identification

"REP.", "DEM.", "IND."

$\underset{\sim}{x}_i$ = income, age, education

numeric

categorical

7 categories

So $n_i = 1$, $i = 1, \dots, 944$

$J = 3$

$\underset{\sim}{y_i} = (y_{i1}, y_{i2}, y_{i3})$

with

$y_{i1} + y_{i2} + y_{i3} = 1$

Exp. family ? GLM ?

variance is matrix

grey area, since response var. is J-1 dimensional

variance function?

Can still think link structure though

multinomial logit model

$$g\begin{pmatrix} p_2 \\ \vdots \\ p_J \end{pmatrix} = \begin{pmatrix} \log\left\{\frac{p_2}{\underbrace{1-(p_2+\ldots+p_J)}_{p_1}}\right\} \\ \vdots \\ \log\left\{\frac{p_J}{1-(p_2+\ldots+p_J)}\right\} \end{pmatrix} = \underbrace{\begin{pmatrix} \underset{\sim}{\beta_2}^T \\ \vdots \\ \underset{\sim}{\beta_J}^T \end{pmatrix}}_{\#\text{ cols} = \#\text{ predictors}} \underset{\sim}{x}_{e\text{ vector}}$$

reduces to logistic regression when $J=2$

and

$p_1 = 1 - [p_2 + \ldots + p_J]$

J-1 rows

\# cols = \# predictors

Interpretation

$$\frac{p_j}{p_1} = e^{\beta_j^T \underset{\sim}{x}}$$

conditional odds?

think constraint $\underset{\sim}{\beta}_1^T = (0, \dots, 0)$

$$\frac{p_j/p_1}{p_k/p_1} = \frac{p_j}{p_k} = e^{(\beta_j - \beta_k)^T \underset{\sim}{x}} \leftarrow \text{Odds}(Y=j \mid Y \in \{j,k\}) = \frac{\frac{p_j}{\cancel{p_j+p_k}}}{\frac{p_k}{\cancel{p_j+p_k}}}$$

inverting the link

$$(p_1, \dots, p_J) = \frac{1}{\left(1 + \sum_{j=2}^{J} e^{\beta_j^T \underset{\sim}{x}}\right)} \left(1, e^{\beta_2^T \underset{\sim}{x}}, \dots, e^{\beta_J^T \underset{\sim}{x}}\right)$$

$$(p_1, \dots, p_J) = \frac{1}{1 + \sum_{j=2}^{J} e^{\beta_{j1}}} \left(1, e^{\beta_{21}}, \dots, e^{\beta_{J1}}\right)$$

Note: typically want

$\underset{\sim}{x} = (1, \text{predictors})$

first col. of $\left(\boxed{}\ \beta\right)$

are intercept params.

so predictors $= 0$

Ex.

$$\hat{\beta} = \begin{array}{ccc} & \text{Intcpt} & \text{Income } (\$1000) \\ \text{"IND"} & -1.18 & 0.0161 \\ \text{"REP"} & -0.95 & 0.0176 \end{array}$$

recall "DEM" is reference level

$$\widehat{p \mid INC=0} \propto \left(e^{0}, e^{-1.18}, e^{-0.95}\right)$$

hypothetical

$$= \left(0.59, 0.18, 0.23\right)$$

"D" "I" "R"

cond. OR

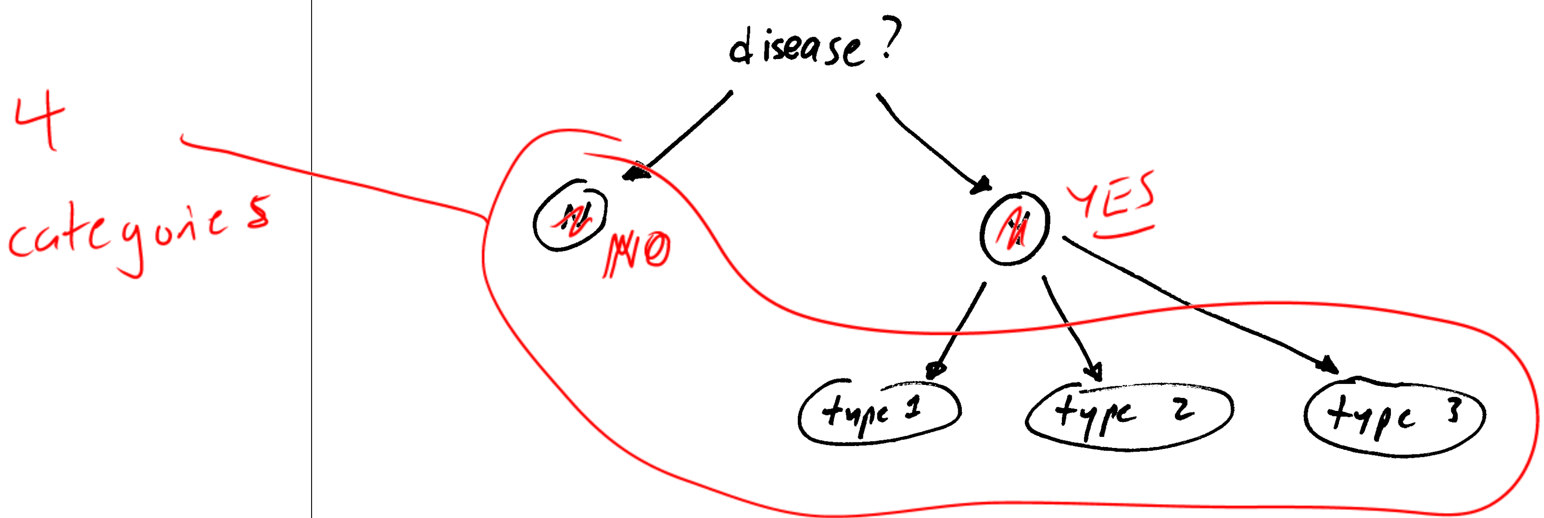
Cond. OR associated with \$ 1000 increase in icome?

$$\frac{\frac{p_2(z+1000)}{p_1(z+1000)}}{\frac{p_2(z)}{p_1(z)}}$$

$$e^{.0161} \approx 1.0162$$

$$e^{.017} \approx 1.0178$$

Hierarchical / Nested



Strategy - separately fit

- Binomial model for DISEASE
- Multinomial model for TYPE / DISEASE = YES

Statistical justification? ✓

Start with multinomial model

$$\text{lik.} \propto \prod_{i=1}^{n} p_1^{y_{i1}} p_2^{y_{i2}} p_3^{y_{i3}} p_4^{y_{i4}}$$

reparameterize $q = \Pr\{\text{Dis.}=Y\} = p_2 + p_3 + p_4$

$1-q = \Pr\{\text{Dis}=\text{No}\}$

$$r_i = \Pr\{\text{Type}=i \mid \text{Dis.}=Y\} = \frac{p_{i+1}}{p_2+p_3+p_4}$$

So $$\text{lik.} \propto \prod_{i=1}^{n} (1-q)^{y_{i1}} (q r_1)^{y_{i2}} (q r_2)^{y_{i3}} (q r_3)^{y_{i4}}$$

$$\propto \underbrace{\left\{ \prod_{i=1}^{n} (1-q)^{y_{i1}} q^{y_{i2}+y_{i3}+y_{i4}} \right\}}_{\text{depends on } q \text{ only}} \underbrace{\left\{ \prod_{i=1}^{n} r_1^{y_{i2}} r_2^{y_{i3}} r_3^{y_{i4}} \right\}}_{\text{depends on } (r_1, r_2, r_3) \text{ only}}$$

stats matches intuition

factors - ∴ can fit separate models

Fundamentally different modelling assumptions by choosing different parameterization for "linking"

e.g. nested model says

$$\log\left(\frac{p_2 + p_3 + p_4}{p_1}\right) \text{ linear in } \underset{\sim}{x}$$

disease

no disease

regular multinomial model says

$$\log\left[\frac{p_2 + p_3 + p_4}{p_1}\right] = \underbrace{\log\left(e^{\underset{\sim}{\beta_2}^T \underset{\sim}{x}} + e^{\underset{\sim}{\beta_3}^T \underset{\sim}{x}} + e^{\underset{\sim}{\beta_4}^T \underset{\sim}{x}}\right)}$$

not linear in $\underset{\sim}{x}$