BST 210
Applied Regression Analysis

Lecture 16
Plan for Today

Extending (binary) Logistic Regression to -

Categorical logistic regression with >2 outcomes:

- Multinomial Logistic Regression (>2 unordered levels)
- Ordinal Logistic Regression (> 2 ordered levels)

Multinomial Outcome Data

- Sometimes we have a categorical outcome variable that can assume ***more than two categories*** (not ordered):

- Choice of occupation
- RCT outcomes: progression-free survival, disease-free survival, death
- 4 modes of operative delivery
- 3 types (epithelial, germ cell, and stromal) of ovarian tumors
- Diagnosis of 3 bacterial infections seen in children
- 5 forms of meningitis
- 3 prognostic outcomes of elderly after hospitalization
- 6 cost-structure models in hospital reimbursement

Multinomial Outcome Data

Multinomial Distribution

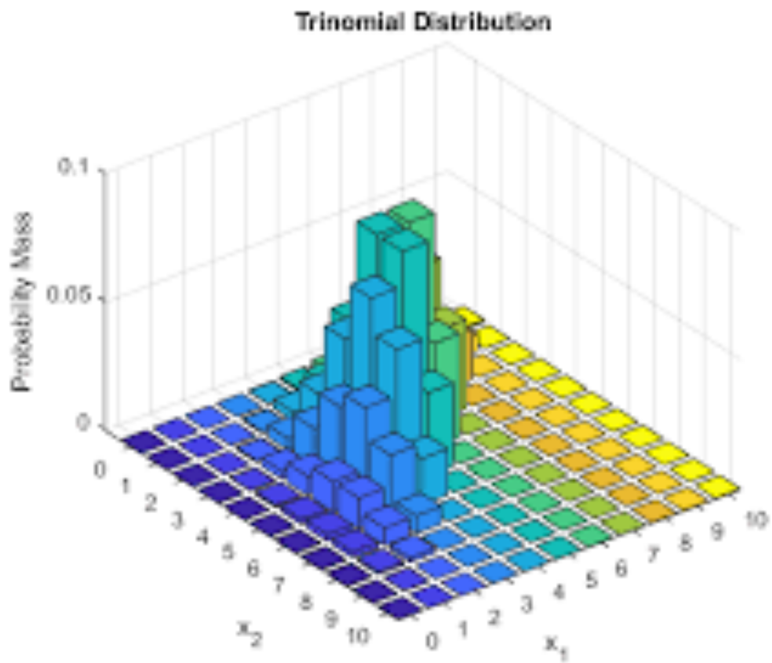
- The Binomial distribution can be extended to describe number of outcomes in a series of independent trials each having more than 2 possible outcomes.

- If a given trail can result in the k outcomes $E_1, E_2, \ldots, E_k$ with probabilities $p_1, p_2, \ldots, p_k$, then the probability distribution of the random variables $X_1, X_2, \ldots, X_k$, representing the number of occurrences for $E_1, E_2, \ldots, E_k$ in n independent trials is

$$p_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

with $\sum_{i=1}^{k} x_i = n$, and $\sum_{i=1}^{k} p_i = 1.$

Multinomial Outcome Data



Recall: (Binary) Logistic Regression

- We have been modeling *binary outcomes* as:

  $$\mathbf{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

  always assuming

  (1) we have the correctly specified model to relate Y_i with X_{ik}

  (2) Y_i are independent

- Same assumptions must hold for modeling *multinomial outcomes*

Multinomial Model

- In the multinomial outcome situation, we can use *multinomial* or *polychotomous logistic regression*
- We now need proportions p_1, p_2, p_3,…p_c for an outcome having c levels, instead of simply p_1 and p_2 (all of which still need to sum to 1)
- AND subjects (Y_i) still must be independent of each other
- We'd also need to have enough observations in each category, or > 10 observations per independent variable (predictor)
- If the above hold, likelihood based methods can be used to estimate model parameters (MLEs), just as in the binary outcome case

Recall: (Binary) Logistic Regression

- For logistic regression with a single covariate X to predict a <u>binary</u> outcome Y, we use the model:

$$\textbf{logit}(p) = \textbf{log}[p\ /\ (1 - p)] = \beta_0 + \beta_1\, x$$

- Here $p = P(Y = 1 \mid X = x)$ and we assume that the relationship between logit(p) and x is linear

- Here we are comparing outcomes for ***<u>Y=1 to Y=0</u>***, for '1-unit changes in X'

- Solving for p, we obtain:

$$p = P(Y{=}1) = \frac{\exp^{(\alpha + \beta X)}}{1 + \exp^{(\alpha + \beta X)}}$$

and

$$1 - p = P(Y = 0) = \frac{1}{1 + \exp^{(\alpha + \beta X)}}$$

Now: Multinomial model

So how do we develop the multinomial model?

- Suppose there are three possible outcomes, say 1, 2, and 3 (or 0, 1, 2 is another possibility)

- One possible approach to this new problem with a multi-level categorical outcome variable would be to run *two separate logistic regression models* such that:

  In model 1: Compare category 3 vs. category 1 and 2 (combined) → ie 0 = {3} and 1 = {1, 2}

  In model 2: Compare category 2 and 3 (combined) vs. category 1 → ie 0 = {2, 3} and 1 = {1}

Multinomial Model

- Fitting *two separate models* is a valid approach *only if* it makes sense to combine the categories like this, that is, if the categories are *ordered* in a logical way (and they may not be)

- Fitting *two separate models* does not allow us to directly compare the coefficients for the same risk factor in the two separate models

- What to do?

Multinomial Logistic Regression

- Instead of fitting two separate models, we will fit a *single model* with *c* outcome categories (c may be 3, 4, 5…not 28!)
- We'll call this a *multinomial logistic regression model*
- Standard errors will be smaller this way; model more stable
- Let group 1 be the reference group say (can be any group), and groups 2 through *c* will be compared to group 1
- Note that c = 2 is simply the special case of a binary logistic regression model
- Let's develop this model →

Multinomial Logistic Regression

- Suppose c = 3 and we have a single covariate x. We can then obtain

$$\text{P(group 1)} = \frac{1}{1+\exp(\alpha_2+\beta_{21}x)+\exp(\alpha_3+\beta_{31}x)}$$

$$\text{P(group 2)} = \frac{\exp(\alpha_2+\beta_{21}x)}{1+\exp(\alpha_2+\beta_{21}x)+\exp(\alpha_3+\beta_{31}x)}$$

$$\text{P(group 3)} = \frac{\exp(\alpha_3+\beta_{31}x)}{1+\exp(\alpha_2+\beta_{21}x)+\exp(\alpha_3+\beta_{31}x)}$$

- Here the probabilities for all c = 3 outcomes sum to 1.

Multinomial Logistic Regression

- How do we derive these probabilities?
- (Remember, we want nice forms for P(1), P(2), and P(3) so that we can begin making interpretations in terms of Risk Ratios, and Relative Risk Ratios.)

Recall: Measures of Effect - Risk Ratio

- The *risk ratio* (RR) is a way of estimating magnitude of association that we discussed earlier in the course

- It is generally defined as p_1/p_2 and can be estimated by

$$\hat{p}_1 / \hat{p}_2.$$

- It is a measure of relative rather than absolute risk

Multinomial Logistic Regression

Now back to our multinomial model setting – let's interpret this model!

- What is the effect for a 1 unit increase in *x*?

- To answer this question, first we need the *risk ratio* of being in outcome category 2 versus the reference category 1, which is given by

P(group 2) / P(group 1) = $\exp(\alpha_2 + \beta_{21} x)$

- Then the *relative risk ratio* of being in outcome category 2 versus the reference category 1 for a one unit increase in *x* is given by

$\exp(\alpha_2 + \beta_{21}(x+1)) / \exp(\alpha_2 + \beta_{21} x) = \exp(\beta_{21})$

Multinomial Logistic Regression

- How do we derive the above interpretations?

Multinomial Logistic Regression

- Similarly, the *relative risk ratio* of being in outcome category 3 versus the reference category 1 for a one unit increase in *x* is given by

$$\exp(\beta_{31})$$

- With a bit more thought we can also get the relative risk ratios comparing category 1 to category 2, or category 2 to category 3, etc.

- Note that multinomial models are not easily interpretable in terms of odds ratios

Multinomial Logistic Regression

- More generally, for *k* covariates and *c* outcomes :

$$\text{P(group 1)} = \frac{1}{1+\sum_{j=2}^{c} \exp(\alpha_j+\sum_{i=1}^{k} \beta_{ji}x_i)}$$

$$\text{P(group } q\text{)} = \frac{\exp(\alpha_q+\sum_{i=1}^{k} \beta_{qi}x_i)}{1+\sum_{j=2}^{c} \exp(\alpha_j+\sum_{i=1}^{k} \beta_{ji}x_i)}$$

for q = 2, 3, …, c.

- Here the probabilities for all *c* outcomes still sum to 1.

Multinomial Logistic Regression

- For each covariate or risk factor x there are $c - 1$ parameters to be estimated (c = outcome group), one for each category of the outcome other than the reference group

- For k covariates, there are $k \times (c - 1)$ regression parameters, plus the $c - 1$ intercept terms

- And as was noted earlier in the single covariate case, when $c = 2$, multinomial logistic regression in the general case reduces to (ordinary/binary) logistic regression with k regression parameters plus 1 intercept term

Multinomial Logistic Regression

- Here, we can generalize the previous results to say that the *relative risk ratio* of being in outcome category *q* versus the reference category 1 for a one unit increase in covariate x_i, i=1,…,k is given by

$$\frac{\boldsymbol{exp[\beta_{qi}(x_i+1)]}}{\boldsymbol{exp(\beta_{qi}x_i)}} = \boldsymbol{exp(\beta_{qi})}$$

adjusting for the remaining covariates included in the model.

Multinomial Logistic Regression

- Probabilities of the outcome(s) can be estimated for subjects with specified combinations of risk factors

- Within each covariate pattern, the estimated probabilities of observing a 1, 2, …, c will sum to 1

- The probabilities themselves differ depending on the risk factors

Multinomial Regression Model Building

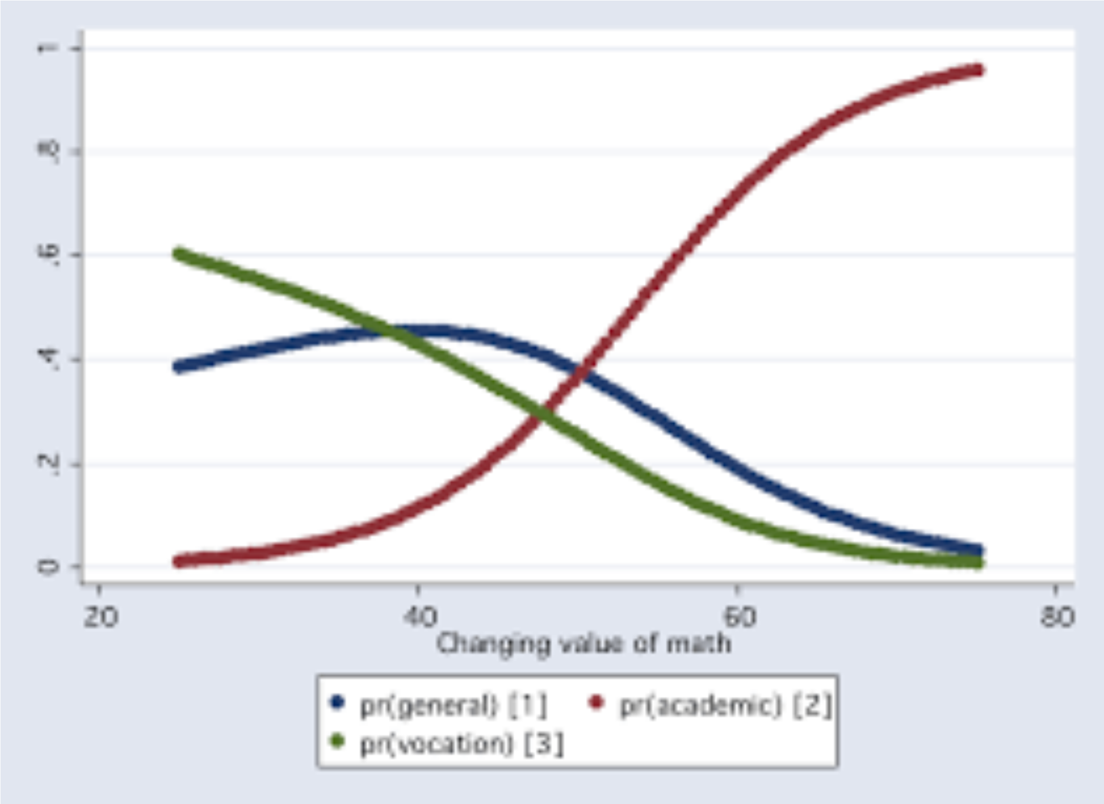
- Once you have a handle on the interpretation of multinomial logistic regression models, assessment of confounding and effect modification as well as model building techniques follow similarly to (ordinary) logistic regression.

- Parameter estimates are found using maximum likelihood methods, and also standard error estimates,

- P-values, and confidence intervals can be obtained.

- Likelihood ratio tests will be especially useful (when testing multiple parameters at a time).

Multinomial Regression Model Building



Example: Hyponatremia

- Data not collected on many (488 out of 14k+ runners); no elite runners
- *Self reported*: gender, run time, body mass index, weight gain, water consumption, urination freq, and more
- Useful to have clinically meaningful cutoff for sodium variable (≤135 mmol/l)

Ordinal Logistic Regression

- Multinomial logistic regression makes no assumptions about the possible ordering of the categories of the outcome variable – the outcome is treated as a nominal variable (and it doesn't even matter which group is chosen to be reference)

- In the hyponatremia example, the outcome is in fact ordinal

- Here we could use *ordinal logistic regression*

Ordinal Logistic Regression

- Suppose that an ordinal outcome *Y* has *c* ordered categories (lowest to highest) labeled as **j** = 1, 2, ... *c*

- The *proportional odds ordinal logistic regression model* is based on the *probability of success* P(*Y* ≥ *j*) *(or cumulative logit)*:

$$\mathbf{logit[P(}Y \geq j\mathbf{)] = log} \frac{\mathbf{P(}Y \geq j\mathbf{)}}{\mathbf{P(}Y < j\mathbf{)}}$$

$$= \boldsymbol{\alpha_j + \beta_1 x_1 + \ \dots + \beta_p x_p}$$

where j = 2, ... C

- The β coefficients do not differ across outcome categories (but the intercept terms do) – much easier to interpret than multinomial!

- And…we once again have an *odds ratio*!

Ordinal Logistic Regression

- Suppose we have two subjects A and B who differ by 1 unit for the variable x_i, with all other covariates taking the same value

- Then $\exp(\beta_i)$ is just the *odds ratio* that $Y \geq j$ versus $Y < j$ for subject A versus subject B

- $\exp(\beta_i)$ is assumed to be the *same* for each possible value of j

- Estimated probabilities do sum to 1 here

Ordinal Logistic Regression

- This is called the *proportional odds* assumption

- If this assumption is valid, there are many fewer parameters to be estimated than in the multinomial logistic regression model

- These betas are also seeming easier to interpret (OR!)

- If c = 2, again the model reduces to (ordinary/binary) logistic regression

Ordinal versus Multinomial

- The ordinal logistic regression model has more assumptions that the multinomial model (ie proportional odds), which may be hard to satisfy (but can be tested)

- Different software packages may have different tests (often approximate tests) of the proportional odds assumption

- This assumption is worth checking, because if the ordinal model is appropriate, model interpretation is simpler (only 1 set of beta coefficients!)

→ *Think Parsimony*

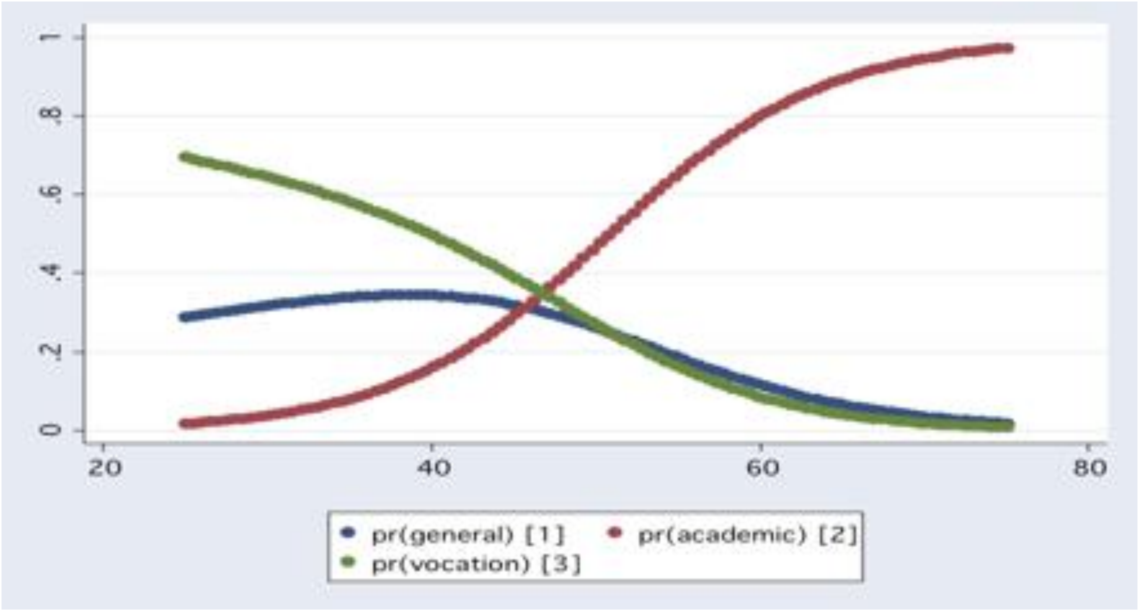
Ordinal versus Multinomial

- Because it has fewer assumptions, the multinomial logistic regression model has more parameters that need to be estimated (so can be more difficult to interpret)
- Because there are so many parameters, it is possible to end up with an over-fitted (not generalizable) multinomial model
- Looking at examples helps!

Note

- Unfortunately, different books and software packages use different notation, and you have to be very careful on how you are interpreting the β coefficients, intercepts, fitted probabilities, and relative risk ratio (for the multinomial model) or the odds ratio (assuming the proportional odds assumption for the ordinal model)…

Recall: Multinomial logistic regression



Ordinal (proportional odds) logistic regression