Intro to GLM – Day 4: Multiple Choices and Ordered Outcomes

Federico Vegetti
Central European University

ECPR Summer School in Methods and Techniques
11 August 2016

Categorical events with more than two outcomes

In social science, many phenomena do not consist of simple yes/no alternatives

1. **Categorical** variables
   - Example: multiple choices
   - A voter in a multiparty system can choose between many political parties
   - A consumer in a supermarket can choose between several brands of toothpaste
2. **Ordinal** variables
   - Survey questions often ask "how much do you agree" with a certain statement
   - You may have 2 options: "agree" or "disagree"
   - You may have more options: e.g. "completely agree", "somewhat agree", "somewhat disagree", "completely disagree"

Categorical dependent variables

- Imagine a country where voters can choose between 3 parties: "A", "B", "C"
- We want to study whether a set of individual attributes affect vote choice
- In theory, we could run several binary logistic regressions predicting the probability to choose between any two parties
- If we have three categories, how many binary regressions do we need to run?

Multiple binary models?

- We need to run only 2 regressions:

$$log\left[\frac{P(A|X)}{P(B|X)}\right] = \beta_{A|B}X; \quad log\left[\frac{P(B|X)}{P(C|X)}\right] = \beta_{B|C}X$$

- Estimating also $log\left[\frac{P(A|X)}{P(C|X)}\right]$ would be redundant:

$$log\left[\frac{P(A|X)}{P(B|X)}\right] + log\left[\frac{P(B|X)}{P(C|X)}\right] = log\left[\frac{P(A|X)}{P(C|X)}\right]$$

- And:

$$\beta_{A|B}X + \beta_{B|C}X = \beta_{A|C}X$$

Multiple binary models? (2)

- However, if we estimated all binary models independently, we would find out that $\beta_{A|B}X + \beta_{B|C}X \neq \beta_{A|C}X$
- Why? Because **the samples would be different**
- The model for *log* $\left[\frac{P(A|X)}{P(B|X)}\right]$ would would include only people who voted for "A" *or* "B"
- The model for *log* $\left[\frac{P(B|X)}{P(C|X)}\right]$ would would include only people who voted for "B" *or* "C"
- We want a model that uses the full sample and estimates the two groups of coefficients simultaneously

Multinomial probability model

- To make sure that the probabilities sum up to *1*, we need to take all alternatives into account in the same probability model
- As a result, the probability that a voter i picks a party m among a set of J parties is:

$$P(Y_i = m|X_i) = \frac{exp(X_i\beta_m)}{\sum_{j=1}^{J} exp(X_i\beta_j)}$$

- **Note**: to make sure the model is identified, we need to set $\beta = 0$ for a given category, called the "baseline category"
- Conceptually, this is the same as running only 2 binary logit models when there are 3 categories

Multinomial probability model (2)

- We can still obtain predicted probabilities for each category
- Assuming that the baseline category is *1*, the probability of $Y = 1$ is:

$$P(Y_i = 1|X_i) = \frac{1}{1 + \sum_{j=2}^{J} exp(X_i\beta_j)}$$

- And the probability of $Y = m$, where m refers to any other category, is:

$$P(Y_i = m|X_i) = \frac{exp(X_i\beta_m)}{1 + \sum_{j=2}^{J} exp(X_i\beta_j)} \text{ for } m > 1$$

- The choice of the baseline category is arbitrary
- However, it makes sense to pick a theoretically meaningful one

Estimation of multinomial logit models

- The likelihood function for the multinomial logit model is:

$$L(\beta_2, \ldots, \beta_j | y, X) = \prod_{m=1}^{J} \prod_{y_j=m} \frac{exp(X_i\beta_m)}{\sum_{j=1}^{J} exp(X_i\beta_j)}$$

- Where $\prod_{y_j=m}$ is the product over the cases where $y_i = m$
- The estimation will work as usual: the software will take the log-likelihood function and it will look for the ML estimates of β iteratively
- For every independent variable, the model will produce $J-1$ parameter estimates

Multinomial logit: interpretation

- Like in binary logit, our coefficients are log-odds to choose category m instead of the baseline category

$$exp(X_i \beta_m) = \frac{\pi_m}{\pi_1}$$

- How do we compare the coefficients between categories that are not the baseline?
- First, again, pick a baseline category that makes sense
- Second, comparing coefficients between estimated categories is straightforward:

$$\frac{\pi_m}{\pi_j} = exp[X_i(\beta_m - \beta_j)]$$

- I.e. the exponentiated difference between the coefficients of two estimated categories is equivalent to the odds to end up in one category instead of the other (given a set of individual characteristics)

Multinomial logit: predicted probabilities

- Predicted probabilities to choose any of the estimated categories are:

$$\pi_{im} = \frac{exp(X_i \beta_m)}{1 + \sum_{j=2}^{J} exp(X_i \beta_j)}$$

- And for the baseline category they are:

$$\pi_{i1} = \frac{1}{1 + \sum_{j=2}^{J} exp(X_i \beta_j)}$$

Multinomial models as choice models

- A way to interpret multinomial models is, more directly, as *choice* models
- This approach is sometimes called "Random Utility Model" and it is quite popular in economics
- This interpretatons is based on two assumptions:
  - *Utility* varies across individuals. Different individuals have different utilities for different options
  - Individual decision makers are *utility maximizers*: they will choose the alternative that yields the highest utility
- Utility: the degree of satisfaction that a person expects from choosing a certain option
- The utility is made of a systematic component μ and a stochastic component *e*

Utility and multiple choice

- For an individual i, the (random) utility for the option m is:

$$U_{im} = \mu_{im} + e_{im} = X\beta_{im} + e_{im}$$

- When there are J options, m is chosen over an alternative $j \neq m$ if $U_{im} > U_{ij}$

$$P(Y_i = m) = P(U_{im} > U_{ij})$$
$$P(Y_i = m) = P(\mu_{im} - \mu_{ij} > e_{ij} - e_{im})$$

- The likelihood function and estimation are identical to the probability model that we just saw

Assumptions

1. The stochastic component follows a Gumbel distribution (AKA "Type I extreme-value distribution")

$$F(e) = exp[-e - exp(-e)]$$

2. Among different alternatives, the errors are identically distributed
3. Among different alternatives, the errors are independent
    - This assumptions is called "independence of the irrelevant alternatives", and it is quite controversial
    - It states that the ratio of choice probabilities for two different alternatives is independent from all the other alternatives
    - In other words, if you are choosing between party "A" and party "B", the presence of party "C" is irrelevant

Conditional logit

- In multinomial logit models, we explain choice beween different alternatives using attributes of the decision-maker
- E.g. education, gender, employment status
- However, it is possible to explain choice using attributes of the alternatives themselves
- E.g. are voters more likely to vote for bigger parties?
- The latter model is called "conditional logit"
- It is not so common in political science, as it requires observing variables that vary between the choice options

Multinomial vs Conditional logit

Multinomial logit

- We keep the values of the predictors constant across alternatives
- We let the parameters vary across alternatives
  - E.g. the gender of a voter is always the same, no matter if s/he's evaluating party "A" or party "B"
  - The effect of gender will be different between party "A" and "B"

Conditional logit

- We let the values of the predictors change across alternatives
- We keep the parameters constant across alternatives
  - The size of party "A" and party "B" is the same for all individuals
  - The effect of size is the same for all parties

Ordinal dependent variables

- Suppose the categories have a natural order
- For instance, look at this item in the World Values Study:
- "*Using violence to pursue political goals is never justified*"
    - Strongly Disagree
    - Disagree
    - Agree
    - Strongly Agree
- Here we can rank the values, but we don't know the distance between them
- We could use a multinomial model, but this way we would ignore the order, losing information

Modeling ordinal outcomes

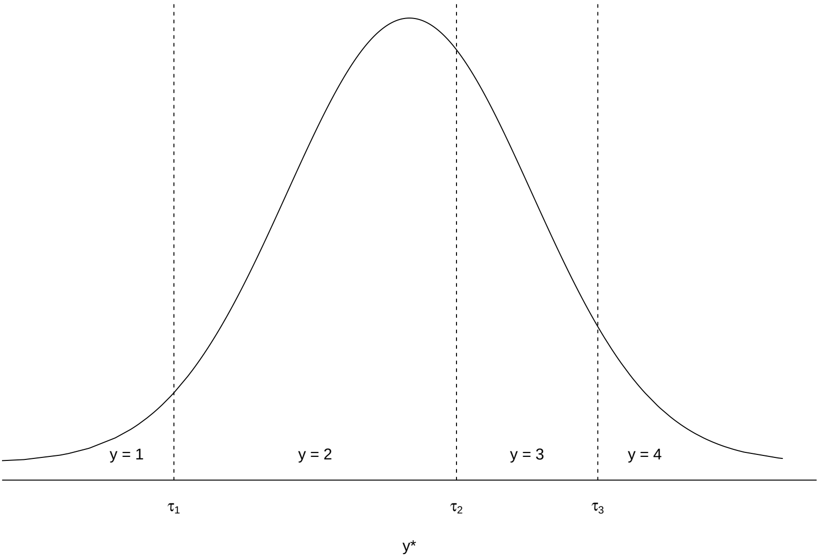
- Two ways of modeling ordered categorical variables:
  - A latent variable model
  - A non-linear probability model
- These two methods reflect what we have seen with binary response models
- In fact, you can think of binary models as special cases of ordered models with only 2 categories
- As with binary models, the estimation will be the same
- However, for ordered models, the latent variable specification is somewhat more common

A latent variable model

- Imagine we have an unobservable latent variable $y*$ that expresses our construct of interest (e.g. endorsement of political violence)
- However, all we can observe is the ordinal variable y with M categories
- $y*$ is mapped into y through a set of cut points τ_m

$$y_i = \begin{cases} 1 & \text{if } -\infty < y_i* < \tau_1 \\ 2 & \text{if } \tau_1 < y_i* < \tau_2 \\ 3 & \text{if } \tau_2 < y_i* < \tau_3 \\ 4 & \text{if } \tau_3 < y_i* < +\infty \end{cases}$$

Cut points



A latent variable model (2)

- Like with the binary model, $y*$ is a function of both a systematic and a stochastic component

$$y_i* = X_i\beta + e_i$$

- Then, the model is essentially a linear regression of $y*$
- To be able to estimate the model we need to:
  - Fix the variance of e to an assumed value
    - Either *1* (then e is normally distributed)
    - Or $\pi^2/3$ (then e il logistically distributed)
  - Exclude the constant term from the estimation of the parameters
    - Instead, estimated values of τ_1, τ_2, ..., τ_{M-1} serve as intercepts
    - Where M is the number of categories

A non-linear probability model

- Ordinal models can be also seen as models of the **cumulative probability** that an outcome y is less than or equal to m
- So, instead of modeling the probability that a certain event happens (like in binary models), here we model the probability of an event *and of all events that are ordered before it*:

$$P(y_i \leq m|X_i) = \sum_{j=1}^{m} P(y_i = j|X_i)$$

- In terms of odds, it is the odds that $y \leq m$ vs $y > m$:

$$\Omega_{im}(X_i) = \frac{P(y_i \leq m|X_i)}{1 - P(y_i \leq m|X_i)} = \frac{P(y_i \leq m|X_i)}{P(y_i > m|X_i)}$$

Probability model

- The cumulative probability to observe an outcome of $y \leq m$ is:

$$P(y_i \leq m|X_i) = F(\tau_m - X_i\beta)$$

- And the probability to observe an outcome of $y = m$ Is:

$$P(y_i = m|X_i) = F(\tau_m - X_i\beta) - F(\tau_{m-1} - X_i\beta)$$

- Where *F()* is either the standard normal or logistic CDF
- Again, the choice of the link function determines whether we estimate an *ordered logit* or an *ordered probit* model

Estimation of ordered models

- The likelihood function for ordered models is:

$$L(\beta, \tau | y, X) = \prod_{j=1}^{J} \prod_{y_i = m} [F(\tau_m - X_i\beta) - F(\tau_{m-1} - X_i\beta)]$$

- Where $\prod_{y_i=m}$ indicates to multiply over the cases where $y = m$
- As usual, the software will plug in the link function, take the log-likelihood function and look for the ML estimates of β and τ

Proportional odds assumption

- In the probability function that we have seen, β is the same regardless which categories we are considering, while τ is different
- This is equivalent to estimate a number of parallel regression lines, where only the intercept changes
- For instance, if y has 4 categories:

$$P(y_i \leq 1|X_i) = F(\tau_1 - X_i\beta)$$

$$P(y_i \leq 2|X_i) = F(\tau_2 - X_i\beta)$$

$$P(y_i \leq 3|X_i) = F(\tau_3 - X_i\beta)$$

- In logit models this is called the "proportional odds assumption"
- It can be tested comparing the β obtained by an ordered regression with a set of βs obtained by a set of binary regressions for each $P(y_i \leq m|X_i)$

Ordered logit: interpretation

- Unlike the multinomial logistic model, we have only one set of βs here
- This is due to the "proportional odds" assumption, which implies that our βs are the same for each cut point τ_m
- As we are accustomed to think, the coefficients are log-odds to choose category m instead of a lower category

$$exp(X_i \beta_m) = \frac{\pi_m}{\pi_{m-1}}$$

- Also the values of τ are on the same scale: they indicate the log-odds to be in a category below the cut point when all predictors are equal to zero

Ordered logit: interpretation (2)

- In ordered models, we can predict two types of probabilities:
  - The *cumulative* probability, i.e. the probability that y will be in the category m or in a lower ranked category
  - The probability that y is in a specific category
- If we use the standard logistic CDF as link function, the formula to get cumulative predicted probabilities is:

$$P(y_i \leq m|X_i) = \frac{exp(\tau_m - X_i\beta)}{1 + exp(\tau_m - X_i\beta)}$$

Ordered logit: interpretation (3)

- To get predicted probabilities for specific categories, we must still take the cumulative probability and subtract the predicted probability for the lower ranked category:

$$P(y_i = m) = \frac{exp(\tau_m - X_i\beta)}{1 + exp(\tau_m - X_i\beta)} - \frac{exp(\tau_{m-1} - X_i\beta)}{1 + exp(\tau_{m-1} - X_i\beta)}$$

- Note that the larger the difference between τ_m and τ_{m-1}, the easier it will be to answer $y_i = m$.
- This is the case in some survey items where many people choose the middle category