

Applied Statistical Modelling: GLMs – ordered logit regression– multinomial regression

Prof.ssa Chiara Seghieri

Laboratorio di Management e Sanità, Istituto di Management, L'EMbeDS

Scuola Superiore Sant'Anna, Pisa

c.seghieri@santannapisa.it

c.tortu@santannapisa.it

Introduction

When a dependent variable is categorical ordinal, with > 2 categories (i.e. being satisfied with a service, likert scale ranging from 1 to 5 where each value is indeed 'higher' than the previous one):

- The categories of the dependent variable have a natural ordering
- In this situation you can use an ordered logistic/probit regression which does take the ordering into account

When the dependent variable is measured on an ordinal scale there can be some options:

- ✓ Treating the variable as continuous, this is widely done when the categories are more than 5 and a large sample size. Other approaches can be used to confirm OLS.
- ✓ Collapsing the categories into 2 but this might increase the probability of making type 1 error, better to use all the information.
- ✓ Using ordered logit/probit models: the approach is equivalent: we simply use for the ordered probit the normal CDF $\Phi()$ and for the ordered logit the logistic CDF.

ORDERED LOGIT MODEL

Let Y_i be an ordinal response variable with C categories for the i -th subject, alongside with a vector of covariates x_i . A regression model establishes a relationship between the covariates and the set of probabilities of the categories $p_{ci} = \Pr(Y_i = y_c | x_i)$, $c=1, \dots, C$.

Usually, regression models for ordinal responses are not expressed in terms of probabilities of the categories, but they refer to convenient one-to-one transformations, such as the cumulative probabilities $g_{ci} = \Pr(Y_i \leq y_c | x_i)$, $c=1, \dots, C$. Note that the last cumulative probability is necessarily equal to 1, so the model specifies only $C-1$ cumulative probabilities.

An ordered logit model for an ordinal response Y_i with C categories is defined by a set of $C-1$ equations where the cumulative probabilities $g_{ci} = \Pr(Y_i \leq y_c | x_i)$ are related to a linear predictor $(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots)$ through the logit function:

$$\text{logit}(g_{ci}) = \log\left(\frac{g_{ci}}{1 - g_{ci}}\right) = \alpha_c + \beta' x_i \quad c=1, 2, \dots, C-1$$

Similar to the binary logit but now we have the cumulative logit

$$\text{logit}(g_{ci}) = \log\left(\frac{g_{ci}}{1 - g_{ci}}\right) = \alpha_c + \beta'x_i \quad c=1,2,\dots,C-1$$

the logit of each cumulative probability is assumed to be a linear function of the covariates with regression coefficients constant across response categories. C-1 equations.

The parameters α_c are called *thresholds* or *cutpoints*. In this model, intercept α_c is the log-odds of falling into or below category c when $x_1=x_2=\dots=x_p=0$.

Intercepts can differ, but that slope for each variable stays the same across different equations thus the effects of the covariates are constant across response categories! One may think of this as a set of parallel lines (or hyperplanes) with different intercepts. The proportional-odds condition forces the lines corresponding to each cumulative logit to be parallel. This feature is called the *parallel regression assumption*. This model is called proportional-odds cumulative-logit model.

Cumulative OR can also be defined similarly to the binary case and interpreted as the odds increase from one category on the response to the next for each increase of X, the odds are considered to be equal across all response categories.

Brant test is used to test the parallel odds assumption of the ordinal logit model.

Generalized ordinal models relax the proportional odds assumption.

Another interpretation: Continuous Latent Response

The proportional-odds cumulative-logit model is connected to the idea of a continuous latent response. Suppose that the categorical outcome is actually a categorized version of an unobservable (latent) continuous variable.

For example, it is reasonable to think that a 5-point Likert scale (1 = strongly disagree, 2 = agree, 3 = neutral, 4 = agree, 5 = strongly agree) is a coarsened version of a continuous variable Y^* indicating degree of approval.

The continuous scale is divided into five regions by four cut-points c_1, c_2, c_3, c_4 which are determined by nature (not by the investigator). If $Y^* \leq c_1$, we observe $Y=1$; if $c_1 < Y^* \leq c_2$, we observe $Y=2$,... and so on.

Example:

Response variable Y: level of agreement with a statement: strongly disagree (1), disagree (2), agree (3), and strongly agree (4)

Explanatory variables:

- yr89: a dummy variable for survey year 1989 (1=1989, 0=1988)
- white: a dummy variable for ethnic group white (1=white, 0=non-white)
- age in years
- ed: years of education
- male: a dummy variable for men (1=Male, 0=Female)

Simultaneously estimates multiple equations. The number of equations it estimates will be the number of categories in the dependent variable minus one. So, for our example, three equations will be estimated.

$$\text{logit}(P(Y \leq 1)) = \log\left(\frac{P(Y \leq 1)}{1 - P(Y \leq 1)}\right) = \alpha_1 + \beta X \quad \text{Strongly disagree VS disagree+agree+ strongly agree}$$

$$\text{logit}(P(Y \leq 2)) = \log\left(\frac{P(Y \leq 2)}{1 - P(Y \leq 2)}\right) = \alpha_2 + \beta X \quad \text{Strongly disagree + disagree VS agree +strongly agree}$$

$$\text{logit}(P(Y \leq 3)) = \log\left(\frac{P(Y \leq 3)}{1 - P(Y \leq 3)}\right) = \alpha_3 + \beta X \quad \text{Strongly disagree + disagree + agree VS strongly agree}$$

Ordered logistic regression

Number of obs = 22
LR chi2(5) = 298.
Prob > chi2 = 0.00
Pseudo R2 = 0.04

Log likelihood = -2846.6132

	warm	Coef.	Std. Err.	z	P> z	[95% Conf. Interva	
	yr89	.5282808	.0798763	6.61	0.000	.3717262	.68483
	white	-.3795009	.1182501	-3.21	0.001	-.6112669	-.14773
	age	-.0207738	.0024195	-8.59	0.000	-.0255159	-.01603
	ed	.0839738	.0131433	6.39	0.000	.0582135	.10973
	male	-.7269441	.0783997	-9.27	0.000	-.8806048	-.57328
	/cut1	-2.443735	.2386412			-2.911463	-1.9760
	/cut2	-.6096001	.2331233			-1.066513	-.15268
	/cut3	1.279352	.2338585			.8209981	1.7377

$\exp(b_{\text{male}}) = 0.48$: the odds of rating 2, 3 or 4 are 52% lower for M vs F and the odds of being rated 3 or 4 (versus 1 or 2) are 52% lower for M vs F etc

$\exp(b_{\text{ed}}) = 1.088$: Controlling for the other explanatory variables, 1 additional year of education is associated with a 8.8% increase in odds of giving a response that indicates higher levels of agreement with the statement

Predicted probabilities after ordinal logit

```
margins, predict(outcome(1)) atmeans post
margins, predict(outcome(2)) atmeans post
margins, predict(outcome(3)) atmeans post
margins, predict(outcome(4)) atmeans post
```

Outcome 1: Pr (opinion==1)							Outcome 2: Pr (opinion==2)						
Model		VCE		OIM			Model		VCE		OIM		
Adjusted predictions				Number of obs = 70			Adjusted predictions				Number of obs = 70		
Expression : Pr (opinion==1), predict (outcome1)							Expression : Pr (opinion==2), predict (outcome2)						
at : x1 = .6480006 (mean)							at : x1 = .6480006 (mean)						
x2 = .1338694 (mean)							x2 = .1338694 (mean)						
x3 = .761851 (mean)							x3 = .761851 (mean)						
		Delta-method							Delta-method				
		Margin	Std. Err.	z	P> z	[95% Conf. Interval]			Margin	Std. Err.	z	P> z	[95% Conf. Interval]
_cons		.2800935	.0541271	5.17	0.000	.1740064 .3861805			.219505	.0502736	4.37	0.000	.1209706 .3180394

The probability of opinion = 1 given that the rest of the variables are at their mean values is 28%

The multinomial logit

When the dependent variable is categorical, with > 2 categories which don't have a natural order.

As example: Question: For which party did you vote in 2010 elections? the possible answers are: Conservatives, Labour, and the Liberal Democrats).

- Cause of death (e.g. different types of cancer)

In these situations you can run a multinomial regression.

Multinomial logistic regression: the basics

Suppose the response variable is categorical and can take values 1, 2,..., K such that $K > 2$

The Multinomial Distribution:

$$P(Y=1)=\pi_1, P(Y=2)=\pi_2, \dots, P(Y=K)=\pi_k$$

$$\sum_{k=1}^K \pi_k = 1$$

If we have an explanatory variable x , then we want to fit a model such that $P(Y=K)=\pi_k$ is a function of x . We choose a baseline category (say $Y=1$) and we have a separate equation for each category of the response relative to the baseline category. If the response has K possible categories, there will be $K-1$ equations as part of the multinomial logistic model.

$$\log\left(\frac{\pi_{ik}}{\pi_{i1}}\right) = \beta_{0k} + \beta_{1k}x_i$$

Suppose we have a response variable that can take three possible outcomes that are coded as "A", "B", "C».

Let "A" be the baseline category. Then:

$$\log\left(\frac{\pi_{iB}}{\pi_{iA}}\right) = \beta_{0B} + \beta_{1B}x_i$$

$$\log\left(\frac{\pi_{iC}}{\pi_{iA}}\right) = \beta_{0C} + \beta_{1C}x_i$$

Multinomial logistic regression: interpretation

- **Interpretation** of any log-odds model depends on which response corresponds to the numerator and the denominator
- The category corresponding to the **denominator** is called the **reference** or **baseline category** and is usually:
 - the first or last category
 - the most ‘meaningful’ category
 - the most frequent category
 - not a ‘rare’ category

Example: the association between economic activity and gender

- **Outcome** economic activity has three categories
 1. Economically inactive
 2. Unemployed
 3. In employment
- Our **explanatory variable** has two categories
 1. Man
 2. Woman

Distribution of the variables

Employment status	N	%
Economically inactive	442	39.32
Unemployed	55	4.89
In employment	627	55.78
Total	1,124	100.0

Gender	N	%
Man	516	45.91
Woman	608	54.09
Total	1,124	100.0

•Data source: University of Manchester. Cathie Marsh Centre for Census and Survey Research. ESDS Government, *ONS Opinions Survey, Well-Being Module, April 2011: Unrestricted Access Teaching Dataset* [computer file]. 2nd Edition. Office for National Statistics. Social Survey Division, [original data producer(s)]. Colchester, Essex: UK Data Archive [distributor], October 2012. SN:

•Example taken from Dr Heini Väisänen University of Southampton

Interpretation of the parameters

- *Model:* $\log(\pi_1/\pi_3) = \alpha_1 + \beta_1 X$ $\log(\pi_2/\pi_3) = \alpha_2 + \beta_2 X$
- β_1 is the effect of X on the log-odds of being in category **1** instead of category **3**, i.e. ‘binary’ outcome *either 1 or 3*.
- β_2 is the effect of X on the log-odds of being in category **2** instead of category **3**, i.e. ‘binary’ outcome *either 2 or 3*.
- Odds ratios can be obtained by calculating $\exp(\beta_1)$ and $\exp(\beta_2)$
- The baseline is category **3** (in employment – the most frequent category).

Economic activity and gender, results

```

Multinomial logistic regression      Number of obs   =      1,124
                                   LR chi2(2)          =      26.42
                                   Prob > chi2          =      0.0000
Log likelihood = -931.25955          Pseudo R2       =      0.0140
    
```

DVIL03a	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
In_Employment	(base outcome)					
ILO_Unemployed						
rsex						
Female	-.5372853	.2914648	-1.84	0.065	-1.108546	.0339752
_cons	-2.203553	.1781179	-12.37	0.000	-2.552658	-1.854449
Economically_Inactive						
rsex						
Female	.5500842	.1267875	4.34	0.000	.3015853	.7985831
_cons	-.6590353	.096188	-6.85	0.000	-.8475604	-.4705103

sig = whether parameter is significant in that pairwise regression (different from whether it is significant in the overall model)

Economic activity and gender, results (odds)

```

Multinomial logistic regression      Number of obs   =    1,124
                                   LR chi2(2)          =    26.42
                                   Prob > chi2         =    0.0000
Log likelihood = -931.25955          Pseudo R2       =    0.0140
  
```

DVIL03a	RRR	Std. Err.	z	P> z	[95% Conf. Interval]	
In_Employment	(base outcome)					
ILO_Unemployed						
rsex						
Female	.5843324	.1703123	-1.84	0.065	.3300385	1.034559
_cons	.1104101	.019666	-12.37	0.000	.0778744	.1565392
Economically_Inactive						
rsex						
Female	1.733399	.2197733	4.34	0.000	1.352	2.22239
_cons	.5173502	.0497629	-6.85	0.000	.4284589	.6246834

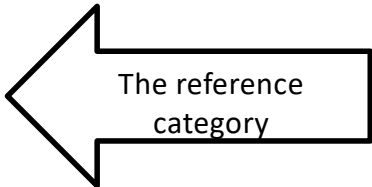
Economic activity and gender, interpretation (odds)

- The odds of being economically inactive rather than in employment are 73% higher for women than for men.
 - $(\exp(0.550)-1)*100 = 73.3\%$
- The odds of being unemployed rather than in employment are 42% lower for women than for men.
 - $(\exp(-0.537)-1)*100 = -41.6\%$

Economic activity and gender, interpretation (probabilities)

$$\pi_1 = \frac{\exp(\alpha_1 + \beta_1 X)}{1 + \exp(\alpha_1 + \beta_1 X) + \exp(\alpha_2 + \beta_2 X)}$$

$$\pi_2 = \frac{\exp(\alpha_2 + \beta_2 X)}{1 + \exp(\alpha_1 + \beta_1 X) + \exp(\alpha_2 + \beta_2 X)}$$

$$\pi_3 = 1 - (\pi_1 + \pi_2)$$


Economic activity and gender, interpretation (probabilities for women)

Economically inactive

$$\log(\pi_1/\pi_3) = -0.659 + 0.550X$$

Unemployed

$$\log(\pi_2/\pi_3) = -2.204 - 0.537X$$

$$\pi_1 = \frac{\exp(-0.659 + 0.550X)}{1 + \exp(-0.659 + 0.550X) + \exp(-2.204 + 0.537X)} = 0.46$$

$$\pi_2 = \frac{\exp(-2.204 + 0.537X)}{1 + \exp(-0.659 + 0.550X) + \exp(-2.204 + 0.537X)} = 0.03$$

$$\pi_3 = 0.51$$

Economic activity and gender, interpretation (probabilities)

- Among women, the probability of being economically inactive was 46%, in employment 51% and unemployed 3%.
- Among men, the probability of being economically inactive was 32%, in employment 61% and unemployed 7%.