Categorical Data Analysis

Hsueh-Sheng Wu

Center for Family and Demographic Research

November 7, 2011

Outline

- Why do we need to learn categorical data analyses?
- A summary of different categorical data analyses
  - Analyses of Contingency tables
  - Binary dependent variable
    - Logistic regression
  - Ordinal dependent variable
    - Ordered logistic regression
  - Nominal dependent variable
    - Multinomial logistic regression
    - Conditional logistic regression
    - Nested logistic regression
- Conclusion

Why Do We Need to Learn Categorical Data Analysis?

- Four measurement levels
  - Nominal (e.g., gender, race)
  - Ordinal (e.g., attitude toward cohabitation)
  - Interval (e.g., temperature)
  - Ratio (e.g., income)
- Categorical variables are those measured at nominal and ordinal levels
- Interval or ratio variables can be transformed into nominal or ordinal variables, but not the other way around.

What Is Special about Categorical Variable?

- The distribution of a categorical variable is described by its frequency and proportion rather than by its mean and variance.
- Statistical methods (i.e., t-test, correlation, OLS regression) designed for continuous dependent variables are not adequate for analyzing categorical dependent variables.
- The decision on how to analyze categorical variables is often based on:
  - The measurement level and number of categories in dependent variables
  - The measurement level and number of categories in independent variables
  - Sample size
  - Number of independent variables

When Do We Need Categorical Data Analysis?

- You have a categorical variable as the dependent variable.
- You have a continuous variable. However, the distribution of this variable is skewed and cannot be analyzed like regular continuous dependent variables.

Different Models for Categorical Dependent Variables

Categorical models address three types of questions:

- Examination of contingency tables
  - Proportions
  - Relative risks
  - Odds ratio
- How the characteristics of individuals affect the choice
  - Binary logistic regression
  - Ordered logistic regression
  - Multinomial logistic regression
- How the characteristics of alternatives affect the choice
  - Conditional logit regression
  - Nested logit regression

Analyzing a Two-way Contingency Table

- Analyzing a 2x2 table

$$\text{Difference of Two Proportions} = \pi_1 - \pi_2 \approx \rho_1 - \rho_2$$

$$SE = \sqrt{\frac{\rho_1(1-\rho_1)}{n_1} + \frac{\rho_2(1-\rho_2)}{n_2}}$$

$$\text{Relative Risk} = \frac{\pi_1}{\pi_2}$$

Analyzing a Two-way Contingency Table (Cont.)

- Odds Ratio

$$\text{Odds Ratio} = \frac{\text{Odds1}}{\text{Odds2}}$$

$$= \frac{\pi 1 / (1-\pi 1)}{\pi 2 / (1-\pi 2)} = \frac{\pi 11 / \lambda 12}{\pi 21 / \pi 22} = \frac{\pi 11 \cdot \pi 22}{\pi 12 \cdot \pi 21}$$

$$SE = \sqrt{\frac{1}{n11} + \frac{1}{n12} + \frac{1}{n21} + \frac{1}{n22}}$$

Example

- Data

Table. Gender and Employment		
	Employed	Unemployed
Male	200	200
Female	200	400

  - Difference of two proportions

    P1 = 200/400 = 0.5

    P2 = 200/600 = 0.33

    P1 - P2 = 0.17

  - Relative risk

    P1/P2 = 1.51

  - Odds Ratio

    (200*400)/(200*200) = 2

BGSU Center for Family and Demographic Research

Analyzing a Three-way Contingency Table

- A three-way contingency table can be viewed as multiple two-way contingency tables created at different levels of a third variable.

- Example:

Table. Relations among Country, Gender, and Employment

	County A		Country B	
	Employed	Unemploye	Employed	Unemployed
Male	180	120	20	80
Female	120	80	80	320

Example

– Difference of proportion

Country A: (180/300) – (120/200)=0

Country B: (20/100) –(80/320)=0

– Relative risk

Country A: (180/300)/(120/200)=0.6/0.6=1

Country B: (20/100) –(80/320)=0.2/0.2=1

– Odds Ratio

Country A: (180*80)/(120*120)=1

Country B: (20*320)*(80*80)=1

Models for Examining How Characteristics of Individuals Affect Choices

Logistic Regression

$$\log(\frac{\pi_1}{\pi_2}) = \log(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta\chi$$

$$\pi(\chi) = \frac{\exp(\alpha+\beta\chi)}{1+\exp(\alpha+\beta\chi)} = \frac{e^{\alpha+\beta\chi}}{1+e^{\alpha+\beta\chi}}$$

Ordered Logistic Regression

$$p(Y \le j) = \pi_1 + \ldots + \pi_{j,}\ j = 1,\ldots, J$$

$$\text{logit}\,[p(Y \le j)] = \log[\frac{p(Y \le j)}{1-p(Y \le j)}] = \log[\frac{\pi_1 + \ldots + \pi_{j,}}{\pi_{j+1} + \ldots + \pi_{J,}}],\ j = 1,\ldots, J$$

Models for Examining How Characteristics of Individuals Affect Choices (Cont.)

Multinomial Logistic Regression

$$\log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j \chi,\ j = 1, \ldots, J-1$$

$$\log\left(\frac{\pi_a}{\pi_b}\right) = \log\left(\frac{\pi_a / \pi_J}{\pi_b / \pi_J}\right) = \log\left(\frac{\pi_a}{\pi_J}\right) - \log\left(\frac{\pi_b}{\pi_J}\right)$$

$$= (\alpha_a + \beta_a \chi) - (\alpha_b + \beta_b \chi)$$

$$= (\alpha_\alpha - \alpha_b) + (\beta_a - \beta_b)\chi$$

Relations among These Three Models

- Ordered logistic regression and multinomial logistic regression are an extension of logistic regression.

- Both ordered and multinomial logistic regression can be treated as models simultaneously estimating a series of logistic regression.

- Ordered logistic regression assumes different intercepts, but the same slope for different categories, while multinomial logistic regression assumes different intercept and slope parameters for different categories.

A list of variables in the Data

1. subject: ID number
2. male: 1 for men and 0 for women
3. income: household income
4. transport: 1=Air, 2=Train, 3=Bus, 4=Car
5. flight: 0=ground and 1 =flight
6. expensive: 1 = very cheap, 2 =cheap, 3 = expensive, and 4= very expensive
7. mode: 1=Air, 2=Train, 3=Bus, 4=Car
8. choice: 1 if the travel mode is chosen
9. time: terminal waiting time, 0 for car
10. cost: cost for taking this mode of transportation
11. air_inc: interaction of air flight and household income, air*income
12. air: 1 for choosing air flight and 0 for other transportations
13. train: 1 for choosing train and 0 for other transportations
14. bus: 1 for choosing bus and 0 for other transportations
15. car: 1 for choosing car and 0 for other transportations

Data for Logistic Regression, Ordered Logistic Regression, and Multinomial Logistic Regression

Table 1. Original data

	subject	male	income	mode	fly	expensive
1	1	1	35	4	0	1
2	2	1	30	4	0	1
3	3	1	40	4	0	2
4	4	1	70	4	0	1
5	5	1	45	4	0	2
6	6	0	20	2	0	1
7	7	0	45	1	1	4
8	8	0	12	4	0	4
9	9	1	40	4	0	1
10	10	1	70	4	0	1
11	11	0	15	4	0	1
12	12	1	35	4	0	1
13	13	1	50	4	0	1
14	14	1	40	4	0	1
15	15	0	26	4	0	4
16	16	0	26	2	0	1
17	17	0	26	2	0	1
18	18	1	6	2	0	1
19	19	0	20	2	0	1
20	20	0	72	2	0	2

Data for Conditional Logistic Regression and Nested Logistic Regression

Table 2. Original data

	subject	male	income	mode	fly	expensive
1	1	1	35	4	0	1
6	6	0	20	2	0	1
106	106	0	30	3	0	3
113	113	0	30	1	1	3

Data for Conditional Logistic Regression and Nested Logistic Regression

Table 3. data for conditional logistic regression and nested logistic regression

	subject	male	income	mode	air	train	bus	car	choice	time	cost	air_inc
1	1	1	35	1	1	0	0	0	0	69	70	35
2	1	1	35	2	0	1	0	0	0	34	71	0
3	1	1	35	3	0	0	1	0	0	35	70	0
4	1	1	35	4	0	0	0	1	1	0	30	0
21	6	0	20	1	1	0	0	0	0	69	70	20
22	6	0	20	2	0	1	0	0	1	40	57	0
23	6	0	20	3	0	0	1	0	0	35	58	0
24	6	0	20	4	0	0	0	1	0	0	43	0
421	106	0	30	1	1	0	0	0	0	69	123	30
422	106	0	30	2	0	1	0	0	0	34	195	0
423	106	0	30	3	0	0	1	0	1	30	114	0
424	106	0	30	4	0	0	0	1	0	0	138	0
449	113	0	30	1	1	0	0	0	1	60	120	30
450	113	0	30	2	0	1	0	0	0	44	84	0
451	113	0	30	3	0	0	1	0	0	53	85	0
452	113	0	30	4	0	0	0	1	0	0	55	0

Logistic Regression

- Stata command:

  logit fly male income

```
. logit fly male income

Iteration 0:   log likelihood = -123.75705
Iteration 1:   log likelihood = -112.27885
Iteration 2:   log likelihood = -111.74298
Iteration 3:   log likelihood = -111.74159
Iteration 4:   log likelihood = -111.74159

Logistic regression                             Number of obs   =        210
                                                LR chi2(2)      =      24.03
                                                Prob > chi2     =     0.0000
Log likelihood = -111.74159                     Pseudo R2       =     0.0971

------------------------------------------------------------------------------
         fly |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |  -1.344985   .3968674    -3.39   0.001    -2.122831   -.5671395
      income |   .0396637   .0102221     3.88   0.000     .0196287    .0596986
       _cons |  -1.934816   .3933682    -4.92   0.000    -2.705804   -1.163829
```

Logistic Regression (Cont.)

SAS command:

Proc Logistic Data = in.logit_ologit_mlogit;

Model fly = male income;

run;

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	248.866	229.248
SC	252.208	239.275
-2 Log L	246.866	223.248

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	23.618	2	<.0001
Score	20.391	2	<.0001
Wald	17.734	2	0.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.9264	0.393	24.0281	<.0001
male	1	1.3312	0.3974	11.2214	0.0008
income	1	-0.0394	0.0102	14.8891	0.0001

Logistic Regression (Cont.)

SAS command:

```
Proc Logistic Descending Data = in.logit_ologit_mlogit;
Model fly = male income;
Run;
```

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	248.866	229.248
SC	252.208	239.275
-2 Log L	246.866	223.248

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	23.618	2	<.0001
Score	20.391	2	<.0001
Wald	17.734	2	0.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.9264	0.393	24.0281	<.0001
male	1	-1.3312	0.3974	11.2214	0.0008
income	1	0.0394	0.0102	14.8891	0.0001

Ordered Logistic Regression

- Stata command:

  ologit expensive male income

Iteration 0: log likelihood = -291.07419					
Iteration 1: log likelihood = -289.30472					
Iteration 2: log likelihood = -289.30331					
Iteration 3: log likelihood = -289.30331					
Ordered logistic regression				Number of obs =	210
				LR chi2(2) =	3.54
				Prob > chi2 =	0.1702
Log likelihood = -289.30331				Pseudo R2 =	0.0061
expensive	Coef.	Std.Err.	z	P>z [95% Conf.	Interval]
male	-0.4442659	0.2580385	-1.7	0.085 -.9500121	0.06148
income	0.0073009	0.0065008	1.12	0.261 -.0054403	0.020042
/cut1	-1.005948	0.2802815		-1.555289	-0.45661
/cut2	0.0641326	0.2702749		-0.4655964	0.593862
/cut3	1.183917	0.2795577		0.6359943	1.73184

Ordered Logistic Regression (Cont.)

SAS command:

```
Proc logistic Data = in.logit_ologit_mlogit desc;
Model expensive =male income /Link=logit;
```

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	585.418	586.268
SC	595.445	602.98
-2 Log L	579.418	576.268

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3.15	2	0.207
Score	3.0561	2	0.217
Wald	3.2164	2	0.2003

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	4	1	-1	0.2789	17.8422	<.0001
Intercept	3	1	-0	0.2659	0.0443	0.8334
Intercept	2	1	1	0.276	13.7104	0.0002
male		1	-0	0.2573	2.6394	0.1042
income		1	0	0.00644	1.1795	0.2775

Multinomial Logistic Regression

Stata command:
mlogit mode male income

```
Iteration 0:   log likelihood = -283.75877
Iteration 1:   log likelihood = -242.52745
Iteration 2:   log likelihood = -241.53038
Iteration 3:   log likelihood = -241.52647
Iteration 4:   log likelihood = -241.52647

Multinomial logistic regression                 Number of obs   =        210
                                                LR chi2(6)      =      84.46
                                                Prob > chi2     =     0.0000
Log likelihood = -241.52647                     Pseudo R2       =     0.1488
```

mode	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1						
male	-.410087	.473168	-0.87	0.386	-1.337479	.5173052
income	.0688465	.0129916	5.30	0.000	.0433835	.0943095
_cons	-2.177179	.4641557	-4.69	0.000	-3.086908	-1.267451
2	(base outcome)					
3						
male	-.3504137	.5190623	-0.68	0.500	-1.367757	.6669297
income	.0265263	.0143836	1.84	0.065	-.0016649	.0547176
_cons	-1.346112	.4768624	-2.82	0.005	-2.280745	-.4114787
4						
male	1.94702	.4420894	4.40	0.000	1.080541	2.8135
income	.0519645	.0117829	4.41	0.000	.0288705	.0750585
_cons	-2.72198	.4803244	-5.67	0.000	-3.663398	-1.780561

Multinomial Logistic Regression (Cont.)

SAS command:

Proc catmod Data = in.logit_ologit_mlogit;
direct male income;
response logits;
Run;

Outcome;

Analysis of Maximum Likelihood Estimates					
Parameter	**Function Number**	**Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**
Intercept	**1**	0.5626	0.5148	1.19	0.2745
	2	2.7261	0.4805	32.19	<.0001
	3	1.3856	0.5518	6.31	0.012
male	**1**	-2.332	0.4608	25.6	<.0001
	2	-1.92	0.4432	18.76	<.0001
	3	-2.27	0.5412	17.59	<.0001
income	**1**	0.0164	0.0114	2.06	0.1513
	2	-0.052	0.0118	19.56	<.0001
	3	-0.026	0.0137	3.55	0.0596

Conditional Logistic Regression

Stata command:

clogit choice air train bus cost time air_inc, group(subject)

Conditional Logistic Regression (Cont.)

Stata Outcome:

```
Iteration 0:   log likelihood =  -205.8187
Iteration 1:   log likelihood = -199.23679
Iteration 2:   log likelihood = -199.12851
Iteration 3:   log likelihood = -199.12837
Iteration 4:   log likelihood = -199.12837

Conditional (fixed-effects) logistic regression   Number of obs   =        840
                                                  LR chi2(6)      =     183.99
                                                  Prob > chi2     =     0.0000
Log likelihood = -199.12837                       Pseudo R2       =     0.3160
```

choice	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
air	5.207443	.7790551	6.68	0.000	3.680523	6.734363
train	3.869043	.4431269	8.73	0.000	3.00053	4.737555
bus	3.163194	.4502659	7.03	0.000	2.280689	4.045699
cost	-.0155015	.004408	-3.52	0.000	-.024141	-.006862
time	-.0961248	.0104398	-9.21	0.000	-.1165865	-.0756631
air_inc	.013287	.0102624	1.29	0.195	-.0068269	.033401

Conditional Logistic Regression (Cont.)

SAS commands

```
PROC MDC DATA=in.clogit_nlogit;
MODEL choice = air train bus cost time air_inc
   /TYPE=CLOGIT NCHOICE=4;
ID subject;
RUN;
```

Parameter Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**
air	1	5.2074	0.7791	6.68	<.0001
train	1	3.869	0.4431	8.73	<.0001
bus	1	3.1632	0.4503	7.03	<.0001
cost	1	-0.016	0.004408	-3.52	0.0004
time	1	-0.096	0.0104	-9.21	<.0001
air_inc	1	0.0133	0.0103	1.29	0.1954

Nested Logistic Regression

- Stata command

  nlogitgen tree = mode(fly: 1, ground: 2 | 3 | 4)

  nlogittree mode tree

  nlogit choice air train bus cost time || tree: air_inc || mode:, case(subject) nonnormalized nolog noconstant notree

Nested Logistic Regression (Cont.)

- Stata Results

```
Nonnormalized nested logit regression          Number of obs      =       840
Case variable: subject                         Number of cases    =       210

Alternative variable: mode                     Alts per case: min =         4
                                                              avg =       4.0
                                                              max =         4

                                                  Wald chi2(6)    =     80.11
Log likelihood = -193.65615                       Prob > chi2     =    0.0000
```

choice	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
mode						
air	6.041827	1.198628	5.04	0.000	3.69256	8.391095
train	5.063954	.6619239	7.65	0.000	3.766607	6.361301
bus	4.095842	.6150907	6.66	0.000	2.890287	5.301398
cost	-.0315757	.0081541	-3.87	0.000	-.0475575	-.0155938
time	-.1126084	.0141277	-7.97	0.000	-.1402981	-.0849187
tree equations						
fly						
air_inc	.0153323	.0093813	1.63	0.102	-.0030548	.0337193
ground						
air_inc	0	(base)				
inclusive-value parameters						
tree						
/fly_tau	.5861148	.1406178			.3105089	.8617207
/ground_tau	.389015	.1236901			.1465869	.6314432

LR test for IIA (tau = 1): chi2(2) = 10.94 Prob > chi2 = 0.0042

Nested Logistic Regression (Cont.)

- SAS command

```
PROC MDC DATA=in.clogit_nlogit;
MODEL choice = air train bus cost time air_inc /TYPE=NLOGIT
CHOICE=(mode);
ID subject;
UTILITY U(1,) = air train bus cost time,
            U(2, 1 2) = air_inc;
NEST LEVEL(1) = (1 @ 1, 2 3 4 @ 2),
         LEVEL(2) = (1 2 @ 1);
RUN;
```

Nested Logistic Regression (Cont.)

SAS outcome

Parameter Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**
air_L1	1	6.0423	1.1989	5.04	<.0001
train_L1	1	5.0646	0.662	7.65	<.0001
bus_L1	1	4.0963	0.6152	6.66	<.0001
cost_L1	1	-0.032	0.008156	-3.87	0.0001
time_L1	1	-0.113	0.0141	-7.97	<.0001
air_inc_L2G1	1	0.0153	0.009381	1.63	0.1022
INC_L2G1C1	1	0.586	0.1406	4.17	<.0001
INC_L2G1C2	1	0.389	0.1237	3.15	0.0017

Conclusion

- If you have categorical dependent variables, you need to choose adequate methods to analyze them.
- You need to choose the regression models that fit your data and research questions.
- If you have event counts (e.g., the number of accidents), you need to use other models such as Poisson regression, Log-linear model, or Negative binomial regression for analyses.
- For additional help with categorical data analysis, feel free to contact me at wuh@bgsu.edu and 372-3119.