Categorical Data Analysis

Hsueh-Sheng Wu

Center for Family and Demographic Research

October 4, 2010

Outline

- What are categorical variables?
- When do we need categorical data analysis?
- Some methods for categorical dependent variable
  - Analysis of a two-way contingency table
  - Analysis of a three-way contingency table
  - Three types of logistic regression
    - Logistic regression
    - Ordered Logistic regression
    - Multinomial Logistic regression
- Conclusion

BGSU

Center for Family and Demographic Research

What Are Categorical Data?

- 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

Analyzing a Two-way Contingency Table

- Analyzing a 2x2 table

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

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

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

BGSU
Center for Family and Demographic Research

Analyzing a Two-way Contingency Table (Cont.)

- Odds Ratio

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

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

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

Examples

- 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 = 0.66

  - Odds Ratio

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

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

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

Three Types of Logistic Regression

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

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

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

BGSU Center for Family and Demographic Research

Three Types of Logistic Regression (Cont.)

Multinomial Logistic Regression

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

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

$$= (\alpha_{\mathrm{a}} + \beta_a \chi) - (\alpha_{\mathrm{b}} + \beta_b \chi)$$

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

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.

Example

```
Contains data from http://www.stata-press.com/data/r11/auto.dta
  obs:            74                          1978 Automobile Data
 vars:            12                          13 Apr 2009 17:45
 size:         3,478 (99.9% of memory free)   (_dta has notes)
-------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------
make            str18  %-18s                  Make and Model
price           int    %8.0gc                 Price
mpg             int    %8.0g                  Mileage (mpg)
rep78           int    %8.0g                  Repair Record 1978
headroom        float  %6.1f                  Headroom (in.)
trunk           int    %8.0g                  Trunk space (cu. ft.)
weight          int    %8.0gc                 Weight (lbs.)
length          int    %8.0g                  Length (in.)
turn            int    %8.0g                  Turn Circle (ft.)
displacement    int    %8.0g                  Displacement (cu. in.)
gear_ratio      float  %6.2f                  Gear Ratio
foreign         byte   %8.0g       origin     Car type
```

Generate New Categorical Variables

- A dichotomous variable (repair2)
  - 1 if rep78 > 3 and rep78 ~= . and 0 if rep78 <= 2
  - 1 indicates the car is very likely to break down, and 0 indicates the car is not.
- A three-category ordinal variable (repair3)
  - 2 if rep78 > 4 and rep78 ~=. 1 if rep78 == 3, and 0 if rep78 <= 2
  - 2 indicates the car is very likely to break down, 1 indicates the car is likely to break down, 0 indicates the car is unlikely to break down

BGSU
Center for Family and Demographic Research

Logistic Regression Results

```
• logit repair2 price mpg gear_ratio foreign

• note: foreign != 0 predicts success perfectly
•       foreign dropped and 21 obs not used

• Iteration 0:   log likelihood = -24.563524
• .
• Iteration 4:   log likelihood = -23.786597

• Logistic regression                               Number of obs   =         48
•                                                   LR chi2(3)      =       1.55
•                                                   Prob > chi2     =     0.6699
• Log likelihood = -23.786597                       Pseudo R2       =     0.0316

• ------------------------------------------------------------------------------
•      repair2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
• -------------+----------------------------------------------------------------
•        price |   .0001677   .0001661     1.01   0.313     -.000158    .0004933
•          mpg |  -.0014679    .099011    -0.01   0.988    -.1955259    .1925901
•   gear_ratio |   1.586402   1.584308     1.00   0.317    -1.518784    4.691588
•      foreign |  (omitted)
•        _cons |  -4.019467   4.685599    -0.86   0.391    -13.20307    5.164138
• ------------------------------------------------------------------------------
```

Ordered Logistic Regression

```
. ologit repair3 price mpg gear_ratio foreign

Iteration 0:   log likelihood = -69.439997
Iteration 1:   log likelihood = -55.714994
Iteration 2:   log likelihood = -55.523055
Iteration 3:   log likelihood =   -55.5227
Iteration 4:   log likelihood =   -55.5227

Ordered logistic regression                       Number of obs   =         69
                                                  LR chi2(4)      =      27.83
                                                  Prob > chi2     =     0.0000
Log likelihood =   -55.5227                       Pseudo R2       =     0.2004

------------------------------------------------------------------------------
     repair3 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |   .0000722   .0001054     0.68   0.493    -.0001344    .0002787
         mpg |   .0811893   .0689252     1.18   0.239    -.0539016    .2162802
  gear_ratio |   -.112221   1.035193    -0.11   0.914    -2.141162     1.91672
     foreign |   2.748854   .9580416     2.87   0.004      .871127    4.626581
-------------+----------------------------------------------------------------
       /cut1 |   .3380733    3.13436                      -5.80516    6.481307
       /cut2 |   2.980449   3.161144                     -3.215279    9.176178
------------------------------------------------------------------------------
```

Multinomial Logistic Regression

```
. mlogit repair3 price mpg gear_ratio foreign, base(0)

Multinomial logistic regression                   Number of obs   =         69
                                                  LR chi2(8)      =      31.30
                                                  Prob > chi2     =     0.0001
Log likelihood = -53.792211                       Pseudo R2       =     0.2253
------------------------------------------------------------------------------
     repair3 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
not_likely~n |  (base outcome)
-------------+----------------------------------------------------------------
likely_to_~n |
       price |   .0001736   .0001711     1.01   0.310    -.0001617    .0005089
         mpg |   -.043372   .1056836    -0.41   0.682     -.250508     .163764
  gear_ratio |   2.017402   1.635181     1.23   0.217    -1.187493    5.222297
     foreign |   14.14439   2122.177     0.01   0.995    -4145.245    4173.534
       _cons |  -4.795311   4.837551    -0.99   0.322    -14.27674    4.686116
-------------+----------------------------------------------------------------
very_likel~n |
       price |   .0001394   .0001895     0.74   0.462    -.0002321    .0005109
         mpg |   .0782156   .1121928     0.70   0.486    -.1416782    .2981094
  gear_ratio |   .5374572   1.852135     0.29   0.772     -3.09266    4.167575
     foreign |   17.38138   2122.177     0.01   0.993    -4142.008    4176.771
       _cons |  -3.775221   5.406854    -0.70   0.485    -14.37246    6.822017
------------------------------------------------------------------------------
```

Conclusion

- If you have categorical dependent variables, you need to choose adequate methods to analyze them.
- There are additional models, including Poisson regression, Log-linear model, Negative binomial regression, and Models for matched pairs.
- For additional help with categorical data analysis, feel free to contact me at wuh@bgsu.edu and 372-3119.

BGSU Center for Family and Demographic Research