Categorical Data Analysis Using SAS and Stata

Hsueh-Sheng Wu

Center for Family and Demographic Research

Mar 3, 2014

Outline

- Why do we need to learn categorical data analyses?
- A summary of different categorical data analyses
  - Analyses of contingency tables
  - Regression models
    - Logistic regression
    - Ordered logistic regression
    - Multinomial logistic regression
- Stata commands
- SAS commands
- Interpreting the results
- Predicted probability
- Conclusions

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

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

Analyzing a Two-way Contingency Table

- Analyzing a 2x2 table

Table. Gender and Employment

	Employed	Unemployed
Male	200	200
Female	200	400

Table. Gender and Employment

	Employed	Unemployed
Male	ρ_1	$1-\rho_1$
Female	ρ_2	$1-\rho_2$

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

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

BGSU

Center for Family and Demographic Research

Analyzing a Two-way Contingency Table (Cont.)

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

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

  P1 = 200/400 = 0.5

  P2 = 200/600 = 0.33

  - Difference of two proportions

    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

BGSU

Center for Family and Demographic Research

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

BGSU Center for Family and Demographic Research

Models for Examining How Characteristics of Individuals Affect Choices

Logistic Regression

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

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

Ordered Logistic Regression

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

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

BGSU

Center for Family and Demographic Research

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

Multinomial Logistic Regression

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

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

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

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

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.

BGSU
Center for Family and Demographic Research

A List of Variables in the Data

variable name	variable label	Label Value	Label Label
aid	ID	57101310 - 99719978	
married	Marital Status	0	Not married
		1	Married
educ	Education	1	Less than High School
		2	High School
		3	Some college
		4	colleges or more
union	Union Status	0	single
		1	cohabiting
		2	married
female	Female	0	Male
		1	Female
age	Age		24-33
agesq	Age squared		576-1089
femaleage	Interaction term of female and age		0-33

BGSU

Center for Family and Demographic Research

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

	aid	married	educ	union	female	age	agesq	femaleage
1	57101310	1	2	2	1	31	961	31
2	57103869	0	1	0	0	32	1024	0
3	57109625	0	1	0	0	27	729	0
4	57111071	0	3	0	0	27	729	0
5	57113943	0	3	1	0	29	841	0
6	57117542	0	1	0	0	28	784	0
7	57118381	1	3	2	1	25	625	25
8	57118943	1	4	2	1	29	841	29
9	57120005	0	4	0	0	26	676	0
10	57120046	1	3	2	0	31	961	0
11	57120371	1	2	2	1	31	961	31
12	57121404	1	4	2	1	28	784	28
13	57121476	0	2	0	1	27	729	27
14	57127241	0	3	1	1	26	676	26
15	57129567	0	3	1	0	27	729	0
16	57131432	1	3	2	0	29	841	0
17	57131909	0	3	1	0	26	676	0
18	57133772	0	3	0	1	26	676	26
19	57134457	0	3	0	1	28	784	28
20	57134967	0	1	1	1	26	676	26

Stata Commands

Logistic Regression

```
logit married female age femaleage
logit married female age femaleage, or
```

Ordered Logistic Regression

```
ologit educ female age femaleage
ologit educ female age femaleage, or
```

Multinomial Logistic Regression

```
mlogit union female age femaleage, base(0)
```

BGSU

Center for Family and Demographic Research

SAS Commands

Logistic Regression

```
Proc Logistic data = in.annotated_3_2;
Format married marriedf. educ educf.;
Model married = educ female age femaleage;
run;
```

SAS and Stata Commands

Ordered Logistic Regression

```
Proc Logistic data = in.annotated descending;
Format educ educf. female femalef.;
Model educ = female age femaleage;
run;

PROC QLIM data = in.annotated;
MODEL educ = female age
femaleage/DISCRETE (DIST=LOGISTIC);
RUN;
```

Multinomial Logistic Regression

```
mlogit union educ female age femaleage, base(0)
```

SAS and Stata Commands

Multinomial Logistic Regression

```
proc logistic data = in.annotated_3_2;
class union (ref = "0");
model union = educ female age femaleage/
link = glogit;
run;
```

BGSU

Center for Family and Demographic Research

Interpreting the Results

- The sample size
- The reference category
- The regression coefficients
- The odds ratio

Predicted Probability

- Predicted probability is useful to describe the results
- Odds = Exp(the sum of coefficients)
- Predicted Probability = Odds/(1+Odds)
- You can present predicted probability with graphs

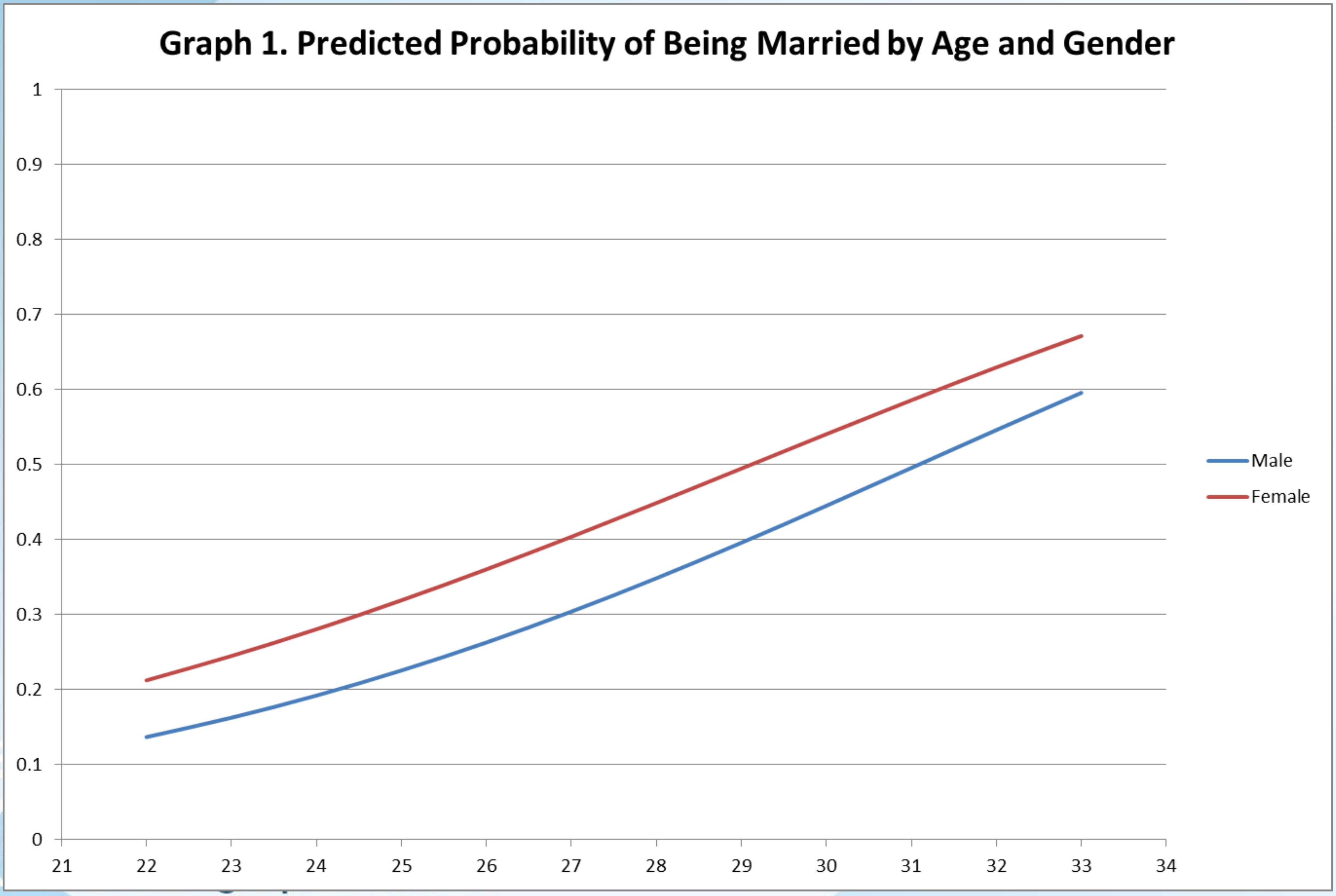
BGSU
Center for Family and Demographic Research

Predicted Probability (continued)

Table 3. Predicated Probability for Male and Female Respondents

Intercept	Female		Age		Age*Female		Sum of coefficents	Odds Ratio	Predicted Probability
coefficent	value	coefficent	value	coefficent	value	coefficent			
-6.295917	0	0.9380865	22	0.2025471	0	-0.0185092	-1.8398808	0.158836358	0.137065391
-6.295917	0	0.9380865	23	0.2025471	0	-0.0185092	-1.6373337	0.194497941	0.162828193
-6.295917	0	0.9380865	24	0.2025471	0	-0.0185092	-1.4347866	0.238166183	0.192353972
-6.295917	0	0.9380865	25	0.2025471	0	-0.0185092	-1.2322395	0.291638721	0.225789701
-6.295917	0	0.9380865	26	0.2025471	0	-0.0185092	-1.0296924	0.357116793	0.263143743
-6.295917	0	0.9380865	27	0.2025471	0	-0.0185092	-0.8271453	0.437295855	0.30424902
-6.295917	0	0.9380865	28	0.2025471	0	-0.0185092	-0.6245982	0.53547654	0.348736386
-6.295917	0	0.9380865	29	0.2025471	0	-0.0185092	-0.4220511	0.655700532	0.396026044
-6.295917	0	0.9380865	30	0.2025471	0	-0.0185092	-0.219504	0.802916946	0.44534328
-6.295917	0	0.9380865	31	0.2025471	0	-0.0185092	-0.0169569	0.983186059	0.495760877
-6.295917	0	0.9380865	32	0.2025471	0	-0.0185092	0.1855902	1.203928789	0.546264832
-6.295917	0	0.9380865	33	0.2025471	0	-0.0185092	0.3881373	1.474232182	0.595834212
-6.295917	1	0.9380865	22	0.2025471	22	-0.0185092	-1.3089967	0.270090903	0.212654781
-6.295917	1	0.9380865	23	0.2025471	23	-0.0185092	-1.1249588	0.324665843	0.245092636
-6.295917	1	0.9380865	24	0.2025471	24	-0.0185092	-0.9409209	0.390268272	0.280714363
-6.295917	1	0.9380865	25	0.2025471	25	-0.0185092	-0.756883	0.469126417	0.319323383
-6.295917	1	0.9380865	26	0.2025471	26	-0.0185092	-0.5728451	0.563918749	0.360580592
-6.295917	1	0.9380865	27	0.2025471	27	-0.0185092	-0.3888072	0.67786495	0.404004476
-6.295917	1	0.9380865	28	0.2025471	28	-0.0185092	-0.2047693	0.814835277	0.448985805
-6.295917	1	0.9380865	29	0.2025471	29	-0.0185092	-0.0207314	0.979482018	0.494817336
-6.295917	1	0.9380865	30	0.2025471	30	-0.0185092	0.1633065	1.177397507	0.540736133
-6.295917	1	0.9380865	31	0.2025471	31	-0.0185092	0.3473444	1.415304072	0.585973455
-6.295917	1	0.9380865	32	0.2025471	32	-0.0185092	0.5313823	1.701282367	0.629805454
-6.295917	1	0.9380865	33	0.2025471	33	-0.0185092	0.7154202	2.04504583	0.671597718

Predicted Probability (continued)



Conclusions

- 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.

BGSU
Center for Family and Demographic Research