Categorical Data Analysis

Hsueh-Sheng Wu

CFDR Workshop Series

September 30, 2019

BGSU
Center for Family and Demographic Research

Outline

- Why do we need to learn categorical data analyses?
- What is special about categorical dependent variables?
- What findings does the analysis of categorical variables provide?
- Select Techniques for categorical data analysis
  - Testing the Association among categorical variables
  - Analyzing Binary Dependent Variables
  - Analyzing Ordered Dependent Variables (Ordered logistic regression and Sequential logistic regression)
  - Analyzing Nominal dependent variable: Multinomial logistic regression, Conditional logistic Regression, Nested Logistic Regression, and Alternative-specific conditional logit model
- Conclusion

What Is Special about Categorical Variable?

- Many social constructs are conceptualized as categorical variables, not continuous variables, for example, marital status, employment status, naturalization status.
- 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.
- For analyzing categorical variables, researchers need to consider
  - How many categories of a dependent variables are
  - Whether these categories are related to each other in a certain way
  - Whether respondents are independent of each other
  - Whether respondents are independent of each other, and whether the attributes of each category will be included in the analysis
  - All the above contribute to the formulation of the research question and the choice of the statistical model

What findings does the analysis of categorical variables provide?

- Chi-Square test helps determine if categorical variables are not independent of each other
- regression analysis for categorical variables identifies important covariates related to how likely an individual move from one response category to another.

Pearson's Chi-Square Test

- Chi-square test can be used for n-dimension tables, meaning that you can tests the independence of more than 2 categorical variables
- The -table- command can generate the n-dimension tables.
- The -tab- command tests independence of two categorical variables, and the -ipf- command tests independence of three or more categorical variables
- The -ipf- command is a user-written ado file created by Dr. Adrian Mander. If this command is not available on lab computers, please let me know.

Pearson's Chi-Square Test (Continued)

Stata Commands for Creating N-Way Tables

```
webuse lbw, clear
des
sum

* Creating  a 2-way table
table low ui, contents(freq)

* Creating  a 3-way table.
table low ui ht, contents(freq)
sort ht
by ht: tab low ui

* Creating a 4-way table.
table low ui ht, by(race)  contents(freq)
 sort race ht
by race ht: tab low us
```

Pearson's Chi-Square Test (Continued)

Stata Commands for Conducting Chi-Square Test

* Chi-Square test for a two-way table

```
tab2 low ui, mis chi lrchi2 exac
```

* Chi-Square test for a three-way table

```
ipf, fit(low+ui+ht)
ipf, fit (low+ui+ht+ low*ui)
```

Binary Dependent Variable

- Binary Dependent Variable only has two categories
- No additional assumption is made about the categories
- Stata command:

```
webuse lbw, clear

*look at the regression coefficients
logit low age lwt i.race smoke ptl ht ui

*Look at the odds ratio
logit low age lwt i.race smoke ptl ht ui,or

*look at the predicted probability for a certain predictor
margins i.race, atmeans
```

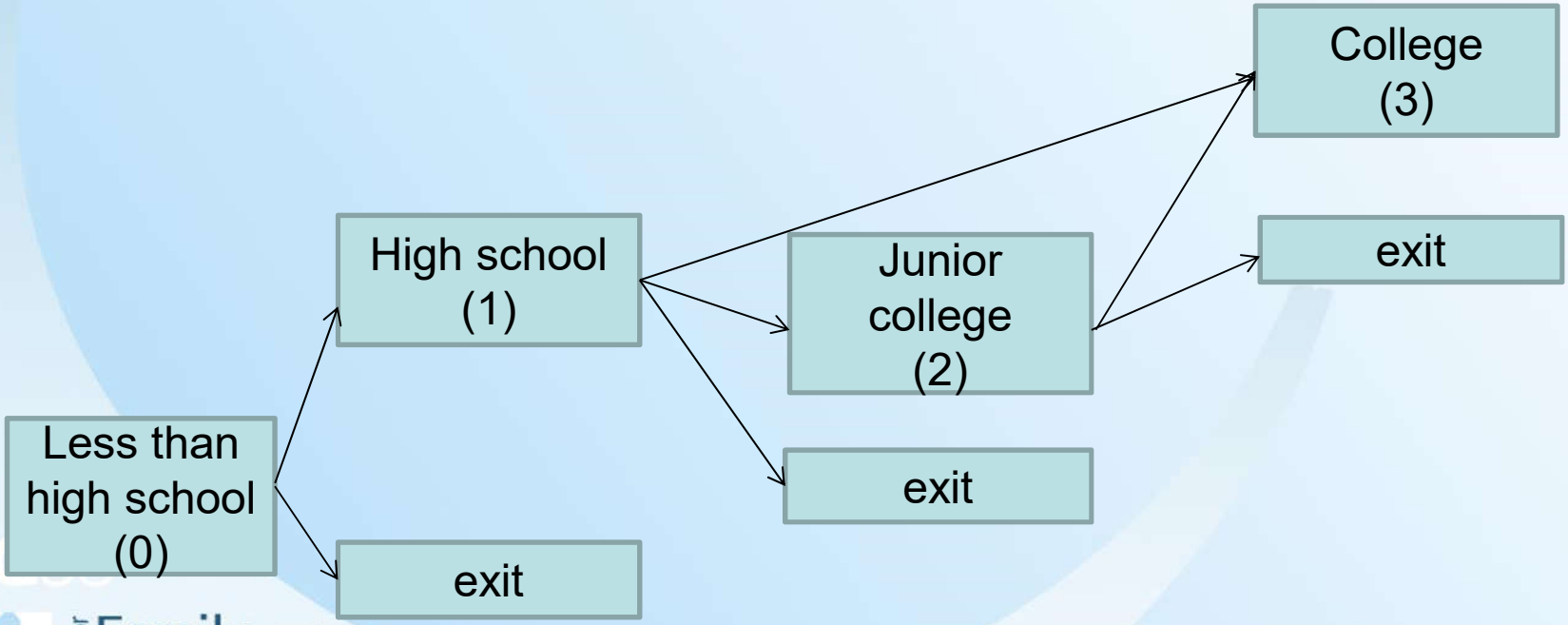
Ordered Categorical Variables (Continued)

- Ordered logistic regression assumes that the categories of the dependent variables follows an incremental fashion and the relation between predictors and each categories remain the same.
- Stata Commands

```
use d:\temp\order.dta, clear
des
list in 1/20

* look at the regression coefficients
ologit degree south c.coh##c.coh i.black paeduc

* look at the odds ratio
ologit degree south c.coh##c.coh i.black paeduc, or

* look at the predicted probability
 margins i.black, atmeans
```

Ordered Categorical Variables (Continued)

- The -seqlogit- module is created by Dr.Maarten L. Buis for sequential logistic regression.
- A diagram of the transition between education categories



Ordered Categorical Variables (Continued)

- Stata comman

```
use d:\temp\order.dta, clear
 des
 list in 1/20
 seqlogit degree south c.coh##c.coh if black == 0 ,       ///
 tree(0 : 1 2 3 , 1 : 2 3 , 2 : 3 )                       ///
 ofinterest(paeduc)  over(c.coh##c.coh)                   ///
 levels(0=9, 1=12, 2=14, 3=16)

 seqlogitdecomp, overat(coh 1.5,  coh 2.5,  coh 3.5,  coh 4.5, coh 5.5, coh 6.5)   ///
 at(south 0 paeduc 12)  line(0) xline(0)                                            ///
 subtitle("1915" "1925" "1935"  "1945" "1955" "1965")                               ///
 eqlabel(`""less than high school" "versus" "high school or more""'                 ///
 `""high school" "versus" "any college""'  `""junior college" "versus" "college""' )
```

Nominal dependent variable

- The categories of a nominal dependent variable does not follow a certain order, e.g., the makes and models of the car that American bought last year.
- Multinomial logistic regression is used If the categories of a nominal dependent variable is independent of each other and respondents are independent of each other, multinomial logistic regression an be used.
- Conditional logistic regression is used if the categories of a nominal dependent variable is independent of each other, but respondents are not independent of each other.
- Nested logistic regression is used If the categories of a nominal dependent variable form a certain decision structure.
- Alternative-Specific Conditional Logistic regression is used to examine whether the attributes of categories and respondents are related to the probability of endorsing a certain category of the dependent variable.

Nominal dependent variable (Continued)

Stata command for multinomial logistic regression

```
use http://www.stata-press.com/data/r13/choice, clear
keep if choice == 1

* obtain regression coefficients
mlogit car i.sex income

* obtain relative-risk ratios,
mlogit car i.sex income, rrr

* obtain predicted probability
margins i.sex, atmeans
margins, over(sex) dydx(income)
```

Nominal dependent variable (Continued)

Stata commands:

```
*Conditional Logistic Regression
 use http://www.stata-press.com/data/r13/union, clear
 des
 sort idcode
 list in 1/20, sepby(idcode)
 clogit union age grade not_smsa south black, group(idcode)
```

Nominal dependent variable (Continued)

*Nested Logistic Regression

webuse restaurant

nlogitgen type = restaurant(fast: Freebirds | MamasPizza, family: CafeEccell | LosNortenos | WingsNmore, fancy: Christophers | MadCows)

nlogittree restaurant type, choice(chosen) case(family_id)

nlogit chosen cost distance rating || type: income kids, base(family) || restaurant:, noconst case(family_id)

Nominal dependent variable (Continued)

```
* Alternative-Specific Conditional Logit
use http://www.stata-press.com/data/r13/choice, clear
sort id car
list in 1/30, sepby(id)
asclogit choice dealer, case(id) alternatives(car) casevars(sex income)
```

Conclusion

- If you have categorical dependent variables, OLS regression is not an adequate method to analyze them.
- Chi-Square test and special regression models are commonly used for analyzing categorical dependent variable
- When deciding your research question and how to analyze the data, you should consider how many response categories the dependent variables has, whether these categories are related to each other in a certain way, whether respondents are independent of each other, and whether the attributes of each category will be included in the analysis
- We did not consider the count data in today's workshop. However, 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.