Week 11: Categorical Data II

Wenhao Jiang

Department of Sociology
New York University

November 18, 2022

Categorical Data as Explanatory Variables

Categorical Data as Explanatory Variable

- We have already learned how to deal with the case when a binary variable like gender is used as a explanatory (independent) variable
- In this case, gender is also a categorical variable, with two possible values
  - In recent years, large national surveys now include more categories in asking gender

Categorical Data as Explanatory Variable

- We have already learned how to deal with the case when a binary variable like gender is used as a explanatory (independent) variable
- In this case, gender is also a categorical variable, with two possible values
  - In recent years, large national surveys now include more categories in asking gender
- We include gender in regression by transforming it into a 0-1 binary, where 1 represents woman, and 0 man (or 1 as man and 0 as woman)
- $Abscale_i = \hat{\beta}_0 + \hat{\beta}_1 edu_i + \hat{\beta}_2 women_i + e_i$
- How do intepret $\hat{\beta}_2$ (suppose this is a positive number)?

Categorical Data as Explanatory Variable

- We have already learned how to deal with the case when a binary variable like gender is used as a explanatory (independent) variable
- In this case, gender is also a categorical variable, with two possible values
  - In recent years, large national surveys now include more categories in asking gender
- We include gender in regression by transforming it into a 0-1 binary, where 1 represents woman, and 0 man (or 1 as man and 0 as woman)
- $Abscale_i = \hat{\beta}_0 + \hat{\beta}_1 edu_i + \hat{\beta}_2 women_i + e_i$
- How do intepret $\hat{\beta}_2$ (suppose this is a positive number)?
- Conditioning on the same level of education, women on average have $\hat{\beta}_2$-unit higher levels abortion attitudes than men do

Categorical Data as Explanatory Variable

- $Abscale_i = \hat{\beta}_0 + \hat{\beta}_1 edu_i + \hat{\beta}_2 women_i + e_i$
- Men do not explicitly appear in regression; instead, it appears as the **reference group** for women to be compared with

Categorical Data as Explanatory Variable

- $Abscale_i = \hat{\beta}_0 + \hat{\beta}_1 edu_i + \hat{\beta}_2 women_i + e_i$
- Men do not explicitly appear in regression; instead, it appears as the **reference group** for women to be compared with
- Including a categorical variable that has multiple variables (e.g., race, region, marital status) is analogous
- For example, if we want to add race (5 categories: White, Black, Hispanic, Asian, Others) in the above equation
- It is incorrect to directly specify $Abscale_i = \hat{\beta}_0 + \hat{\beta}_1 edu_i + \hat{\beta}_2 women_i + \hat{\beta}_3 race_i + e_i$
- As race is not ordered and additive

Categorical Data as Explanatory Variable

- Instead, we transform race into 5 "dummy variables" (dummy means a binary variable created from categorical variables), `white`, `black`, `hispanic`, `asian`, `others`
- If the respondent is a black, $black = 1$ and other other dummy variables $= 0$
- If the respondent is a hispanic, $hispanic = 1$ and other other dummy variables $= 0$

Categorical Data as Explanatory Variable

- Instead, we transform race into 5 “dummy variables” (dummy means a binary variable created from categorical variables), `white`, `black`, `hispanic`, `asian`, `others`
- If the respondent is a black, $black = 1$ and other other dummy variables $= 0$
- If the respondent is a hispanic, $hispanic = 1$ and other other dummy variables $= 0$
- We put all these dummy variables but one group in regression; the group that is omitted serves as the **reference** group
- $Abscale_i = \hat{\beta}_0 + \hat{\beta}_1 edu_i + \hat{\beta}_2 women_i + \hat{\beta}_3 black_i + \hat{\beta}_4 hispanic_i + \hat{\beta}_5 asian_i + \hat{\beta}_6 others_i + e_i$

Categorical Data as Explanatory Variable

- Instead, we transform race into 5 “dummy variables” (dummy means a binary variable created from categorical variables), `white`, `black`, `hispanic`, `asian`, `others`
- If the respondent is a black, $black = 1$ and other other dummy variables $= 0$
- If the respondent is a hispanic, $hispanic = 1$ and other other dummy variables $= 0$
- We put all these dummy variables but one group in regression; the group that is omitted serves as the **reference** group
- $Abscale_i = \hat{\beta}_0 + \hat{\beta}_1 edu_i + \hat{\beta}_2 women_i + \hat{\beta}_3 black_i + \hat{\beta}_4 hispanic_i + \hat{\beta}_5 asian_i + \hat{\beta}_6 others_i + e_i$
- Which group to omit is up to you, but it is related to the interpretation of the results

Categorical Data as Explanatory Variable

- $Abscale_i = \hat{\beta}_0 + \hat{\beta}_1 edu_i + \hat{\beta}_2 women_i + \hat{\beta}_3 black_i + \hat{\beta}_4 hispanic_i + \hat{\beta}_5 asian_i + \hat{\beta}_6 others_i + e_i$
- How do we interpret $\hat{\beta}_3$ (suppose this is a positive number)?

Categorical Data as Explanatory Variable

- $Abscale_i = \hat{\beta}_0 + \hat{\beta}_1 edu_i + \hat{\beta}_2 women_i + \hat{\beta}_3 black_i + \hat{\beta}_4 hispanic_i + \hat{\beta}_5 asian_i + \hat{\beta}_6 others_i + e_i$
- How do we interpret $\hat{\beta}_3$ (suppose this is a positive number)?
- Conditioning on the same level of education and gender, blacks on average have $\hat{\beta}_3$-unit higher levels of abortion attitudes than whites do
- Whites do not explicitly appear in regression; instead, it appears as the **reference group** for other racial groups to be compared with

Practice

- We are now interested in the association between marital status and religious identification, controlling for gender and education
- There are five categories of marital status, Married, Widowed, Divorced, Separated, Never married

Practice

- $Relig_i = 2.56 + 0.013 edu_i + (-0.094) women_i + (-0.355) widowed_i + 0.139 divorced_i + 0.376 separated + 0.635 nevmar_i + e_i$

Table 1: The association between education, sex, marital status and religion, GSS 2021

	Dependent variable:
	Religious identification
Education	-0.007^{***} (0.003)
Women	0.061^{***} (0.014)
Widowed	0.076^{***} (0.028)
Divorced	-0.048^{**} (0.020)
Separated	-0.062 (0.048)
Never married	-0.193^{***} (0.018)
Intercept	0.832^{***} (0.041)
Observations	3,878
Adjusted R^2	0.041
Note:	*p<0.1; **p<0.05; ***p<0.01

Data Memo and R Operations

Logistics

- Research Memo: 5-page memo presenting an interesting statistical result and discussing its social science implications
- Due **16 December**

General Expectations

- Not required to go very advanced and substantively innovative (but encouraged!), but we generally expect to see
- (1) An introduction with some literature review
  - A clear-defined topic (e.g., what is the association between education and gender ideologies/attitudes)
- (2) A brief description of the data you intend to use
  - For example, if you want to use GSS, you may describe which year of data you use, how many individuals are included, what is their gender and race composition, etc.

General Expectations

- (3) Clear-stated dependent (e.g., gender ideologies) and independent (years of education) variables
- (4) Start with a two-way cross-table
- (5) Proceed to a bivariate regression
- (6) Adding a set of other controlling variables in a multivariate regression setting, such as gender, age, and region
  - Try to test for interactions. For example, does the effect of years of education in gender ideologies depend on which gender the respondent is?
  - Or, is the variable a mediator, a moderator, or a multiple cause?
- (7) Visualize the relationship, possibly after step (3) and step (6)
- (8) Conclusion and implications

Logistics

- It is almost certain that you will encounter difficulties in cleaning the data
- Come to my office hour and I can help!

Data Sources

- General Social Survey
- Current Population Survey
- American Community Survey
- American Decennial Census
- American Time Use Survey
- National Longitudinal Survey of Youth 1979 and 1997

Read Data

```r
## set your working directory - you should set your own unique one!
setwd("~/Dropbox/Teaching/SOCUA-302/Week 11")

## read csv data - this is 2021 GSS data
gss <- read.csv("GSS_SOCUA_W11.csv")
```

run a new line of code: library(dplyr)

Cross Table

- The frequency of men and women who have and have no religious identification

```
## recode religion
gss <- gss %>%
  mutate(relig=ifelse(relig==4,0,1))

## cross table
source("http://pcwww.liv.ac.uk/~william/R/crosstab.r")
crosstab(gss, row.vars = "sex",
         col.vars = "relig",
         type = "f") ## "f" represents frequency
```

```
##      relig    0    1  Sum
## sex
## 0           552 1172 1724
## 1           556 1623 2179
## Sum        1108 2795 3903
```

Cross Table

- The proportion of people who have and have no religious identification among men and women

```
## cross table
crosstab(gss, row.vars = "sex",
         col.vars = "relig",
         type = "r") ## "r" represents row-wise proportion
```

```
##      relig       0      1    Sum
## sex
## 0            32.02  67.98 100.00
## 1            25.52  74.48 100.00
```

Cross Table

- The `crosstab` function by default returns proportion by row
- We can also change it to column-wise proportion
- The proportion of men and women among the people who have and have no religious identification

```
## cross table
crosstab(gss, row.vars = "sex",
         col.vars = "relig",
         type = "c") ## "c" represents column-wise proportion
```

```
##      relig      0      1
## sex
## 0           49.82  41.93
## 1           50.18  58.07
## Sum        100.00 100.00
```

Cross Table

- We can also include multiple categories as rows

```r
## cross table
crosstab(gss, row.vars = c("marital","sex"),
         col.vars = "relig",
         type = "r")
```

```
##              relig      0      1    Sum
## marital sex
## 1       0           25.53  74.47 100.00
##         1           22.26  77.74 100.00
## 2       0           23.17  76.83 100.00
##         1           10.63  89.37 100.00
## 3       0           34.05  65.95 100.00
##         1           23.96  76.04 100.00
## 4       0           28.57  71.43 100.00
##         1           30.16  69.84 100.00
## 5       0           47.20  52.80 100.00
##         1           38.63  61.37 100.00
```

- Women who are widowed have the highest proportion of religious believers among all categories.

Cross Table

- The proportion can also be joint percentages

```
## cross table
crosstab(gss, row.vars = c("marital","sex"),
         col.vars = "relig",
         type = "j")
```

```
##              relig     0     1    Sum
## marital sex
## 1       0          12.55 36.62  49.17
##         1          11.31 39.51  50.83
##         Sum        23.86 76.14 100.00
## 2       0           6.57 21.80  28.37
##         1           7.61 64.01  71.63
##         Sum        14.19 85.81 100.00
## 3       0          12.32 23.87  36.19
##         1          15.29 48.52  63.81
##         Sum        27.61 72.39 100.00
## 4       0           8.79 21.98  30.77
##         1          20.88 48.35  69.23
##         Sum        29.67 70.33 100.00
## 5       0          21.54 24.09  45.63
##         1          21.00 33.37  54.37
##         Sum        42.54 57.46 100.00
```

- Although the function `crosstab` is user-written, it is pretty flexible
- Take a look at the original instructions here if you want more ways of cross-tabulation to explore the data

Regression

- The basic format of regression in R is `lm(y ~ x1 + x2 + ... xn, data)`
- Save the regression model by some name, e.g., `model1 <- lm(y ~ x1 + x2 + ... xn, data)`
- You can check the regression output by `summary(model1)`

Regression

- The basic format of regression in R is `lm(y ~ x1 + x2 + ... xn, data)`
- Save the regression model by some name, e.g., `model1 <- lm(y ~ x1 + x2 + ... xn, data)`
- You can check the regression output by `summary(model1)`
- There is a more elegant way to present results by calling `stargazer()`

```
library(stargazer)
model1 <- lm(relig~educ,gss)
stargazer(model1, type = "text",
          header=FALSE,
          title = "The association between education and religious identification",
          digits = 3,
          omit.stat = c("rsq","f","ser"))
```

Regression

Table 2: The association between education and religious identification

	Dependent variable:
	relig
educ	-0.006^{**}
	(0.003)
Constant	0.807^{***}
	(0.039)
Observations	3,925
Adjusted R^2	0.001
Note:	*p<0.1; **p<0.05; ***p<0.01

Regression

- There is a more elegant way to present results by calling stargazer()
- stargazer() can present multiple results of regressions

```
model1 <- lm(relig~educ,gss)
model2 <- lm(relig~educ+sex,gss)
stargazer(model1,model2, type = "text",
          header=FALSE,
          title = "The association between education, sex and religious identification",
          digits = 3,
          omit.stat = c("rsq","f","ser"))
```

Regression

Table 3: The association between education, sex and religious identification

install.packages("stargazer")

	Dependent variable:	
	relig	
	(1)	(2)
educ	-0.006^{**}	-0.005^{**}
	(0.003)	(0.003)
sex		0.064^{***}
		(0.015)
Constant	0.807^{***}	0.758^{***}
	(0.039)	(0.040)
Observations	3,925	3,885
Adjusted R^2	0.001	0.006
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$	

Regression

- We can also include an interaction term to examine whether the association between education and religious identification depends on gender

Regression

- We can also include an interaction term to examine whether the association between education and religious identification depends on gender
- In R, the product of two terms is controlled by the * sign

```
model1 <- lm(relig~educ,gss)
model2 <- lm(relig~educ+sex,gss)
model3 <- lm(relig~educ+sex+educ*sex,gss)
stargazer(model1,model2,model3, type = "text",
          header=FALSE,
          title = "The association between education, sex and religious identification",
          digits = 3,
          omit.stat = c("rsq","f","ser"))
```

Regression

Table 4: The association between education, sex and religious identification

	Dependent variable:		
	relig		
	(1)	(2)	(3)
educ	−0.006**	−0.005**	−0.001
	(0.003)	(0.003)	(0.004)
sex		0.064***	0.174**
		(0.015)	(0.079)
educ:sex			−0.007
			(0.005)
Constant	0.807***	0.758***	0.697***
	(0.039)	(0.040)	(0.059)
Observations	3,925	3,885	3,885
Adjusted R^2	0.001	0.006	0.006
Note:			*p<0.1; **p<0.05; ***p<0.01

Regression

- We can change the dependent variable from religious identification to income

```
model1 <- lm(rincome~educ,gss)
model2 <- lm(rincome~educ+sex,gss)
model3 <- lm(rincome~educ+sex+educ*sex,gss)
stargazer(model1,model2,model3, type = "text",
          header=FALSE,
          title = "The association between education, sex and income",
          digits = 1,
          omit.stat = c("rsq","f","ser"))
```

Regression

Table 5: The association between education, sex and income

	Dependent variable:		
	rincome		
	(1)	(2)	(3)
educ	893.4***	886.5***	737.1***
	(73.4)	(73.2)	(105.2)
sex		−1,545.4***	−5,924.5***
		(390.7)	(2,249.4)
educ:sex			289.3**
			(146.4)
Constant	12,261.1***	13,184.8***	15,455.1***
	(1,127.3)	(1,148.0)	(1,623.4)
Observations	2,501	2,501	2,501
Adjusted R^2	0.1	0.1	0.1
Note:			*p<0.1; **p<0.05; ***p<0.01

Plot Fitted Regression Line

- We can use the function `predict.lm` to apply the estimated regression equation to any data
- Here we want to show the predicted values by men and women separately

```
## create a hypothetical data for men
predict_men <-
  data.frame(educ=seq(0,20,1),
             sex=rep(0,21))

## make predictions
y_hat <-
  predict.lm(model3,
             predict_men)

## store the data
predict_men$y_hat <- y_hat
```

```r
## look at the stored data
head(predict_men)
```

```
##   educ sex    y_hat
## 1    0   0 15455.12
## 2    1   0 16192.20
## 3    2   0 16929.29
## 4    3   0 17666.37
## 5    4   0 18403.45
## 6    5   0 19140.53
```

- We can symmetrically create predicted values for women

```
## create a hypothetical data for women
predict_women <-
  data.frame(educ=seq(0,20,1),
             sex=rep(1,21))

## make predictions
y_hat <-
  predict.lm(model3,
             predict_women)

## store the data
predict_women$y_hat <- y_hat
```

```r
## look at the stored data
head(predict_women)
```

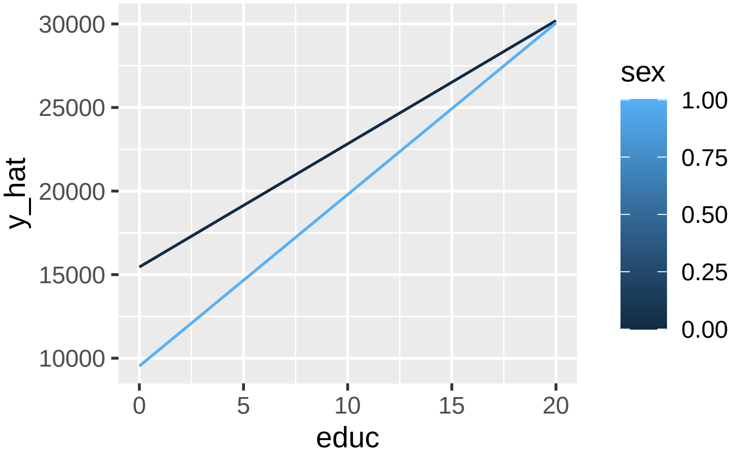
```
##   educ sex     y_hat
## 1    0   1  9530.589
## 2    1   1 10557.005
## 3    2   1 11583.421
## 4    3   1 12609.836
## 5    4   1 13636.252
## 6    5   1 14662.668
```

- We combine these two datasets into one by calling `rbind()` that represents row-wise binding
- To make `rbind()` work, the two datasets have to have same column names

```
predict <- rbind(predict_men,predict_women)
```

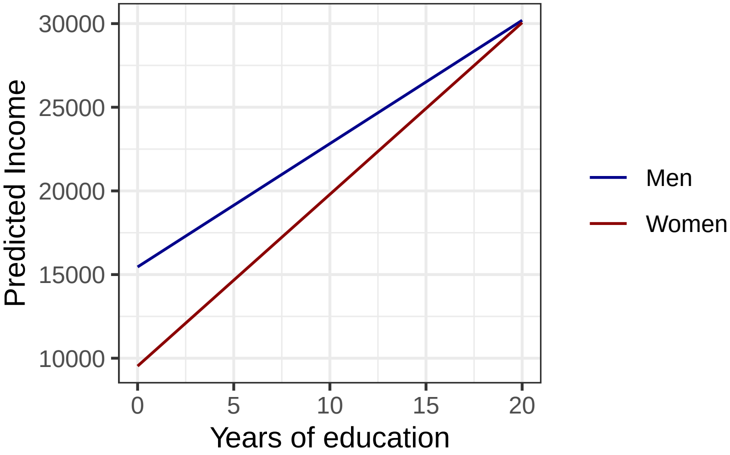
- Plot the predicted value by calling ggplot()

```
library(ggplot2)
ggplot(predict,
       aes(x=educ,y=y_hat,group=sex,color=sex)) +
  geom_line()
```



- We want to adjust
- 1. Sex from a continuous variable to a categorical variable
- 2. Manually defined colors
- 3. x-axis and y-axis labels
- 4. Omit the legend title
- 5. Change the

```
ggplot(predict,
       aes(x=educ,y=y_hat,group=factor(sex),color=factor(sex))) +
  geom_line() +
  scale_color_manual(labels = c("Men", "Women"),
                     values = c("darkblue", "darkred")) +
  xlab("Years of education") +
  ylab("Predicted Income") +
  theme_bw() +
  theme(legend.title= element_blank())
```


Predicted Income
30000
25000
20000
15000
10000
0
5
10
15
20
Years of education
Men
Women

Adding more categorical variables

- When we want to add categorical variables that can take multiple values, we use `factor()`
- R automatically omits one group as reference

```
model4 <- lm(relig~educ+sex+factor(marital),gss)
stargazer(model4, type = "text",
          header=FALSE,
          title = "The association between education, sex, marital status and religion",
          digits = 3,
          single.row = T,
          omit.stat = c("rsq","f","ser"))
```

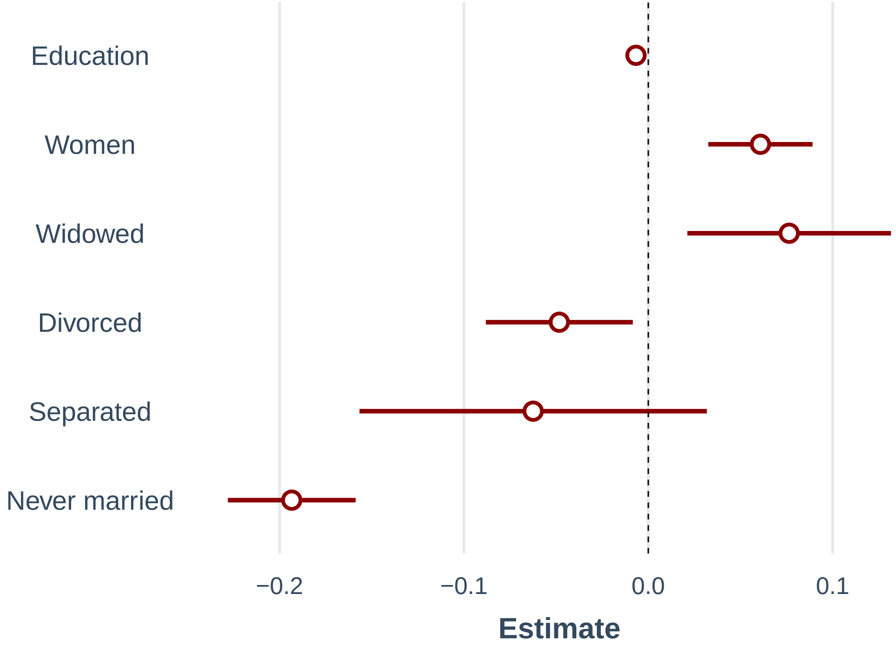
Table 6: The association between education, sex, marital status and religion

	Dependent variable:
	relig
educ	-0.007^{***} (0.003)
sex	0.061^{***} (0.014)
factor(marital)2	0.076^{***} (0.028)
factor(marital)3	-0.048^{**} (0.020)
factor(marital)4	-0.062 (0.048)
factor(marital)5	-0.193^{***} (0.018)
Constant	0.832^{***} (0.041)
Observations	3,878
Adjusted R^2	0.041
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Plot the Coefficients

- We use function `plot_summs` in R to create a coefficient plot
- Take a look at what `plot_summs` can do and its flexibilities here

```
library(jtools)
plot_summs(model4,
           colors="darkred",
           coefs = c("Education" = "educ", "Women" = "sex",
                     "Widowed" = "factor(marital)2",
                     "Divorced" = "factor(marital)3",
                     "Separated" = "factor(marital)4",
                     "Never married" = "factor(marital)5"))
```


Education
Women
Widowed
Divorced
Separated
Never married
−0.2
−0.1
0.0
0.1
Estimate