Week 10: Categorical Data

Wenhao Jiang

Department of Sociology
New York University

November 11, 2022

Quiz 3

Common Mistakes

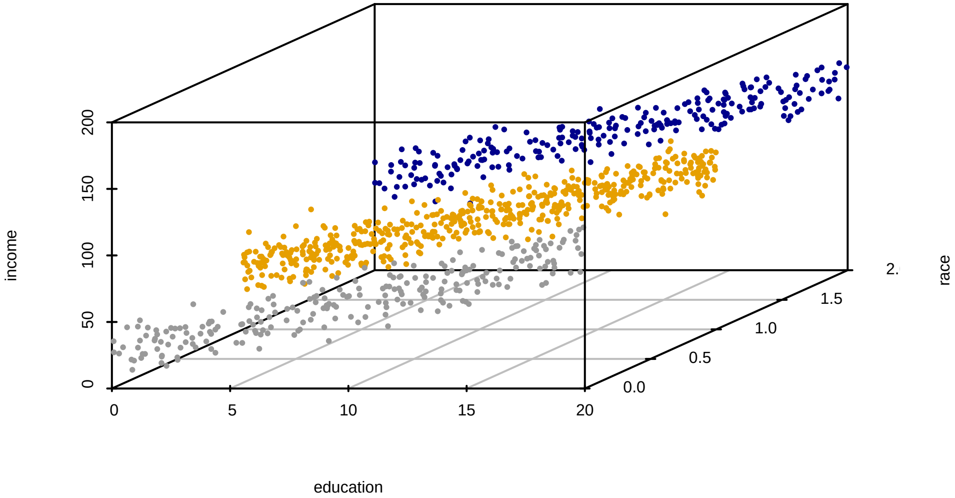
- True or False
- A significant regression coefficient for an independent variable X indicates that X is a cause of the dependent variable y

Read STATA output

- Look at the STATA output

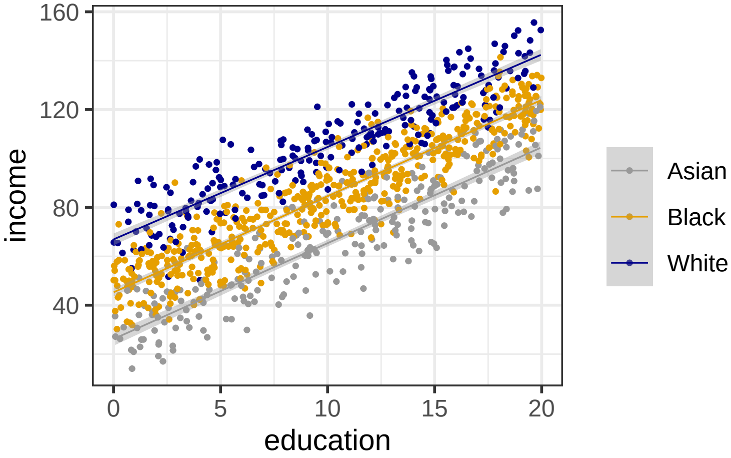
Intuition of Multivariate Regression

- When the second independent variable is a categorical variable with three possible values; slopes do not differ by the second independent variable



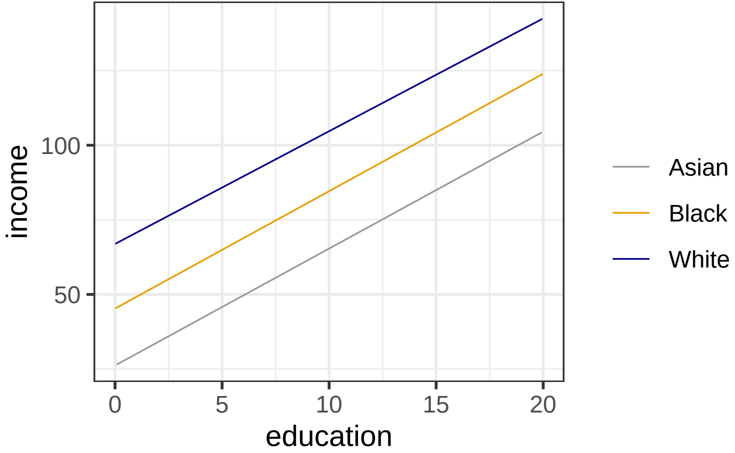
Intuition of Multivariate Regression

- We can visualize the 3-D plot by a 2-D scatterplot



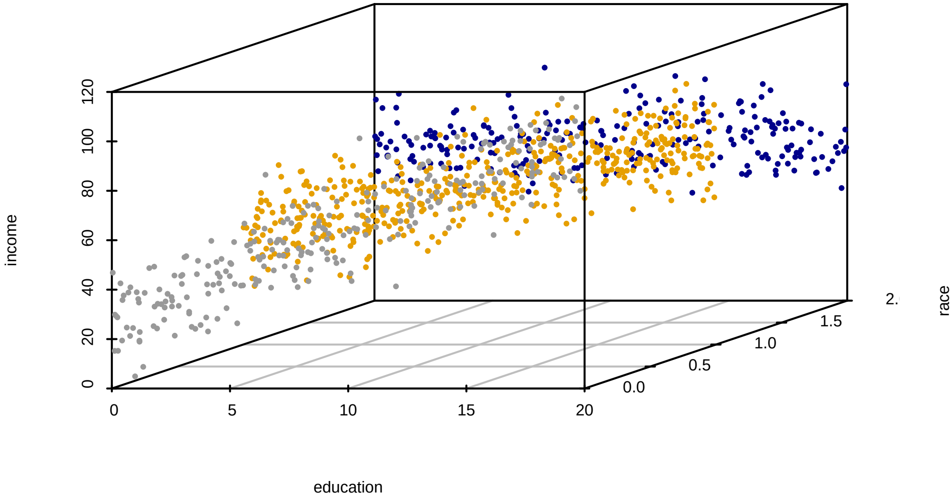
Intuition of Multivariate Regression

- Without the hypothetical points, we get similar lines as in the quiz



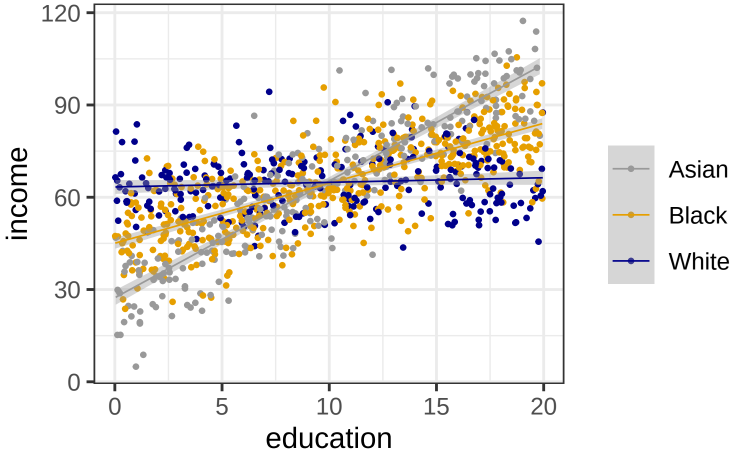
Intuition of Multivariate Regression

- When the second independent variable is a categorical variable with three possible values; slopes differ by the second independent variable



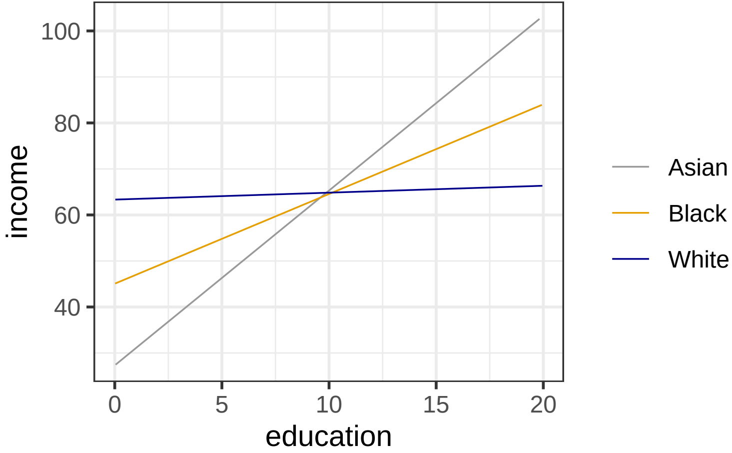
Intuition of Multivariate Regression

- We can visualize the 3-D plot by a 2-D scatterplot



Intuition of Multivariate Regression

- Without the hypothetical points, we get similar lines as in the quiz



Categorical Data

Basics

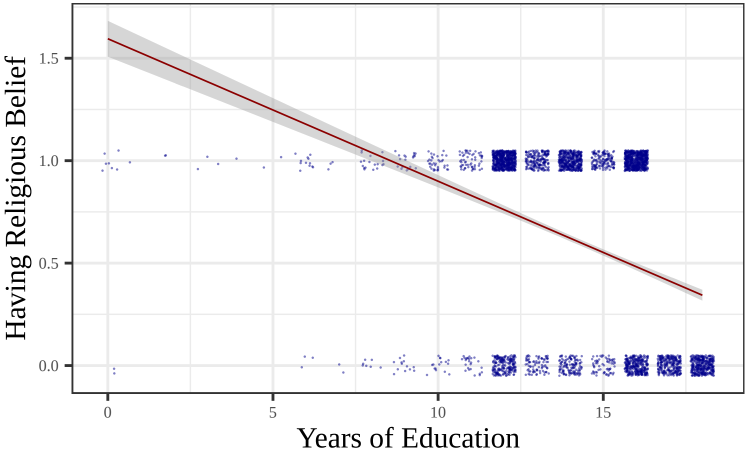
- While we already talked much about quantitative/numeric/continuous data (e.g., income), we have not discussed much about categorical data
- In statistics, a categorical variable is a variable that can take on one of a limited, and usually fixed, number of possible values, assigning each unit of observation to a particular group or nominal category on the basis of some qualitative property
  - Typical examples are gender, race, class (working class, middle class, upper class), and religious preferences
  - Sometimes these variables are the core of contemporary sociology

Basics

- While we already talked much about quantitative/numeric/continuous data (e.g., income), we have not discussed much about categorical data
- In statistics, a categorical variable is a variable that can take on one of a limited, and usually fixed, number of possible values, assigning each unit of observation to a particular group or nominal category on the basis of some qualitative property
  - Typical examples are gender, race, class (working class, middle class, upper class), and religious preferences
  - Sometimes these variables are the core of contemporary sociology
- why do we specifically care about categorical data beyond numeric data?

Basics

- We may be interested in the association between years of education and religious belief (1=having some belief; 0=no religious belief)



Basics

- We may be interested in the association between years of education and religious belief
- When we treat religious belief, a categorical variable, as a continuous variable, we may have fitted values that do not make sense in reality (e.g., $\hat{y} > 1$, or $\hat{y} < 0$)
- We want predictions to be bounded within 0 and 1
- We will go beyond OLS in the following weeks to address these particular scenarios stemming from categorical data

Basics

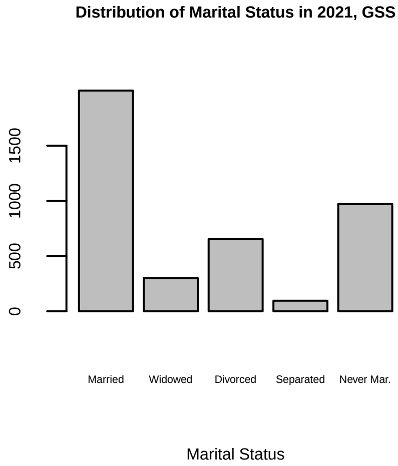
- How do we describe categorical data?
- For a single variable
  - For numeric/quantitative data, we use histogram

Basics

- How do we describe categorical data?
- For a single variable
  - For numeric/quantitative data, we use histogram
  - For categorical data, we use **barplot**
- We use **barplot** to describe the distribution of categorical values (e.g., gender, race, marital status)

Barplot

- Barplots describe the distribution of categorical data



Basics

- How do we describe categorical data?
- For a single variable
  - For numeric/quantitative data, we use histogram

Basics

- How do we describe categorical data?
- For a single variable
  - For numeric/quantitative data, we use histogram
  - For categorical data, we use **barplot**
- We use **barplot** to describe the distribution of categorical values (e.g., gender, race, marital status)
- For two categorical variables (e.g., gender current religious belief)
  - We use a cross-table to summarize the relationship

Basics

- The proportion of men and women who have and have no religious belief

```
gss %>%
  filter(!is.na(sex)) %>%
  group_by(sex) %>%
  summarize(religious = mean(relig,na.rm=T),
            nonreligious=mean(1-relig,na.rm=T))
```

```
## # A tibble: 2 x 3
##     sex religious nonreligious
##   <int>     <dbl>        <dbl>
## 1     1     0.680        0.320
## 2     2     0.745        0.255
```

t-test

- Suppose we want to know the gender difference in religious belief
- This is a typical problem of testing whether **two samples** differ in a proportion

t-test

- Suppose we want to know the gender difference in religious belief
- This is a typical problem of testing whether **two samples** differ in a proportion
- The point estimate is $\hat{p}_w - \hat{p}_m$
- Under the null hypothesis $H_0 : p_m = p_w = p$, the estimated $SE_{\hat{p}_m - \hat{p}_w} = \sqrt{\frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_m} + \frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_w}}$, where
- $\hat{p}_{pooled} = \frac{\hat{p}_m \times n_m + \hat{p}_w \times n_w}{n_m + n_w}$

t-test

- Now in 2021 GSS, there are 1736 men and 2204 women
- $\hat{p}_{pooled} = \frac{\hat{p}_m \times n_m + \hat{p}_w \times n_w}{n_m + n_w} = \frac{0.680 \times 1736 + 0.745 \times 2204}{1736 + 2204} = 0.716$
- $SE_{\hat{p}_m - \hat{p}_w} = \sqrt{\frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_m} + \frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_w}} = 0.014$
- The point estimate $\hat{p}_w - \hat{p}_m = 0.745 - 0.680 = 0.065$

t-test

- Now in 2021 GSS, there are 1736 men and 2204 women
- $\hat{p}_{pooled} = \frac{\hat{p}_m \times n_m + \hat{p}_w \times n_w}{n_m + n_w} = \frac{0.680 \times 1736 + 0.745 \times 2204}{1736 + 2204} = 0.716$
- $SE_{\hat{p}_m - \hat{p}_w} = \sqrt{\frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_m} + \frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_w}} = 0.014$
- The point estimate $\hat{p}_w - \hat{p}_m = 0.745 - 0.680 = 0.065$
- $t = \frac{0.065 - 0}{0.0145} = 4.64$
- Do we reject the null hypothesis?

Regression

- The t-test produces the exact same estimates as OLS regression
- $relig_i = \hat{\beta}_0 + \hat{\beta}_1 gender_i + e_i$

Table 1: The association between gender and religious belief

	Dependent variable:
	relig
sex	0.065^{***}
	(0.014)
Constant	0.615^{***}
	(0.024)
Observations	3,903
Adjusted R^2	0.005
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

t=(0.065-0)/0.014=4.4
Do you reject the null? Yes, because t>1.96

compare z with 1.96
z=1.96, p=0.05, so z> 1.96, p<0.05

R Operations

Read Data

```r
## set your working directory - you should set your own unique one!
setwd("~/Dropbox/Teaching/SOCUA-302/Week 8")

## read csv data - this is 2021 GSS data
gss <- read.csv("GSS_SOCUA_W8.csv")
```

Barplot

```r
## create a count summary in each category by function `table`
counts <- table(gss$marital)

## barplot
barplot(counts, main="Distribution of Marital Status in 2021, GSS",
   xlab="Marital Status",
   names.arg=c("Married", "Widowed", "Divorced", "Separated",
               "Never Mar."),
   cex.lab=0.5, cex.axis=0.5, cex.main=0.5, cex.sub=0.5, cex.names=0.32)
```

Cross Table

- The proportion of men and women who have and have no religious belief

```
## recode religion
gss[which(gss$relig!=4),
    "relig"] <- 1
gss[which(gss$relig==4),
    "relig"] <- 0
## cross table
gss %>%
  filter(!is.na(sex)) %>%
  group_by(sex) %>%
  summarize(religious = mean(relig,na.rm=T),
            nonreligious=mean(1-relig,na.rm=T))
```

```
## # A tibble: 2 x 3
##     sex religious nonreligious
##   <int>     <dbl>        <dbl>
## 1     1     0.680        0.320
## 2     2     0.745        0.255
```

Regression

- The t-test produces the exact same estimates as OLS regression
- $relig_i = \hat{\beta}_0 + \hat{\beta}_1 gender_i + e_i$

```r
library(stargazer)
model <- lm(relig~sex,gss)
stargazer(model, out = "text",
          single.row = F,
          header=FALSE,
          title = "The association between gender and religious belief",
          digits = 3,
          omit.stat = c("rsq","f","ser"))
```