# Week 13: Categorical Data III

Wenhao Jiang

Department of Sociology
New York University

December 2, 2022

# Categorical Data as Dependent Variable

# Categorical Data as Independent Variable - Basics

- We talked about the case where categorical data are independent variables two weeks ago
- Instead of (incorrectly) assuming an additive model for categorical data (e.g. race) in a regression model e.g., $Abscale_i = \hat{\beta}_0 + \hat{\beta}_1 edu_i + \hat{\beta}_2 women_i + \hat{\beta}_3 race_i + e_i$, we transform race into 5 "dummy variables" (dummy means a binary variable created from categorical variables), `white`, `black`, `hispanic`, `asian`, `others`, and omit one dummy group as the reference group

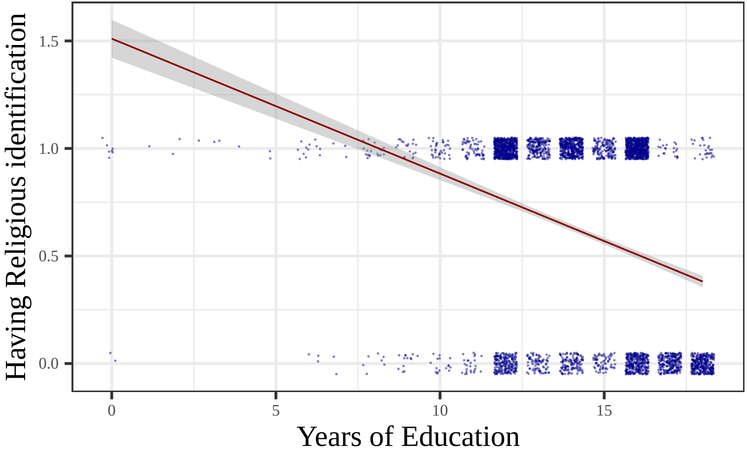
# Categorical Data as Dependent Variable - Basics

- In many cases, categorical data are dependent variables, such as a binary categorical variable indicating whether the respondent supports legal abortion or has any religious identification

# Categorical Data as Dependent Variable - Basics

- In many cases, categorical data are dependent variables, such as a binary categorical variable indicating whether the respondent supports legal abortion or has any religious identification
- We may be interested in the association between years of education and religious identification (1=having some identification; 0=no religious identification)
- It is however incorrect to use the OLS model $relig_i = \hat{\beta}_0 + \hat{\beta}_1 edu_i + e_i$ to estimate the association
- Why?

# Categorical Data as Dependent Variable - Basics



# Categorical Data as Dependent Variable - Motivation

- When we treat religious identification, a categorical variable, as a continuous variable and apply an OLS model in estimation, we may have fitted values that do not make sense in reality (e.g., $\hat{y} > 1$, or $\hat{y} < 0$)
- We want predictions/fitted values to be bounded within 0 and 1, i.e., for a given set of attributes (education), the probability for the respondent to have some religious identification is within 0 and 1

# Categorical Data as Dependent Variable - Motivation

- When we treat religious identification, a categorical variable, as a continuous variable and apply an OLS model in estimation, we may have fitted values that do not make sense in reality (e.g., $\hat{y} > 1$, or $\hat{y} < 0$)
- We want predictions/fitted values to be bounded within 0 and 1, i.e., for a given set of attributes (education), the probability for the respondent to have some religious identification is within 0 and 1
- This desired property of "boundedness" motivates **logistic** models and **logit** transformations
- (From Lecture) The assumptions of OLS requires that the error terms are normally distributed with a mean of zero and a constant variance, which is impossible when the dependent variable is e.g. dichotomous

# Categorical Data as Dependent Variable - Logit Transformation

- Before any transformation, we will need to first think of the predicted/fitted dependent variable as the **probabilities** of having religious identification ($P_{relig}$), given some observed characteristic such as education
  - This is analogous to OLS, where the estimated linear line models the predicted "trend"/"value" of some dependent variable

# Categorical Data as Dependent Variable - Logit Transformation

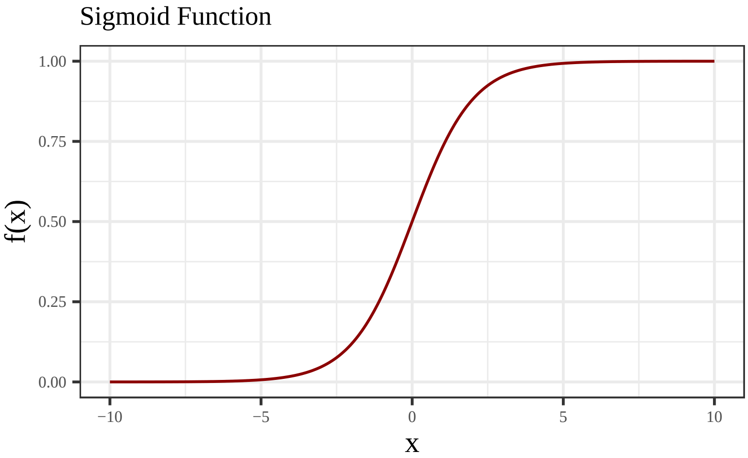
- Before any transformation, we will need to first think of the predicted/fitted dependent variable as the **probabilities** of having religious identification ($P_{relig}$), given some observed characteristic such as education
  - This is analogous to OLS, where the estimated linear line models the predicted "trend"/"value" of some dependent variable
- Logit transformation: $\log(\frac{P_{relig}}{1-P_{relig}}) = \hat{\beta}_0 + \hat{\beta}_1 edu_i$

# Categorical Data as Dependent Variable - Logit Transformation

- Before any transformation, we will need to first think of the predicted/fitted dependent variable as the **probabilities** of having religious identification ($P_{relig}$), given some observed characteristic such as education
  - This is analogous to OLS, where the estimated linear line models the predicted "trend"/"value" of some dependent variable
- Logit transformation: $\log(\frac{P_{relig}}{1-P_{relig}}) = \hat{\beta}_0 + \hat{\beta}_1 edu_i$
- But why in this specific form?

# Categorical Data as Dependent Variable - Sigmoid Function

- The logit transformation originates from the Sigmoid Function
- $f(x) = \frac{1}{1+e^{-x}}$



# Categorical Data as Dependent Variable - Sigmoid Function

- The logit transformation originates from the Sigmoid Function
- $f(x) = \frac{1}{1+e^{-x}}$
- The Sigmoid Function has two desired properties
  - 1. $f(x)$ is bounded within 0 and 1
  - 2. $x$ has no limit

# Categorical Data as Dependent Variable - Sigmoid Function

- The logit transformation originates from the Sigmoid Function
- $f(x) = \frac{1}{1+e^{-x}}$
- The Sigmoid Function has two desired properties
  - 1. $f(x)$ is bounded within 0 and 1
  - 2. $x$ has no limit
- The Sigmoid Function is therefore a good candidate to model the **probabilities** of some categorical dependent variable (e.g., having religious identification $P_{relig}$), given some observed characteristic such as education

# Categorical Data as Dependent Variable - Sigmoid Function

- The logit transformation originates from the Sigmoid Function
- $f(x) = \frac{1}{1+e^{-x}}$
- The Sigmoid Function has two desired properties
  - 1. $f(x)$ is bounded within 0 and 1
  - 2. $x$ has no limit
- The Sigmoid Function is therefore a good candidate to model the **probabilities** of some categorical dependent variable (e.g., having religious identification $P_{relig}$), given some observed characteristic such as education
- We borrow the idea from OLS and estimate $P_{relig_i} = \frac{1}{1+e^{-(\hat{\beta_0}+\hat{\beta_1}edu_i)}}$

# Categorical Data as Dependent Variable - Sigmoid Function

- The logit transformation originates from the Sigmoid Function
- $f(x) = \frac{1}{1+e^{-x}}$
- The Sigmoid Function has two desired properties
  - 1. $f(x)$ is bounded within 0 and 1
  - 2. $x$ has no limit
- The Sigmoid Function is therefore a good candidate to model the **probabilities** of some categorical dependent variable (e.g., having religious identification $P_{relig}$), given some observed characteristic such as education
- We borrow the idea from OLS and estimate $P_{relig_i} = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 edu_i)}}$
- With some algebra
- $1 - P_{relig_i} = 1 - \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 edu_i)}} = \frac{e^{-(\hat{\beta}_0+\hat{\beta}_1 edu_i)}}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 edu_i)}}$
- $\frac{P_{relig_i}}{1-P_{relig_i}} = \frac{1}{e^{-(\hat{\beta}_0+\hat{\beta}_1 edu_i)}}$
- $\log(\frac{P_{relig_i}}{1-P_{relig_i}}) = \log(\frac{1}{e^{-(\hat{\beta}_0+\hat{\beta}_1 edu_i)}}) = \hat{\beta}_0 + \hat{\beta}_1 edu_i$

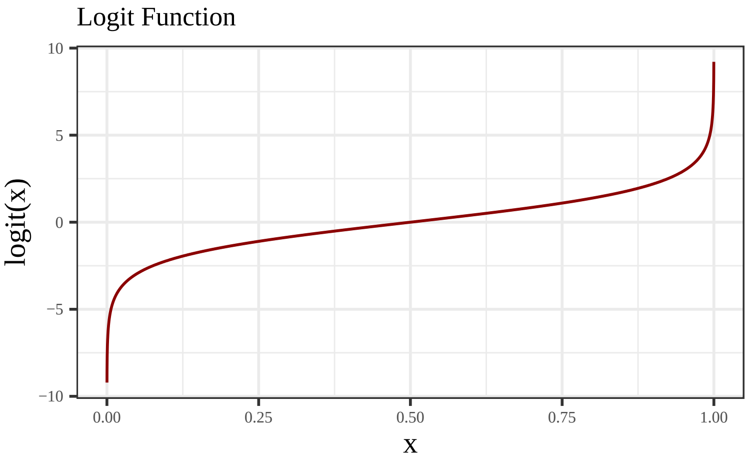
# Categorical Data as Dependent Variable - Sigmoid Function

- The logit transformation originates from the Sigmoid Function
- $f(x) = \frac{1}{1+e^{-x}}$
- The Sigmoid Function has two desired properties
  - 1. $f(x)$ is bounded within 0 and 1
  - 2. $x$ has no limit
- The Sigmoid Function is therefore a good candidate to model the **probabilities** of some categorical dependent variable (e.g., having religious identification $P_{relig}$), given some observed characteristic such as education
- We borrow the idea from OLS and estimate $P_{relig_i} = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 edu_i)}}$
- With some algebra
- $1 - P_{relig_i} = 1 - \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 edu_i)}} = \frac{e^{-(\hat{\beta}_0+\hat{\beta}_1 edu_i)}}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 edu_i)}}$
- $\frac{P_{relig_i}}{1-P_{relig_i}} = \frac{1}{e^{-(\hat{\beta}_0+\hat{\beta}_1 edu_i)}}$
- $\log(\frac{P_{relig_i}}{1-P_{relig_i}}) = \log(\frac{1}{e^{-(\hat{\beta}_0+\hat{\beta}_1 edu_i)}}) = \hat{\beta}_0 + \hat{\beta}_1 edu_i$
- This is the logit transformation!

# Categorical Data as Dependent Variable - Sigmoid and Logit

- Indeed, Sigmoid function and logit function are inverse functions for each other
- Sigmoid function: $y = \frac{1}{1+e^{-x}}$
- Inverse of Sigmoid function: $x = \frac{1}{1+e^{-y}} \rightarrow y = \log(\frac{x}{1-x})$
- $\frac{1}{1+e^{-x}}$ is bounded within 0 and 1. Inversely, the $x$ in $\log(\frac{x}{1-x})$ is bounded within 0 and 1
- Note that the boundedness corresponds to $P_{relig_i}$. $\log(\frac{P_{relig_i}}{1-P_{relig_i}})$ is unbounded

# Categorical Data as Dependent Variable - Logit



# Categorical Data as Dependent Variable - Odds and Odds Ratio

- We call the term $\frac{P_{relig_i}}{1-P_{relig_i}}$ in the log() function "odds" (probability of "event" divided by probability of no "event")
- Odds $\frac{P_{relig_i}}{1-P_{relig_i}} = exp(\hat{\beta}_0 + \hat{\beta}_1 edu_i)$

# Categorical Data as Dependent Variable - Odds and Odds Ratio

- We call the term $\frac{P_{relig_i}}{1-P_{relig_i}}$ in the log() function "odds" (probability of "event" divided by probability of no "event")
- Odds $\frac{P_{relig_i}}{1-P_{relig_i}} = exp(\hat{\beta}_0 + \hat{\beta}_1 edu_i)$
- Odds ratio describes the (multiplicative) change of the odds when the independent variable of interest changes by 1 unit (with other independent variables remain constant in the case of multivariate regression)
- Odds ratio $= \frac{Odds_{edu_i+1}}{Odds_{edu_i}} = \frac{exp(\hat{\beta}_0+\hat{\beta}_1(edu_i+1))}{exp(\hat{\beta}_0+\hat{\beta}_1 edu_i)} = exp(\hat{\beta}_1)$
- You may find it analogous to OLS, where $\hat{\beta}_1$ describes the (additive) change of the dependent variable when the independent variable change by 1 unit

# Categorical Data as Dependent Variable - Interpret Results

- We estimate the logistic regression model
- $\log(\frac{P_{relig_i}}{1-P_{relig_i}}) = \hat{\beta}_0 + \hat{\beta}_1 edu_i$

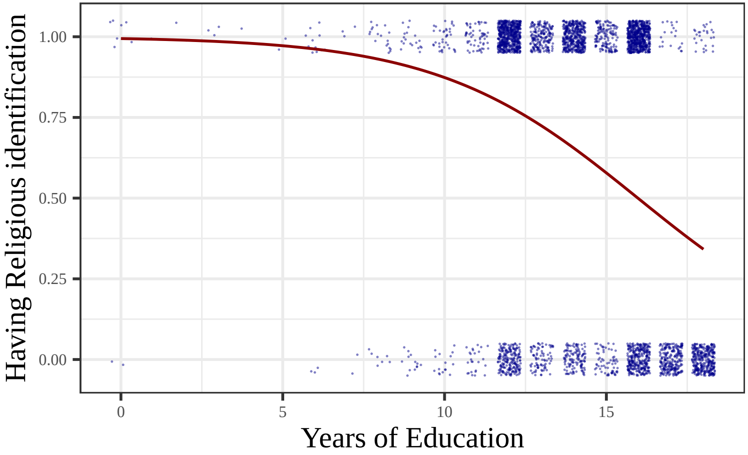
Table 1: The association between education and religious identification

| | *Dependent variable:* |
|---|---|
| | relig |
| educ | $-0.324^{***}$ |
| | (0.017) |
| Constant | $5.170^{***}$ |
| | (0.259) |
| Observations | 3,601 |
| Log Likelihood | $-2,197.671$ |
| Akaike Inf. Crit. | 4,399.342 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

- $\log(\frac{P_{relig_i}}{1-P_{relig_i}}) = \hat{\beta}_0 + \hat{\beta}_1 edu_i = 4.94 - 0.31 \times educ$
- A year increase in education is associated with 0.31-unit decrease in $\log(\frac{P_{relig_i}}{1-P_{relig_i}})$
- The associated change of $P_{relig_i}$ is non-linear

# Categorical Data as Dependent Variable - Predicted Probabilities

- After logit transformation $\frac{P_{relig_i}}{1-P_{relig_i}} = exp(\hat{\beta}_0 + \hat{\beta}_1 edu_i)$, the predicted $P_{relig}$ from Maximum Likelihood Estimation relative to years of education becomes



# Categorical Data as Dependent Variable - R Operations

- $\log(\frac{P_{relig_i}}{1-P_{relig_i}}) = \hat{\beta}_0 + \hat{\beta}_1 edu_i$

```
logit <- glm(relig ~ educ, data = gss, family = "binomial")
library(stargazer)
stargazer(logit, type = "text",
          single.row = F,
          header=FALSE,
          title = "The association between education and religious identification",
          digits = 3)
```

# Quiz 4 Reviews

# Exercise for Quiz 4

- How to plot predicted values from OLS when the dependent variable is categorical?
  - Note that in formal statistics, we do not use OLS when estimating categorical dependent variable. We use it here for simplicity

# Exercise for Quiz 4

- How to calculate Odds, Logit, and probability?