# Dichotomous Predictors, Non-Linearity, and Data Transformations

Dr. Michael Fix
mfix@gsu.edu

Georgia State University

23 February 2023

# Variable Types Revisited

- Four types of variables:
  1. Nominal ("Factors")
  2. Ordinal
  3. Interval
  4. Ratio
- In the context of OLS: Which work as DVs? Which work as IVs?

# Dummy Variables

- A term that gets used a lot to mean many things...
- Naturally dichotomous things
- Simplified categorizations
- "Factor" variables
- Ordinal variables (treated as "factors")

# Dummy Variable Coding

- The term "dummy" variable is associate with a $\{0,1\}$ coding scale
- e.g.

$$\texttt{woman} = \begin{cases} 0 \text{ if man} \\ 1 \text{ if woman} \end{cases}$$

- Why $\{0,1\}$?

# Dummy Variable Coding

- Two reasons:
  1. Math (will talk about this in a minute)
  2. Software
- Theoretically, as this variables have no meaningful ordering among their values, the assigned numbers do not matter
- **However**, you should always *name* the variable to correspond outcome of interest and set that outcome equal to 1.

# Bivariate Regression with Dichotomous $X$s

## The Math

- For

$$Y_i = \beta_0 + \beta_1 D_i + u_i$$

- we have

$$\mathrm{E}(Y|D=0) = \beta_0$$

- and

$$\mathrm{E}(Y|D=1) = \beta_0 + \beta_1.$$

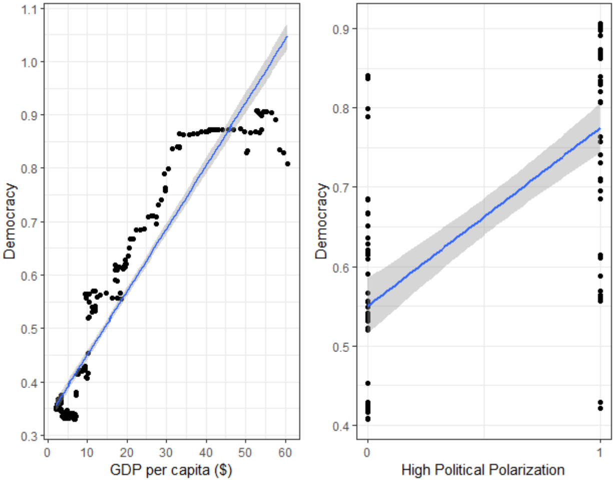
# Bivariate Regression with Dichotomous $X$s

## The Intuition

- Intuitively, we think of OLS as "fitting a line"
- This breaks down with a dummy IV:

# Bivariate Regression with Dichotomous $X$s

## The Intuition



# Regression with Dichotomous and Continuous $X$

## The Math

- For,

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + u_i$$

- we have

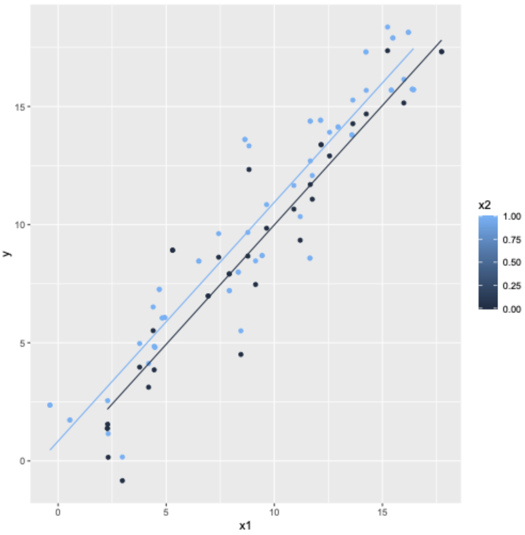
$$\mathrm{E}(Y|X, D = 0) = \beta_0 + \beta_2 X_i$$

- and

$$\mathrm{E}(Y|X, D = 1) = (\beta_0 + \beta_1) + \beta_2 X_i$$

# Regression with Dichotomous and Continuous $X$

## The Intuition



# Regression with Dichotomous and Continuous $X$

## The Intuition

- As the prior slide shows, effectively the dummy variable represents an intercept shift.
- The estimated effect of $X_i$ on $Y_i$ ($\beta_2$) determines the slope of the regression line and is unchanged based on the value of $D_i$.
- BUT, the intercept of the regression line shifts based on the value of $D_i$
  - When $D_i = 0$, the intercept is $\beta_0$
  - When $D_i = 1$, the intercept is $(\beta_0 + \beta_1)$

# Multiple Dummies

## The Math

- For

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + ... + \beta_\ell D_{\ell i} + u_i$$

- We have

$$\mathrm{E}(Y|D_k = 0)\,\forall\, k \in \ell = \beta_0$$

- Otherwise,

$$\mathrm{E}(Y) = \beta_0 + \sum_{k=1}^{\ell} \beta_k \,\forall\, k\, s.t.\, D_k = 1$$

# Multiple Dummies

## An Important Note

- Where the $D_\ell$ are *mutually exclusive and exhaustive*:
    - This is usually the case for so called "factor" variables
    - The expected values are the same as the within-group means.
    - Identification requires that we either
        - omit a "reference category," or
        - omit $\beta_0$.

# Multiple Dummies

## Ordinal Variables: A Special Case

- Suppose we have:

$$\texttt{party} = \begin{cases} -2 = \text{Strong Democrat} \\ -1 = \text{Weak Democrat} \\ 0 = \text{Independent} \\ 1 = \text{Weak Republican} \\ 2 = \text{Strong Republican} \end{cases}$$

# Multiple Dummies

## Ordinal Variables: A Special Case

- We could estimate:

$$Y_i = \beta_0 + \beta_1(\texttt{party}_i) + u_i$$

- Effectively treating an ordinal variable as if it was continuous

# Multiple Dummies

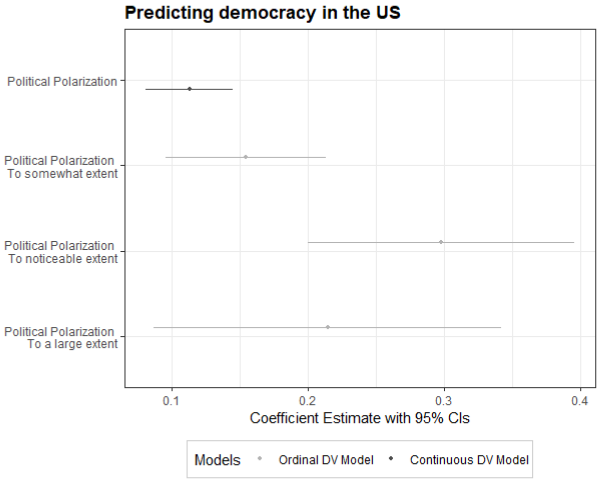
## Ordinal Variables: A Special Case

- Alternatively, we could convert it to a series of dummies

$$Y_i = \beta_0 + \beta_1(\texttt{strongdem}_i) + \beta_2(\texttt{weakdem}_i) + \\ \beta_3(\texttt{weakgop}_i) + \beta_4(\texttt{stronggop}_i) + u_i$$

- Note the excluded "reference category" as the outcomes are mutually exclusive and exhaustive

# Ordinal Variables: A Comparison



# Why Transform Variables?

- Normality (of $u_i$s)
- Linearity
- Additivity
- Interpretation / Model Specification

Note: John Fox has some really helpful slides online that you might find useful for more depth on various transformations.

# Monotonic Transformations

"Family of Powers and Roots"

| Transformation | $p$ | $f(X)$ | Fox's $f(X)$ |
|---|---|---|---|
| Cube | 3 | $X^3$ | $\frac{X^3-1}{3}$ |
| Square | 2 | $X^2$ | $\frac{X^2-1}{2}$ |
| (None/Identity) | (1) | $(X)$ | $(X)$ |
| Square Root | $\frac{1}{2}$ | $\sqrt{X}$ | $2(\sqrt{X}-1)$ |
| Cube Root | $\frac{1}{3}$ | $\sqrt[3]{X}$ | $3(\sqrt[3]{X}-1)$ |
| Log | 0 (sort of) | $\ln(X)$ | $\ln(X)$ |
| Inverse Cube Root | -$\frac{1}{3}$ | $\frac{1}{\sqrt[3]{X}}$ | $\frac{\left(\frac{1}{\sqrt[3]{X}}-1\right)}{-\frac{1}{3}}$ |
| Inverse Square Root | -$\frac{1}{2}$ | $\frac{1}{\sqrt{X}}$ | $\frac{\left(\frac{1}{\sqrt{X}}-1\right)}{-\frac{1}{2}}$ |
| Inverse | -1 | $\frac{1}{X}$ | $\frac{\left(\frac{1}{X}-1\right)}{-1}$ |
| Inverse Square | -2 | $\frac{1}{X^2}$ | $\frac{\left(\frac{1}{X^2}-1\right)}{-2}$ |
| Inverse Cube | -3 | $\frac{1}{X^3}$ | $\frac{\left(\frac{1}{X^3}-1\right)}{-3}$ |

## A General Rule

**Using higher-order power transformations (e.g. squares, cubes, etc.) "inflates" large values and "compresses" small ones; conversely, using lower-order power transformations (logs, etc.) "compresses" large values and "inflates" (or "expands") smaller ones.**

# Nonmonotonicity

Simple solution: Polynomials

- Second-order / quadratic:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

- Third-order / cubic:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i$$

- $p$th-order:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + ... + \beta_p X_i^p + u_i$$

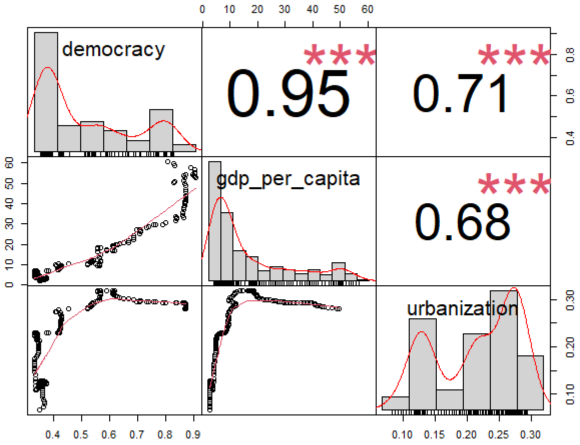
# How Do You Know?

**Plots are your best friend!**

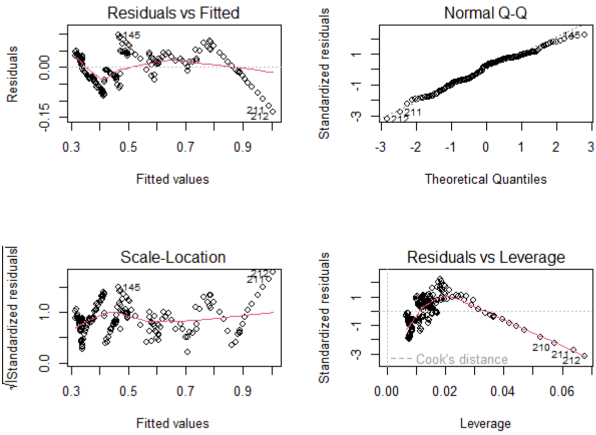
# How Do You Know? Toy Model Example

```r
## Load your data ----
my_data <- readRDS("data/vdem12.rds")

# Let's change names of some of these variables for the sake of simplicity
# I am also subsetting it to only US
us_data <- my_data |>
  filter(country_name == "United States of America") |>
  rename(democracy = v2x_polyarchy,
         gdp_per_capita = e_gdppc,
         urbanization = e_miurbani,
         regime = v2x_regime,
         polarization = v2cacamps,
         polarization_ordinal = v2cacamps_ord) |>
  mutate(regime_binary = ifelse(regime %in% c(2,3), 1, 0),
         high_polarization = ifelse(polarization >= -1, 1, 0))

# Use correlation matrix to see the relationship between variables
chart.Correlation(us_data |> select(democracy, gdp_per_capita, urbanization))

# This is our toy model
multiple <- lm(democracy ~ gdp_per_capita + urbanization, data = us_data)

# Use plot() to get diagnostics
plot(multiple)
```

# First, check your variables



# Model diagnostics using *plot()*



# Residual distribution and density

```r
# Residual plot with histogram
hist(multiple$residuals, freq = F, xaxt = "n", xlab = "", ylab = "", main = "")
par(new = T) # sets graphical parameters so that I can plot histogram and density plots
plot(density(resid(multiple)))
```