Applied Statistical Analysis I

Multiple linear regression

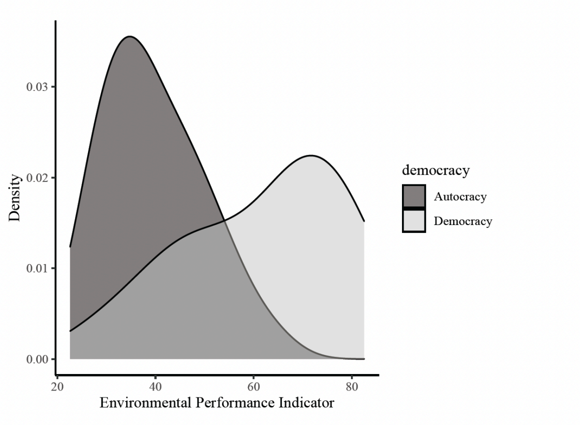
Elena Karagianni, PhD Candidate
karagiae@tcd.ie

November 12, 2025

Department of Political Science, Trinity College Dublin

Categorical Independent variables

What is the reference category?



$$\text{Environmental Performance}_i = \alpha + \beta_1 \times \text{Regime Type}_i$$

How to Create a Categorical Variable

```r
# Load Quality of Government data
qog_data <- read_csv("https://www.qogdata.pol.gu.se/data/qog_bas_cs_jan21.csv")

# Generate a dummy variable for regime type as factor variable ("democracy")
# vdem_polyarchy ranges between 0 and 1: cutoff at 0.7
# Countries with score equal or above 0.7 are democracies, those below autocracies
qog_data$democracy <- factor(ifelse(qog_data$vdem_polyarchy >= 0.7, 1, 0))

# Define levels of democracy in factor variable
levels(qog_data$democracy) <- c("Autocracy", "Democracy")

# Summarize generated dummy variable
summary(qog_data$democracy)
```

```
summary(qog_data$democracy)
Autocracy Democracy      NA's
      119        54        21
```

How to Create a Categorical Variable

```r
# Generate dummy variable for regime type as factor variable ("autocracy")
qog_data$autocracy <- factor(ifelse(qog_data$vdem_polyarchy < 0.7, 1, 0))

# Define levels of autocracy is factor variable
levels(qog_data$autocracy) <- c("Democracy", "Autocracy")

# Print first 10 rows in dataset
head(qog_data[c("democracy", "autocracy")], 10)
```

```
 1 Autocracy Autocracy
 2 Autocracy Autocracy
 3 Autocracy Autocracy
 4 NA        NA
 5 Autocracy Autocracy
 6 NA        NA
 7 Autocracy Autocracy
 8 Democracy Democracy
 9 Democracy Democracy
10 Democracy Democracy
```

What is a Reference Category?

- What happens if we run the model with both dummy variables?

```
lm(formula = epi_epi ~ democracy + autocracy, data = qog_data)

Residuals:
    Min      1Q  Median      3Q     Max
-34.107  -8.860  -0.610   9.293  26.190

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   39.610      1.138   34.80   <2e-16 ***
democracy1    22.098      2.002   11.04   <2e-16 ***
autocracy1        NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- It violates the assumption of no perfect multicollinearity. One category needs to be excluded $\rightarrow$ the reference category.
- The interpretation of the model is relative to the reference category.

Binary Independent Variables

```
## Call:
## lm(epi_epi ~ democracy + income, data = qog_data)

## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.563  -6.502   0.498   6.773  20.198

## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         35.3027     1.1269  31.327  < 2e-16 ***
## democracyDemocracy  16.5270     1.8409   8.978 9.08e-16 ***
## income               3.5793     0.4266   8.390 2.92e-14 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 9.982 on 154 degrees of freedom
##   (37 observations deleted due to missingness)
## Multiple R-squared:  0.6175,Adjusted ## R-squared:  0.6126
## F-statistic: 124.3 on 2 and 154 DF,  p-value: < 2.2e-16
```

In comparison to autocracies (=reference category), democracies have a 16.52 scale point higher on the Environmental Performance Index, on average, under control of income.

Binary Independent Variables

$$\hat{Y}_i = \alpha + \beta_1 \text{Regime Type}_i + \beta_2 \text{Income}_i$$

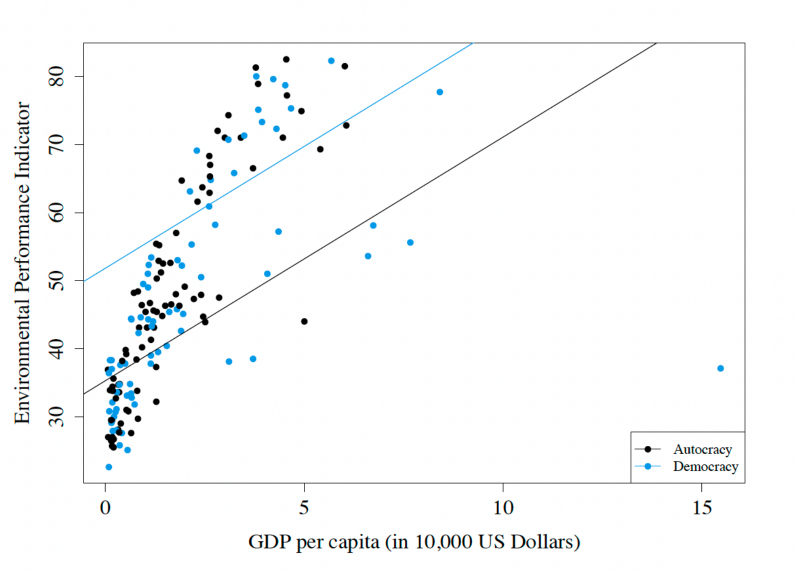
Model for Autocracies:

$$\begin{aligned}\hat{Y}_i &= 35.303 + (16.527 \times \text{Regime Type}_i) + (3.579 \times \text{Income}_i) \\ &= 35.303 + (16.527 \times 0) + (3.579 \times \text{Income}_i) \\ &= 35.303 + 3.579 \times \text{Income}_i\end{aligned}$$

Model for Democracies:

$$\begin{aligned}\hat{Y}_i &= 35.303 + (16.527 \times \text{Regime Type}_i) + (3.579 \times \text{Income}_i) \\ &= 35.303 + (16.527 \times 1) + (3.579 \times \text{Income}_i) \\ &= 51.83 + 3.579 \times \text{Income}_i\end{aligned}$$

Plotting the example



Using Dummy Variables for Multiple Categories

- Split a k-category into k-1 binary dummies.
- Interpretation is always **relative** to the baseline category.
- Suppose you analyze the effect of different social classes (lower, middle, upper) on income using $\hat{Y} = \hat{\beta}_1 D_1 + \hat{\beta}_2 D_2$:

	Dummy Variables		
Social Class	D_1	D_2	
lower	0	0	$\hat{Y} = \hat{\beta}_0$
middle	1	0	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1$
upper	0	1	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_2$

Using Dummy Variables for Multiple Categories

- What if we want to test the difference between middle and upper class?
- Cleverly construct dummy variables such that an estimated coefficient identifies this difference.

	Dummy Variables		
Social Class	$\tilde{D}_1$	$\tilde{D}_2$	
lower	0	0	$\hat{Y} = \hat{\beta}_0$
middle	1	0	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1$
upper	1	1	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2$

- When estimating $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1\tilde{D}_1 + \hat{\beta}_2\tilde{D}_2$ then the estimated coefficient of the second dummy, $\hat{\beta}_2$, represents (by design!) the difference between middle and upper class.

Interpretation of Multiple Categories

- α: expected value of Y when $X_k = 0$.
- β: expected average change in Y for $X = 1$, in comparison to reference category.