INSTITUTE
OF LAW,
POLITICS AND
DEVELOPMENT

Sant'Anna
School of Advanced Studies – Pisa

Introduction to R

Linear Regression

Prof. Kevin Koehler
kevin.koehler@santannapisa.it

What is a linear regression?

Sant'Anna
School of Advanced Studies – Pisa

What is a linear regression?

- Regression is a way to understand **relationships** between variables.

What is a linear regression?

- Regression is a way to understand **relationships** between variables.
- Example: How does **economic development** affect **democracy**?

What is a linear regression?

- Regression is a way to understand **relationships** between variables.
- Example: How does **economic development** affect **democracy**?
- We want to **predict** or **explain** one variable using another.

Key concepts

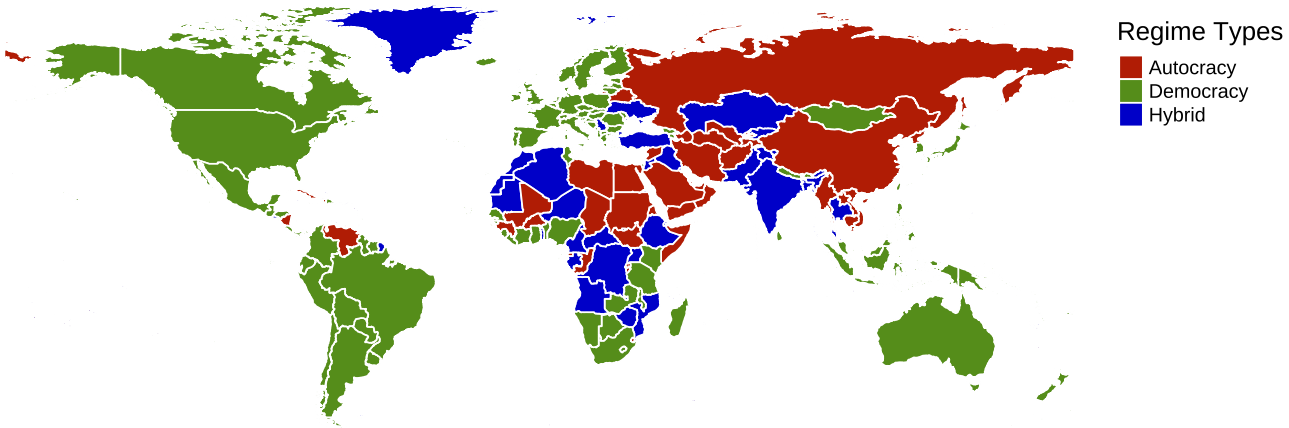
- **Dependent variable (Y)**: what you want to explain or predict (e.g., levels of democracy)
- **Independent variable (X)**: what you think influences Y (e.g., economic development)
- We write this as:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

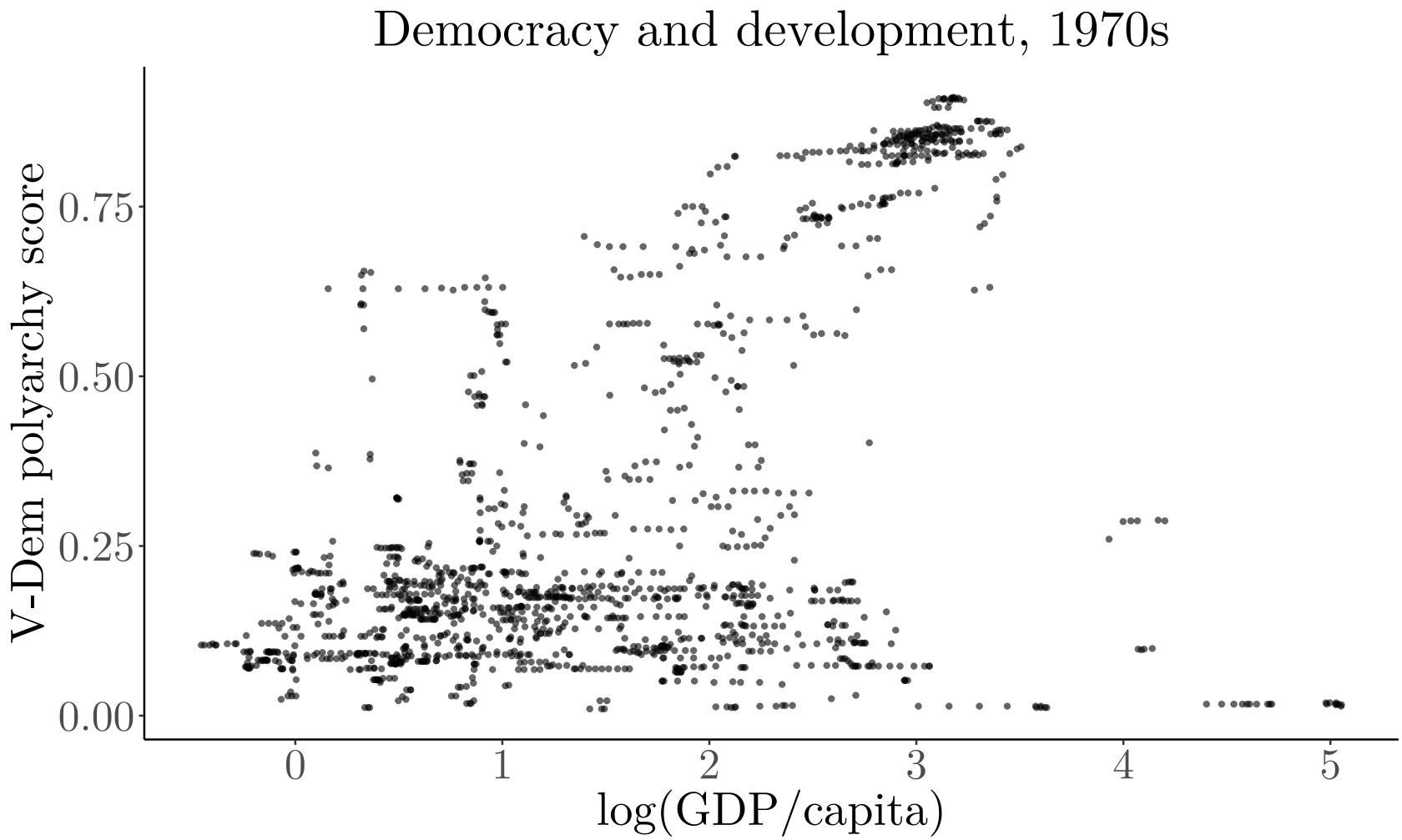
The usual example: Economic development and democracy

Why are some countries democratic while others are not?

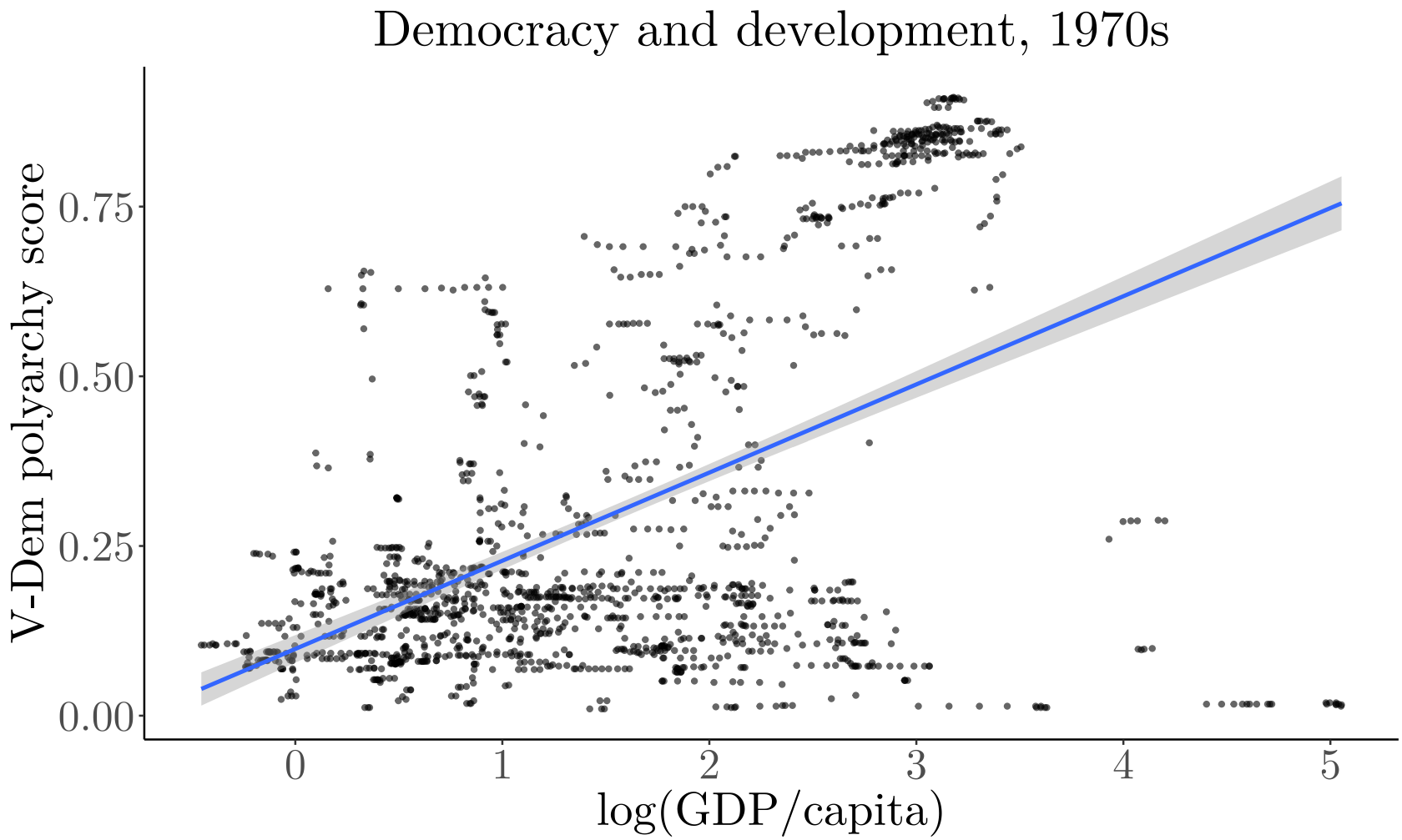
Regime types in 2023 (VDem)



Is Modernization Theory right?



Is Modernization Theory right?



How can we know?

Two basic aims

We need to know two things:

1. Whether economic development (GDP/capita) and democracy (V-Dem polyarchy scores) are related (and how strong this relationship is)

Two basic aims

We need to know two things:

1. Whether economic development (GDP/capita) and democracy (V-Dem polyarchy scores) are related (and how strong this relationship is)
2. How sure we can be that this relationship is not due to chance

Two basic aims

Ordinary Least Squares (OLS) regression allows us to answer these questions.

Two basic aims

Ordinary Least Squares (OLS) regression allows us to answer these questions.

In more technical language, we can calculate **point estimates** which tell us whether, how, and how strongly two variables are related, and **measures of (un)certainty** telling use how likely we are to find this relationship purely due to chance.

How can we calculate point estimates (coefficients)?

- **Goal**: Find the line that best fits the data.
- This line minimizes the distance between the observed values and the predicted values.
- The predicted value for observation i, $\hat{Y}_i$, is given by:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

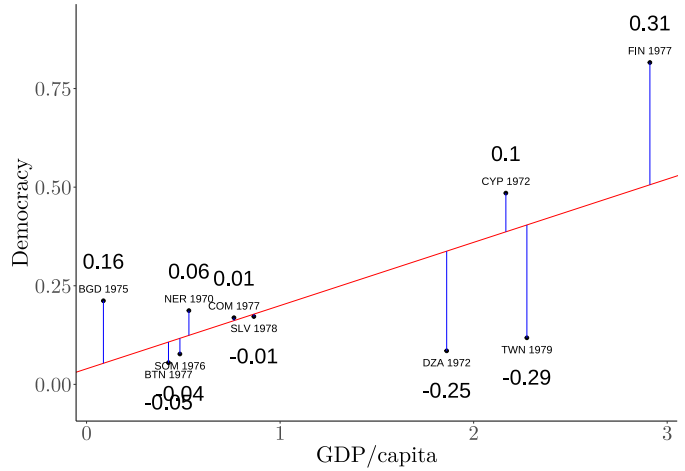
- The **residual** is the distance between our prediction ($\hat{Y}_i$) and the actual value (Y_i)

Graphical representation of residuals

```
set.seed(123)
plot <- data[sample(nrow(data), 10),]

model <- lm(democracy ~ gdp, data=plot)
plot$pred <- predict(model)
plot$dev <- plot$democracy - plot$pred

ggplot(plot, aes(x = gdp, y = democracy)) +
  geom_point() +
  geom_abline(intercept = coef(model)[1],
              slope = coef(model)[2],
              color = "red") +
  geom_segment(aes(xend = gdp,
                   yend = pred),
               color = "blue") +
  geom_text(aes(label = round(dev, 2),
                y = ifelse(dev>0,pred+dev+0.1,
                           pred+dev-0.1)),
            size=8) +
  geom_text(aes(label = paste(country_text_id,
                              year,sep=" "),
                y = ifelse(dev>0,
                           democracy+0.03,
                           democracy-0.03)),
            size = 4) +
  labs(x = "GDP/capita", y = "Democracy") +
  custom_theme
```



- Regression line
- residual

How can we calculate point estimates (coefficients)?

- For each observation, the **residual** is:

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

- OLS chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the **sum of squared residuals**:

$$\text{SSR} = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Why the sum of squared residuals (SSR)?

- Squaring penalizes large deviations more heavily
- Ensures the errors do not cancel out (positive + negative)
- Makes the math tractable — allows us to use calculus to find the minimum

Some maths

Deriving the OLS equations

Step 1: First-order derivative w.r.t. $\hat{\beta}_0$:

$$\frac{\partial SSR}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$$

Set equal to zero:

$$\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \qquad (1)$$

Step 2: First-order derivative w.r.t. $\hat{\beta}_1$:

$$\frac{\partial SSR}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$$

Set equal to zero:

$$\sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \qquad (2)$$

Step 3: Solve the system
Rearranging (1), we get:

$$\sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} X_i = 0 \qquad (1a)$$

Taking $n\hat{\beta}_0$ to the left and dividing by n, we obtain:

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} X_i \qquad (1b)$$

since $\frac{1}{n} \sum_{i=1}^{n} Y_i$ and $\frac{1}{n} \sum_{i=1}^{n} X_i$ are just the sample means of X and Y, respectively:

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$$

So:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Step 4: Plug into (2) and simplify

$$\sum_{i=1}^{n} X_i \left(Y_i - \bar{Y} + \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i \right) = 0$$

Algebra $\rightarrow$ yields:

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \qquad (3)$$

Final OLS formula (simple regression)

Step 5: Note that

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \qquad (3)$$

includes the covariance of X and Y in the numerator and the variance of X in the denominator.

$$Cov(X, Y) = \frac{1}{n-1}\sum(X_i - \bar{X})(Y_i - \bar{Y})$$

and

$$Var(X) = \frac{1}{n-1}\sum(X_i - \bar{X})^2$$

if we plug these to into (3), the two $\frac{1}{n-1}$ terms cancel out. We can therefore also write the formula of the slope as:

$$\hat{\beta}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

- **Slope**:

$$\boxed{\hat{\beta}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}}$$

- **Intercept**:

$$\boxed{\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}}$$

Linear regression in R

```
library(stargazer)
data <- data %>%
  filter(year>1969 & year<1980) %>%
  rename(democracy=v2x_polyarchy) %>%
  mutate(gdp=log(e_gdppc))

model1 <- lm(democracy~gdp, data=data)
stargazer(model1,
          type="latex",
          style = "apsr",
          header = F,
          font.size = "tiny",
          title="Basic regression",
          no.space = T,
          omit.stat = c("adj.rsq","f","ser"),
          dep.var.caption = "",
          dep.var.labels = "Polyarchy score",
          covariate.labels = "GDP/capita (log)")
```

Table 1: Basic regression

	Polyarchy score
GDP/capita (log)	0.130***
	(0.006)
Constant	0.098***
	(0.010)
N	1,544
R^2	0.265

*p < .1; **p < .05; ***p < .01

$$\text{Democracy}_i = 0.098 + 0.130 \cdot \text{GDP}_i$$

Interpretation

So far we know that, in the 1970s, there was a positive relationship between **GDP/capita** and **democracy**

Interpretation

So far we know that, in the 1970s, there was a positive relationship between **GDP/capita** and **democracy**

In substantive terms, this means that multiplying GDP/capita by approximately 2.718 is associated with a 0.13-point increase in the V-Dem polyarchy score.

Interpretation

So far we know that, in the 1970s, there was a positive relationship between **GDP/capita** and **democracy**

In substantive terms, this means that multiplying GDP/capita by approximately 2.718 is associated with a 0.13-point increase in the V-Dem polyarchy score.

But how can we know whether this relationship is real (i.e., not by chance)?

Standard errors

In conceptual terms, we can go back to the **frequentist interpretation** we already discussed to understand standard errors.

Standard errors

In conceptual terms, we can go back to the **frequentist interpretation** we already discussed to understand standard errors.

Assume we draw repeated random samples from our data and calculate the relationship between GDP/capita and democracy for each sample. If we do this frequently enough, we can use the variance of the coefficient as an estimate of uncertainty.

Standard errors

In conceptual terms, we can go back to the **frequentist interpretation** we already discussed to understand standard errors.

Assume we draw repeated random samples from our data and calculate the relationship between GDP/capita and democracy for each sample. If we do this frequently enough, we can use the variance of the coefficient as an estimate of uncertainty.

More formally, the standard error is defined as the square root of the variance of the coefficient estimates:

$$\text{SE}(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum(X_i - \bar{X})^2}}$$

In calculus

Step 1: We start from the standard regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

We have derived the OLS estimators above:

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Step 2: Variance of $\hat{\beta}_1$:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$

This formulation is appropriate under specific assumptions:

- **Exogeneity**: $E[u_i|X_i] = 0$
- **Homoscedasticity**: $\text{Var}(u_i|X_i) = \sigma^2$

Step 3: Estimating σ^2
The true variance of the error, σ^2, is unknown. We estimate it using the residuals:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_i X_i$$

The estimated variance of the error is thus given by:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

This is what we use in the formula for the standard error:

$$\text{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum(X_i - \bar{X})^2}}$$

A quick excursus on exogeneity

What is exogeneity?

A quick excursus on exogeneity

What is exogeneity?

Imagine the following situation: You examine whether highly educated persons earn more in professional life. You do so by regressing education on income. Now consider what happens if there is a third, unobserved variable, say: **ability**. Since it is not included in the regression, it goes into the error term. But ability is likely correlated with both, educational achievement and income. Hence the exogeneity assumption is violated (or $E[u_i|X_i] \neq 0$).

A quick excursus on exogeneity

What is exogeneity?

Imagine the following situation: You examine whether highly educated persons earn more in professional life. You do so by regressing education on income. Now consider what happens if there is a third, unobserved variable, say: **ability**. Since it is not included in the regression, it goes into the error term. But ability is likely correlated with both, educational achievement and income. Hence the exogeneity assumption is violated (or $E[u_i|X_i] \neq 0$).

You can only hope to identify a causal relationship if you can establish exogeneity!

A quick excursus on heteroscedasticity

What is heteroscedasticity?

A quick excursus on heteroscedasticity

What is heteroscedasticity?

We continue in the same scenario. Now consider that variation in income might be much higher for highly educated people than for less educated people. While high educational achievment is a precondition for high earnings, it is not a guarantee. Hence, the variance of the error might depend on the level of X_i, in our example education.

A quick excursus on heteroscedasticity

What is heteroscedasticity?

We continue in the same scenario. Now consider that variation in income might be much higher for highly educated people than for less educated people. While high educational achievment is a precondition for high earnings, it is not a guarantee. Hence, the variance of the error might depend on the level of X_i, in our example education.

The presence of heteroscedasticity makes it impossible to estimate the variance of $\widehat{\beta}_1$. We need heteroscedasticity robust standard errors.

Panel data

Sant'Anna
School of Advanced Studies – Pisa

What is panel data?

The data we use in political science very frequently comes in the form of panel data.

What is panel data?

The data we use in political science very frequently comes in the form of panel data.

For example, we might observe countries across various years as in the GDP/capita and democracy analysis.

What is panel data?

The data we use in political science very frequently comes in the form of panel data.

For example, we might observe countries across various years as in the GDP/capita and democracy analysis.

Formally, in panel data we observe units $i = 1, \dots, N$ over time $t = 1, \dots, T$. An OLS model might thus look like this:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it}$$

What is panel data

In standard OLS, we assume that observations are

1. **Independent**: No observation is related to any other
2. **Identically distributed**: All observations come from the same distribution with the same variance.

Formally:

$$u_i \sim \text{i.i.d}(0, \sigma^2)$$

All errors u_i come from a distribution with mean 0 and variance σ^2.

What is panel data

In standard OLS, we assume that observations are

1. **Independent**: No observation is related to any other
2. **Identically distributed**: All observations come from the same distribution with the same variance.

Formally:

$$u_i \sim \text{i.i.d}(0, \sigma^2)$$

All errors u_i come from a distribution with mean 0 and variance σ^2.

This assumption is routinely violated in panel data.

Panel data and the i.i.d. assumption

1. **Violation of independence** (serial correlation or clustering): Panel data by definition contains repeated observations of the same units over time. This means that residuals within these units might be correlated over time. For example, a country's GDP/capita in 1970 is likely related to that country's GDP/capita in 1971. Formally, this means that $\text{Cov}(u_{it}, u_{is}) \neq 0$ for $t \neq s$.

Panel data and the i.i.d. assumption

1. **Violation of independence** (serial correlation or clustering): Panel data by definition contains repeated observations of the same units over time. This means that residuals within these units might be correlated over time. For example, a country's GDP/capita in 1970 is likely related to that country's GDP/capita in 1971. Formally, this means that $\text{Cov}(u_{it}, u_{is}) \neq 0$ for $t \neq s$.
2. **Violation of identical distribution** (heteroscedasticity across units): Different units might have different variance in their error terms. For example, some countries might have higher levels of economic volatility. As a result, they would have systematically higher variance. Formally, this means that $\text{Var}(u_{it})$ is not constant over time.

Clustered standard errors

- Treat **each group/cluster** as the key unit of analysis.
- Allow:
  - **Any correlation within a group**
  - **Independence across groups**
- This gives **more honest** standard errors.

Clustered standard errors are usually larger — reflecting the reality of shared influences.

Clustered standard errors in R

```r
library(stargazer)
library(sandwich)
library(lmtest)

model1 <- lm(democracy ~ gdp, data = data)
clustered_se <- vcovCL(model1,
                       cluster = ~ country_id)

stargazer(model1,
          type = "latex",
          style = "apsr",
          header = FALSE,
          font.size = "tiny",
          title = "Basic regression",
          subtitle = "SEs clustered in country",
          no.space = TRUE,
          omit.stat = c("adj.rsq", "f", "ser"),
          dep.var.caption = "",
          dep.var.labels = "Polyarchy score",
          covariate.labels = "GDP/capita (log)",
          se = list(sqrt(diag(clustered_se))))
```

Table 2: Basic regression

	Polyarchy score
GDP/capita (log)	0.130^{***}
	(0.025)
Constant	0.098^{***}
	(0.029)
Clustered SEs	country
N	1,544
R^2	0.265

*p $<$ .1; **p $<$ .05; ***p $<$ .01

$$\text{Democracy}_i = 0.098 + 0.130 \cdot \text{GDP}_i$$