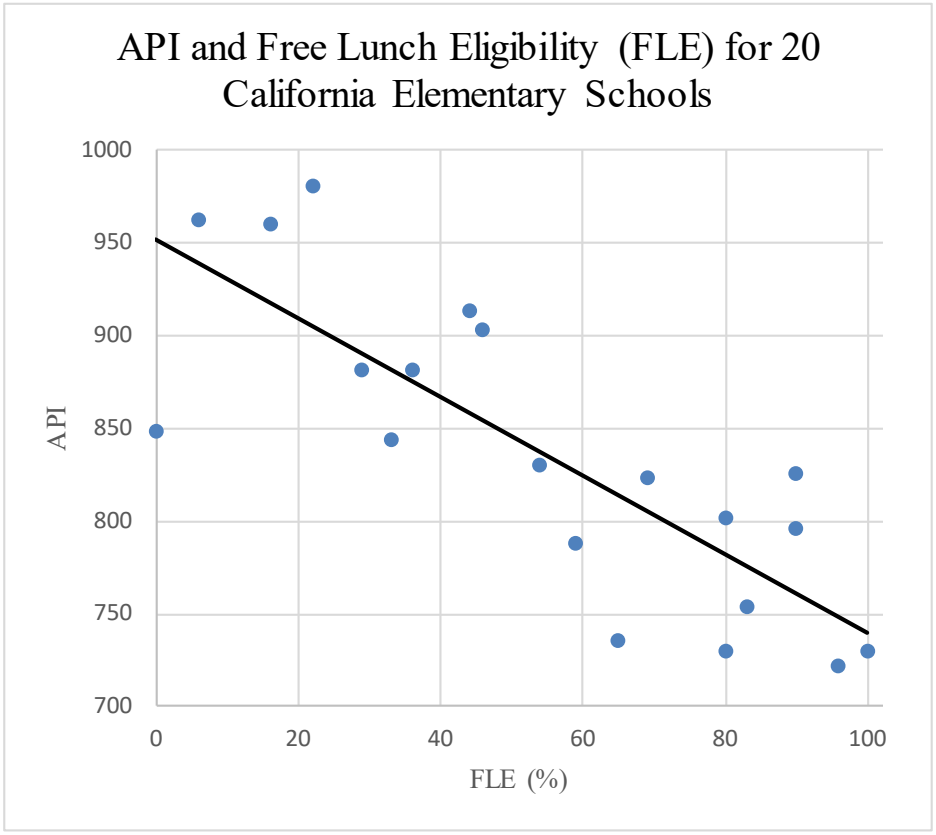
Chapter 4

Generalizing from a Sample

Learning Objectives

- Articulate the difference between sample coefficients (*b*) and population coefficients (β)
- Develop the three steps to generalize from a sample to a population
- Explain the classical regression assumptions
- Connect the data type to potential assumption failures and potential relevant populations

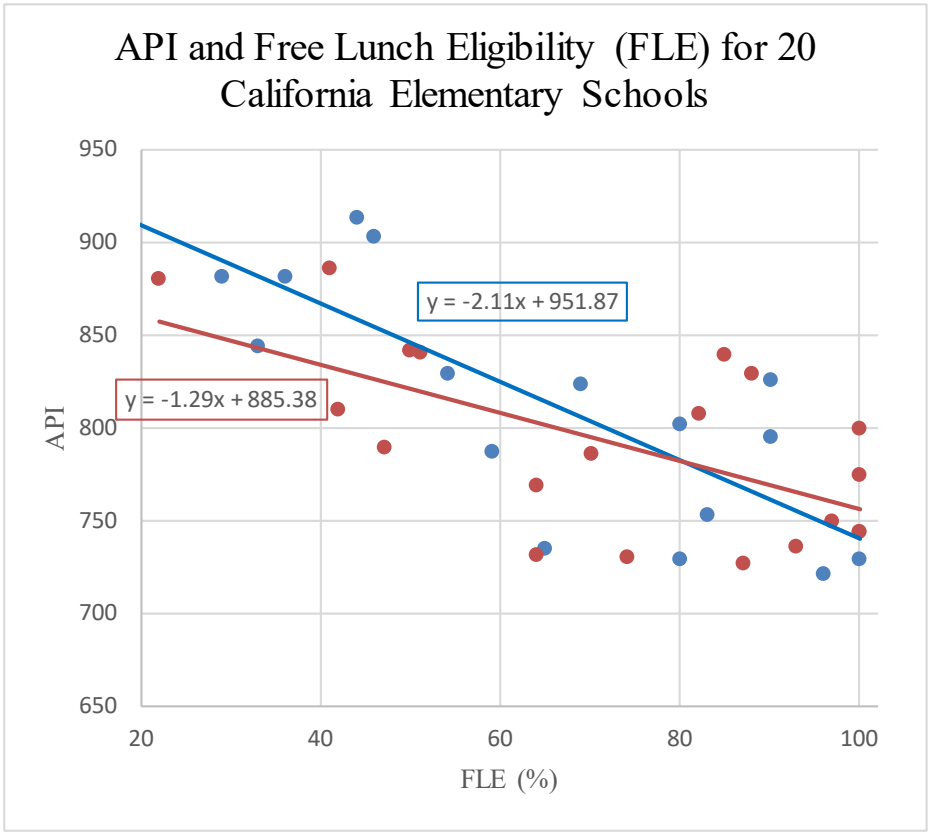
We estimated a regression model. What now?



$$API_i = 951.87 - 2.11FLE_i + e_i$$

- Would this relationship be the same for a different 20 schools?
- Would it be the same for the same schools in a different year?
- If we gave more kids free lunch, would API decrease?

20 More Schools



$$API_i = 951.87 - 2.11FLE_i + e_i$$

- **Would this relationship be the same for a different 20 schools?**
- Would it be the same for the same schools in a different year?
- If we gave more kids free lunch, would API decrease?

Three Steps to Generalizing from a Sample

1. Define Your **Population** and **Research Goal**

2. Make **Assumptions** about Your Population (and How Your Sample Represents It)

3. **Compute** Statistics to Measure OLS Accuracy

Two Models

1. Sample Model

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_K X_{Ki} + e_i$$

2. Population Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki} + \varepsilon_i$$

$$\text{cov}[X_{1i}, \varepsilon_i] = \ldots = \text{cov}[X_{Ki}, \varepsilon_i] = 0$$

The covariance conditions imply that population model is the best linear prediction of Y using the X's.

The best forecasting model by definition has zero correlation between the X's and errors

Step 1: Define Your Population and Research Goal

What is your population?

- the same 20 schools in a different year?
- all schools in California?
- all schools in the United States?
- all schools in CA if we removed free lunch?

Research goal is often best phrased as a question

- what is the predicted API of this school next year?
- what would happen to API if we restricted free lunch eligibility?

This seems like an easy step, but it is often the hardest part of the analysis

Step 1: Define Your Population and Research Goal

- **Scope**
  - e.g., all individuals in a state, all countries in the world, all companies in a country, all houses in a county.
- **Time**
  - e.g., next year, any future year, the same year as my sample.
- **Assignment of X variables**
  - e.g., exactly as in my sample, through a new government policy or project, an experiment, or a different state of nature (like weather).

Step 1: Define Your Population and Research Goal

Examples:

- Do good teachers produce better student outcomes?
- Does the law of demand hold for electricity?
- Is it possible to forecast stock returns?

Step 2: Make Assumptions about Your Population (and How Your Sample Represents It)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki} + \varepsilon_i$$

CR1: Representative sample. Our sample is representative of the population we want to say something about.

CR2: Homoscedastic Errors. $Var[\varepsilon_i] = \sigma^2$
The variance of the error is constant over all of our data, i.e., it is homoskedastic.

CR3: Uncorrelated Errors. $Cov[\varepsilon_i, \varepsilon_j] = 0$
The errors are uncorrelated across observations.

CR4: Normally distributed errors. This assumption is only required of small samples.

Sometimes we need one more assumption (Ch 11-12)

CR5: Exogenous *X*. The values of the explanatory variable are exogenous, or given (there are no errors in the X-direction).

Example: Google Flu trends

Step 3: Compute Statistics to Measure OLS Accuracy

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_K X_{Ki} + e_i$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki} + \varepsilon_i$$

The main statistic for measuring accuracy is the **standard error:**

$$\sqrt{V[b_j]} = \sqrt{E\left[(b_j - \beta_j)^2\right]}$$

Recap:
Three Steps to Generalizing from a Sample

1. Define Your **Population** and **Research Goal**

2. Make **Assumptions** about Your Population (and How Your Sample Represents It)

3. **Compute** Statistics to Measure OLS Accuracy

Data Types

- **Cross section** – observe multiple units at a point in time
- **Time series** – observe the same unit at different points in time
- **Panel** – observe multiple units over time

Question: Which assumptions may fail in each case

What We Learned

- Three steps to generalizing from a sample:
  1. Define your population and research goal.
  2. Make assumptions about your population (and how your sample represents it).
  3. Compute statistics to measure OLS accuracy.

- When defining the population, consider scope, time period, and how the *X* variables are determined.

- Three critical assumptions are required to generalize to a population:
  - CR1: representative sample
  - CR2: homoscedastic errors
  - CR3: uncorrelated errors
- A fourth assumption matters only in small samples:
  - CR4: Normally distributed errors
- A fifth assumption is only relevant if doing causality analysis:
  - CR5: The values of the right-hand-side variables are exogenous

- Knowing your data type tells you which assumptions are likely to fail and which populations are potentially relevant.