Causal Inference

Dr. Syed Badruddoza

Texas Tech

May 2, 2025

Chapter 3

Violation of Zero Conditional Mean (Exogeneity)

Understanding $E(\varepsilon|X) = 0$

"All factors in the unobserved error term are uncorrelated with the explanatory variables. The functional relationship between the response and predictors is correctly specified."

Recall that, $p(\varepsilon|X) = \frac{p(X,\varepsilon)}{p(X)}$ where the marginal density is: $p(X) = \int p(X,\varepsilon)\, d\varepsilon$. Thus, the conditional expectation is:

$$E(\varepsilon|X) = \int \varepsilon \frac{p(X,\varepsilon)}{p(X)}\, d\varepsilon.$$

Covariance formula $\text{Cov}(X,\varepsilon) = E(X\varepsilon) - E(X)E(\varepsilon)$.
If $E(\varepsilon) = 0$, then: $\text{Cov}(X,\varepsilon) = E(X\varepsilon)$.

By the Law of Iterated Expectations: $E(X\varepsilon) = E[XE(\varepsilon|X)]$.

The OLS assumption $E(\varepsilon|X) = 0$ (strict exogeneity) suffices to ensure $E(X\varepsilon) = 0$ (population orthogonality).

Strict exogeneity (ε is mean-independent of X) implies population orthogonality (ε and X are not linearly correlated), but the reverse is not necessarily true.

Violation of $E(\varepsilon|X) = 0$

- **Omitted Variable Bias** (under-specifying): If a relevant variable Z is missing in the estimated model where the true model is:

$$Y = X\beta + Z\gamma + \varepsilon$$

If X and Z are correlated, then $E(\varepsilon|X) \neq 0$, biasing $\hat{\beta}$.

- **Measurement Error in** X: Suppose we observe $X^* = X + u$, where u is a measurement error:

$$Y = X^*\beta + (\varepsilon - u\beta)$$

Since u is part of the new error term, it can correlate with X^*, violating $E(\varepsilon|X) = 0$.

- **Simultaneity**: If X and Y are jointly determined, then:

$$X = Y\pi + \nu$$

Since Y is a function of ε, so is X, leading to $E(\varepsilon|X) \neq 0$.

Omitted Variable Bias

- Suppose we estimate a misspecified model that omits Z:

$$Y = X\beta + \tilde{\varepsilon}$$

where $\tilde{\varepsilon} = Z\gamma + \varepsilon$ (the omitted variable is absorbed into the error).

- The OLS estimator for β is:

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

- Substituting the true model:

$$\hat{\beta} = (X'X)^{-1}X'(X\beta + Z\gamma + \varepsilon).$$

- Expanding:

$$\hat{\beta} = \beta + (X'X)^{-1}X'Z\gamma + (X'X)^{-1}X'\varepsilon.$$

Bias Expression

- Taking expectations:

$$E[\hat{\beta}] = \beta + E[(X'X)^{-1}X'Z]\gamma.$$

  Note: The second part affects all $\hat{\beta}$s we are estimating. [Wooldridge Ch.3]

- The bias term is:

$$\text{Bias}(\hat{\beta}) = E[(X'X)^{-1}X'Z]\gamma.$$

- If X and Z are correlated, then $E[(X'X)^{-1}X'Z] \neq 0$, meaning:

$$E[\hat{\beta}] \neq \beta.$$

- Hence, the omission of Z biases the estimate of β.
- $E[\hat{\beta}] > \beta \Rightarrow$ Upward bias. $E[\hat{\beta}] < \beta \Rightarrow$ Downward bias.
- Biased towards zero: $E[\hat{\beta}]$ is closer to 0 than β.

Direction of Omitted Variable Bias

- The sign of the bias depends on:
  1. $\text{Cov}(X, Z) \rightarrow$ how strongly X and Z are correlated.
  2. $\gamma \rightarrow$ the effect of Z on Y.

- If $\text{Cov}(X, Z) > 0$ and $\gamma > 0$, then:

$$\text{Bias}(\hat{\beta}) > 0.$$

- If $\text{Cov}(X, Z) < 0$ and $\gamma > 0$, then:

$$\text{Bias}(\hat{\beta}) < 0.$$

- The bias overstates or understates the true effect depending on the direction of correlation.

Omitted Variable Bias and Consistency

- Definition of Consistency: An estimator $\hat{\beta}$ is consistent if:

$$\hat{\beta} \xrightarrow{p} \beta \quad \text{as } n \to \infty.$$

- From our previous derivation:

$$\hat{\beta} = \beta + (X'X)^{-1}X'Z\gamma + (X'X)^{-1}X'\varepsilon.$$

- Taking the probability limit:

$$\text{plim}\hat{\beta} = \beta + \text{plim}[(X'X)^{-1}X'Z]\gamma.$$

  Note: The second part affects all $\hat{\beta}$s we are estimating.

- If $\text{Cov}(X, Z) \neq 0$, then: $\text{plim}\hat{\beta} \neq \beta$.
- This means $\hat{\beta}$ is inconsistent.

Omitted Variable Bias and Efficiency

- The OLS estimator in the misspecified model is: $\hat{\beta}^* = (X'X)^{-1}X'Y$.
- Substituting Y: $\hat{\beta}^* = (X'X)^{-1}X'(X\beta + \tilde{\varepsilon})$.
- Where $\tilde{\varepsilon} = Z\gamma + \varepsilon$.

$$\hat{\beta}^* = \beta + (X'X)^{-1}X'\tilde{\varepsilon}.$$

- Variance of $\hat{\beta}^*$:

$$\text{Var}(\hat{\beta}^*) = (X'X)^{-1}X'\text{Var}(\tilde{\varepsilon})X(X'X)^{-1}.$$

- Since $\tilde{\varepsilon} = Z\gamma + \varepsilon$, the variance of $\tilde{\varepsilon}$ is:

$$\text{Var}(\tilde{\varepsilon}) = \gamma^2\text{Var}(Z) + \sigma^2 I.$$

- Omitting Z increases the standard errors of $\hat{\beta}^*$, making estimation less efficient.

$$\text{Var}(\hat{\beta}^*) > \text{Var}(\hat{\beta}).$$

- Important detail: Omitting Z when $\gamma \neq 0$ practically changes the composition of $(X'X)^{-1}$ since Z is not included in X anymore (Explained further below with simpler notations).

Bias-Variance trade-off when a Variable is Omitted

Suppose we run $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$ instead of $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ where $\beta_2 \neq 0$ (x_2 is relevant). Then $\tilde{\beta}_1$ is biased from omitting x_2, and $\hat{\beta}_1$ is unbiased.

- From the previous slide we know that error variance increases $\tilde{\sigma}^2 > \hat{\sigma}^2$.
- But for a simple model like above it is likely that, $\mathrm{Var}(\tilde{\beta}_1) < \mathrm{Var}(\hat{\beta}_1)$: that is, $\frac{\tilde{\sigma}^2}{SST_1} < \frac{\hat{\sigma}^2}{SST_1(1-R_1^2)}$, where R_1^2 is obtained from regressing x_1 on x_2.
- A simpler (misspecified and biased) model can have lower $Var(\tilde{\beta}_1)$, depending on partial correlations of x_1 with all other variables.
- Bias-variance trade-off [Wooldridge Ch.3]: Correct model ($\hat{\beta}_1$) eliminates bias but has larger $Var(\hat{\beta})$, if x_1 is correlated with x_2 ($R_1^2 > 0$).
- Both $Var(\tilde{\beta}_1), Var(\hat{\beta}_1) \to 0$ as $n \to \infty$. In large samples, we prefer $\hat{\beta}_1$.
- In practice, we want an unbiased estimator of β_1 (even with higher variance) because consistency is crucial—especially for large samples where the variance cost diminishes.

The opposite: Inclusion of An Irrelevant Variable

Inclusion of an irrelevant variable does not affect the unbiasedness of OLS estimators but increases their variance. Consider two models:

$$\text{True model:} \quad \tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

$$\text{We run:} \quad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2.$$

That is, $\beta_2 = 0$ (x_2 is irrelevant), then:

- Both $\tilde{\beta}_1$ and $\hat{\beta}_1$ are unbiased.
- But $Var(\tilde{\beta}_1) < Var(\hat{\beta}_1)$ meaning the model with the irrelevant variable has higher variance.
- Note: $Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_1(1-R_1^2)} > Var(\tilde{\beta}_1) = \frac{\sigma^2}{SST_1}$.
- Where R_1^2 is from regressing x_1 on x_2.
- Including an irrelevant variable adds a parameter without reducing residual variance, so the standard error of $\hat{\beta}_1$ goes up.

An experiment with omitted variable

The true model is: $y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + \epsilon$
where $\epsilon \sim N(0, 1)$ and $X_1 = \log(X_3)$.

The estimated model omits X_3:

$$y = B_0 + B_1 X_1 + B_2 X_2 + u$$

where $u = B_3 X_3 + \epsilon$.

When X_3 is omitted, the estimated coefficients $\hat{B}_0$, $\hat{B}_1$, and $\hat{B}_2$ will be biased.

Empirically, the bias arises because X_3 is correlated with X_1 and X_2, and its effect is absorbed into the error term u. In this example, the formula for the bias in the coefficients due to omitted variable bias is:

$$\text{Bias}(\hat{B}_j) = B_3 \cdot \frac{\text{Cov}(X_j, X_3)}{\text{Var}(X_j)}$$

where $j = 1, 2$.

Experiment (Contd.)

The bias in the intercept B_0 is:

$$\text{Bias}(B_0) = B_3 \cdot (E[X_3] - \gamma_1 E[X_1] - \gamma_2 E[X_2])$$

where γ_1 and γ_2 are the coefficients from the auxiliary regression of X_3 on X_1 and X_2:

$$X_3 = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \text{error}.$$

The bias in B_1 is:

$$\text{Bias}(B_1) = B_3 \cdot \frac{\text{Cov}(X_1, X_3)}{\text{Var}(X_1)}.$$

The bias in B_2 is:

$$\text{Bias}(B_2) = B_3 \cdot \frac{\text{Cov}(X_2, X_3)}{\text{Var}(X_2)}.$$

The omission of X_3 also affects the variance of the estimated coefficients. Specifically:

- The variance of $\hat{B}_1$ and $\hat{B}_2$ will generally increase because the omitted variable X_3 contributes to the error term u, increasing the overall noise in the model.
- The exact formula for the variance depends on the correlation structure between X_1, X_2, and X_3.

Why don't we check this empirically!
github.com/Badruddoza/AAEC6311/blob/main/Omitted_variable_bias

Conclusion: Omitted Variable Bias Implications

- Omitting a relevant variable generally makes the OLS estimator $\hat{\beta}$ biased and inconsistent. The bias arises when the omitted variable Z is correlated with at least one included regressor X.
- The bias can "smear" across all coefficients in the model, even for variables uncorrelated with Z, due to the correlations among the included variables [See Greene Ch.8]. The partial correlations between the endogenous variable x_1 and the other predictors in the model can be measured using auxiliary regressions (e.g., regressing x_1 on the other predictors to compute R^2).
- The direction of the omitted variable bias depends on the strength of the correlation between Z and the included regressors X, and the relevance of the omitted variable (γ).
- The change in $Var(\hat{\beta})$ depends on how much error variance has changed due to omitting Z and the correlation between included regressors. Using the incorrect $Var(\hat{\beta})$ affects hypothesis testing, leading to incorrect inferences.

Measurement Error in X

We'll discuss a simpler model for this to avoid complexity.
Consider the true model:

$$y = \beta_0 + \beta_1 x^* + u,$$

where u is the error term with $E(u) = 0$ and $\text{Var}(u) = \sigma_u^2$.

However, suppose we do not observe the true regressor x^*, but observe a mismeasured version x.

The measurement error in the population:

$$e = x - x^*$$

$$\Rightarrow x = x^* + e$$

We maintain the assumption that $E(e) = 0$ and u is uncorrelated with both x^* and x.

Measurement Error (Contd.)

Case 1: $Cov(x, e) = 0$

- The error e is uncorrelated with the *observed* x.
- Example: Survey participants report their weights rounded to the nearest integer (e.g., 70 kg instead of 70.3 kg). Here, the measurement error e is small and uncorrelated with the true weight x^*.
- The model becomes:

$$y = \beta_0 + \beta_1(x - e) + u = \beta_0 + \beta_1 x + \underbrace{(u - \beta_1 e)}_{\text{New Error}}.$$

- Note: $E(u - \beta_1 e) = 0$ and $Cov(u - \beta_1 e, x) = 0$.
- Also, $Var(u - \beta_1 e) = \sigma_u^2 + \beta_1^2 \sigma_e^2$.
- Measurement error increases error variance, but does not affect any of the OLS properties and will produce a consistent estimator of β_0 and β_1.

Measurement Error (Contd.)

Case 2: $Cov(x^*, e) = 0$

- The error e is uncorrelated with the *unobserved* x^*.
- Example: Suppose x^* is actual income, but respondents report their income with error e (e.g., due to recall bias). Here, e is uncorrelated with x^*, but $x = x^* + e$ is correlated with e.
- This is called the classical errors-in-variables assumption. Where the covariance of the observed x and the measurement error is just the variance of the measurement error σ_e^2.
- $Cov(x, e) = E(xe) = E[(x^* + e)e] = E(x^*e) + E(e^2) = 0 + \sigma_e^2 = \sigma_e^2$
- Then the covariance between the observed x and the new error:

$$Cov(x, u - \beta_1 e) = Cov(x, u) - Cov(x, \beta_1 e) = -\beta_1 Cov(x, e) = -\beta_1 \sigma_e^2$$

- This makes $\hat{\beta}_1$ inconsistent

$$\text{plim}(\hat{\beta}_1) = \beta_1 + \frac{Cov(x, u - \beta_1 e)}{Var(x)} = \beta_1 - \frac{\beta_1 \sigma_e^2}{\sigma_{x^*}^2 + \sigma_e^2}$$

- Since $\hat{\beta}_1 = Cov(x, y)/Var(x)$ and $Var(x) = Var(x^*) + Var(e)$.

Measurement Error (Contd.)

Case 2: $Cov(x^*, e) = 0$ (Contd.)

- This makes $\hat{\beta}_1$ inconsistent

$$\text{plim}(\hat{\beta}_1) = \beta_1 - \frac{\beta_1 \sigma_e^2}{\sigma_{x^*}^2 + \sigma_e^2} = \beta_1 \left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2} \right) = \beta_1 \left[\frac{Var(x^*)}{Var(x)} \right]$$

- Note: $Var(x^*) < Var(x)$ so the ratio is < 1.
- The estimated effect will be *attenuated* (biased towards zero).
- In case of multiple predictor variables, all the coefficients will be biased and inconsistent, in unknown directions, even though the other variables are not measured with error. [Greene Ch.4, Wooldridge Ch.9]
- This happens because x is usually related with other predictor variables.

Check: github.com/Badruddoza/AAEC6311/blob/main/Measurement_error

Solutions to Measurement Error in X

- **Proxy Variable**:
  - Use an external variable Z that is correlated with the true X^* but uncorrelated with the measurement error e. Example: If X^* is true income and X is reported income, Z could be tax records.
- **Repeat Measurements**:
  - Collect multiple observations of X and average them to reduce the effect of measurement error. Example: Repeated surveys or multiple measurements of the same variable.
- **Errors-in-Variables Models**:
  - Model the measurement error structure. Example: Use MLE (assuming distribution for e), or IV regression to account for measurement error.
- **Bounding Strategies**:
  - Use sensitivity analysis to assess the potential impact of measurement error on the results. Example: Estimate bounds for the true coefficient β under different assumptions about the magnitude of measurement error.

Causal Inference and the Notion of Ceteris Paribus

Models for endogeneity concerns: Causal Inference

- This subsection discusses models for endogeneity problems.
- Note: Instead of separately discussing simultaneity, we start the discussion of causal models here. Simultaneity will be discussed in the simultaneous equations part.
- Causal inference aims to estimate the cause-and-effect relationship between a treatment and an outcome under the *ceteris paribus* assumption, i.e., holding everything else fixed.
- Challenge: Correlation is not causation. The missing counterfactual outcome is never observed as we only observe one potential outcome per unit.
- Goal: Identify and estimate the true causal effect while addressing potential biases.
- We start with a binary treatment as it is the simplest form to understand the assumptions.

A framework with binary treatment (D)

Consider the true model:

$$Y = \tau D + X\beta + \varepsilon$$

where:

- Y = outcome variable.
- D = treatment variable (possibly endogenous).
- X = exogenous control variables (with the intercept).
- τ = true causal effect of D.
- ε = unobserved factors (error term).

Example: Effect of a drug (D) in reducing blood cholesterol levels (Y).

Can we just subtract the average cholesterol of people who received the drug and people who did not to find the effect?

[Read Wooldridge2 Ch.21; Greene Ch.19; Angrist and Pischke Ch.2]

Identification of τ

- Let $Y_i(1)$ be the outcome if unit i receives treatment.

$$Y(1) = \tau + X\beta + \varepsilon \quad \text{(if treated, } D = 1\text{)}$$

- Let $Y_i(0)$ be the outcome if unit i does not receive treatment.

$$Y(0) = X\beta + \varepsilon \quad \text{(if untreated, } D = 0\text{)}$$

- The true causal effect is: $\tau_i = Y_i(1) - Y_i(0)$
- But the observed outcome is $Y_i(1)$ and $Y_j(0)$ where $i \neq j$.
- Problem: We observe either $Y_i(1)$ or $Y_i(0)$, but not both.
- Assume person i received the drug, while person j did not. We do not observe how i's cholesterol level would have been had they not received the drug, nor do we observe how j's cholesterol level would have been had they received the drug.
- Fundamental problem of causal inference [Holland, 1986]: Missing counterfactual outcomes.

Average Treatment Effect (ATE)

The Average Treatment Effect (ATE) is the expected difference in potential outcomes:

$$ATE = E[Y(1) - Y(0)]$$

$$\Rightarrow ATE = E[\tau + X\beta + \varepsilon - (X\beta + \varepsilon)] = E[\tau] = \tau$$

For the ATE to be valid, the following assumptions must hold:

- **Consistency**: The observed outcome corresponds to the potential outcome under the actual treatment.

$$Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$$

- **Ignorability**: Treatment assignment is independent of potential outcomes (no unmeasured confounding=ignoring X is justified).

$$(Y(1), Y(0)) \perp D$$

- **Overlap**: Every individual has a chance of receiving or not receiving the treatment.

$$0 < P(D = 1|X) < 1$$

Selection Bias

The observed difference in outcomes (Naïve Estimator) is:

$$E[Y|D=1] - E[Y|D=0]$$

$$= E[\tau + X\beta|D=1] - E[X\beta|D=0]$$

$$= \tau + E[X\beta|D=1] - E[X\beta|D=0]$$

Taking naive averages ignores Xs. If treatment assignment D is correlated with X, then

$$E[X\beta|D=1] \neq E[X\beta|D=0]$$

$$\text{Selection Bias} = E[X\beta|D=1] - E[X\beta|D=0]$$

- Selection bias arises when treated and untreated groups differ systematically in their characteristics.
- Example: if only health-conscious individuals choose to take the drug, the estimated treatment effect may not generalize to those who are not health-conscious.
- Here, health awareness is a confounding variable because it affects both treatment assignment (taking the drug) and the outcome (cholesterol levels), potentially biasing the estimated effect of the drug.

Average Treatment Effect on the Treated (ATT)

The Average Treatment Effect on the Treated (ATT or ATET) measures the treatment effect for those who actually receive the treatment:

$$ATT = E[Y(1) - Y(0)|D = 1]$$

ATT estimates the causal effect by comparing treated individuals to their counterfactual outcomes (what would have happened if they had not been treated).

Substituting the potential outcomes:

$$ATT = E[\tau + X\beta + \varepsilon - (X\beta + \varepsilon)|D = 1] = E[\tau|D = 1] = \tau$$

Thus, under no selection bias, ATT is equal to the true treatment effect τ.

Average Treatment Effect on the Untreated (ATU)

The Average Treatment Effect on the Untreated (ATU) measures the expected treatment effect for individuals who do not receive the treatment:

$$ATU = E[Y(1) - Y(0)|D = 0]$$

ATU estimates the causal effect by comparing untreated individuals to their counterfactual outcomes (what would have happened if they had been treated).

Substituting the potential outcomes:

$$ATU = E[\tau + X\beta + \varepsilon - (X\beta + \varepsilon)|D = 0] = E[\tau|D = 0] = \tau$$

Thus, under no selection bias, ATU is also equal to the true treatment effect τ.

Example: Effect of a Job Training Program on Wages

- **ATE:** Measures the average effect of the training program on all workers (both participants and non-participants).
- **ATU:** Measures the effect on workers who did not participate in the program, estimating what their wages would have been had they received training.
- **ATT:** Measures the effect on workers who actually participated in the program. Assume we are interested in ATT. Then,

$$ATT = E[Y|D = 1] - E[Y|D = 0, \text{matched}]$$

where:

  - $E[Y|D = 1]$ is the average wage for trained workers.
  - $E[Y|D = 0, \text{matched}]$ is the counterfactual wage for trained workers had they not participated, estimated using a matched comparison group.

- **Propensity Score:** The probability of receiving training given observed characteristics. Used in matching techniques to reduce selection bias and ensure fair comparisons between treated and untreated workers.

Hypothetical example with observed counterfactuals

The ideal situation is where counterfactuals are observable. In this example, the effect of the treatment does not vary between the treated and control groups, so $ATE = ATT = ATU$.

Individual	$Y(1)$	$Y(0)$	D	$\tau_i = Y(1) - Y(0)$
1	10	5	1	5
2	8	6	1	2
3	7	4	0	3
4	9	5	0	4

$$ATE = E[Y(1) - Y(0)] = \frac{5+2+3+4}{4} = 3.5$$

$$ATT = E[Y(1) - Y(0)|D = 1] = \frac{5+2}{2} = 3.5$$

$$ATU = E[Y(1) - Y(0)|D = 0] = \frac{3+4}{2} = 3.5$$

Unfortunately, the counterfactuals in red are not observable. All we can do is try to predict average counterfactual outcomes using various methods.

Randomized Treatment Assignment

Assume that the distribution of Y is identical between the treated and control groups, except for differences due to the treatment D. Then the difference in mean outcomes identifies the ATE. Specifically, under Completely Randomized Treatment Assignment we can just use the observed values:

Individual	$Y(1)$	$Y(0)$	D
1	10	5	1
2	8	6	1
3	7	4	0
4	9	5	0

- Mean outcome for treated ($D = 1$): $(10 + 8)/2 = 9$.
- Mean outcome for control ($D = 0$): $(4 + 5)/2 = 4.5$.
- $ATE = \text{Mean}(Y|D = 1) - \text{Mean}(Y|D = 0) = 9 - 4.5 = 4.5$.
- Why does this work? Randomization ensures that *the treatment and control groups are comparable*.

Relationship Between ATE, ATT, and ATU

Decomposition of the Average Treatment Effect (ATE):

$$ATE = ATT \cdot P(D = 1) + ATU \cdot P(D = 0)$$

- $P(D = 1)$: Proportion of treated individuals.
- $P(D = 0) = 1 - P(D = 1)$: Proportion of untreated individuals.

Example: Out of 100 patients, 45 randomly receive a cholesterol-lowering drug ($P(D = 1) = 0.45$). Suppose $ATT = -20$ (treated group) and $ATU = -10$ (untreated group). Then:

$$ATE = (-20) \cdot (0.45) + (-10) \cdot (0.55) = -14.5$$

Notes:

- ATE is a weighted average of ATT and ATU.
- Under random treatment assignment and homogeneous treatment effects, we have: $ATE = ATT = ATU$.
- Treatment effects may vary (heterogeneous effects) both between and within groups.

Related concept: Confounders

- Confounder: A variable that affects both treatment assignment D and the outcome Y, potentially biasing causal inference.
- Example: Effect of a Job Training Program D on Wages Y. Let $X =$ "Education Level." If more educated workers are more likely to participate in the training program and also tend to earn higher wages, then X is a confounder.
- If a confounder affects both D and Y, controlling for it helps isolate the treatment effect.
- Including confounders minimizes differences between treated and untreated groups, improving comparability.
- Conditional Independence: If all confounders are accounted for, treatment assignment can be considered as good as random.
- Omitting important confounders still leads to bias.

Related concept: Reverse Causality

Reverse causality occurs when the causal direction between two variables is reversed. For example, instead of D causing Y, Y causes D. This can create spurious associations in observational studies.

- **E.g.**: Studying the effect of diet soda consumption (D) on obesity (Y).
  - Reverse causality: People who are gaining weight (Y) may switch to diet soda (D) to cut calories.
  - This creates a spurious association between diet soda consumption and obesity.

Related concept: Mediation

Mediation refers to a causal structure where a treatment (or independent variable) affects the outcome indirectly through an intermediate variable, called a mediator.

- Example: Studying the effect of education (D) on income (Y).
  - Work experience (X) is a mediator: Education (D) increases work experience (X), which in turn affects income (Y).
  - This follows the causal chain: $D \to X \to Y$.
  - If the goal is to estimate the *total* effect of education on income, controlling for X would block part of the effect and lead to an underestimation.
  - If the goal is to estimate the *direct* effect of education on income (regardless of experience), controlling for X is appropriate to remove the mediated pathway.

Mediation: Regression Setup

To study how a treatment D affects outcome Y through a mediator X:

- **Total effect (no mediator)**: θ: Total effect of D on Y

$$Y_i = \beta_0 + \theta D_i + \varepsilon_i$$

- **Mediator equation**: γ: Effect of D on mediator X

$$X_i = \delta_0 + \gamma D_i + u_i$$

- **Direct and indirect effects (no interaction)**:

$$Y_i = \beta_0 + \tau D_i + \beta_1 X_i + \varepsilon_i$$

  - τ: Direct effect of D on Y
  - $\gamma \cdot \beta_1$: Indirect (mediated) effect
  - $\theta = \tau + \gamma \cdot \beta_1$: Total effect decomposition

- **Heterogeneous effects (with interaction)**:

$$Y_i = \beta_0 + \tau D_i + \beta_1 X_i + \beta_2 (D_i \cdot X_i) + \varepsilon_i$$

  - Allows effect of X on Y to vary by D
  - Total effect of D: $\tau + \beta_2 X_i$. That is, the effect of D varies by X.

Related concept: Collider Bias

Collider bias occurs when conditioning on a variable (V) that is influenced by both the treatment (D) and the outcome (Y). This creates a spurious association between D and Y.

- Example: Studying the effect of exercise (D) on heart health (Y).
- Collider: Doctor visits (V) are influenced by:
  - Exercise (D): People who exercise more may visit the doctor less.
  - Heart health (Y): People with poor heart health may visit the doctor more. Thus, $D \rightarrow V \leftarrow Y$.
- Bias: Conditioning on V opens a non-causal path between D and Y.
  - Among frequent doctor visitors, exercisers may appear to have worse heart health because they are more likely to have underlying health issues.
  - This creates a spurious negative association between exercise and heart health.

Key Takeaway: Avoid conditioning on variables influenced by both D and Y, as this can introduce collider bias and distort causal estimates.

Stable Unit Treatment Value Assumption (SUTVA)

SUTVA: The outcome of each unit depends only on its own treatment and not on the treatment received by others.

- A drug only affects the person taking it, assuming no biological spillovers.
- A training program influences only the participant's performance, not their coworkers'.

SUTVA consists of two primary assumptions:

1. **No Interference:** The treatment of one unit does not affect the outcome of another unit. Each unit's outcome depends only on its own treatment.

$$Y_i(D_i, D_{-i}) = Y_i(D_i)$$

2. **No Hidden Variations of Treatment:** The treatment must be uniquely defined, with no unobserved variations in how it is administered.

   D_i is well-defined and applied consistently across all treated units.

   For example, if a drug is considered the treatment, all treated units must receive the same formulation and dosage.

where D_{-i} represents treatments assigned to all units other than i.

Estimation Methods

1. **Experimental Data:**
   - Randomized Controlled Trials (RCTs)
   - Gold standard for estimating ATE.
   - Randomization ensures $E[Y(0)|D = 1] = E[Y(0)|D = 0]$.

2. **Observational Data:**
   - Use methods like difference-in-differences (DID), propensity score matching, regression adjustment, regression discontinuity design (RDD), instrumental variables, double machine learning, synthetic control etc.

Randomized Controlled Trials (RCT)

Randomized Controlled Trials (RCTs) in Economics

- **Randomized Controlled Trials (RCTs)** are experiments where subjects are randomly assigned to treatment and control groups.
- **Goal**: Estimate the causal effect of a policy, intervention, or treatment by comparing outcomes between the two groups.
- **Key Advantage**: Randomization eliminates selection bias, ensuring groups are comparable on average.
- **Economic Applications**: Evaluating the effects of cash transfer programs, microfinance, education subsidies, and job training programs.
- **Placebo Control**: Optional but recommended. Sometimes people may change their outcomes just because they believe they're receiving treatment — this is called the placebo effect. If you don't use a placebo in the control group, any observed improvement in the treatment group could be due to belief, not the actual treatment itself.
- **Design of experiments**: Optional reading: Michael Sullivan, Statistics: Informed Decisions Using Data, 7th edition.

RCT Assumptions

For RCTs to provide valid causal estimates, the following assumptions must hold:

- **Random Assignment**: Treatment assignment is random and independent of potential outcomes.
- **Excludability**: The treatment affects outcomes only through the assigned intervention (no spillover effects).
- **Stable Unit Treatment Value Assumption (SUTVA)**:
  - The outcome of one unit is unaffected by the treatment status of other units.
  - No hidden variations of the treatment.
- **Non-Interference**: No interference between treatment and control groups.

RCT Properties

- **Unbiasedness**: Randomization ensures that, on average, the treatment and control groups are identical except for the treatment.
- **Consistency**: The estimated treatment effect converges to the true causal effect as the sample size grows.
- **Efficiency**: RCTs provide precise estimates of treatment effects, especially with large sample sizes.
- **Transparency**: The randomization process is clear and replicable.

RCT Calculations

- Let $Y_i(1)$ be the potential outcome for unit i under treatment and $Y_i(0)$ under control.
- The **individual treatment effect** is:

$$\tau_i = Y_i(1) - Y_i(0).$$

- The **average treatment effect (ATE)** is:

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)].$$

- In RCTs, the ATE is estimated as:

$$\hat{\tau} = \frac{1}{N_T} \sum_{i \in T} Y_i - \frac{1}{N_C} \sum_{i \in C} Y_i,$$

where T is the treatment group, C is the control group, and N_T, N_C are their respective sizes.

RCT using OLS to Estimate Treatment Effects

- Randomization ensures that a simple regression can estimate the ATE without omitted variable bias.
- Consider the regression model:

$$Y_i = \beta_0 + \tau D_i + \varepsilon_i,$$

where:

  - Y_i is the outcome (e.g., test scores, wages, attendance).
  - D_i is the treatment indicator (1 if treated, 0 if control).
  - τ represents the ATE.

- In RCTs, OLS provides an unbiased estimate of β_1 because D_i is uncorrelated with omitted variables.
- Controlling for baseline characteristics (covariates) can improve precision.
- Heterogeneous treatment effects can be explored with interaction terms.

[Python code]

RCT Example: Microcredit in India

Study: Banerjee et al. (AER, 2015) examined the impact of microcredit on poverty alleviation in India.

Design: Randomized controlled trial in 104 slums across Hyderabad, India.

- Treatment group: Microcredit institutions opened branches in 52 randomly selected slums.
- Control group: No microcredit access in the remaining 52 slums.
- Outcomes tracked over 18 months, including household consumption, business creation, and women's empowerment.

Findings:

- Microcredit increased business creation by 1.7 percentage points.
- No significant impact on average household consumption or women's empowerment.
- Heterogeneous effects: Entrepreneurs with existing businesses benefited more.
- Policy: Microcredit is not a "miracle" solution for poverty but can help specific groups.

RCT Limitations

- **Ethical Concerns**: Randomization may not be ethical in some cases (e.g., withholding a known effective treatment).
- **External Validity**: Results may not generalize to other populations or settings.
- **Cost and Feasibility**: RCTs can be expensive and time-consuming to conduct.
- **Non-Compliance**: Subjects may not adhere to their assigned treatment, complicating the analysis.

Instrumental Variable (IV) Regression
Two Stage Least Squares (2SLS)
and Control Function Approach

[Read Greene Ch.8, Ch.13; Wooldridge Ch.15; Wooldridge2 Ch.5]

Instrumental Variable (IV) Regression

If the strict exogeneity condition $E(\epsilon|X) = 0$ or the weak exogeneity condition $E(X'\epsilon) = 0$ is violated, we can use a variable Z to instrument X such that:

- **Relevance**: $E(Z'X) \neq 0$ (the IV is correlated with X).
- **Exogeneity**: $E(Z'\epsilon) = 0$ (the IV is uncorrelated with the error).

But this is not part of the true model $y = X\beta + \epsilon$ and the above equations may differ across samples. So we can use the asymptotic versions [Greene, Ch. 8]:

$$\text{plim}\left(\frac{1}{n}Z'\epsilon\right) = 0 \Rightarrow \text{plim}\left(\frac{1}{n}Z'(y - X\beta)\right) = 0.$$

$$\Rightarrow \text{plim}\left(\frac{1}{n}Z'y\right) = \beta\,\text{plim}\left(\frac{1}{n}Z'X\right).$$

Solving for β, we obtain the IV estimator:

$$\hat{\beta}_{IV} = \left(Z'X\right)^{-1}\left(Z'y\right).$$

We have just proved that the estimator is consistent under the given conditions.

IV (Contd.)

Important: The dimension of Z must conform with X.

- Number of instruments in Z = Number of endogenous variables in X (The IV system is **exactly identified**).

Example: A simple version of the wage equation:

$$\text{wage} = \beta_0 + \beta_1 \text{education} + \beta_2 \text{experience} + \epsilon,$$

- Education is endogenous (correlated with ϵ because *ability* is omitted).
- Experience is exogenous (assumed uncorrelated with ϵ).

A valid instrument (e.g., proximity to a college) must satisfy:

- Relevance: the instrument is correlated with education.
- Exogeneity: the instrument is uncorrelated with ϵ.

Thus, both Z and X are $n \times K$.

- X has $K = 3$: intercept, education, and experience.
- Z also has $K = 3$: intercept, proximity to a college, experience.

In a funny way, the intercept and exogenous variable are IVs of themselves!

Asymptotic Variance of IV

The estimator of the error variance, which is asymptotically consistent, is:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - X_i\hat{\beta}_{IV})^2.$$

This estimator is asymptotically unbiased because $\hat{\beta}_{IV}$ is a consistent estimator of β.

Substituting $\hat{\beta}_{IV} = (Z'X)^{-1}Z'y$, the asymptotic variance of the IV estimator is:

$$\text{Asy. Var}(\hat{\beta}_{IV}) = \hat{\sigma}^2(Z'X)^{-1}(Z'Z)(X'Z)^{-1}.$$

Think: What happens if instruments are "weak"?

IV Regression: Cigarette Demand and Taxation

How do cigarette prices (real price) affect cigarette consumption (packs sold)?

- Cigarette price (`log_rprice`) is endogenous. So use IV: (`salestax`).
- IV Conditions: Sales tax affects cigarette sales only through prices. Sales tax changes are set by policymakers and are unlikely to be directly influenced by individual smoking behaviors.

IV Estimates

Variable	Coefficient	Std. Error	P-value
Constant	9.7745	0.5797	0.0000
Log Income (`log_rincome`)	0.2639	0.1447	0.0682
Log Price (`log_rprice`)	-1.2412	0.1640	0.0000

Interpretation:

- Price Elasticity: A 1% increase in real cigarette prices reduces cigarette consumption by 1.24% ($p < 0.01$)

Python code from scratch

Two-Stage Least Squares (2SLS) Motivation

- So far we have:

$$y = X\beta + \varepsilon,$$
$$\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y.$$

- OLS is biased if X is endogenous (i.e., X is correlated with ε).
- IV estimator is restrictive depending on the dimensions of Z and X:

$$\hat{\beta}_{\text{IV}} = (Z'X)^{-1}Z'y,$$

- The idea of 2SLS: Use a version of X, call it $\hat{X}$, that is uncorrelated with ε, capturing only the exogenous variation in X. But how?
- Stage 1: Regress X on instruments Z where $E(X'Z) \neq 0$ and $E(Z'\varepsilon) = 0$:

$$X = Z\gamma + \nu,$$

- Stage 2: Use the predicted $\hat{X} = Z\hat{\gamma}$ in place of X in the original equation.
- Now it is possible to have more instruments than endogenous variables (over-identified case).

Two-Stage Least Squares (2SLS)

First stage regression:

$$X = Z\gamma + \nu,$$

where $\gamma = (Z'Z)^{-1}Z'X$. Obtain the predicted values:

$$\hat{X} = Z(Z'Z)^{-1}Z'X = P_Z X,$$

where $P_Z = Z(Z'Z)^{-1}Z'$ is the projection matrix.

Second stage regression:

$$y = \hat{X}\beta + \epsilon.$$

The 2SLS estimator is:

$$\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'y.$$

Combined Formula: Substitute $\hat{X} = P_Z X$:

$$\hat{\beta}_{2SLS} = (X'P_Z X)^{-1}X'P_Z y.$$

This is equivalent to:

$$\hat{\beta}_{2SLS} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y.$$

2SLS Example: Education's Impact on Wages

Consider estimating

$$\log(\texttt{wage}) = \beta_0 + \beta_1 \texttt{educ} + \texttt{other Xs...} + u$$

Education may be endogenous due to omitted `ability`. Use IVs like mother's education (`motheduc`) and father's education (`fatheduc`).

2SLS Procedure:

1. **First Stage:** Regress `educ` on IVs and exogenous variables:

   $$\texttt{educ} = \pi_0 + \pi_1 \texttt{motheduc} + \pi_2 \texttt{fatheduc} + \texttt{other Xs...} + v$$

   Obtain fitted values: $\widehat{\texttt{educ}}$. Check instrument strength using the first-stage F-statistic.

2. **Second Stage:** Regress $\log(\texttt{wage})$ on $\widehat{\texttt{educ}}$ and exogenous variables:

   $$\log(\texttt{wage}) = \gamma_0 + \gamma_1 \widehat{\texttt{educ}} + \texttt{other Xs...} + \varepsilon$$

Key Insight: Using parents' education as instruments helps isolate the exogenous variation in `educ`, providing a consistent estimate of its effect on wages. If the model is over-identified, test instrument validity using an over-identification test.

Relevant Concept: Structural vs. Reduced Form

Structural Model:

- Causal relationship between variables derived from economic theory.
- Often includes endogenous variables. For example, X is endogenous (correlated with ε)

$$y = \beta_0 + \beta_1 X + \varepsilon,$$

Reduced Form Model:

- Expresses endogenous variables as functions of exogenous variables only.
- Eliminates endogeneity by substituting out endogenous variables.

$$y = \pi_0 + \pi_1 Z + u,$$

where Z is an instrument (exogenous) and u is the error term.

Connection to 2SLS: From the first stage, $\hat{X}$ is a function of Z, so plugging in $\hat{X}$ in the structural model for X gives the reduced form equation

$$\log(\texttt{wage}) = \lambda_0 + \lambda_1 \texttt{motheduc} + \lambda_2 \texttt{fatheduc} + \texttt{other Xs...} + \eta$$

Interpretation: The reduced form shows how the IVs affect the outcome variable directly.

Key Differences: IV vs. 2SLS

- **IV**: Most straightforward when the model is *just-identified* (one endogenous variable per instrument). The estimator is $\hat{\beta}_{IV} = (Z'X)^{-1}Z'y$. While IV can be extended to overidentified cases, 2SLS is typically preferred.
- **2SLS**: A specific IV method that efficiently handles *overidentified* models (multiple instruments or endogenous variables). Estimated in two stages: (1) regress endogenous variables on instruments, (2) use fitted values in the main equation.

When to Use What?

- Use **IV** if: The model is just-identified (instruments = endogenous variables) and a direct estimator suffices.
- Use **2SLS** if: The model is overidentified (more instruments than endogenous variables) or has multiple endogenous variables.

Both require instruments to be **relevant** *and* **exogenous**.

IV Regression Example: Returns to Schooling

Using Geographic Variation in College Proximity to Estimate the Return to Schooling [David Card (NBER, 1995)]

Research Question: Does additional education causally increase wages?

- **Endogenous Variable:** Years of education (*Schooling*)
- **Instrumental Variable:** Proximity to a 4-year college (*Distance*)
  - *Rationale:* Living near college reduces education costs (relevance) but distance doesn't directly affect wages (exclusion restriction)
- **Data:** 1976 National Longitudinal Survey (NLS) of Young Men
- **First Stage:** Strong instrument (F-stat > 10)
  Each 10 miles farther from college $\Rightarrow$ 0.3 fewer years of schooling
- **IV Estimate:** 13% higher wages per additional year of schooling
  (vs. 7% in OLS, suggesting ability bias)

Reducing college access barriers (e.g., distance, tuition) could significantly raise earnings.

GMM Estimation for IV Regression

The formulas for IV and 2SLS works when instruments are valid but depends on assumptions like homoskedasticity, so we need a more flexible approach.

Moment/Orthogonality Conditions: IV requires valid instruments Z such that:

$$E[Z'(y - X\beta)] = 0.$$

GMM Estimation Steps:

1. Define moment conditions: $g(\beta) = Z'(y - X\beta)$.
2. Use a weighting matrix W to minimize the quadratic form:

$$\hat{\beta}_{GMM} = \arg\min_{\beta} \; g(\beta)'.W.g(\beta).$$

3. The optimal weighting matrix $W = (Z'\Omega Z)^{-1}$, where $\Omega = E[\epsilon\epsilon']$, gives the efficient GMM estimator:

$$\hat{\beta}_{GMM} = (X'ZWZ'X)^{-1}X'ZWZ'y.$$

GMM Estimation (Contd.)

But is this feasible? We cannot observe Ω because ϵ is from the structural model. So we use a matrix that is feasible to obtain from the data $\hat{\Omega}$.

- Minimize the quadratic form using weighting matrix W:

$$\hat{\beta}_{GMM} = \arg\min_{\beta} \; g_n(\beta)' W g_n(\beta).$$

- Optimal W exploits the variance structure of moments. For efficient GMM:

$$W_{opt} = \left(\frac{1}{n} Z'\hat{\Omega} Z\right)^{-1}, \quad \hat{\Omega} = \text{diag}(\hat{\epsilon}_i^2) \text{ (heteroskedasticity-robust).}$$

  Yielding the estimator:

$$\hat{\beta}_{GMM} = \left(X'Z(Z'\hat{\Omega}Z)^{-1}Z'X\right)^{-1} X'Z(Z'\hat{\Omega}Z)^{-1}Z'y.$$

Python code from scratch

Why Use GMM Instead of the Simple IV Formula?

GMM Advantages:

- **Efficiency**: Accounts for heteroskedasticity/ cluster-wise variance using an optimal weighting matrix.
- **Overidentified Models**: Handles cases with more instruments than endogenous variables by minimizing moment conditions.
- **Weak Instruments**: Less bias in the presence of weak instruments.
- **Robust Testing**: Hansen's J-test helps assess instrument validity.
- **Flexible Moment Conditions**: Can adapt to autocorrelation, dynamic panels, and other complex settings.

If the model is homoskedastic and just-identified, IV and GMM give the same result. For real-world settings with heteroskedasticity and multiple instruments, GMM is more efficient.

Things to remember

IVs must affect Y only through X

OLS Unbiased and consistent, IV Consistent but biased ($\text{Var}(\beta_{IV}) > \text{Var}(\beta_{OLS})$). Check for instrumental validity.

1. Relevance (Rank Condition)
   - Check first-stage F-statistic: $F > 10$ (weak instrument test)
   - Stata `weakivtest` (Olea & Pflueger, 2015)
2. Exogeneity (Exclusion Restriction) [See Greene Ch.8]
   - Hausman and Wu specification test
   - A test for over-identification (Sargan/Hansen J-test)
   - Placebo/Falsification Test: Check if Y affects Z (should not)

Robustness Checks

- Alternative Instruments (Angrist-Krueger: QOB→schooling→wage)
- Use JIVE (Jackknife IV) if $F < 10$ ($\hat{X}_i$ using Z_{-i})
- Machine Learning for out-of-sample validation

Python code 2sls tests

Hausman and Wu Specification Test

Purpose: Tests whether an endogenous regressor needs IV correction.

Hypotheses:

- H_0 : OLS and IV estimates are similar (no endogeneity, OLS is efficient)
- H_A : OLS and IV estimates differ (endogeneity present, OLS is inconsistent)

Procedure:

1. Estimate $\hat{\beta}_{\text{OLS}}$ and $\hat{\beta}_{\text{IV}}$.
2. Compute test statistic:

$$H = (\hat{\beta}_{\text{OLS}} - \hat{\beta}_{\text{IV}})' V^{-1} (\hat{\beta}_{\text{OLS}} - \hat{\beta}_{\text{IV}})$$

   where V is the difference in covariance matrices.
3. Under H_0, $H \sim \chi^2_k$.

Interpretation:

- If H is small (high p-value), use OLS (no endogeneity).
- If H is large (low p-value), IV is necessary.

Durbin-Wu-Hausman (DWH) Test

Purpose: Compare OLS and IV estimates to detect endogeneity.

1. Estimate 1st-stage regression for the suspected endogenous regressor X_k:

$$X_k = Z\gamma + X_{-k}\delta + \nu$$

2. Save residuals $\hat{\nu}$ and add them to the original OLS model:

$$y = X\beta + \hat{\nu}\lambda + \epsilon$$

3. Test $H_0 : \lambda = 0$:
   - Reject H_0: Evidence of endogeneity (use IV).
   - Fail to reject: OLS is consistent.

Advantages:

- Simple implementation.
- Directly tests endogeneity.

Limitations:

- Requires valid instruments (exogeneity + relevance).
- Sensitive to weak instruments.

A relevant approach: Control Function Approach (CFA)

What if we try to "control" for the endogenous part of X_k.

1. Estimate the first-stage regression (same as DWH):

$$X_k = Z\gamma + X_{-k}\delta + \nu$$

2. Save residuals $\hat{\nu}$ and include them in the second-stage OLS:

$$y = X\beta + \hat{\nu}\lambda + \epsilon$$

3. Interpret $\hat{\beta}$ as the corrected estimate:
   - If $\lambda \neq 0$, X_k is endogenous.
   - Standard errors must be adjusted for two-step estimation.

Advantages:

- More efficient than 2SLS if errors are heteroskedastic.
- Flexible (*works for nonlinear models*). Example: Amin et al. (2021).

Limitations:

- Requires correct first-stage specification.
- Residuals $\hat{\nu}$ must be independent of Z.

Remember: Variances Change After IV

- Don't use regular OLS standard errors after using instruments!
- **IV Estimator:**

$$\text{Var}(\widehat{\beta}_{IV}) = (X'Z)(Z'Z)^{-1}\Omega(Z'Z)^{-1}(Z'X)$$

- **2SLS Estimator:**

$$\text{Var}(\widehat{\beta}_{2SLS}) = (X'P_ZX)^{-1}X'P_Z\Omega P_ZX(X'P_ZX)^{-1}$$

- **GMM Estimator:**

$$\text{Var}(\widehat{\beta}_{GMM}) = (G'WG)^{-1}G'WSWG(G'WG)^{-1}$$

G is the sensitivity of the moments to the parameters. S is the variance of the sample moments. W is the weighting matrix (ideally S^{-1}). If you use the optimal weight $W = S^{-1}$ the simplifies to $\text{Var}(\widehat{\beta}_{GMM}) = (G'S^{-1}G)^{-1}$

- **Control Function Approach (CFA):**

$$\text{Var}(\widehat{\beta}) = A^{-1}BA^{-1}$$

where A is the curvature of the second-stage loss (Hessian) and B captures the joint uncertainty from both stages (outer product of scores).

Test for Over-Identification (Sargan/Hansen J-Test)

Purpose: Checks if instruments are valid (uncorrelated with errors).

Hypotheses:

- H_0 : Instruments are valid (uncorrelated with ε)
- H_A : At least one instrument is invalid (correlated with ε)

Procedure:

1. Estimate 2SLS and obtain residuals $\hat{\varepsilon}$.
2. Regress $\hat{\varepsilon}$ on all exogenous variables (including instruments) and obtain R^2 from that regression.
3. Compute test statistic:
$$J = nR^2$$
where R^2 is from the residual regression.
4. Under H_0, $J \sim \chi^2_{(L-K)}$ (where L = instruments, K = endogenous variables).

Interpretation:

- If J is small (high p-value), instruments are valid.
- If J is large (low p-value), at least one instrument is invalid.

Relevant concept: Compliance in IV Framework

IV gives local average treatment effects (LATE). In an instrumental variables (IV) framework, individuals can be classified into compliance groups:

- **Compliers**: Individuals who take treatment ($D = 1$) if assigned to treatment and ($D = 0$) if not assigned.
- **Never-Takers**: Individuals who never take the treatment ($D = 0$) regardless of assignment.
- **Always-Takers**: Individuals who always take the treatment ($D = 1$) regardless of assignment.
- **Defiers**: Individuals who take treatment when not assigned and avoid treatment when assigned (typically ruled out by the monotonicity assumption).

The Local Average Treatment Effect (LATE) measures the causal effect for compliers:

$$LATE = E[Y(1) - Y(0)|\text{Compliers}]$$

LATE is identified when an instrument Z influences treatment assignment but is uncorrelated with potential outcomes beyond its effect through treatment.

Detecting Endogeneity Without IVs: Recent Developments

1. Sensitivity Analysis (e.g., Oster (2019) Bounds)
If treatment effects shift drastically with added controls, hidden bias may be present. Stable coefficients suggest robustness; instability hints endogeneity.

2. Heteroskedasticity-Based Tests (Klein & Vella 2010)
Use (known) heteroskedastic errors as a signal for omitted variables. Variance patterns can indirectly reveal correlation with unobservables. Similar to Lewbel (2012), Rigobon (2003).

3. Gaussian Copula (Park & Gupta; Rutz & Watson, 2019)
Joint distribution of error and endogenous X follows Gaussian copula. Error is normal, endogenous regressor is non-normal. This is similar to the Latent Structure Approach (Ebbes et al. 2005; Sonnier et al. 2011).

4. Moment-Based Tests
Instruments constructed from higher-order moments. Uses population moments and functional forms to estimate parameters. Relies on distributional assumptions and maximizes the likelihood (Wooldridge, 2012)

See Eckert, C. and Hohberger, J. (2022) for a survey.

The Difference-in-Differences (DID) Framework

[Read Baker et al. 2025]

DiD Motivation

Individual	Year	Cholesterol (Y)	Drug (D)	Post
1	2000	220	0	0
1	2001	215	0	1
2	2000	225	0	0
2	2001	220	0	1
3	2000	**230**	**1**	0
3	2001	**200**	**1**	1
4	2000	**235**	**1**	0
4	2001	**205**	**1**	1

Step 1: Mean change in post $-$ pre for each group:

- Control group (D=0): $\frac{215+220}{2} - \frac{220+225}{2} = 217.5 - 222.5 = \mathbf{-5}$
- **Treated group (D=1)**: $\frac{200+205}{2} - \frac{230+235}{2} = 202.5 - 232.5 = \mathbf{-30}$

Step 2: Difference-in-Differences estimate:

$$\text{DiD} = -30 - (-5) = \boxed{-25}$$

Note: The part of Y that depends on time-invariant, individual-level characteristics is canceled out through differencing, making the treated and control groups more comparable.

DiD as a Regression

We can estimate the DiD using a linear regression with an interaction term:

$$Y_{it} = \alpha + \beta \cdot \text{Post}_t + \gamma \cdot \text{Treated}_i + \delta \cdot (\text{Treated}_i \times \text{Post}_t) + \epsilon_{it}$$

- $\text{Post}_t = 1$ if year is 2001, 0 otherwise
- $\text{Treated}_i = 1$ if individual received the drug, 0 otherwise
- δ captures the **Difference-in-Differences estimate**

From previous slide:

Control change: $217.5 - 222.5 = -5$ Treated change: $202.5 - 232.5 = -30$

$$\Rightarrow \boxed{\delta = -30 - (-5) = -25}$$

Interpretation: The drug is associated with a 25-point additional reduction in cholesterol levels compared to the control group.

DiD via Two-Way Fixed Effects (TWFE)

We can also estimate DiD using a TWFE specification:

$$Y_{it} = \alpha_i + \lambda_t + \delta \cdot D_{it} + \epsilon_{it}$$

- α_i = individual fixed effect (for time-invariant differences between units)
- λ_t = time fixed effect (for shocks common to all units in each period)
- D_{it} = treatment indicator: 1 if treated **and** post, 0 otherwise
- δ = Difference-in-Differences estimate

Calculate the DiD to see individual Fixed Effects disappear. In our setup:

- Treated group: individuals 3 and 4
- Post period: year 2001
- So, $D_{it} = 1$ only for 3 and 4 in 2001
- Here as well, $\delta = -25$.

Note: In this simple case with two periods and binary treatment, TWFE and DiD with interaction are numerically identical.

[Python Code]

Estimating Causal Effects in a 2-Period Panel

Several equivalent ways to estimate DiD effects in a balanced panel with treatment and control groups:

1. **Group-Time Average DiD Estimator (Classic Formula):**

$$\hat{\delta}_{DiD} = (\bar{Y}_{Treated,post} - \bar{Y}_{Treated,pre}) - (\bar{Y}_{Control,post} - \bar{Y}_{Control,pre})$$

2. **Two-Way Fixed Effects (TWFE) Regression:**

$$Y_{it} = \alpha_i + \lambda_t + \delta D_{it} + \varepsilon_{it}$$

3. **First-Difference Regression:** (Where $\Delta x_t = x_t - x_{t-1}$)

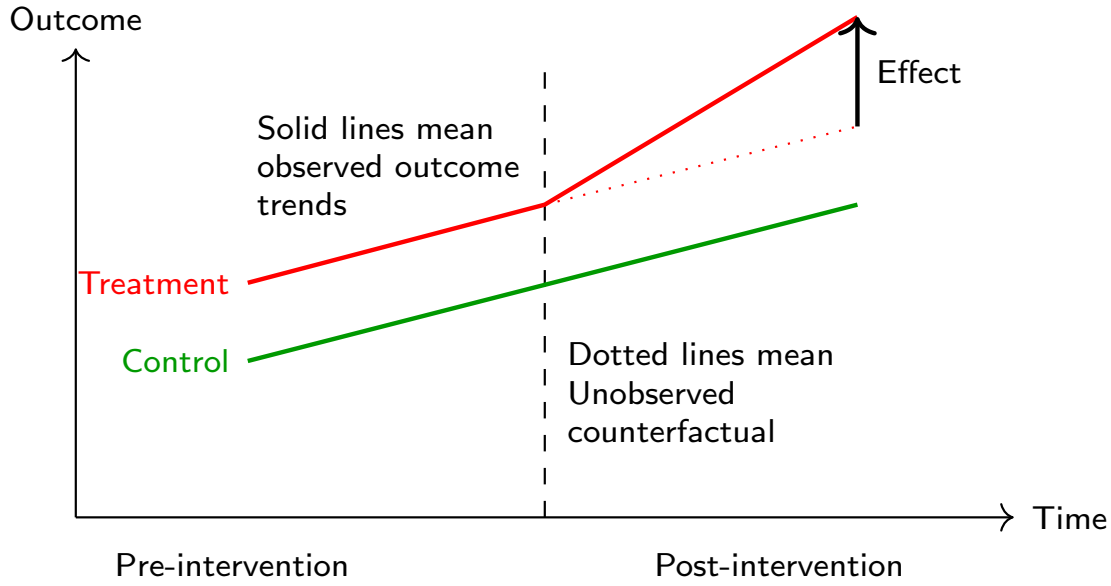
$$\Delta Y_i = \delta \cdot \Delta D_i + \Delta \varepsilon_i$$

4. **Interaction Regression (No Fixed Effects Form):**

$$Y_{it} = \beta_0 + \beta_1 \text{Treat}_i + \beta_2 \text{Post}_t + \delta(\text{Treat}_i \times \text{Post}_t) + \varepsilon_{it}$$

Note: All estimators rely on the parallel trends assumption. With only two periods, TWFE and interaction models are numerically equivalent.

Parallel Trends Assumption



Definition: In the absence of treatment, the treated and control groups would have followed the *same trend* over time.

Interpretation: The dotted red line represents what would have happened to the treated group if they hadn't received treatment. This is the core identifying assumption in DiD.

Canonical 2x2 DiD Design

- 2 Groups (Treated & Control), 2 Periods (Pre & Post)
- Define potential outcomes $Y_{it}(0)$, $Y_{it}(1)$
- ATT in period t:

$$ATT_t = \mathbb{E}[Y_{it}(1) - Y_{it}(0)|D_i = 1]$$

- Under Parallel Trends (PT), identify ATT via:

$$ATT_t = (\bar{Y}_{T,post} - \bar{Y}_{T,pre}) - (\bar{Y}_{C,post} - \bar{Y}_{C,pre})$$

- Complexities arise in extended settings, e.g., more than two groups, more than two times, heterogeneous treatment effects, staggered treatment.
- Relevant concept: **Staggered treatment** Units are exposed to a treatment (policy/intervention) at different points in time, and once treated, they remain treated thereafter.

Assumptions for Identification

1. **No anticipation of treatment:**

$$Y_{it}(1) = Y_{it}(0) \quad \text{for all } t < G_i$$

where G_i typically denotes the period in which unit i receives treatment (i.e., the treatment cohort or first treated period).

2. **Parallel Trends:** Without treatment, both treated and control group could have the same slope.

$$\mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0) \mid D_i = 1] = \mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0) \mid D_i = 0]$$

3. **Stable Unit Treatment Value Assumption (SUTVA):** No interference across units, and treatment is homogeneous.

4. **Homogeneous treatment effects:** $\delta_i = \delta$.
Modern methods allow group-specific treatment effects based on treatment timing.

- Under these assumptions, DiD identifies the Average Treatment Effect on the Treated (ATT).

DID Example: Minimum Wage Effects

Minimum Wage and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania [Card & Krueger (AER, 1994)]

Research Question: Does increasing the minimum wage reduce employment?

- **Treatment:** New Jersey (min. wage increased from 4.25 to 5.05)
- **Control:** Eastern Pennsylvania (no minimum wage change)
  - *Rationale:* Similar fast-food markets but different policy changes Parallel trends assumption: Employment trends would be similar absent treatment
- **Data:** Survey of 410 fast-food restaurants before/after policy change
- **DID Estimate:** Employment in NJ *increased* by 0.6 FTE relative to PA (Contradicting simple competitive model predictions)
- **Robustness:** Matched sample + placebo tests supported parallel trends

Policy: Moderate minimum wage hikes may not reduce employment.

Things to Remember: DiD and TWFE

- For time-varying predictors, consider **conditional parallel trends**:

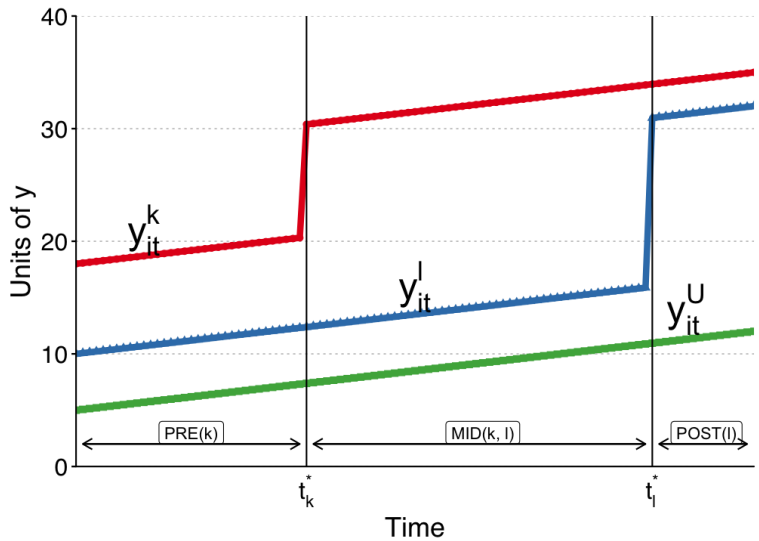
$$\mathbb{E}[\Delta Y_{it}(0) \mid X_i, D_i = 1] = \mathbb{E}[\Delta Y_{it}(0) \mid X_i, D_i = 0]$$

- Conditioning on X_i (e.g., through regression or weighting) can improve the plausibility of the parallel trends assumption.
- In that case you are assuming an **overlap**: That the probability of receiving treatment is similar in both groups after controlling for X.
- You can use: Inverse Probability Weighting (IPW), Regression adjustment, or Doubly robust methods for controlling for selection into treatment (see Propensity Score Matching literature).
- Always cluster standard errors at the unit or group level.
- How? Three perspectives on error structure (hence inference)
  - *Design-based:* Treatment is as random as in a lab experiment.
  - *Sampling-based:* Units are randomly sampled from a population.
  - *Model-based:* Your structural model generates assumptions about the error structure.

Why TWFE Was Not Enough?

Limitations of Traditional DiD / TWFE Models:

- TWFE average over all possible 2x2 comparisons in staggered settings.
- When treatment effects are heterogeneous across groups or over time, TWFE estimates can be **biased** and **hard to interpret**.
- Some comparisons may involve **already-treated units as controls**, which can generate **negative or misleading weights** [Goodman-Bacon 2021].



Recent methods correct for TWFE biases and offer valid event study designs under heterogeneity.

Recent Developments in DID

Problems with TWFE under Staggered Timing:

- Traditional TWFE estimators rely on a weighted average of all possible 2x2 DiD comparisons (Goodman-Bacon, 2021).
- Under staggered adoption, TWFE implicitly uses:
  - Early-treated units as controls for later-treated units (violating the "no treatment reversal" assumption).
  - Negative weights on some comparisons, leading to biased ATT estimates if treatment effects are heterogeneous (e.g., dynamic effects).

Key issue: TWFE compares early-treated ($G_i = g$) units to not-yet-treated units (where $G_i > t$) rather than to **never-treated** or **proper controls**.

Recent Developments (Contd.)

Heterogeneous Treatment Effects (HTE) Exacerbate the Problem:

- If treatment effects vary across cohorts or over time (e.g., fading effects), TWFE weights may:
  - Overweight early-treated units with smaller effects.
  - Assign **negative weights** to certain comparisons (e.g., late-treated vs. early-treated).
- Result: TWFE estimates may be **attenuated** or even **opposite in sign** to the true ATT (Roth et al., 2023).

Recent methods address these issues by explicitly modeling heterogeneity:

- **Sun & Abraham (2021):** Propose an event-study design robust to heterogeneous treatment effects.
- **Callaway & Sant'Anna (2021):** Estimate group-time average treatment effects ($\text{ATT}_{g,t}$).
- **Borusyak, Jaravel, & Spiess (2024):** Use imputation-based estimators to recover ATT.

Recent Dev: Callaway & Sant'Anna (2021)

- Callaway & Sant'Anna (2021) propose estimating treatment effects separately for each group, based on when they first receive treatment.
- The key quantity is the group-time average treatment effect (ATT):

$$ATT_{g,t} = \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid G_i = g,\ t \geq g]$$

- $G_i = g$ indicates that unit i first received treatment in period g, and $t \geq g$ ensures we are considering post-treatment periods for that group.
- Since we observe $Y_{it}(1)$ but not $Y_{it}(0)$ once treated, they estimate the counterfactual $Y_{it}(0)$ using outcomes from not-yet-treated or never-treated groups.
- This allows estimation of treatment effects for each group g in each period $t \geq g$, respecting treatment timing and allowing heterogeneity.
- These group-time effects are then aggregated:

$$ATT = \sum_{g} w_g \cdot ATT_{g,t}$$

- Where, w_g is the share of treated units that first received treatment in period g.

Example of CS2021: Estimating $ATT_{2,2}$ with Toy Data

Toy panel: 6 units, 4 periods, staggered treatment timing.

Unit	Group G_i	Time t	Treated?	Y_{it}
A	2	2	Yes	7
B	3	2	No	4.5
C	Never	2	No	6.5

Goal: Estimate $ATT_{2,2}$ (effect in period 2 for group 2)

- Unit A is treated in $t = 2$, so we observe $Y_{A,2}(1) = 7$.
- But $Y_{A,2}(0)$ (the untreated outcome) is not observed.
- Estimate $Y_{A,2}(0)$ using outcomes from units not yet treated:

$$\hat{Y}_{A,2}(0) = \frac{Y_{B,2} + Y_{C,2}}{2} = \frac{4.5 + 6.5}{2} = 5.5$$

- Thus, estimated treatment effect:

$$ATT_{2,2} = Y_{A,2}(1) - \hat{Y}_{A,2}(0) = 7 - 5.5 = \boxed{1.5}$$

[Python code]

Event Study

- Event studies estimate how outcomes evolve relative to the timing of a treatment or event (Dynamic Effects).
- They allow visualization and formal testing of the parallel trends assumption using pre-treatment periods.
- Useful for capturing dynamic treatment effects before and after the event, including potential anticipation or delayed effects.
- Particularly helpful in staggered adoption settings to track effect heterogeneity across cohorts and over time.
- Widely applied in policy evaluation, industrial organization, labor economics, and health economics.

[Read Miller 2023]

Event Study: Intuition

Estimate the dynamic causal effect of a treatment (e.g., policy change) relative to event time.

- Organizes outcomes relative to treatment timing (leads and lags).
- Controls for unit-specific (α_i) and time-specific (γ_t) effects.
- Tests for **parallel pre-trends** (critical identification assumption).
- Basic model (with reference period omitted for identification):

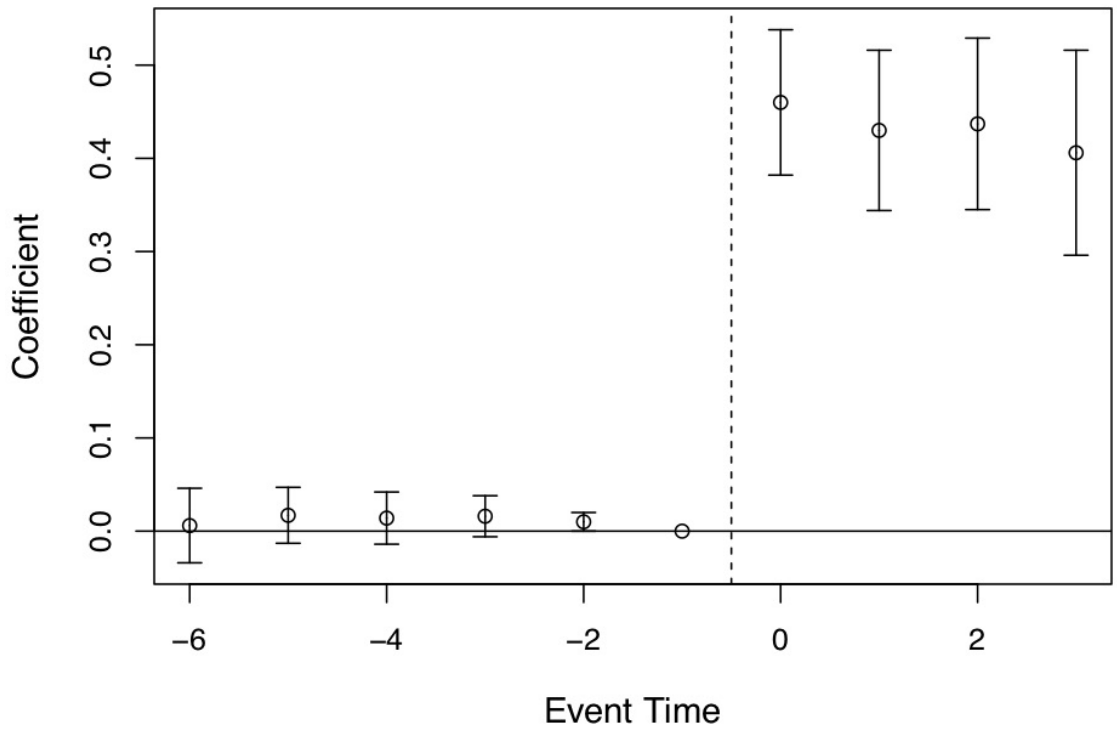
$$Y_{it} = \alpha_i + \gamma_t + \sum_{k \neq -1} \beta_k D_{it}^k + \varepsilon_{it}$$

- Example: 6-period model

$$Y_{it} = \alpha_i + \gamma_t + \beta_{-3} D_{it}^{-3} + \beta_{-2} D_{it}^{-2} + \beta_0 D_{it}^0 + \beta_1 D_{it}^1 + \beta_2 D_{it}^2 + \varepsilon_{it}$$

- Here, D_{it}^{-3}, D_{it}^{-2}, D_{it}^0, D_{it}^1, D_{it}^2 are dummy variables.
- The period -1 (just before treatment) is omitted as the reference group.
- $\beta_{-3}, \beta_{-2} = 0$? $\rightarrow$ Parallel trends hold (Use F-test).
- $\beta_0, \beta_1, \beta_2 \rightarrow$ Treatment effects.

Event Study: Estimated Effects Over Time



Notes: Each point shows the estimated treatment effect ($\hat{\beta}_k$) at event time k. The dashed vertical line at $k = 0$ marks treatment onset. The omitted reference period is $k = -1$. Confidence intervals (95%) indicate uncertainty around estimates.

Event Study: Assumptions & Application

Key Assumptions

- **Parallel Trends**:

$$\mathbb{E}[Y_{it}(0) - Y_{it-1}(0) \mid D_{it}^{k}] \text{ is constant across } k.$$

- **No Anticipation**: Only post-treatment effects ($\beta_{k<0} = 0$).
- **Staggered Adoption**: Units may be treated at different times.
- **Testing**: Examine whether pre-treatment β_k estimates are jointly zero.

[Python code Event Study]

Recommended Practices

- Define event time carefully (e.g., months since treatment).
- Create binned dummies if the event window is wide (e.g., $t < -6$).
- Use fixed effects regression (e.g., unit and time FE via `PanelOLS`).
- Interpret coefficients relative to the reference period (e.g., $k = -1$).

Event Study: **Do's**

1. **Balance the event window**: Ensure all units have sufficient pre- and post-event observations.
2. **Use a clear reference period**: Normalize coefficients by omitting one pre-treatment dummy (e.g., $t = -1$).
3. **Check parallel trends**: Use an F-test to jointly test equality of pre-treatment coefficients.
4. **Cluster standard errors**: Usually by unit (e.g., store, individual) to account for autocorrelation.
5. **Visualize dynamics**: Plot event-time coefficients with confidence intervals.
6. **Interpret with caution**: Especially near the event time where effects may be anticipatory or noisy.

Event Study: **Don'ts**

1. **Don't include late-treated units without proper adjustment**: This may contaminate post-treatment periods.
2. **Don't use post-treatment dummies as pre-trend tests**: Only use pre-treatment dummies for that.
3. **Don't forget control variables**: Include trends, seasonality, or fixed effects as needed.
4. **Don't extrapolate beyond the window**: Effects outside the event range are not reliable.
5. **Don't center or scale coefficients for hypothesis testing**: Use raw coefficients for F-tests.