Causal Analysis

Impact Evaluation and Causal Machine Learning with Applications in R

Chapter 3: Social Experiments and Linear Regression (1)

Table of Contents

Intuition of Experiments

- Subjects are randomized into treatment and control groups (e.g., by coin flip), see e.g. Fisher (1935).
- This ensures that background characteristics U are comparable across the two groups (at least in large samples).
- Therefore, the causal effect of the treatment D can be assessed by comparing the outcomes Y of both groups.
- Graphical illustration:

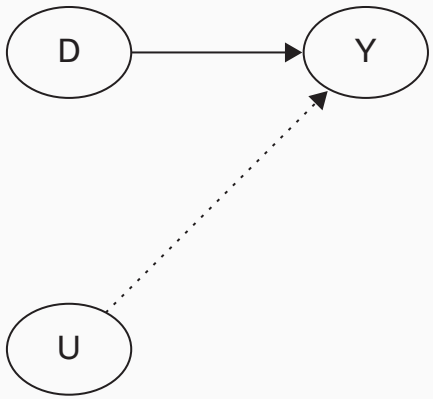


Figure 2.3: No treatment selection bias

Independence Assumption in Experiments

- Potential outcomes $Y(1)$, $Y(0)$ are statistically independent of the treatment D:

$$\{Y(1), Y(0)\} \perp D, \tag{3.1}$$

  where $\perp$ denotes statistical independence.

- As a consequence, means of the potential outcomes are comparable across treatment and control groups. Therefore,

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

- The ATE corresponds to the mean difference in the outcomes of treated and nontreated observations (no selection bias):

$$\Delta = E[Y(1)] - E[Y(0)] = E[Y|D = 1] - E[Y|D = 0] \tag{3.2}$$

Sample Version

- The average outcomes $E[Y|D = 1]$ and $E[Y|D = 0]$, and the ATE on the previous slide, refer to the total population.
- However, experiments are typically conducted in a sample drawn randomly from the population of interest for representativeness.
- The ATE in the sample can be estimated by:

$$\frac{\sum_{i=1}^{n} Y_i \cdot D_i}{\sum_{i=1}^{n} D_i} - \frac{\sum_{i=1}^{n} Y_i \cdot (1 - D_i)}{\sum_{i=1}^{n} (1 - D_i)} \tag{3.3}$$

  - n is the sample size, and $i \in \{1, 2, \ldots, n\}$ is the index of a specific observation in the sample.
  - The first term refers to the mean outcome among the treated in the sample.
  - The second term refers to the mean outcome among the nontreated in the sample.

Table of Contents

Representation of Causal Effects by Linear Regression

- ATE evaluation under a randomized treatment can also be expressed as a linear regression problem (Gauss, 1809).
- Conditional mean of outcome Y given treatment D:

$$\begin{aligned} E[Y|D] &= E[Y(0) + (Y(1) - Y(0)) \cdot D|D] \\ &= E[Y(0)|D] + \{E[Y(1)|D] - E[Y(0)|D]\} \cdot D \end{aligned} \tag{3.4}$$

- Under the independence assumption (3.1):

$$E[Y|D] = \underbrace{E[Y|D=0]}_{\alpha} + (\underbrace{E[Y|D=1] - E[Y|D=0]}_{\beta}) \cdot D$$

  - β equals the ATE, $\Delta = E[Y(1) - Y(0)]$.
  - α equals the mean potential outcome under nontreatment, $E[Y(0)]$.
  - $\alpha + \beta$ equals the mean potential outcome under treatment, $E[Y(1)]$.

Error Term

Error term ε

Difference between the observed outcome Y and its respective conditional mean in a specific treatment group, $E[Y|D]$

$$\varepsilon = Y - \underbrace{(\alpha + \beta D)}_{E[Y|D]} \tag{3.5}$$

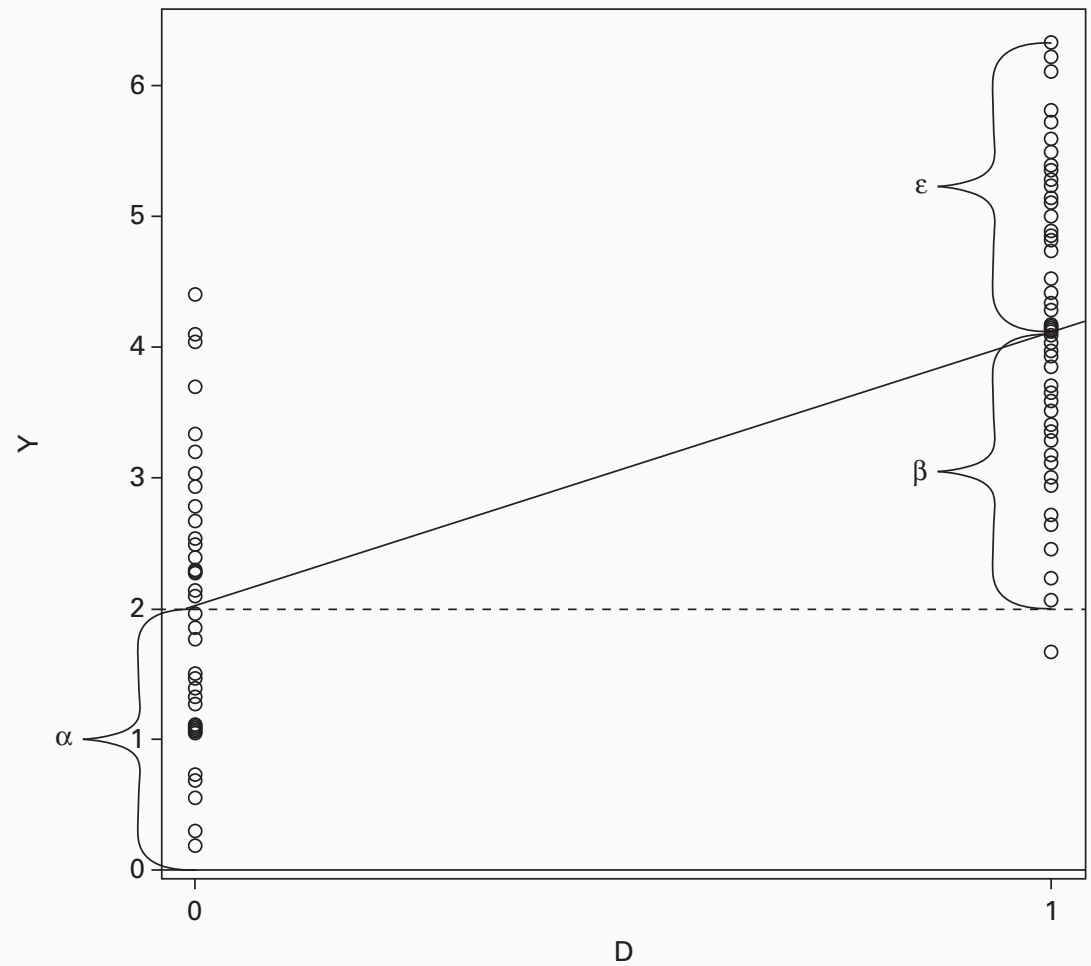
- Rearranging terms shows that Y can be expressed as:

$$Y = \underbrace{\alpha + \beta D}_{E[Y|D]} + \varepsilon \tag{3.6}$$

  - $E[Y|D]$: Average outcome in a specific treatment state.
  - ε: Deviation of Y from the average outcome.

Graphical Illustration

Figure 3.1: Linear regression



Moment Conditions (1)

- To compute coefficients α and β, linear regression is based on exploiting two specific properties of error term ε.
- These properties are known as **moment conditions**, as they refer to the first moments (means) of the distribution of a variable.

First moment condition

Deviations from a variable's mean must average to zero.

$$E[\varepsilon] = E[Y] - E[E[Y|D]] = E[Y] - E[Y] = 0 \tag{3.7}$$

- The second equality in (3.7) follows from the law of iterated expectations.

Law of iterated expectations

The mean of a variable corresponds to the mean of the conditional means of that variable.

Moment Conditions (2)

- Derivation of the second moment condition:

$$\begin{aligned}\varepsilon &= Y - E[Y|D] \\ &= Y - E[Y(0)] - \{E[Y(1)|D] - E[Y(0)|D]\} \cdot D \\ &= Y - E[Y(1)] \cdot D - E[Y(0)] \cdot (1 - D) \\ &= (Y(1) - E[Y(1)]) \cdot D + (Y(0) - E[Y(0)]) \cdot (1 - D) \end{aligned} \tag{3.8}$$

  - Independence assumption (3.1) permits replacing $E[Y|D]$ in the first line with functions of mean potential outcomes.
  - The last equality follows from the definition of the observed outcome, $Y = Y(1) \cdot D + Y(0) \cdot (1 - D)$.

- Since $E[Y(1) - E[Y(1)]] = 0$ and $E[Y(0) - E[Y(0)]] = 0$:

$$E[\varepsilon|D] = 0$$

Moment Conditions (3)

Second moment condition

$$E[D \cdot \varepsilon] = 0 \tag{3.9}$$

Follows from the law of iterated expectations: $E[D \cdot \varepsilon] = E[D \cdot E[\varepsilon|D]]$.

- It can be shown that:

$$E[D \cdot \varepsilon] = E[(D - E[D]) \cdot \varepsilon] = Cov(\varepsilon, D)$$

  where Cov denotes covariance.

- Independence assumption (3.1) implies that the covariance of the treatment and the error term in the population is zero:

$$Cov(\varepsilon, D) = 0$$

ATE Identification (1)

- Solve the first and second moment conditions for α and β.
- The first moment condition implies:

$$\begin{aligned} E[\varepsilon] &= E[Y - \alpha - \beta D] = 0 \\ \Leftrightarrow \alpha &= E[Y] - \beta E[D] \end{aligned} \tag{3.10}$$

- The second moment condition implies:

$$\begin{aligned} E[D \cdot \varepsilon] &= E[D \cdot (Y - \alpha - \beta D)] = 0 \\ &= E[D \cdot (Y - E[Y] - \beta(D - E[D]))] = 0 \\ \Leftrightarrow \beta &= \frac{E[D \cdot (Y - E[Y])]}{E[D \cdot (D - E[D])]} = \frac{E[(D - E[D]) \cdot (Y - E[Y])]}{E[(D - E[D]) \cdot (D - E[D])]} \\ &= \frac{Cov(D, Y)}{Var(D)} \end{aligned} \tag{3.11}$$

  - The second equality follows from the definition of α above.

ATE Identification (2)

- Linear regression identifies the ATE of a binary treatment in experiments as:

$$\beta = \frac{Cov(D, Y)}{Var(D)}$$

- This covariance-variance-ratio thus corresponds to $E[Y|D = 1] - E[Y|D = 0]$

- Plugging β into equation (3.10) identifies:

$$\alpha = E[Y|D = 0] = E[Y(0)]$$

(i.e., the mean potential outcome under nontreatment)

Alternative Approach for ATE Identification

- Express linear regression as the following optimization problem to identify the coefficients α and β:

$$\alpha, \beta = \arg \min_{\alpha^*, \beta^*} E[\underbrace{Y - \alpha^* - \beta^* D}_{\varepsilon}]^2 \tag{3.12}$$

  - α^* and β^* represent a range of candidate values for α and β.
  - Select the values that minimize the expectation of the squared error terms.
- First-order conditions with respect to α^* and β^*:

$$E[\varepsilon] = 0 \quad \text{and} \quad E[D \cdot \varepsilon] = 0$$

  - Correspond to the moment conditions on the previous slides.
  - Choose α^* and β^* such that these derivatives are zero.
- Note that even though we have considered linear regression, no linear relationship is imposed when the treatment is binary.

Table of Contents

Estimating the ATE in a Sample

- Estimation of the ATE based on linear regression in the sample:

$$\hat{\alpha}, \hat{\beta} = \arg \min_{\alpha^*, \beta^*} \sum_{i=1}^{n} (Y_i - \alpha^* - \beta^* D_i)^2 \tag{3.13}$$

  - $\hat{\alpha}, \hat{\beta}$: Sample estimates of the population parameters α, β.
  - α^*, β^*: Candidate values chosen such that the sum of squared residuals is minimized.
  - Residuals: Deviations between observed outcomes and the estimated regression line.

- Minimizing the mean squared residuals (instead of the sum) yields the same result, as the mean is the sum divided by n:

$$\hat{\alpha}, \hat{\beta} = \arg \min_{\alpha^*, \beta^*} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \alpha^* - \beta^* D_i)^2$$

- Linear regression is also known as **ordinary least squares (OLS)**.

Sample-Based Estimates β and α

- Estimate $\hat{\beta}$ in the minimization problem on the previous slide:

$$\hat{\beta} = \frac{\widehat{Cov}(Y_i, D_i)}{\widehat{Var}(D_i)}, \text{ where} \tag{3.14}$$

$$\widehat{Cov}(Y_i, D_i) = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right) \left(D_i - \bar{D}\right) \text{ and}$$

$$\widehat{Var}(D_i) = \frac{1}{n-1} \sum_{i=1}^{n} \left(D_i - \bar{D}\right)^2.$$

- $\hat{\beta}$ corresponds to the mean difference in outcomes between treated and nontreated groups in the sample.
- Estimate $\hat{\alpha}$ in the minimization problem on the previous slide:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{D} \tag{3.15}$$

- $\bar{D} = \frac{1}{n} \sum_{i=1}^{n} D_i$ and $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ denote the sample averages of the treatment and the outcome.

Desirable Properties of Estimators (1)

- $\hat{\beta}$ and $\hat{\alpha}$ may differ from their true values in a given sample, as the sample may not be fully representative of the population.
- In this case, the independence assumption (3.1) might not hold exactly in the sample.

Three properties are desirable for sample-based estimates:

Unbiasedness

The estimates hit their true values on average when applied to infinitely many randomly drawn samples.

$$E[\hat{\beta}] = \beta, \quad E[\hat{\alpha}] = \alpha \tag{3.16}$$

- The expectations of estimates $\hat{\beta}$ and $\hat{\alpha}$ correspond to the true parameters β and α.

Desirable Properties of Estimators (2)

Consistency

As the sample size increases, $\hat{\beta}$ is more and more likely to be close to the true β.

$$\Pr(|\hat{\beta} - \beta| > \epsilon) \to 0 \text{ for any } \epsilon > 0 \text{ when } n \to \infty \tag{3.17}$$

- The probability of obtaining an estimate $\hat{\beta}$ that is different to the true effect β by more than ϵ goes down in a larger sample.
- When the sample size goes to infinity, $\hat{\beta}$ collapses to β.

Equivalent statement: β is the probability limit of $\hat{\beta}$:

$$plim(\hat{\beta}) = \beta \tag{3.18}$$

Desirable Properties of Estimators (3)

Asymptotic normality

The pooled estimates $\hat{\beta}$ and $\hat{\alpha}$ obtained from many randomly drawn, sufficiently large samples follow a normal distribution.

- Enables approximating the distribution of an estimate across many samples, even with only a single sample at hand.
- Useful for statistical inference (e.g., confidence intervals and hypothesis testing).

- The OLS-based estimates $\hat{\beta}$ and $\hat{\alpha}$ of the ATE and the mean potential outcomes satisfy all three desired properties.
- Other estimators may also satisfy some or all of these properties under particular identifying assumptions.

Unbiasedness of $\hat{\beta}$ (1)

To show the unbiasedness of $\hat{\beta}$, replace Y_i in $\frac{\widehat{Cov}(Y_i, D_i)}{\widehat{Var}(D_i)}$ with $Y_i = \alpha + \beta D_i + \varepsilon$.

$$
\begin{aligned}
\hat{\beta} &= \frac{\widehat{Cov}(\alpha + \beta D_i + \varepsilon_i, D_i)}{\widehat{Var}(D_i)} \\
&= \beta \frac{\widehat{Var}(D_i)}{\widehat{Var}(D_i)} + \frac{\widehat{Cov}(\varepsilon_i, D_i)}{\widehat{Var}(D_i)} \\
&= \beta + \frac{\widehat{Cov}(\varepsilon_i, D_i)}{\widehat{Var}(D_i)} \\
&= \beta + \frac{\frac{1}{n-1} \sum_{i=1}^{n} \left(\varepsilon_i - \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i\right) (D_i - \bar{D})}{\frac{1}{n-1} \sum_{i=1}^{n} (D_i - \bar{D})^2} \\
&= \beta + \frac{\sum_{i=1}^{n} \varepsilon_i \cdot (D_i - \bar{D})}{\sum_{i=1}^{n} (D_i - \bar{D})^2} \qquad (3.19)
\end{aligned}
$$

- **Line 2:** α has a covariance of zero (as α is a constant). The covariance of D_i with itself equals its variance.
- **Line 3:** $\hat{\beta}$ corresponds to β plus the sample covariance of D_i and ε_i divided by the sample variance of D_i.
- **Line 4:** Provides the formulas for the covariance and variance terms.
- **Line 5:** Follows from the fact that $\frac{1}{n-1} \sum_{i=1}^{n} (\varepsilon_i - \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i)(D_i - \bar{D}) = \frac{1}{n-1} \sum_{i=1}^{n} \varepsilon_i \cdot (D_i - \bar{D})$.

  $\frac{1}{n-1}$ cancels out in the numerator and denominator.

Unbiasedness of $\hat{\beta}$ (2)

- Taking expectations of the terms in equation (3.19) yields:

$$\begin{aligned} E[\hat{\beta}] &= \beta + E\left[\frac{\sum_{i=1}^{n} \varepsilon_i \cdot (D_i - \bar{D})}{(D_i - \bar{D})^2}\right] \\ &= \beta + E\left[\frac{\sum_{i=1}^{n} \overbrace{E[\varepsilon_i|D_i]}^{=0} \cdot (D_i - \bar{D})}{(D_i - \bar{D})^2}\right] \\ &= \beta \end{aligned} \tag{3.20}$$

- The second line follows from the law of iterated expectations.
- Unbiasedness holds because the errors have an expectation of zero in either treatment group: $E[\varepsilon_i|D_i] = 0$.

Unbiasedness of $\hat{\alpha}$

- To show the unbiasedness of $\hat{\alpha}$, take expectations in $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{D}$:

$$\begin{aligned} E[\hat{\alpha}] &= E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i - \hat{\beta}\frac{1}{n}\sum_{i=1}^{n} D_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[Y_i] - \frac{1}{n}\sum_{i=1}^{n} E[\hat{\beta} D_i] \\ &= \frac{n}{n}E[Y] - \beta\frac{n}{n}E[D] = E[Y] - \beta E[D] = \alpha \end{aligned} \tag{3.21}$$

- The third line follows from the fact that:
  - $E[\hat{\beta}] = \beta$, which has been shown on the previous slide.
  - The sum of n identical averages is n times the average: $\sum_{i=1}^{n} E[Y_i] = n \cdot E[Y]$ and $\sum_{i=1}^{n} E[D_i] = n \cdot E[D]$.

Consistency of $\hat{\beta}$

- Use the *plim* operator to verify to which expressions the parameters converge as the sample size goes to infinity.
- To show the consistency of $\hat{\beta}$, consider the probability limits of the third line of equation (3.19):

$$plim(\hat{\beta}) = plim(\beta) + plim\left(\frac{\widehat{Cov}(\varepsilon_i, D_i)}{\widehat{Var}(D_i)}\right)$$

$$= \beta + \frac{Cov(\varepsilon, D)}{Var(D)} = \beta \tag{3.22}$$

  - Since β is a constant (the ATE in the population), its probability limit is β itself.
  - Sample covariances and variances converge to the covariances and variances in the population by the weak law of large numbers.
  - The last equality follows from $Cov(\varepsilon, D) = 0$ (see slide 12).

Consistency of $\hat{\alpha}$

- To show the consistency of $\hat{\alpha}$, consider the probability limits of $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{D}$:

$$plim(\hat{\alpha}) = plim\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \hat{\beta}\sum_{i=1}^{n} D_i\right) = E[Y] - \beta E[D] = \alpha \qquad (3.23)$$

- The last equality follows from the definition of α in equation (3.10).

Asymptotic Normality of $\hat{\beta}$ (1)

- To show the asymptotic normality of $\hat{\beta}$, reconsider the fifth line of equation (3.19) and bring β to the left side:

$$\begin{aligned}
\hat{\beta} - \beta &= \frac{\sum_{i=1}^{n} \varepsilon_i \cdot (D_i - \bar{D})}{\sum_{i=1}^{n} (D_i - \bar{D})^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \cdot (D_i - \bar{D})}{\frac{1}{n}\sum_{i=1}^{n} (D_i - \bar{D})^2} \\
\Leftrightarrow \sqrt{n}(\hat{\beta} - \beta) &= \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varepsilon_i \cdot (D_i - \bar{D})}{\frac{1}{n}\sum_{i=1}^{n} (D_i - \bar{D})^2} \qquad (3.24)
\end{aligned}$$

  - **Line 2:** Expand the right expression with $\frac{1}{n}$.
  - **Line 3:** Multiply both sides by $\sqrt{n}$. Note that $\frac{\sqrt{n}}{n} = \frac{\sqrt{n}}{\sqrt{n}\sqrt{n}} = \frac{1}{\sqrt{n}}$.

- Based on this expression, asymptotic normality can be shown by the **central limit theorem** (De Moivre, 1738, Lyapunov, 1901, Lindeberg, 1922, Lévy, 1937).

Central Limit Theorem

Central limit theorem

For any randomly sampled variable, W, that has a mean of zero ($E[W] = 0$) and a bounded variance, it holds that:

$$\sum_{i=1}^{n} W_i \rightarrow^d N(0, n \cdot Var(W_i)),$$

$$\Leftrightarrow \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \rightarrow^d N(0, Var(W_i)) \tag{3.25}$$

As n increases, the sum of W converges to a normal distribution with a zero mean and a variance given by n times the variance of W.

- The second line follows by multiplying by $\frac{1}{\sqrt{n}}$.
- Note that this fraction enters the variance formula in squared form, $\frac{1}{n}$, such that $\frac{n}{n} \cdot Var(W_i)$ becomes $Var(W_i)$.

Asymptotic Normality of $\hat{\beta}$ (2)

- By the central limit theorem, the numerator in equation (3.24) converges in distribution to a normal distribution:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \cdot \left(D_i - \bar{D}\right) \to^d N(0, E[\varepsilon^2 \cdot (D - E[D])^2]) \tag{3.26}$$

  - $\varepsilon_i \cdot (D_i - \bar{D})$ corresponds to W_i on the previous slide.
  - $E[\varepsilon_i \cdot (D_i - \bar{D})] = 0$ because $E[\varepsilon_i | D_i] = 0$ and the law of iterated expectations (see slide 23).
  - For a variable with zero mean, $Var(W_i) = E[W_i^2]$. Therefore,

$$Var(W_i) = E[W_i^2] = E\left[\varepsilon_i^2 \cdot \left(D_i - \bar{D}\right)^2\right] = E[\varepsilon^2 \cdot (D - E[D])^2] \tag{3.27}$$

- The probability limit of the denominator in equation (3.24) is:

$$plim \left(\frac{1}{n} \sum_{i=1}^{n} (D_i - \bar{D})^2 \right) = Var(D)$$

Asymptotic Normality of $\hat{\beta}$ (3)

- By *Slutsky's theorem*, it holds that $E\left[\frac{W_i}{Var(D_i)}\right] = 0$ because $E[W_i] = 0$ and $Var\left(\frac{W_i}{Var(D_i)}\right) = E\left[\frac{W_i^2}{(Var(D_i))^2}\right]$.
- Therefore, $\sqrt{n}(\hat{\beta} - \beta)$ converges to a normal distribution with a zero mean and a specific variance:

$$\sqrt{n}(\hat{\beta} - \beta) \to^d N\left(0, \frac{E[\varepsilon^2 \cdot (D - E[D])^2]}{(Var(D))^2}\right) \tag{3.28}$$

- The difference between the estimate $\hat{\beta}$ and the true effect β converges to zero, with a convergence rate of $\frac{1}{\sqrt{n}}$ as n increases.
- Put differently, the estimate $\hat{\beta}$ converges to the true ATE β with a convergence rate of $\frac{1}{\sqrt{n}}$.
- This so-called $\sqrt{n}$-consistency is the fastest convergence rate that estimators of causal effects can attain.

Heteroscedasticity-Robust Variance

- Divide the expression on the previous slide by $\sqrt{n}$ and add the true effect β:

$$\hat{\beta} \to^d N\left(\beta, \frac{E[\varepsilon^2 \cdot (D - E[D])^2]}{n \cdot (Var(D))^2}\right) \tag{3.29}$$

- The estimate $\hat{\beta}$ converges to a normal distribution whose mean is the true effect β and whose variance is:

$$Var(\hat{\beta}) = \frac{E[\varepsilon^2 \cdot (D - E[D])^2]}{n \cdot (Var(D))^2} \tag{3.30}$$

- This equation is the heteroscedasticity-robust variance formula.
- Robustness to heteroscedasticity allows the variance of the error term ε to vary across treatment states.
- This is a plausible scenario in many empirical contexts.

Homoscedasticity

Homoscedasticity

The variance of ε is the same under both treatment states.

- The variance of ε, $E[\varepsilon^2]$, does not depend on D under homoscedasticity $\Rightarrow$ equation on the previous slide simplifies to:

$$\frac{E[\varepsilon^2 \cdot (D - E[D])^2]}{n \cdot (Var(D))^2} = \frac{E[\varepsilon^2] \cdot \overbrace{E[(D - E[D])^2]}^{Var(D)}}{n \cdot (Var(D))^2} = \frac{E[\varepsilon^2]}{n \cdot Var(D)} \tag{3.31}$$

- Homoscedasticity assumes that the treatment does not affect the dispersion of the outcome around its mean.
- The heteroscedasticity-robust variance formula on the previous slide is more universal as it does not impose this restriction.
- Thus, relying on equation (3.30) rather than (3.31) when assessing the variance of ATE estimation appears more appropriate.

Asymptotic Normality of $\hat{\alpha}$

- To show the asymptotic normality of $\hat{\alpha}$, start with $\alpha = E[Y|D = 0]$.
- Apply the central theorem to the subsample of nontreated observations only:

$$\frac{1}{\sqrt{n_0}} \sum_{i:D_i=0} (Y_i - \alpha) \to^d N(0, Var(Y|D = 0)) \tag{3.32}$$

  - $Y_i - \alpha$ corresponds to W_i on slide 28 and has an expectation of zero under nontreatment, i.e., $E[Y_i - \alpha|D_i = 0] = 0$.
  - n_0 is the sample size of nontreated observations.
- $\frac{1}{\sqrt{n_0}} \sum_{i:D_i=0}(Y_i - \alpha) = \frac{\sqrt{n_0}}{n_0} \sum_{i:D_i=0}(Y_i - \alpha) = \sqrt{n_0}(\hat{\alpha} - \alpha)$.
- Sample average outcome of the nontreated: $\hat{\alpha} = \frac{1}{n_0} \sum_{i:D_i=0} Y_i$
- $\hat{\alpha}$ converges to a normal distribution with mean α and a variance of $\frac{Var(Y|D=0)}{n_0}$:

$$\hat{\alpha} \to^d N\left(\alpha, \frac{Var(Y|D = 0)}{n_0}\right) \tag{3.33}$$

Mean Squared Error (1)

Mean squared error (MSE)

The mean squared error (MSE) measures the overall accuracy of a method by using the average squared difference between $\hat{\beta}$ and β:

$$\begin{aligned} MSE(\hat{\beta}) &= E[(\hat{\beta} - \beta)^2] \\ &= E[(\hat{\beta} - E[\hat{\beta}] + E[\hat{\beta}] - \beta)^2] \\ &= E[(\hat{\beta} - E[\hat{\beta}])^2] + 2 \cdot \underbrace{(E[\hat{\beta}] - E[\hat{\beta}])}_{=0} \cdot (E[\hat{\beta}] - \beta) + E[(E[\hat{\beta}] - \beta)^2] \\ &= \underbrace{E[(\hat{\beta} - E[\hat{\beta}])^2]}_{\text{variance}} + \underbrace{(E[\hat{\beta}] - \beta)^2}_{\text{squared bias}} \end{aligned} \tag{3.34}$$

- The MSE can be decomposed into an estimate's variance and its squared bias.

- For the OLS estimate $\hat{\beta}$, the variance corresponds to $\frac{E[\varepsilon^2 \cdot (D - E[D])^2]}{n \cdot (Var(D))^2}$.

Mean Squared Error (2)

- In the context of social experiments, the squared bias equals zero, because $\hat{\beta}$ is an unbiased estimate.
- For other estimators, unbiasedness may not hold, at least not in small samples (but only when the sample size is infinitely large).
- In such cases, the MSE is a useful concept for evaluating trade-offs between variance and bias.
- Some methods can be adjusted to increase bias while reducing variance or vice versa.
- The key question is how these adjustments affect the overall MSE.
- The goal is to have an overall MSE that is as small as possible to minimize the estimation error.