Causal Analysis

Impact Evaluation and Causal Machine Learning with Applications in R

Chapter 3: Social Experiments and Linear Regression (2)

Table of Contents

Motivation of Variance Estimation

- Even if unbiasedness and consistency hold, the estimate of the ATE in a sample typically differs from the true ATE in the population.
- This is due to the variance of the ATE estimates across samples.
- Knowing the variance and distribution is useful for quantifying the precision with which the true ATE is estimated in a sample.
- This permits, for instance, answering the following questions relevant to statistical inference:
  - With which error probability can we rule out that the ATE is equal to zero in the population, given the ATE estimate in our sample?
  - What is the range or interval of values that likely includes the ATE in the population, given the findings in our sample?

Estimating the Variance of $\hat{\beta}$

- Directly using $Var(\hat{\beta}) = \frac{E[\varepsilon^2 \cdot (D - E[D])^2]}{n \cdot (Var(D))^2}$ is infeasible, as it contains (unobserved) population parameters, such as $E[D]$ and ε.
- However, we may estimate these parameters in the sample.
- The residual $\hat{\varepsilon}_i$ is the estimate of the true error term ε_i:

$$\hat{\varepsilon}_i = Y_i - \underbrace{(\hat{\alpha} + \hat{\beta} D_i)}_{\hat{E}[Y_i|D_i]} \tag{3.35}$$

  - $\hat{\varepsilon}_i$: Difference between observation i's outcome and the conditional sample average of the outcome given the treatment ($\hat{E}[Y|D_i]$).
  - Prediction $\hat{E}[Y|D_i]$ is an estimate of $E[Y|D = D_i]$ in the population.

- The variance estimator corresponds to:

$$\widehat{Var}(\hat{\beta}) = \frac{\frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \cdot (D_i - \bar{D})^2}{n \cdot (\widehat{Var}(D_i))^2} \tag{3.36}$$

  - $E[D]$ has been replaced by the sample mean $\bar{D}$.

Hypothesis Testing (1)

Hypothesis testing

Assesses whether the true ATE in the population is likely different, smaller, or larger than a specific value, given $\hat{\beta}$ and $Var(\hat{\beta})$.

- Reconsider the asymptotically normal distribution of $\hat{\beta}$:

$$\hat{\beta} \to^d N(\beta, Var(\hat{\beta})), \text{with } Var(\hat{\beta}) = \frac{E[\varepsilon^2 \cdot (D - E[D])^2]}{n \cdot (Var(D))^2}$$

- Normalize this to obtain a standard normal distribution:

$$\hat{\beta} \to^d N(\beta, Var(\hat{\beta})),$$

$$\Leftrightarrow \frac{\hat{\beta} - \beta}{sd(\hat{\beta})} \to^d N(0,1) \tag{3.37}$$

$$\text{with} \quad sd(\hat{\beta}) = \sqrt{\frac{E[\varepsilon^2 \cdot (D - E[D])^2]}{n \cdot (Var(D))^2}} \tag{3.38}$$

Hypothesis Testing (2)

- In large enough samples, the z-statistic $\frac{\hat{\beta}-\beta}{sd(\hat{\beta})}$ closely follows a standard normal distribution.
- This result can be used for checking the plausibility of hypothesized values of β.
- To test whether a treatment has an effect, use:

$$H_0 : \beta = 0, \quad H_1 : \beta \neq 0 \tag{3.39}$$

  - Null hypothesis H_0: The treatment has no effect.
  - Alternative hypothesis H_1: The ATE is different to zero.

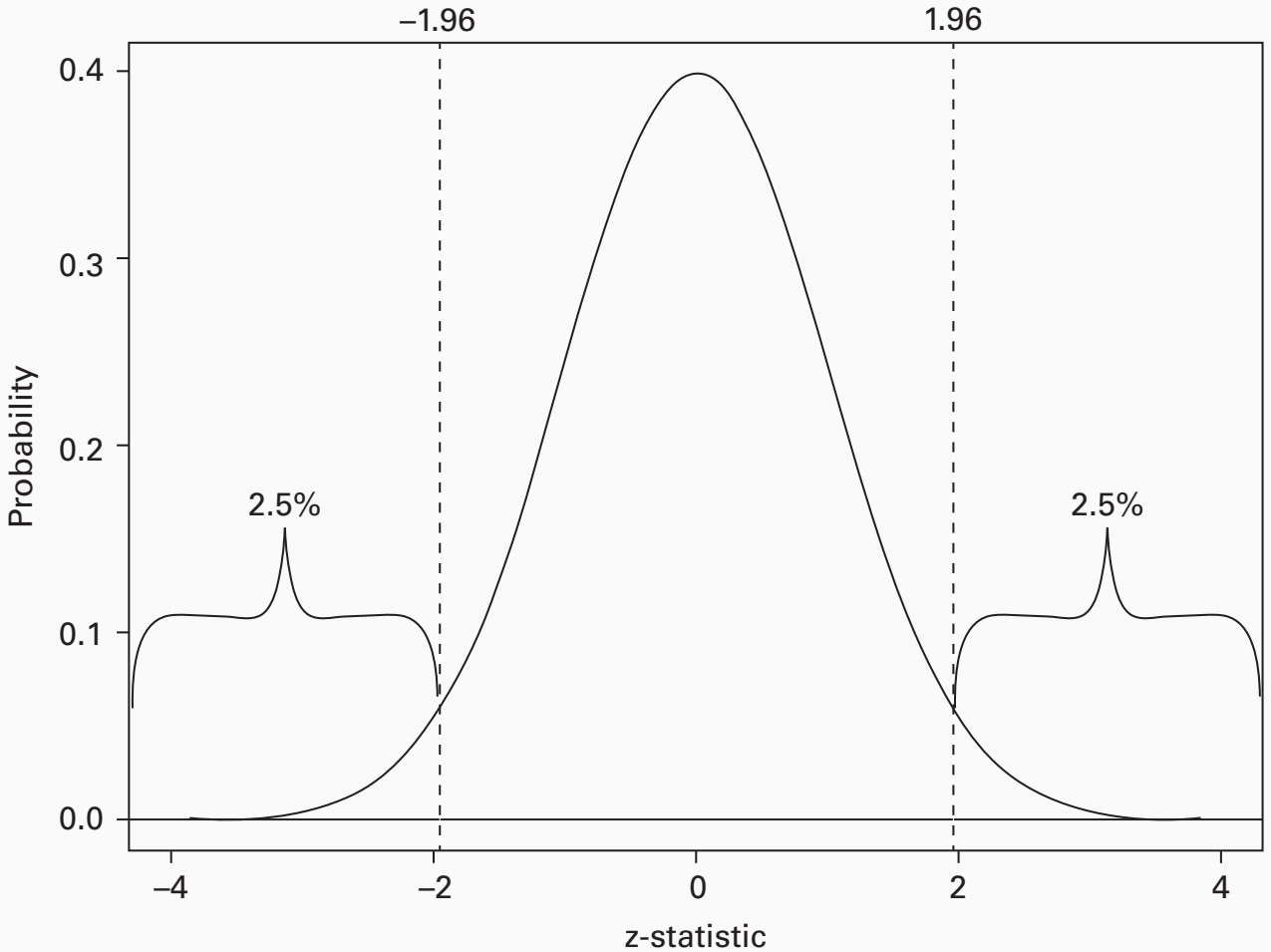
- If $H_0 : \beta = 0$, the z-statistic simplifies to $\frac{\hat{\beta}}{sd(\hat{\beta})}$.
- $\frac{\hat{\beta}}{sd(\hat{\beta})}$ measures the size of the estimated ATE $\hat{\beta}$ normalized by the standard deviation $sd(\hat{\beta})$ as unit of estimation uncertainty.
- Permits assessing how likely it is that the ATE in the population is different from zero, given the value of $\frac{\hat{\beta}}{sd(\hat{\beta})}$ in the sample.

Hypothesis Testing (3)

- If the true ATE in the population is zero, the probability of observing a value of $\left|\frac{\hat{\beta}}{sd(\hat{\beta})}\right| > 1.96$ in a sample is just 5%:
  - The probability of obtaining $\frac{\hat{\beta}}{sd(\hat{\beta})} > 1.96$ is 2.5%.
  - The probability of obtaining $\frac{\hat{\beta}}{sd(\hat{\beta})} < -1.96$ is also 2.5%.
- Thus, if such an extreme value is observed, $H_0 : \beta = 0$ is rejected (and $H_1 : \beta \neq 0$ is accepted) with an error probability below 5%.
- This error probability describes the probability of incorrectly rejecting H_0 (type I error).
- The lower the error probability is, the more confident we are in rejecting H_0.
- A maximum admissible error probability for rejecting H_0, such as 5%, is conventionally predefined.

Hypothesis Testing (4)

Figure 3.2: Standard normal distribution



Hypothesis Testing (5)

- The asymptotic standard deviation $sd(\hat{\beta})$ is typically unknown, as it relies on population parameters, such as $E[D]$ and ε.
- Therefore, $sd(\hat{\beta})$ is replaced with an estimate obtained in the sample, the standard error $se(\hat{\beta})$:

$$se(\hat{\beta}) = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 \cdot (D_i - \bar{D})^2}{n \cdot (\widehat{Var}(D_i))^2}} \tag{3.40}$$

- Replacing $sd(\hat{\beta})$ with $se(\hat{\beta})$ yields the t-statistic which converges to a standard normal distribution as the sample size n increases:

$$\frac{\hat{\beta} - \beta}{se(\hat{\beta})} \rightarrow^{d} N(0,1) \tag{3.41}$$

  - Note that in smaller samples, the t-statistic follows a t-distribution.
  - The t-distribution converges to a normal one as n increases.
  - For samples with $n \geq 120$, a t-distribution is practically indistinguishable from a normal distribution.

Hypothesis Testing (6)

- Under $H_0 : \beta = 0$, the t-statistic simplifies to: $\frac{\hat{\beta}}{se(\hat{\beta})}$
- The type I error probability associated with ATE estimate $\hat{\beta}$ is the probability of values that are at least as large as the t-statistic.
- This error probability of incorrectly rejecting the null hypothesis based on the estimate in the sample - i.e., the significance level implied by using $\left|\frac{\hat{\beta}}{se(\hat{\beta})}\right|$ as threshold value - is the p-value:

$$\text{p-value} = \Pr\left(|A| \geq \left|\frac{\hat{\beta} - \beta}{se(\hat{\beta})}\right|\right), \tag{3.42}$$

where A denotes a random variable following a t-distribution.

p-value

Is the probability of observing a test statistic at least as extreme as the value $\left|\frac{\hat{\beta}}{se(\hat{\beta})}\right|$ observed in the sample under the satisfaction of the null hypothesis $\beta = 0$.

Hypothesis Testing (7)

Comments on hypothesis testing

- Hypothesis testing can only reject the validity of a null hypothesis, but never confirm it.
- Nonrejection of a null hypothesis means we cannot rule out its correctness based on the data, not that it is definitively correct.

Procedure for Hypothesis Testing

1. Define the null and alternative hypotheses H_0 and H_1

- To test the presence of an ATE: $H_0 : \beta = 0$, $H_1 : \beta \neq 0$.
- To test if β differs from another value, e.g., 1: $H_0 : \beta = 1$, $H_1 : \beta \neq 1$.

2. Set the significance level α

- α is the maximally accepted type I error probability of incorrectly rejecting H_0.
- Typical values: $\alpha = 0.05$ (5%), 0.01 (1%), or 0.1 (10%).

3. Compute the critical value c

- c is the value in the standard normal or t-distribution that corresponds to α.
- For $\alpha = 0.05$, $c = 1.96$ in a standard normal distribution.

4. Evaluate the t-statistic and statistical significance

- Reject H_0 if $\left|\frac{\hat{\beta}-\beta}{se(\hat{\beta})}\right| \geq c$ or if the p-value $\leq \alpha$; otherwise, keep H_0.
- If H_0 is rejected, $\hat{\beta}$ is statistically significantly different from the β hypothesized under H_0 at the α level of significance.

One-Sided Hypothesis Testing (1)

Right-tailed hypothesis test

- Tests whether the ATE estimated in the sample is statistically significantly larger than zero (or another value of interest):

$$H_0 : \beta \leq 0, \quad H_1 : \beta > 0 \tag{3.43}$$

- The p-value corresponds to:

$$\text{p-value} = \mathbf{Pr}\left(A \geq \frac{\hat{\beta} - \beta}{se(\hat{\beta})}\right) \tag{3.44}$$

- The condition for a rejection of the null hypothesis is:

$\frac{\hat{\beta} - \beta}{se(\hat{\beta})} \geq c$, where c is a suitable threshold value for one-sided tests, e.g., $c = 1.64$ for $\boldsymbol{\alpha} = 0.05$

- An equivalent condition is p-value $\leq \boldsymbol{\alpha}$.

One-Sided Hypothesis Testing (2)

Left-tailed hypothesis test

- Tests whether the ATE estimated in the sample is statistically significantly smaller than zero (or another value of interest):

$$H_0 : \beta \geq 0, \quad H_1 : \beta < 0 \tag{3.45}$$

- The p-value corresponds to:

$$\text{p-value} = \mathbf{Pr}\left(A \leq \frac{\hat{\beta} - \beta}{se(\hat{\beta})}\right) \tag{3.46}$$

- The condition for a rejection of the null hypothesis is:

$\frac{\hat{\beta} - \beta}{se(\hat{\beta})} \leq c$, where c is a suitable threshold value for one-sided tests, e.g., $c = 1.64$ for $\boldsymbol{\alpha} = 0.05$

- An equivalent condition is p-value $\leq \boldsymbol{\alpha}$.

Confidence Intervals

Confidence interval (CI)

Provides a range of ATE values such that the true ATE β is included with probability/proportion $1 - \boldsymbol{\alpha}$, when constructing CIs in (infinitely) many samples.

- The confidence interval is constructed as follows:

$$\begin{aligned} CI =& [\underline{\beta}, \bar{\beta}], \text{ with} \\ \underline{\beta} =& \hat{\beta} - c \cdot se(\hat{\beta}), \quad \bar{\beta} = \hat{\beta} + c \cdot se(\hat{\beta}) \end{aligned} \tag{3.47}$$

  - $\underline{\beta}$ and $\bar{\beta}$ denote the lower and upper bound of the CI.
  - c is the critical value of a two-sided hypothesis test.
- For $\boldsymbol{\alpha} = 0.05$ (and thus $c = 1.96$), the CI includes the true β with 95% probability.
- Whenever $\hat{\beta}$ is not statistically significantly different from zero, the corresponding CI includes the zero.

Introduction to Bootstrapping

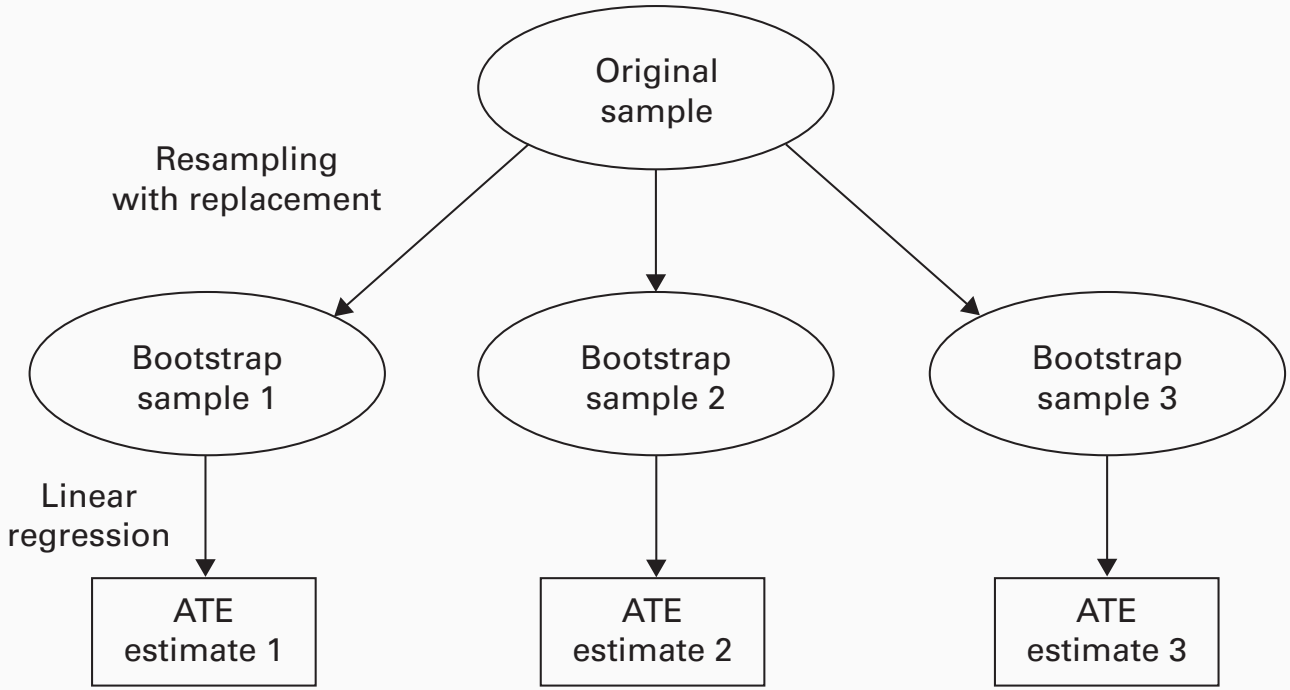
Bootstrapping (Efron, 1979)

Repeatedly draw samples of n observations from the original data (with replacement) to compute p-values, confidence intervals etc.

- Bootstrapping is an alternative method for computing standard errors that does not rely on the asymptotic formula (3.40).
- Drawing observations with replacement implies that some subjects may appear several times or not at all.
- For this reason, bootstrap samples differ from the original data and one another; however, they match the data on average.
- This mimics the fact that the original data is a random sample from the population and a new sample may differ from it.
- Therefore, this quite cleverly approximates the approach of randomly drawing many samples from the population.

Bootstrapping Procedure

Figure 3.3: Bootstrapping



- Generate bootstrap samples, reestimate the ATE in each of them and then use these estimates to calculate the standard error.

Bootstrap-Based Standard Error

- The standard error is computed as the standard deviation of the ATE estimates across all bootstrap samples:

$$se(\hat{\beta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}^b - \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}^b\right)^2} \tag{3.48}$$

  - B is the number of bootstrap samples (B should ideally be large, preferably $B > 999$).
  - b is the index of a specific bootstrap sample with $b \in \{1, 2, \ldots, B\}$.
  - $\hat{\beta}^b$ is the ATE estimate in the bootstrap sample b.

- The bootstrap-based standard error can then be used in the t-statistic (equation (3.41)) to proceed with statistical inference.

Bootstrap-Based p-Value

- An alternative bootstrap approach is to directly compute the p-value without using the t-statistic.
- For a two-sided hypothesis test, the p-value can be calculated by counting how often $|\hat{\beta}^b - \hat{\beta}|$ is larger than $|\hat{\beta}|$:

$$\text{p-value} = \frac{1}{B}\sum_{b=1}^{B} I\{|\hat{\beta}^b - \hat{\beta}| > |\hat{\beta}|\} \tag{3.49}$$

  - $I\{|\hat{\beta}^b - \hat{\beta}| > |\hat{\beta}|\}$ is an indicator function that is equal to 1 whenever $|\hat{\beta}^b - \hat{\beta}| > |\hat{\beta}|$ holds and zero otherwise.

- The distribution of $\hat{\beta}^b - \hat{\beta}$ has a mean of zero, which mimics the distribution under the null hypothesis of no effect.
- If $\hat{\beta}$ appears rather extreme compared to this distribution, the null hypothesis is rejected.

Goodness of Fit and R^2 (1)

- Goodness of Fit: Assessing the relative importance of the treatment in explaining the outcome compared to other characteristics.
- The outcome Y_i can be decomposed into two components:

$$Y_i = \hat{E}[Y_i|D_i] + \hat{\varepsilon}_i \tag{3.50}$$

  - Prediction $\hat{E}[Y_i|D_i]$: The part of the outcome explained by the treatment.
  - Residual $\hat{\varepsilon}_i$: The part of the outcome explained by other (possibly unobserved) characteristics.

- The variance of Y_i is the sum of the variances of these components:

$$Var(Y_i) = Var(\hat{E}[Y_i|D_i]) + Var(\hat{\varepsilon}_i) \tag{3.51}$$

Goodness of Fit and R^2 (2)

- The variances of the parts of the outcome explained by the treatment and the residuals sum to 1:

$$1 = \underbrace{\frac{Var(\hat{E}[Y_i|D_i])}{Var(Y_i)}}_{R^2} + \frac{Var(\hat{\varepsilon}_i)}{Var(Y_i)} \tag{3.52}$$

- The goodness of fit (R^2) can be judged by the share in the variation of Y_i caused by the treatment.
- Interpretation:
  - R^2 close to 1: Treatment causes most of the variation in Y_i.
  - R^2 close to 0: Other characteristics cause most of the variation in Y_i.
- R^2 is different from the magnitude of the ATE:
  - A treatment may have a large ATE but still explain only little of the variation in the outcome relative to other characteristics.

Table of Contents

Discrete Treatments (1)

- So far, the discussion has focused on binary treatments (0 or 1).
- In many empirical applications, however, the interest lies in the effects of several, potentially competing treatments.
- First, consider treatments that are discrete, i.e., can take only a limited number of different values.
- Formally, a treatment can take values $D \in \{0, 1, 2, \ldots, J\}$, where J denotes the number of treatments.
- Covers both ordered (e.g., 1 week vs. 2 weeks of training) and unordered treatments (e.g., IT course vs. sales training).
- If nontreatment and all the various treatments $1, \ldots, J$ are randomized, the independence assumption extends to:

$$\{Y(0), Y(1), Y(2), \ldots, Y(J)\} \perp D \tag{3.53}$$

Discrete Treatments (2)

- To analyze the ATEs for each nonzero treatment, create binary variables: $D_1 = I\{D = 1\},\ D_2 = I\{D = 2\},\ \ldots,\ D_J = I\{D = J\}$.
- The regression model in the population is:

$$E[Y|D] = \underbrace{\alpha}_{E[Y|D=0]} + \underbrace{\beta_1}_{E[Y|D=1]-E[Y|D=0]} D_1 + \underbrace{\beta_2}_{E[Y|D=2]-E[Y|D=0]} D_2 + \cdots + \underbrace{\beta_J}_{E[Y|D=J]-E[Y|D=0]} D_J \tag{3.54}$$

  - $\beta_1, \beta_2, \ldots, \beta_J$ correspond to the ATEs of the various treatments vs. no treatment, i.e., $E[Y(1) - Y(0)]$, $E[Y(2) - Y(0)], \ldots, E[Y(J) - Y(0)]$.
- This model makes pairwise comparisons of the average outcomes between any treatment group and the control group.
- Note that no linear relationship between Y and D is imposed.

Continuous Treatments (1)

- Now consider a treatment D that is continuously distributed, i.e., may take infinitely many values that respect cardinality.
- The independence assumption is adapted to:

$$Y(d) \perp D \quad \text{for any value } d \text{ that treatment } D \text{ might take.} \tag{3.55}$$

- **Approach one:** Discretize the continuous treatment
  - Generate binary indicators for specific brackets of values (e.g., $D_1 = I\{D \leq 1000\}$, $D_2 = I\{1000 < D \leq 2000\}$,...) and use (3.54).
  - Permits analyzing the ATEs of the various brackets but not the average effect of a marginal increase in D on Y.
- **Approach two:** Directly include D in the linear regression

$$Y = \alpha + \beta D + \varepsilon \tag{3.56}$$

  - Under the independence assumption, β represents the average effect of a marginal increase in D on Y.

Continuous Treatments (2)

- Denote the conditional mean of Y given a specific value d of treatment D as $\mu_d = E[Y|D = d]$.
- By the independence assumption, $\mu_d = E[Y(d)]$, ruling out treatment selection bias across values of d.
- $\nabla\mu_d = \frac{\partial\mu_d}{\partial d}$ indicates how much $E[Y(d)]$ changes in reaction to a marginal change in treatment D at treatment value d.
- If Y is a linear function of D, β corresponds to the average marginal effect:

$$E[\nabla\mu_D] = \beta \tag{3.57}$$

- Under certain conditions (namely if D is normally distributed), $E[\nabla\mu_D] = \beta$ holds even if Y is not a linear function of D (i.e., the marginal effect $\nabla\mu_d$ may differ across treatment values of d).

Continuous Treatments (3)

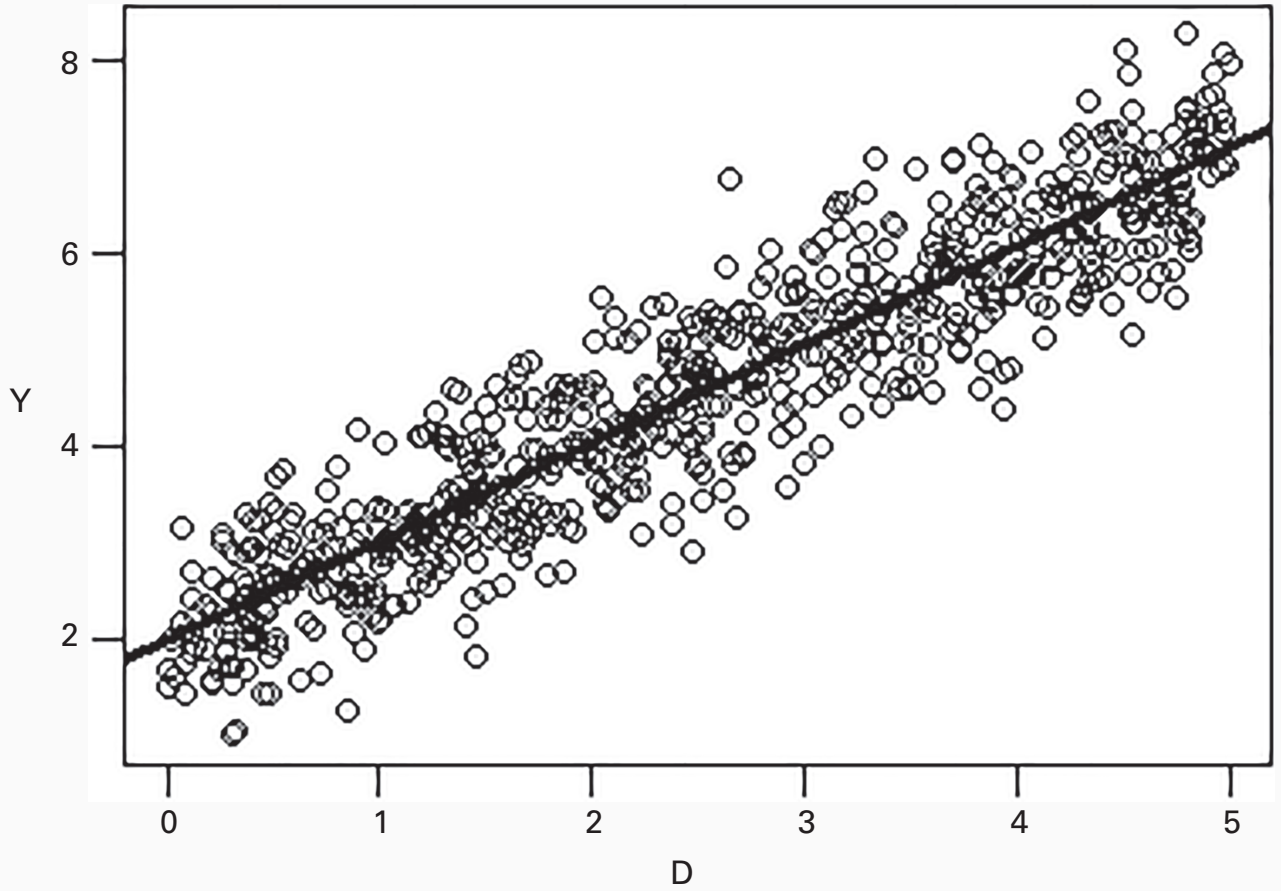
- In general, the marginal effect $\nabla\mu_d$ differs from the average marginal effect $E[\nabla\mu_D]$.
- Exception: When the marginal effect is homogeneous, so that:

$$\frac{\partial E[Y(d')]}{\partial d'} = \frac{\partial E[Y(d)]}{\partial d} \text{ for any values } d' \neq d \text{ that } D \text{ might take.} \quad (3.58)$$

- Therefore, under homogeneous effects, the conditional mean outcome $E[Y|D]$ is truly linear in D.
- Only in this case does linear regression permit identifying the marginal effect at d, since $\frac{\partial E[Y(d)]}{\partial d} = \frac{\partial E[Y(d')]}{\partial d'} = \beta$ for any d and d'.

Continuous Treatments (4)

Figure 3.4: Linear association of the outcome and treatment



Continuous Treatments (5)

- In many empirical settings, the causal relation of the outcome Y and a continuous treatment D might be nonlinear.
- This implies that marginal effects are heterogeneous; they differ depending on the values of the treatment.
- To allow for nonlinearities, the regression model can be made more flexible by including higher-order terms of D, e.g., D^2:

$$E[Y|D] = \underbrace{\alpha}_{E[Y|D=0]} + \beta_1 D + \beta_2 D^2 \tag{3.59}$$

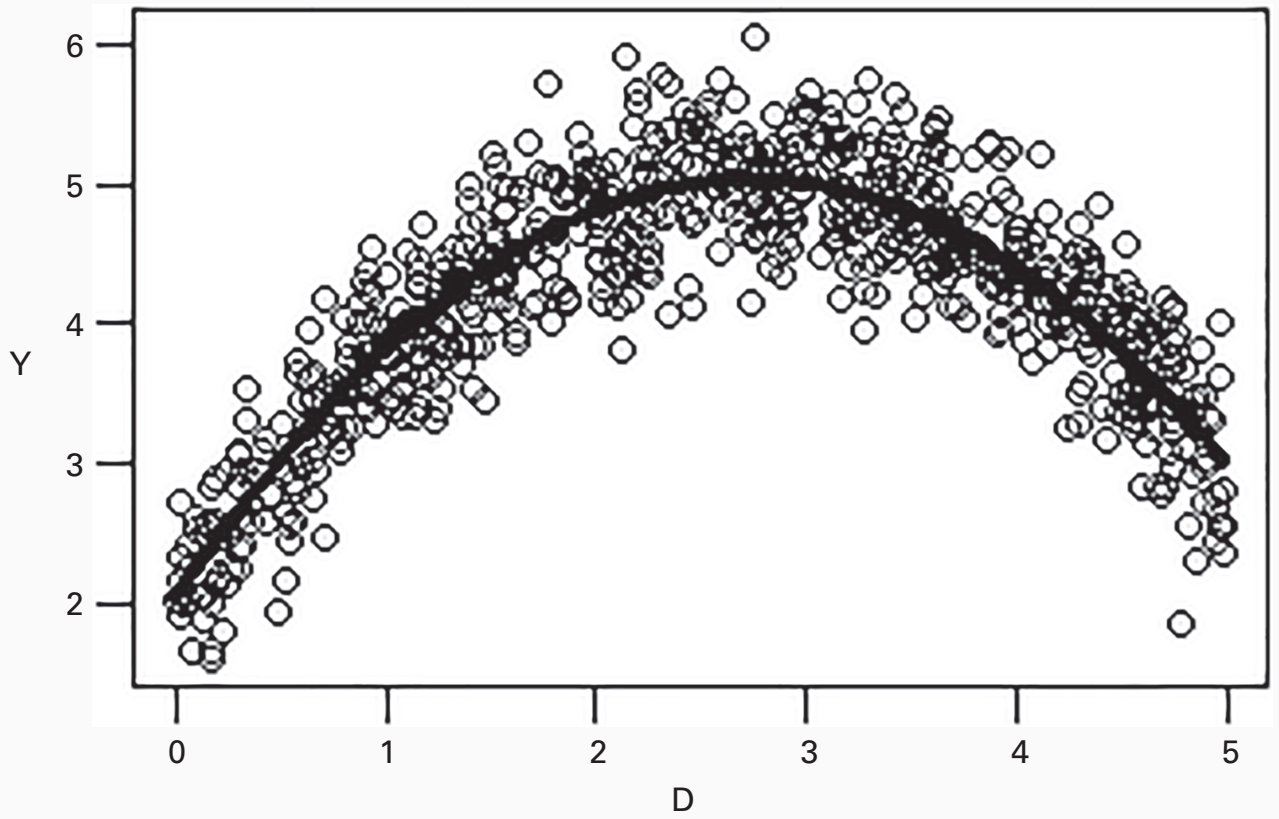
- Taking the first derivative of $E[Y(d)]$ with respect to d yields:

$$\frac{\partial E[Y(d)]}{\partial d} = \beta_1 + 2\beta_2 d \tag{3.60}$$

- Thus, the marginal effect coming from the nonlinear model now depends on the treatment value d.

Continuous Treatments (6)

Figure 3.5: Nonlinear association of the outcome and treatment



- Here, the outcome-treatment relation is even nonmonotonic.

Continuous Treatments (7)

- Including additional higher-order terms (e.g., D^3, D^4, etc.) further increases the model's flexibility to incorporate nonlinearities.
- However, too many higher-order terms increase the variance of effect estimation, particularly in small samples.
- Ideally, the optimal number of higher-order terms is chosen in a way that minimizes the overall estimation error (MSE).
- This can be done in a data-driven way, e.g., based on cross-validation (Stone, 1974).
- For continuous treatments, nonparametric methods (e.g., series or kernel regression) offer a flexible alternative to linear regression.
- Increased flexibility comes at the cost of a higher variance.
- In large samples, the gain in flexibility from nonparametric approaches often outweighs this increase in variance.

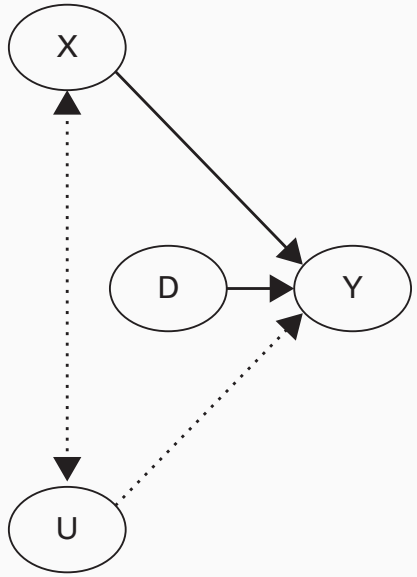
Table of Contents

Why Control for Covariates?

- Under the independence assumption $\{Y(1), Y(0)\} \perp D$, treated and control groups are comparable in background characteristics.
- Therefore, observed characteristics are not needed to obtain an unbiased and consistent ATE estimate.
- Nevertheless, controlling for such covariates $X = (X_1, X_2, \ldots, X_K)$ can reduce the variance of treatment effect evaluation.
- Since covariates are measured prior to treatment, we assume $X(1) = X(0) = X$, ruling out any influence of D on X.

Graphical Illustration

Figure 3.8: Pretreatment covariates



- Pretreatment covariates X may influence Y, but (due to random treatment assignment) neither influence, nor are influenced by D.
- X may influence/be influenced by unobserved characteristics U that affect Y (but not D due to treatment randomization).

Regression Model with Covariates

- To control for covariates, we now include $X = (X_1, \ldots, X_K)$ on the right-hand side of the regression:

$$Y = \alpha + \beta_D D + \beta_{X_1} X_1 + \cdots + \beta_{X_K} X_K + \varepsilon \tag{3.61}$$

- This models the conditional mean outcome given the treatment and the covariates as:

$$E[Y|D, X] = \alpha + \beta_D D + \beta_{X_1} X_1 + \cdots + \beta_{X_K} X_K \tag{3.62}$$

- The definition of the outcome in the sample is provided by:

$$Y_i = \underbrace{\hat{\alpha} + \hat{\beta}_D D_i + \hat{\beta}_{X_1} X_{i1} + \cdots + \hat{\beta}_{X_K} X_{iK}}_{\hat{E}[Y_i|D_i, X_i]} + \hat{\varepsilon}_i \tag{3.63}$$

- Some of the variation in Y is now captured by X; therefore, the residuals $\hat{\varepsilon}_i$ tend to decrease in absolute magnitude.

Covariates and the Goodness of Fit

- Including covariates yields the following definition of R^2:

$$R^2 = \frac{Var(\hat{E}[Y_i|D_i, X_i])}{Var(Y_i)} \tag{3.64}$$

- Whenever X partly explains Y, the variation in Y explained by D and X—and thus R^2—is larger than when D is the only regressor.
- The estimated variance of $\hat{\beta}_D$ is reduced, leading to a smaller standard error and (if $\hat{\beta}_D \neq 0$) a higher t-statistic/lower p-value.
- As a result, estimation uncertainty goes down, and statistical power to detect nonzero ATEs in the population goes up.

Misspecification of the X–Y Relationship

- Even if the influence of X on Y is not linear, $\hat{\beta}_D$ remains a consistent estimate of the ATE.
- This result comes from the fact that D is not associated with X due to randomization: $D \perp X$.
- Thus, the error of incorrectly assuming a linear association between Y and X does not spill over to the evaluation of the ATE.
- $\hat{\beta}_D$ remains asymptotically unbiased in this case.
- In small samples, misspecification may cause bias, but this bias vanishes as sample size increases.
- This would not hold if D were not fully randomized, but rather associated with X.

Bad Controls

- Consider the case where X is affected by D, such that $X(1) \neq X(0)$.
- In this case, controlling for X does not allow for assessing the causal effect of D for two reasons:
  (1) Part of the causal effect of D on Y may operate via $X \Rightarrow$ controlling for X conditions part of the effect away.
  (2) If both D and U (which also affects Y) have a causal effect on X, controlling for X introduces a statistical association between D and U that would not exist otherwise. $\Rightarrow$ *collider bias*

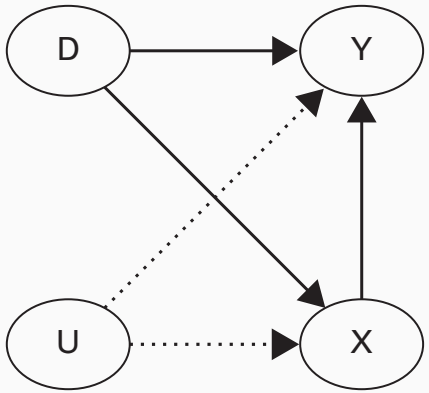


Figure 3.9: Posttreatment covariates that are bad controls.

Example: Birth Weight as a Bad Control

Example

Consider the effect of mothers' smoking during pregnancy (D) on children's postnatal health (Y) using birth weight (X) as a control.

(1) Birth weight may already reflect part of the negative effects of smoking on a child's health.
   - Controlling for X (e.g., comparing only low-birth-weight children) conditions away part of the effect of smoking on postnatal health.

(2) Collider bias: Low-birth-weight children of smoking ($D = 1$) and nonsmoking mothers ($D = 0$) are not comparable.
   - Low birth weight in newborns of nonsmoking mothers is caused by other characteristics (U), such as birth defects, which also affect Y.
   - Thus, the effect of smoking is mixed with that of birth defects.

$\Rightarrow$ May lead to paradoxical findings, such as smoking appearing to reduce mortality among newborns with a low birth weight.