

Impact Evaluation in Firms and Organizations

With Applications in R and Python

Chapter 4: Selection on Observables: Aim to Compare Apples with Apples

4.1 Making Groups Comparable in Observed Characteristics

4.2 Behavioral Assumptions

4.3 Methods for Impact Evaluation

Short evaluation window

- Experiments are often conducted over a short period of time and thus primarily capture short-term effects
- Long-term effects may not be captured

Representativeness and scope

- Participants may not represent the broader population
- Focus on narrow changes in highly controlled environments

Practical constraints

- Financial, ethical, or organizational barriers may limit experimentation
- Some interventions may be impossible to manipulate experimentally

Observational Data for Impact Evaluation

Use of observational data

- Impact evaluations often rely on nonexperimental data
- Data sources include surveys, company records, and administrative data

Covariates in observational settings

- Data include intervention, outcomes, and additional covariates
- Covariates cover characteristics like age, gender, income, or past behavior

Limits of observational data

- Interventions are typically not randomly assigned
- Treatment and control groups are not directly comparable
- Simple outcome comparisons are insufficient for impact evaluation

Impact evaluation with observational data

- Causal analysis is feasible even without experiments
- Causal analysis requires specific methods, typically relying on observed covariates
- The main goal is to compare apples with apples

Ensuring comparability

- Selection bias may be present: Covariates may influence both the intervention and the outcome
- Causal methods aim to compare units with the same or very similar covariates
- The impact is then evaluated within these comparable units

Comparing Apples with Apples

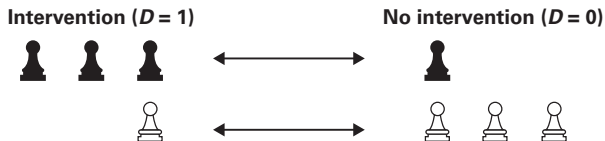


Figure 4.1: Comparing apples with apples

Selection-on-Observables Assumption

Limits of covariate adjustment

- Comparability is assumed based on observed covariates
- Groups may still differ in unobserved characteristics

Presence of hidden factors

- Unobserved traits may affect outcomes
- Apples-to-apples comparisons may fail due to such hidden factors

Behavioral assumption

- Selection-on-observables assumption: Covariates must be sufficiently rich to ensure comparability
- The intervention is treated as quasi-random conditional on the observed covariates

Evaluating a Discount with Observational Data

Discounts in online marketplaces

- Evaluating the impact of discounts on customer purchasing decisions
- Use of observational data without random assignment

Role of observed covariates

- Covariates such as past purchasing behavior affect both discount receipt and purchases
- Conditional on these covariates, discount receipt is assumed to be as good as random

Requirements for plausibility

- Key drivers of discount assignment are observed in the data
- No unobserved or hidden factors

Selection on Observables When Evaluating a Discount

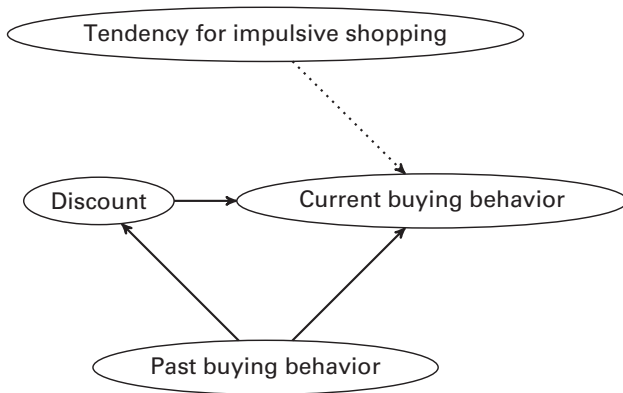


Figure 4.2: Selection on observables when evaluating a discount

Including pre-intervention covariates

- Groups should be made comparable based on characteristics measured *prior* to the intervention
- Covariates measured after the intervention may reflect part of the intervention's impact

Example

Evaluating the impact of a discount on buying behavior

- Compare customers with similar buying behavior before the discount
- Post-discount buying behavior may already reflect the intervention's impact
- Conditioning on it would bias the estimated impact

When is the selection-on-observables assumption unlikely to hold?

- Not all drivers of intervention and outcomes are observed
- For example, training participation and worker productivity may both be affected by unobserved motivation or personality traits

Implications for causal analysis

- Comparability based on observed covariates may be insufficient
- Unobserved characteristics may lead to apples-and-oranges comparisons

Observational versus experimental estimates

- Observational analyses often rely on rich sets of covariates
- If the selection-on-observables assumption holds, estimated effects should match experimental results
- In practice, substantial differences may arise and thus indicate unobserved confounding

4.1 Making Groups Comparable in Observed Characteristics

4.2 Behavioral Assumptions

4.3 Methods for Impact Evaluation

Causal Graph for Selection on Observables

Causal structure

- X denotes observed characteristics or covariates
- Covariates X affect both the intervention D and the outcome Y
- No unobserved variables jointly affect D and Y given X
- Sets of unobserved variables may still affect D or Y separately
- However, these sets must not be associated with each other

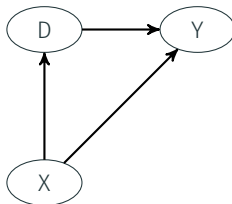


Figure 4.3: Selection on observables

Common support

- In addition to selection on observables, common support must hold
- Common support requires sufficient overlap in covariate distributions across groups
- In other words, comparable subjects must exist across groups for all relevant covariate values
- For example, individuals of similar ages must be present in both treatment and control groups if age is included as a covariate

Consequences of lack of common support

- No valid comparison is possible if comparable units are missing
- Impact cannot be evaluated for those covariate values

Selection on observables

- Treatment is as good as random conditional on covariates X
- Formally

$$\{Y(1), Y(0)\} \perp D \mid X \quad (4.1)$$

Common support

- Requires overlap of treatment and control for all values of X
- Treated and untreated units exist for each covariate value
- Formally

$$0\% < \Pr(D = 1 \mid X) < 100\% \quad (4.2)$$

4.1 Making Groups Comparable in Observed Characteristics

4.2 Behavioral Assumptions

4.3 Methods for Impact Evaluation

Matching approach

- Intuitive method within the selection-on-observables framework
- Matches treated and untreated observations with identical or very similar covariate values
- Treatment and control units differ only in intervention status

Estimating causal effects

- Compare average outcomes across matched treatment and control units
- Matching all units, i.e., finding a matching observation for each treated and untreated unit, yields the ATE
- Matching only treated units, i.e., finding a matching observation only for each treated unit, yields the ATET

Pair Matching Based on Age

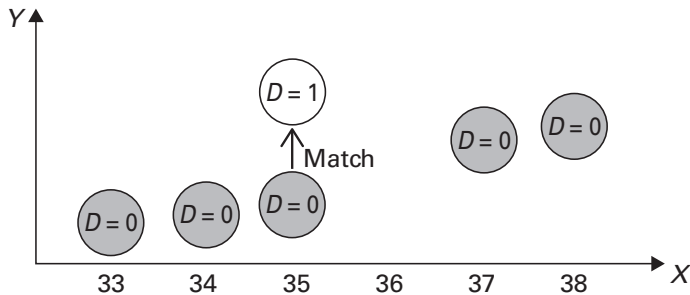


Figure 4.4: Pair matching based on age

Example

Employee training evaluation

- Match each trained employee with an untrained employee with the same or similar age
- Compare outcomes of trained employees to matched untrained employees
- Average outcome differences estimate the ATET

Matching on multiple covariates

- Real-world matching uses several characteristics simultaneously
- Similarity can be assessed using distance measures like the Mahalanobis distance
- Matches based on this measure ensure comparability across all relevant covariates

Matching methods

- Pair (1:1) matching: Match each treated unit to exactly one control unit with similar covariates X
- One-to-many (1:M) matching: Match each treated unit to multiple control units with similar covariates X
- Radius or caliper matching: Match each treated unit with all control units within a certain distance in covariates
- ATET is estimated by comparing average outcomes

Pair Matching

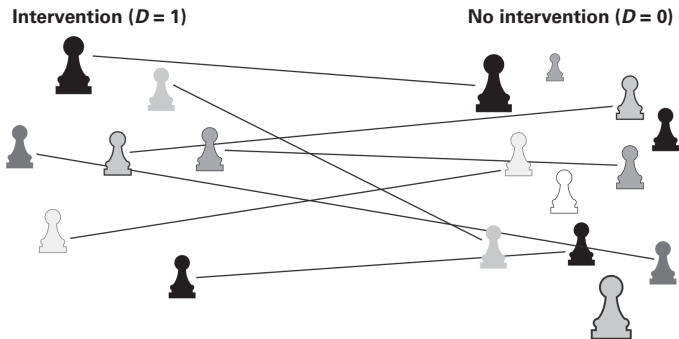


Figure 4.5: Pair matching

1:M Matching

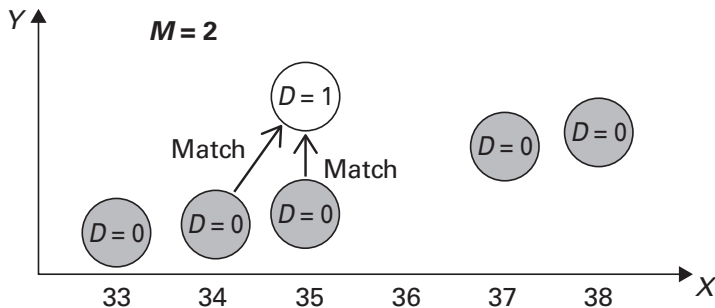


Figure 4.6: 1:M matching

Limits of direct matching

- Requires similarity across all covariates X
- Difficult to find exact matches with increasing number of characteristics

Propensity score matching

- Units are matched based on the propensity score $\Pr(D = 1 | X)$, i.e., the probability of receiving the intervention given X
- The propensity score is estimated using a probit or logit regression
- Matching is based on a single propensity score instead of multiple covariates

Propensity Score Matching

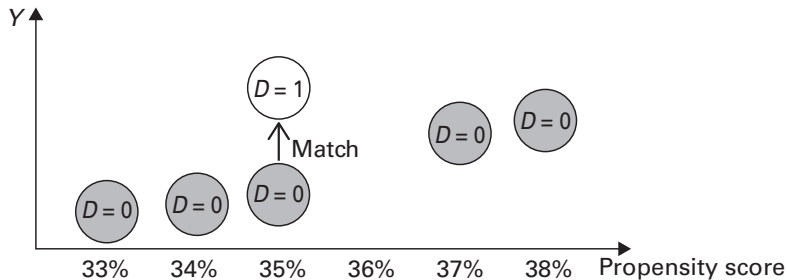


Figure 4.7: Propensity score matching

Regression as an alternative

- Regression can account for observed covariates X
- Aims to make treatment and control groups comparable

Restrictions of linear regression

- Assumes constant effect of covariates on the outcome
- Effect does not vary with the level of the covariate
- May be overly restrictive in practice

Flexible regression approaches

- Nonparametric regression allows varying covariate effects
- Less restrictive but requires larger datasets
- Choice depends on research question and available data

ATE via regression

- ATE corresponds to the average of conditional outcome differences
- Uses regression estimates for treated and untreated outcomes
- ATE is defined as

$$\Delta = E[E[Y | D = 1, X] - E[Y | D = 0, X]] \quad (4.3)$$

ATET via regression

- Focuses on the subpopulation with intervention $D = 1$
- Averages conditional differences among treated units
- ATET is defined as

$$\Delta_{D=1} = E[E[Y | D = 1, X] - E[Y | D = 0, X] | D = 1] \quad (4.4)$$

Weighting based on propensity scores

- Observations are weighted using the propensity scores
- Aim: balance covariates X between treatment and control groups

Inverse probability weighting expressions

- ATE is defined as

$$\Delta = E \left[\frac{Y \cdot D}{\Pr(D = 1 | X)} - \frac{Y \cdot (1 - D)}{1 - \Pr(D = 1 | X)} \right] \quad (4.5)$$

- ATET is defined as

$$\Delta_{D=1} = E \left[\frac{Y \cdot D}{\Pr(D = 1)} - \frac{Y \cdot (1 - D) \cdot \Pr(D = 1 | X)}{(1 - \Pr(D = 1 | X)) \cdot \Pr(D = 1)} \right] \quad (4.6)$$

- $\Pr(D = 1 | X)$ is the propensity score and $\Pr(D = 1)$ is the unconditional probability to be treated

Inverse Probability Weighting

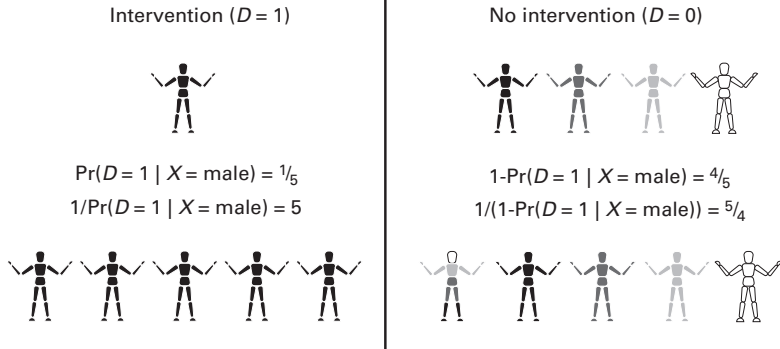


Figure 4.8: Inverse probability weighting

Advantages and Limitations of Inverse Probability Weighting

Advantages of inverse probability weighting

- Computationally fast even in large datasets
- No need to choose matching parameters

Limitations of inverse probability weighting

- Can be unstable when propensity scores are close to 0 or 1

Trimming and overlap

- Drop, or trim, observations with very low or high propensity scores
- Keep only observations with sufficient overlap between groups

Combining methods

- Combines outcome regression and inverse probability weighting
- Doubly robust approach linked to Neyman orthogonality

Doubly robust property

- Consistent if either the outcome model or the propensity scores are correct
- Robust to minor errors in both models

Extensions to multiple and continuous treatments

- Regression, inverse probability weighting, and doubly robust methods are applicable to multiple and/or continuous treatments
- Require large datasets with different interventions and/or diverse intervention values

Doubly robust expression of the ATE

$$\Delta = E \left[E[Y | D = 1, X] - E[Y | D = 0, X] \right. \\ \left. + \frac{(Y - E[Y | D = 1, X]) \cdot D}{\Pr(D = 1 | X)} - \frac{(Y - E[Y | D = 0, X]) \cdot (1 - D)}{1 - \Pr(D = 1 | X)} \right]$$

Doubly robust expression of the ATET

$$\Delta_{D=1} = E \left[\frac{(Y - E[Y | D = 0, X]) \cdot D}{\Pr(D = 1)} \right. \\ \left. - \frac{(Y - E[Y | D = 0, X]) \cdot (1 - D) \cdot \Pr(D = 1 | X)}{(1 - \Pr(D = 1 | X)) \cdot \Pr(D = 1)} \right] \quad (4.7)$$