

Impact Evaluation in Firms and Organizations

With Applications in R and Python

Chapter 3: Experiments (A/B Testing)

3.1 Comparing Apples to Apples

3.2 Behavioral Assumptions and Methods for Analyzing Experiments

3.3 Multiple Interventions

Experiments in impact evaluation

- Also known as A/B testing
- Used in actual business cases such as marketing research for coupons and discounts

Core principle of experiments

- Intervention is assigned randomly
- In other words, the treatment and control groups are formed by chance
- Assignment of the intervention is thus comparable to a coin flip

Experiments address selection bias

- Intervention is independent of background characteristics due to its random assignment
- Background characteristics U do not affect intervention D

Why Experiments Identify the ATE

Identification of the ATE in experiments

- Treatment and control groups are comparable in background characteristics
- Difference in average outcomes between treatment and control groups reflects only the intervention $\Rightarrow$ corresponds to the ATE

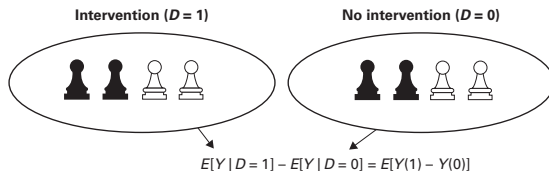


Figure 3.1: Comparable treatment and control groups

Example

Website design experiments

- E-commerce websites test different versions of web pages
- Control and treatment versions are randomly shown to customers, e.g., differing in buy button placement
- Evaluation of effects on sales or click-through rates by comparison of groups

Online advertising experiments

- Marketing teams test different online advertisements
- Comparison of alternative versions of advertisements being randomly shown
- Evaluation of effects on purchases, page views, or downloads

A/B Testing

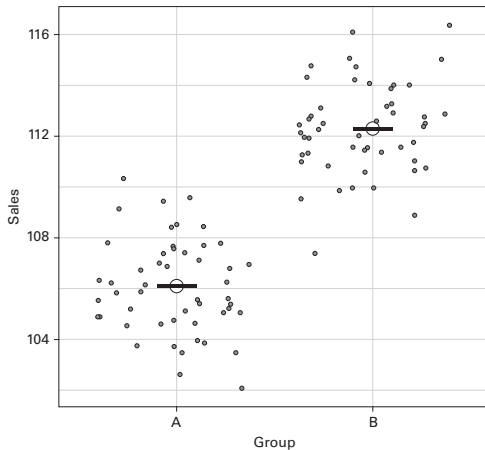


Figure 3.2: A/B testing

Example

Testing algorithms

- Streaming services apply A/B testing to evaluate the impact of an alternative algorithm on user engagement
- The control group uses the existing recommendation algorithm, while the treatment group is exposed to a modified algorithm

Offline experiments

- A/B testing may be applied in an offline business context
- For example, randomized interventions in stores such as alternative store layouts or product displays
- Evaluation of the impact on sales and other business goals comparing treatment and control groups

3.1 Comparing Apples to Apples

3.2 Behavioral Assumptions and Methods for Analyzing Experiments

3.3 Multiple Interventions

Random assignment

- Intervention D is assigned randomly to subjects, e.g., stores or customers
- Assignment is thus independent of background characteristics U

Comparability of groups

- Treatment and control groups are comparable in U
- Potential outcomes $Y(1)$ and $Y(0)$ comparable across these groups

Statistical implications of experimentation

- Background characteristics U are the only other determinants of Y
- D is statistically independent of $Y(1)$ and $Y(0)$

Independence Assumption in Experiments

Formal definition of statistical independence

- Random assignment implies:

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

where $\perp$ denotes statistical independence

Implications for selection bias

- Potential outcomes are comparable across treatment and control groups
- Consequently, no selection bias is present

Identification of the ATE

- Both groups are representative of the population in terms of background characteristics (comparing apples with apples)
- This holds under properly randomized interventions

Sample-Based Versus Population Effects

Population versus sample

- ATE refers to the entire population of interest
- Experiments are conducted on a sample from that population

Sample estimate of the ATE

- Sample averages are denoted using a hat symbol
- $\hat{E}[Y|D = 1]$ and $\hat{E}[Y|D = 0]$ are sample group averages among the treated and control groups
- Difference defines the sample-based ATE estimate

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

Reasons why $\hat{\Delta}$ may not exactly match the true Δ in the population

- Sample may not perfectly represent the population
- Background characteristics may be imbalanced by chance

Sampling variation

- Random assignment may still yield small group imbalances
- Example: gender share may differ between treatment and control

Role of large samples

- Larger samples improve representativeness of the population
- Noncomparability tends to be small or negligible

Convergence to the true effect

- Sample-based estimate $\hat{\Delta}$ approaches the true ATE Δ with an increase in size
- Accuracy increases with the number of participants according to the law of large numbers

Regression as an analysis tool

- Linear regression is used to analyze outcome differences
- The average outcome can be expressed as a linear function of the intervention:

$$\hat{E}[Y | D] = \underbrace{\hat{E}[Y | D = 0]}_{\hat{\alpha}} + \underbrace{(\hat{E}[Y | D = 1] - \hat{E}[Y | D = 0])}_{\hat{\beta}} \cdot D \quad (3.3)$$

Interpretation of a linear regression

- Coefficients, denoted by $\hat{\alpha}$ and $\hat{\beta}$, represent sample-based estimates
- $\hat{\alpha}$ represents the average outcome without the intervention
- $\hat{\beta}$ equals the difference in average outcomes across treatment and control groups, i.e., it represents the estimated ATE

Linear Regression

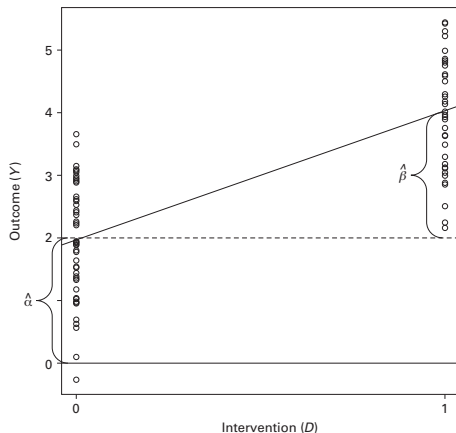


Figure 3.3: Linear regression

Sampling uncertainty

- Lack of full representativeness introduces uncertainty
- Estimated ATE in a sample may differ from the ATE in the population

Role of variance

- Regression provides an estimate of the variance of the effect
- Low variance implies less uncertainty about the estimated effect
- High variance implies greater risk of large estimation errors

p-value

- Probability of detecting a non-zero impact as extreme as $\hat{\beta}$ when the true effect is zero
- Reflects the likelihood of detecting an impact in the sample despite no effect in the population

Type I error probability

- Probability of incorrectly claiming a nonzero population impact when true effect is zero
- Significance level is the maximum acceptable type I error probability, commonly set to 5%
- Example: a 5% significance level allows a 5% chance the true effect is actually zero despite a non-zero estimate

Confidence intervals

- Interval reflects plausible values of the population impact
- Derived from the variance of the estimated impact
- Example: true effect lies within an interval with 95% probability

3.1 Comparing Apples to Apples

3.2 Behavioral Assumptions and Methods for Analyzing Experiments

3.3 Multiple Interventions

Motivation for multiple interventions

- Real-world decisions often involve competing interventions
- Interest in comparing effectiveness across interventions

Experimental setup

- Random assignment of different campaigns to users
- Some users receive no ad and thus form a control group
- Outcomes measured using click rates

Comparative evaluation

- Compare each campaign group to the control group
- Estimate average impact of each intervention separately
- Identify the most effective campaign

Motivation for varying intervention intensities

- Interest in how intervention intensity, or doses, affects outcomes
- Example of offering discounts to increase off-peak ticket sales

Experimental design

- Random assignment of different discount levels, e.g., ranging from 10% to 50%, to customers
- Comparison of ticket sales across discount intensities

Decision support

- Evaluate relative impact of different discount levels
- Assess whether larger discounts yield higher sales
- Allows trade-off between increased sales and increased loss due to discounts

Regression with continuous interventions

- Intervention can take many values
- Regression is used to analyze small or marginal effects

Homogeneous impact assumption

- Marginal increases assumed to have the same effect
- Effect does not depend on the initial intervention level
- Example: increasing a discount from 4% to 5% has the same impact on sales as increasing a discount from 14% to 15%

Linear association

- Outcome changes linearly with the intervention level
- Corresponds to a homogeneous impact assumption

Average Outcomes Under Homogeneous Impacts

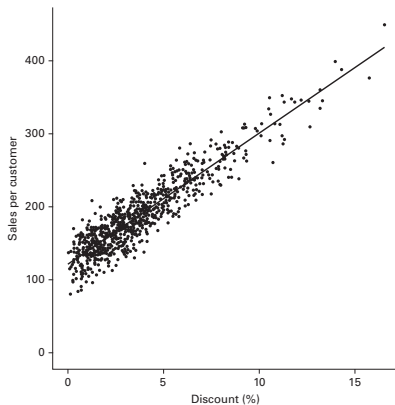


Figure 3.4: Average outcomes under homogeneous impacts

Limits of Linear Regression for Continuous Interventions

When linear regression is appropriate

- Suitable for binary interventions taking values 0 or 1
- Applicable when intervention effects are homogeneous

Limitations in other scenarios

- Linear regression assumes constant marginal effects
- Assumption may not hold for many-valued interventions

Need for flexible methods

- Marginal effects may vary across intervention levels
- Example: increasing a discount from 4% to 5% has a different impact on sales than increasing a discount from 14% to 15%
- Requires alternative approaches allowing heterogeneous effects

Average Outcomes Under Heterogeneous Impacts

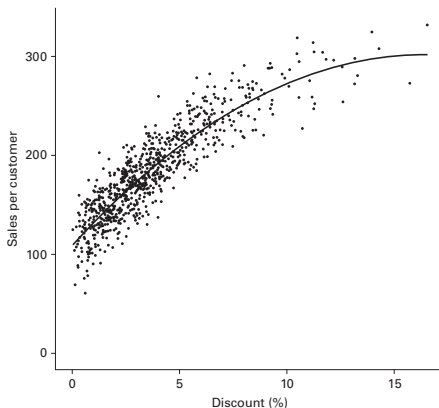


Figure 3.5: Average outcomes under heterogeneous impacts

Nonparametric methods

- Allow effects to vary across intervention levels
- Suitable for multi-valued interventions
- Require sufficiently large samples

Intuition of the approach

- Compare average outcomes for nearby intervention values
- Measure changes from small increases in the intervention at different value levels of the intervention

Kernel regressions

- Assesses marginal effects across different levels of the intervention
- Assigns greater weight to observations closer to the evaluation point and lower weight to more distant observations

Estimation of the Conditional Mean Outcome

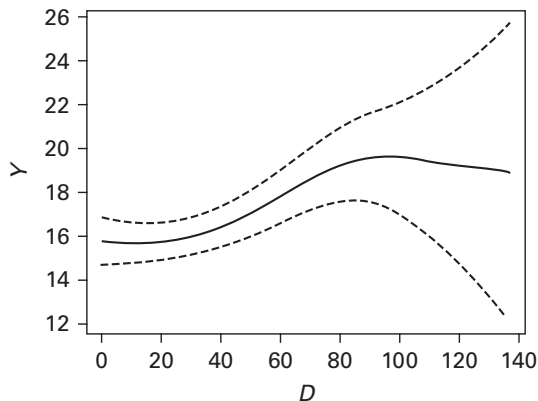


Figure 3.6: Estimation of the conditional mean outcome

Estimation of the Marginal Effects

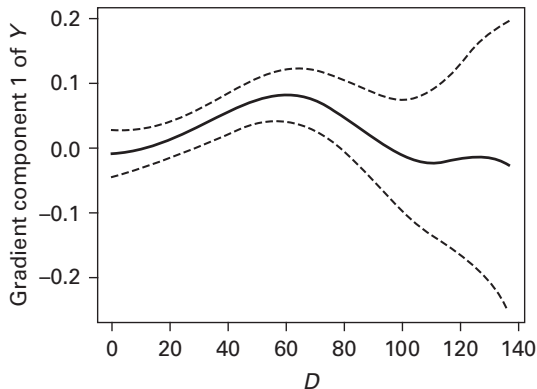


Figure 3.7: Estimation of the marginal effects