Methods for Causal Inference in Marketing
Other titles in *Foundations and Trends*® in Marketing

*Aesthetics in Marketing*
Henrik Hagtvedt
ISBN: 978-1-63828-286-0

*Artificial Intelligence in Marketing and Consumer Behavior Research*
TaeWoo Kim, Umair Usman, Aaron Garvey and Adam Duhachek

*Leveraging Online Search Data as a Source of Marketing Insights*
Rex Yuxing Du and Tsung-Yiou Hsieh
ISBN: 978-1-63828-262-4

*The Rise of Mobile Marketing: A Decade of Research in Review*
Zeynep Aydin-Gokgoz, M. Berk Ataman and Gerrit van Bruggen

*Marketing and Firm Value*
Shuba Srinivasan and Dominique M. Hanssens
ISBN: 978-1-63828-044-6
Methods for Causal Inference in Marketing

Zezhen (Dawn) He
University of Rochester
Zezhen.He@simon.rochester.edu

Vithala R. Rao
Cornell University
vrr2@cornell.edu

Full text available at: http://dx.doi.org/10.1561/1700000080
Foundations and Trends® in Marketing
Volume 18, Issue 3–4, 2024
Editorial Board

Editor-in-Chief
Jehoshua Eliashberg
University of Pennsylvania

Editors
Dawn Iacobucci
Vanderbilt University
Leonard Lee
National University of Singapore
Sharon Ng
National University of Singapore
Koen Pauwels
Northeastern University
Stefano Puntoni
University of Pennsylvania
William Rand
North Carolina State University
Bernd Schmitt
Columbia University
Gerrit van Bruggen
Erasmus University
Hema Yoganarasimhan
University of Washington
Juanjuan Zhang
Massachusetts Institute of Technology
Editorial Scope

Foundations and Trends® in Marketing publishes survey and tutorial articles in the following topics:

- B2B Marketing
- Bayesian Models
- Behavioral Decision Making
- Branding and Brand Equity
- Channel Management
- Choice Modeling
- Comparative Market Structure
- Competitive Marketing Strategy
- Conjoint Analysis
- Customer Equity
- Customer Relationship Management
- Game Theoretic Models
- Group Choice and Negotiation
- Discrete Choice Models
- Individual Decision Making
- Marketing Decisions Models
- Market Forecasting
- Marketing Information Systems
- Market Response Models
- Market Segmentation
- Market Share Analysis
- Multi-channel Marketing
- New Product Diffusion
- Pricing Models
- Product Development
- Product Innovation
- Sales Forecasting
- Sales Force Management
- Sales Promotion
- Services Marketing
- Stochastic Model

Information for Librarians

Foundations and Trends® in Marketing, 2024, Volume 18, 4 issues. ISSN paper version 1555-0753. ISSN online version 1555-0761. Also available as a combined paper and online subscription.
## Contents

1 Introduction 3

2 Origins and Developments in Causal Inference Methods 9  
   2.1 Introduction ................................................. 9  
   2.2 Randomized Experiments ..................................... 9  
   2.3 Potential Outcomes .......................................... 10  
   2.4 Illustrations of Marketing Experiments ...................... 11  
   2.5 Treatment Effects .......................................... 13  
   2.6 A Taxonomy of Data Types ................................... 14  
   2.7 Summary .................................................... 14

3 Data and Methods of Causal Inference 15  
   3.1 Introduction .................................................. 15  
   3.2 Directed Acyclic Graphs (DAGs) .............................. 16  
   3.3 Types of Data ............................................... 17  
   3.4 Methods for Estimating Causal Effects ....................... 18  
   3.5 Summary ..................................................... 36

4 Recently Developed CI Methods 37  
   4.1 Introduction .................................................. 37  
   4.2 Factor Model or Generalized Synthetic Control .......... 38  
   4.3 Augmented DID .............................................. 39
Methods for Causal Inference in Marketing

Zezhen (Dawn) He and Vithala R. Rao

1 Simon Business School, University of Rochester, USA; Zezhen.He@simon.rochester.edu
2 Johnson Graduate School of Management, Cornell University, USA; vrr2@cornell.edu

ABSTRACT

Establishing causal relationships between the marketing variables under the control of a firm and outcome measures such as sales and profits is essential for the successful operation of a business. The goal is generally hindered by a lack of suitable experimental data owing to the costs and feasibility of conducting randomized experiments. Accordingly, researchers have employed observational and quasi-experimental data for causal inference. The evolutionary trajectory of causal inference is closely intertwined with advancements in business and technology, particularly as we enter the digital era characterized by big data and multi-channel marketing.

Against this background, this monograph is a systematic review of recent developments in causal inference methods and their applications within the marketing field. For each causal inference method, five recently published academic papers in marketing research that employ these methods are discussed.

Supplementary Material is available at http://dx.doi.org/10.1561/1700000080_supp
©2024 Z. He and V. R. Rao
In addition, this monograph provides simplified code for developing simulated data (using Python) and hypothetical examples of data analysis (using Stata). This addition will enable marketing researchers to practice several methods of causal analysis.

Sections 1–5 elucidate the fundamental principles of causal inference. Subsequent sections (beginning from Section 6) delve into the details of a selection of papers that utilize various methods. These encompass: (i) well-established techniques, such as Differences-In-Differences, Instrumental Variable, Regression Discontinuity, Synthetic Control, and Propensity Score Methods, and (ii) emerging methodologies of Factor Model and Augmented Differences-In-Differences, Forward Differences-In-Differences, and Bayesian methods for causal inference. Further, this monograph reviews how machine learning methods enhance causal inference. This monograph includes several important and useful references not reviewed in the monograph. We expect this monograph to serve as a useful resource both to current and future researchers in marketing.
Marketing is a business function that relates the firm to its customers and end-consumers. Undoubtedly, it is an essential and important function of businesses and other organizations (see Kotler and Keller, 2012). On the business side, the function involves activities such as product design, sales forecasting, design of advertising strategy for a product and its execution, and sales and distribution activities. The marketing function is similar in other types of organizations, even though the terminologies may differ.

Concurrent with the marketing developments in business, the academic field of marketing has blossomed over the last 80 years or so. This discipline is thriving considering the high visibility of academic marketing associations and academic journals. Three associations, the American Marketing Association (AMA), the European Marketing Academy (EMAC), and the Institute for Operations Research and Management Science (INFORMS) played a significant role in the discipline’s growth. AMA’s premier journals, *Journal of Marketing* (JM) is in its 88th year of publication and the *Journal of Marketing Research* (JMR) is in its 61st year of publication. The *International Journal of Research in Marketing* (IJRM) of EMAC is in its 40th year of publication.
while Marketing Science of INFORMS is in its 43rd year of publication. Several other academic journals (e.g., the Journal of the Academy of Marketing Science, the Journal of Consumer Research, and the Journal of Retailing) cover the subject matter of marketing in several ways. In a similar vein, universities have established doctoral degree programs in marketing and are thriving well. The demand for faculty in marketing is on the rise and is met by doctoral programs in marketing and other subjects such as economics, psychology, and computer/information science. All these point to the assertion that marketing is a viable field of academic endeavor.

Figure 1.1 shows the external factors that affect a firm along with the connections to intermediaries and consumers. It also shows the effect of competitors. The intermediary firms include not only those firms used for distributing the firm’s products but also other firms that collect data on purchase and viewing behavior of end-consumers. The external factors are categorized as Political, Economic, Social, Technological, and Legal (or PESTLE for short).

While the field of marketing does not have a unified theory as such, several important theories, paradigms, or frameworks have evolved over the years, which are essential to both the practice of marketing and research or theory building. The extant theories and paradigms include the theory of buyer behavior (Howard and Sheth, 1969), Information
Processing (Bettman, 1979), Hierarchy of effects of advertising (Lavidge and Steiner, 1961), Customer lifetime value (Gupta and Lehmann, 2005; Kumar, 2013) are used quite frequently. Further, the idea of “Marketing Science” has taken hold; some of the research in this direction utilizes game theory. There is an opportunity in marketing science to develop a sound theoretical understanding of marketing issues.

On the applied side, several frameworks have appeared over the years. These include the Four Ps (Product, Price, Place, and Promotion, also called marketing-mix), STP (Segmentation, Targeting, and Positioning), consumer choice process, and market response. These frameworks are utilized to develop and implement a marketing plan for a firm’s product (brand). Firms use several pieces of data developed from both internal and external sources in the process of developing marketing plans. External sources are typically focus groups, consumer surveys and consumer panels, advertising test data, and occasionally experiments, etc. The goal here is to come up with a plan that enables the firm to reach a level of sales target for an existing product as well as a new product (new brand). The research methods employed for this purpose include several data analytic techniques (including multivariate methods). Conceptually, the attempt here is to determine a market response function for the product or brand in question that shows the relationship between the sales outcome and variables that are in control of the firm (i.e., the Four Ps). In general, the data sources within the firm are not available to academic researchers.

In the early days of marketing research, the tendency was to report summaries of data collected. Currently, the emphasis is on the development of appropriate models and estimation using advanced and appropriate methods.

Over the years, academic researchers have utilized data that they collected on their own and also have utilized data sources that are publicly available. Several new types of data have become available for academic research as can be seen from Table 1.1. An example is the data on online reviews of services like hotels (collected by companies like Trip Advisor or Yelp). Another example is genetic data (Daviet et al., 2022) that shed light on consumer behavior based on genetic variants.
Introduction

Table 1.1: Emerging types of data in marketing

<table>
<thead>
<tr>
<th>A. Standard Marketing Data</th>
<th>B. Unstructured Data</th>
<th>C. Social Network Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>A3. Archival data</td>
<td>B3. Videos, pictures</td>
<td>C3. Postings to social media</td>
</tr>
<tr>
<td>A5. Media ratings data</td>
<td>B5. Data on physiological measurements (e.g., eye tracking)</td>
<td></td>
</tr>
</tbody>
</table>

Note: *A special case of these data is conjoint analysis data (ratings or choices).

Academic researchers employ various methods in their research. These methods are either drawn from statistics, econometrics, and psychometrics or specially developed newer techniques. A few developments in this area are the use of structural models, Hierarchical Bayesian (HB) methods, and the application of newer methods drawn from AI (such as deep learning). It is worth noting that the HB methods enable the estimation of parameters at the individual unit (person, firm, or other) level. One should underscore that the basic premise of sound research is a viable theory translated into a model (preferably a mathematical model). Whether explicit or not, academic research involves estimating some model parameters, which are intended to measure the impact of a predictor variable on the outcome variable (such as sales or choice). If the researcher develops a system of equations to be estimated with some econometric techniques, the coefficients of the model measure such impacts. Depending on the theoretical basis for the research, the estimate can be construed as a “causal effect.” As an example, consider the Koyck model of dynamic effects of advertising specified as \( S_t = \alpha + \beta \times A_t + \gamma \times S_{t-1} \), where \( A_t \) is the advertising expenditures in period \( t \), and \( S_t \) is the sales in period \( t \). If this theory is accepted, then \( \beta \) is the causal effect of current advertising, and \( \gamma \) is the carry-over effect.

Also, we may note that real-life experiments or randomized controlled experiments out of the lab (Fisher, 1935) are expensive to carry
out in marketing, even when researchers collaborate with industry. Even though most marketing data are not based on designed experiments, marketing researchers have employed data from quasi-experiments (also called observational data\(^1\)) for estimating causal effects. During the last 10 years or so, novel methods have emerged to estimate “causal effects.” Researchers have also adopted methods developed for text analysis as well; see Feder et al. (2022) for a comprehensive review of these methods.

Two examples of studies that measure causal effects will be relevant to mention here. One is the study by Wang et al. (2022) that uses data from a natural experiment to determine the effect of the Black Lives Matter movement on consumer responses (likes). They employ the Differences-In-Differences (DID) method and estimate the causal effect of the BLM movement on firms’ empirical strategy. The authors exploit Blackout Tuesday as a natural experiment in which BLM support occurred on Instagram (treated platform) but not on Twitter (control platform) to perform a within-brand cross-platform Differences-In-Differences (DID) analysis.

In a different context, Manchanda et al. (2015) measure the incremental economic impact (expenditures) of customer communities (called social dollars) in a multichannel retail environment. They measure this by estimating a regression model. Because a well-defined control group is missing, the authors test the effect of joining the online community on economic activity using the Regression Discontinuity Analysis with the time of joining the community as the threshold.

Against this background, the objective of this monograph is to provide a comprehensive review of the methods for estimating causal effects in marketing along with a review of the applications where these methods have been applied. This monograph consists of seven sections. In the next (and second) section, we offer an overview of the estimation of causal effects attempting to synthesize the large literature in various disciplines. The third section discusses various methods in brief; the

---

\(^1\) Although there are slight differences between observational data and quasi-experimental data, we use them interchangeably in this monograph.
methods discussed are: Directed Acyclic Graphs,\textsuperscript{2} Analysis of Variance and Covariance, Differences-In-Differences regression methods, regression with Instrumental Variables, Regression Discontinuity methods, Synthetic Control Methods, as well as Sub-classification and Matching methods (including Propensity Score Methods). In this section, we attempt to provide the theoretical basis for each method along with numerical examples and estimation codes. Section 4 describes four emerging methods (i.e., Factor Model, Augmented DID, Forward DID, and Bayesian Synthetic Control). The fifth section describes the role and development of machine learning methods in causal estimation with some examples. Section 6 reviews five applications of each of the methods in marketing (i.e., Differences-In-Differences (DID), Instrumental Variable (IV), Regression Discontinuity (RD), Synthetic Control Method (SCM), Propensity Score Method (PSM)), as well as emerging methods and machine learning-related methods. Section 7 provides a summary and a discussion of future directions in this rapidly growing area.

\textsuperscript{2}This monograph will not cover the Directed Acyclic Graphs (DAGs) developed by Pearl and his colleagues with an exclusive focus on econometrics-based methods. See Pearl (2009a) for the DAG approaches. But we will briefly describe the debate between Rubin and Pearl (Pearl, 2009b; Rubin, 1974).
Appendices
Python Code for Generating Simulated Data

Regression of Treatment Effects

Example Context: Assume a firm is interested in testing the impact of a new television ad compared to its existing television ad, and the firm will be airing the current ad in TV areas 1, . . . , 20 and airing the new add in TV areas 21, . . . , 40. We define each area in terms of artificial area codes $i$, where $i = 1, . . . , 40$, and denote each outcome measure as $Y(i, t)$ for $t = 1, . . . , 10$, where $t$ is month.

The variables describing the TV areas are the average age of people in the area, average income per household, the percentage of females in the area, and the percentage of days in the period the brand was sold on promotion.

The two outcome variables are the percentage of households buying the brand during the period, and the percentage of buyers (households) buying for the first time during the period.

The data is generated with the assumed treatment effects of 2 for the percentage of households buying the brand during the period and 3 for the percentage of buyers (households) buying for the first time during the period.

\[^1\text{To enhance the replicability of the code, we have made the code available on GitHub https://github.com/zhesimon/Methods-for-Causal-Inference-in-Marketing.}\]
import numpy as np
import pandas as pd
import random
from scipy.stats import norm
from sklearn.preprocessing import MinMaxScaler

np.random.seed(1)
TV_areas, n_periods = 40, 10
age_mean, age_std = 38, 5
income_household_mean, income_household_std = 70000, 10000
female_rate, female_std = 50, 10
pct_days_promo_mean, pct_days_promo_std = 50, 15

data = []
for i in range(TV_areas):
    TV_area = i+1
    age = int(np.random.normal(age_mean, age_std))
    income = int(np.random.normal(income_household_mean, income_household_std))
    female = round(np.random.normal(female_rate, female_std),2)
    pct_days_promo = round(np.random.normal(pct_days_promo_mean, pct_days_promo_std),2)

    for j in range(n_periods):
        TV_area_data = {
            'period': j+1,
            'TV_areas': TV_area,
            'avg_age': age,
            'avg_income': income,
            '%female': female,
            'pct_days_promo': pct_days_promo,
        }
        data.append(TV_area_data)

do = pd.DataFrame(data)

# outcome var, %households buying the brand
df['purchase_rate'] = (0.02 * df['avg_age'] + 0.5/10000 * df['avg_income'] + 0.1 * df['%female'] + 0.05 * df['pct_days_promo'] + np.random.uniform(-2, 2, size = len(df)))

# outcome var, %buyers (households) buying for the first time
df['pct_buyer_1sttime'] = (0.01 * df['avg_age'] + 0.3/10000 * df['avg_income'] + 0.05 * df['%female'] + 0.1 * df['pct_days_promo'] + np.random.uniform(-2, 2, size = len(df)))

# trim values to between 0 and 100
df[['purchase_rate', 'pct_buyer_1sttime']] = df[['purchase_rate', 'pct_buyer_1sttime']].round(2).clip(0, 100)

# treatment group: TV_areas 21-40
df['treatment'] = [0 if i < TV_areas/2 * 10 else 1 for i in range(df.shape[0])]

df['purchase_rate_t'] = df['purchase_rate'] + 2 * df['treatment']

df['pct_buyer_1sttime_t'] = df['pct_buyer_1sttime'] + 3 * df['treatment']
# treatment effect = 3
df.to_csv('simulated_TV_areas.csv', index=False)

Nearest Neighbor Matching and Propensity Score Matching

Example Context: Here we used a similar dataset except that we generated the TV areas 21, . . . , 40 as close neighbors of TV areas 1, . . . , 20.

To do that, we duplicated the first 200 rows (thus, duplicated TV_areas 1–20 that are control units) and added noise to the average age of people living in the area, the average income per household, the percentage of female in the area, and the percentage of days in the period the brand was sold on promotion.

The covariates are the same with noise added, so the areas from the duplicated rows are close neighbors of control areas.

# controls
new_df1 = df.iloc[:200, :6]

# new_df2 to be modified to close neighbors of controls
new_df2 = new_df1.copy()

new_df2['TV_areas'] = new_df2['TV_areas'] + 20

# error ~ N(0, 2) for age
new_df2['avg_age'] = df.groupby('TV_areas')['avg_age'].transform(lambda x: x + np.random.normal(0, 2)).round(2)

# error ~ N(0, 1000) for income
new_df2['avg_income'] = df.groupby('TV_areas')['avg_income'].transform(lambda x: x + np.random.normal(0, 1000)).round(2)

# error ~ N(0, 2) for percent female
new_df2['%female'] = df.groupby('TV_areas')['%female'].transform(lambda x: x + np.random.normal(0, 2)).round(2)

# error ~ N(0, 2) for percent of days the brand was sold on promotion
new_df2['pct_days_promo'] = df.groupby('TV_areas')['pct_days_promo'].transform(lambda x: x + np.random.normal(0, 2)).round(2)

df = pd.concat([new_df1, new_df2]).reset_index(drop=True)

# outcome var, %households buying the brand
df['purchase_rate'] = (0.02 * df['avg_age'] + 0.5/10000 * df['avg_income'] + 0.1 * df['%female'] + 0.05 * df['pct_days_promo'] + np.random.uniform(-2, 2, size = len(df)))

# outcome var, %buyers (households) buying for the first time
df['pct_buyer_1sttime'] = (0.01 * df['avg_age'] + 0.3/10000 * df['avg_income'] +
0.05 * df['%female'] + 0.1 * df['pct_days_promo'] + np.random.uniform(-2, 2, size=len(df))

# trim values to between 0 and 100
df[['purchase_rate', 'pct_buyer_1sttime']] = df[['purchase_rate', 'pct_buyer_1sttime']].round(2).clip(0, 100)

# treatment for TV_areas 21-40
df['treatment'] = [0 if i < TV_areas/2 * 10 else 1 for i in range(df.shape[0])]
df['purchase_rate_t'] = df['purchase_rate'] + 2 * df['treatment']
# treatment effect = 2
df['pct_buyer_1sttime_t'] = df['pct_buyer_1sttime'] + 3 * df['treatment']
# treatment effect = 3
df.to_csv('simulated_TV_areas_nnmatch.csv', index=False)

Instrumental Variable (IV) Method

Example Context: We used the IV method to estimate the causal effect of advertising on sales. For this purpose, we generated one unobserved variable that affects both advertising expenditure and sales at the same time, so there is endogeneity if we directly regress sales on advertising expenditure. The instrumental variable is set as advertising costs. This variable is generated to affect advertising expenditure but not sales directly. We also assumed a linear relationship between advertising costs and advertising expenditure, and between advertising expenditure and sales.

We generated data for 100 periods, for each period there are different advertising costs, advertising expenditure, and sales. The treatment effect of advertising expenditure on sales is set to be 2.5.

np.random.seed(42)
n = 100  # periods
p0 = 30  # intercept1
p1 = -2  # effect of advertising costs on advertising expenditure
beta0 = 35  # intercept2
beta1 = 2.5  # effect of advertising expenditure on sales
omit_exp = 2  # effect of omitted variable on advertising expenditure
omit_sales = 3  # effect of omitted variable on sales
omitted_variable = np.random.randint(1, 3, size=n)

# IV (Z): advertising costs
advertising_costs = np.random.randint(1, 11, size=n)

# X: advertising expenditure
advertising_expenditure = p0 + p1 * advertising_costs + omit_exp * omitted_variable + np.random.uniform(-2, 2, size=n)
Python Code for Generating Simulated Data

```python
# Y: sales
sales = beta0 + beta1 * advertising_expenditure + omit_sales * omitted_variable + np.random.uniform(-5, 5, size=n)

data = pd.DataFrame({'Ad_Costs': advertising_costs, 'Ad_Expenditure': advertising_expenditure, 'Omitted_variable': omitted_variable, 'Sales': sales})
data.to_csv('IV_Data.csv', index=False)
```

Regression Discontinuity Method

Example Context: Assume a firm owns an online platform where hotel guests provide ratings on three variables (brand image, price, and service) after staying at a hotel.

Based on the overall feedback on the three variables (i.e., the average score of brand image of the hotel, the average perceived price level, and the average score of service performance), the firm develops a rating of each hotel listed on the platform.

Further, the firm assigns symbols to the hotels based on a threshold. If the rating is below 3, the platform assigns a Symbol B, and if the rating is equal or above 3, Symbol A is assigned. Relative to Symbol B, the treatment effect of being assigned Symbol A is 8.

The firm is interested in estimating the impact of being assigned to Symbol A compared to Symbol B on hotel sales based on 1000 hotels. We define each hotel in terms of hotel identification code $i$, where $i = 1, \ldots, 1000$, and denote each Sales measure as $Y(i)$.

```python
random.seed(0)
num_samples = 1000
beta1 = 8  # effect of symbol A on sales
beta2 = 5  # effect of ratings on sales

data = []
for _ in range(num_samples):
    brandimage = random.uniform(1, 5)
    price = random.uniform(1, 5)
    service = random.uniform(1, 5)
    rating = brandimage * 0.3 + price * -0.2 + service * 0.4

data.append({
    'Brandimage': brandimage,
    'Price': price,
    'Rating': rating,
})
data.to_csv('IV_Data.csv', index=False)
```
Synthetic Control Method (and Differences-In-Differences)

Example Context: Assume that a country consists of 51 geographically identified States and that these States have a sales tax for products sold within the State. Assume further that for a specific product category, one State reduced tax in period 100 while other States did not. We are interested in estimating the effect of the tax reduction on the State’s product category sales using data available for 200 units of time (e.g., weeks). We define each State in terms of index $i$, where $i = 0, 1, \ldots, 50$, where State 0 is the treated unit, and States 1–50 are control States. We denote each outcome measure as $Y(i, t)$ for $t = 1, \ldots, 200$, where $t$ is period. Periods 1–100 are pre-treatment periods, the treatment happened in period 100, and periods 101–200 are post-treatment periods. We assumed that the treatment effect is 10.

Initially, we generated the treatment unit using only two States i.e., State 1 (weight 0.2) and State 2 (weight 0.8) and random noise. Later, we performed the Synthetic Control Method based on all 50 States, States 1–25, States 1–10, and States 1–5. The pre-treatment periods in our simulation were set as periods 1–100, periods 51–100, periods 81–100, or periods 91–100.

After obtaining the synthetic control, we used DID to estimate the treatment effect.
# Python Code for Generating Simulated Data

```python
# creates variables beta1, ..., beta50 and assigns them values from betas
for i in range(1, 51):
    globals()[f'beta{i}'] = betas[i - 1]
print(beta1, beta2)
print(betas)

# generates a dictionary containing 50 key-value pairs, keys are "mu1", "mu2", ..., "mu50", values are random integers in range (5, 15)
mu_gen = {}
for i in range(1, 51):
    mu_gen[f"mu{i}"{}] = random.randint(5, 15)
print(mu_gen)

# generates a list containing values of all mu's
mus = list(mu_gen.values())
print(mus)

# creates 50 variables mu1, ... , mu50, corresponding to the values stored in dictionary mu_gen
for i in range(1, 51):
    globals()[f'mu{i}'] = mu_gen[f'mu{i}']
# print(mu1, mu2, mu3)

control_units = ['Y' + str(i) for i in range(1, 51)]

# for i, Y in enumerate (control_units):
# print(control_units[i], mus[i], betas[i])

n = 200
treated_period = int(n/2)
random_state = 100
data = pd.DataFrame(index=range(n))

# control unit Y[i] ~ N (mu[i], 5)
for i, Y in enumerate (control_units):
    random_state += 1
    data[control_units[i]] = norm.rvs(loc=mus[i], scale=5, size=data.shape[0], random_state=random_state)

# error term for the synthetic control ~ N (0, 1)
random_state = random_state + 1
data['error'] = norm.rvs(loc=0, scale=1, size=data.shape[0], random_state=random_state)

# data_w holds the weighted control units Y[i]_w, weighted by their corresponding beta values
data_w = pd.DataFrame(index=range(n))
for i, Y in enumerate (control_units):
    data_w[control_units[i] + '_w'] = betas[i] * data[control_units[i]]
data_w['error'] = data['error']
data_w.to_csv('Data_w.csv', index=False)

# treated outcome in the absence of treatment: weighted sum plus noise
```

Full text available at: http://dx.doi.org/10.1561/1700000080
data['Y0_treated'] = data_w[list(data_w.columns)].sum(axis = 1)

# treatment effect for the treated units (post-treatment periods 101-200), which equals 10 + eps ~Uniform (-0.1,+0.1)
data['eps_te'] = np.concatenate((np.zeros(treated_period,dtype = int),
                               np.random.uniform(-0.1,0.1,treated_period) + 10))

# treated outcome in the presence of treatment
data['Y0_te'] = data['Y0_treated'] + data['eps_te']
data['period'] = np.arange(1,n + 1)

# check that the weighted Ys are Y*beta
for i in range(data.shape[0]):
    assert(data['Y1'][i] * betas[0] == data_w['Y1w'][i])
data.to_csv('Synthetic_Control_Data.csv',index=False)
B

Stata Code for Analysis of Data

Regression of Treatment Effects
import delimited "/simulated_TV_areas.csv", clear
rename female pct_female
*Dependent var: purchase rate
reg purchase_rate_t avg_age avg_income pct_female pct_days_promo treatment
*Dependent var: percent of first-time buyer
reg pct_buyer_1sttime_t avg_age avg_income pct_female pct_days_promo treatment

Nearest-Neighbor Matching
import delimited "/simulated_TV_areas_nnmatch.csv", clear
rename female pct_female
*Dependent var: purchase rate
teffects nnmatch (purchase_rate_t avg_age avg_income pct_female pct_days_promo)(treatment), nneighbor(1)
*Dependent var: percent of first-time buyer
teffects nnmatch (pct_buyer_1sttime_t avg_age avg_income pct_female pct_days_promo)(treatment), nneighbor(1)

Propensity Score Matching
import delimited "/simulated_TV_areas_nnmatch.csv", clear
rename female pct_female
*Dependent var: purchase rate
teffects psmatch (purchase_rate_t) (treatment avg_age avg_income pct_female pct_days_promo)
*Dependent var: percent of first-time buyer
teffects psmatch (pct_buyer_1sttime_t) (treatment avg_age avg_income pct_female pct_days_promo)
**Instrumental Variable**

import delimited "/IV_Data.csv", clear

*Manually run 2sls
reg ad_expenditure ad_costs

gen constructed_ad_expenditure = _b[_cons] + _b[ad_costs] * ad_costs
reg sales constructed_ad_expenditure

*IV
iivregress 2sls sales (ad_expenditure = ad_costs)

*Test of endogeneity: check whether ad_expenditure is endogenous
estat endog

*Check whether the instrument is weak
estat firststage

**Regression Discontinuity Method**

import delimited "/RD_ratings.csv", clear

*RD plot, Cut-off c = 3
rdplot sales rating, c(3) graph_options(title(RD Plot) xtitle("Rating") ytitle("Sales"))

*Sharp RD estimates
rdrobust sales rating, c(3) all

**Synthetic Control Method**

import delimited "/Synthetic_Control_Data.csv", clear

reshape long y, i(period) j(State)
rename period Period
rename y Sales
save "/Synth_Panel.dta", replace

*Performing synthetic control method using all 50 control states and all pre-treatment periods
use "/Synth_Panel.dta", clear
tset state Period
synth Sales Sales(1(1)100), trunit(0) trperiod(101) fig keep(100period_50state)
graph export "/Graph_100period_50state.png", replace

*Performing synthetic control method using partial (25) control states and all pre-treatment periods
use "/Synth_Panel.dta", clear
keep if State<26
tset state Period
synth Sales Sales(1(1)100), trunit(0) trperiod(101) fig keep(100period_25state)
graph export "/Graph_100period_25state.png", replace

*Performing synthetic control method using all 50 control states and 50 pre-treatment periods
use "/Synth_Panel.dta", clear
keep if Period>50
tset State Period
synth Sales Sales(51(1)100), trunit(0) trperiod(101) fig keep(50period_50state)
graph export "/Graph_50period_50state.png", replace
*Calculating treatment effects using DID (reshape data from wide format to long format)
use "/100period_50state.dta", clear
drop _Co_Number _W_Weight
reshape long _Y_, i(_time) j(State, string)
gen treatment = 0
replace treatment = 1 if _time >=101 & State == "treated"
encode State, generate(nState)
xtset nState
xtdidregress (_Y_)(treatment), group(nState) time(_time)

Differences-In-Differences

use "/Synth_Panel.dta", clear
*DID result using State2 as control (which has weight of beta 0.8 when generating the treatment unit in our simulation)
keep if State ==2 | State ==0
gen treatment = 0
replace treatment = 1 if Period >=101 & State == 0
xtset State
xtdidregress (Sales)(treatment), group(State) time(Period)
In this appendix, we briefly summarize the models behind the ADID and its alternatives (DID, SC, MSC, and HCW) for estimating the average treatment effect in quasi-experimental studies, and some notes about Double Machine Learning.

**ADID and Alternative Methods for ATT Estimation**

After introducing the notation, we will describe the specific methods and compare them with ADID.

The problem is to estimate the average treatment effect on the treated units (or ATT) using data on different units, some of which are treated for a number of time periods. For simplicity, we will assume that the first unit is treated at time $T_1; 1 < T_1 < T$ and all other $(N - 1)$ units do not receive treatment.

**Notation:**

$y_{it}^{(1)} = \text{Outcome measure after treatment for the } i\text{-th unit at time } t; i = 1, \ldots, N; \text{ and } t = 1, \ldots, T.$

$y_{it}^{(0)} = \text{Outcome measure before treatment for the } i\text{-th unit at time } t; i = 1, \ldots, N; \text{ and } t = 1, \ldots, T.$
\[ \hat{\Delta}_{1t} = y^{(1)}_{1t} - y^{(0)}_{1t} \] is Treatment effect for the first unit at time \( t \), with \( t > T_1 + 1 \).

We can describe the observed data as:

\[ y_{it} = d_{it} y^{1}_{it} + (1 - d_{it}) y^{0}_{it}, \]

where \( d_{it} = 1 \) if the \( i \)-th unit receives treatment at time \( t \) and 0 otherwise.

Given the above assumption on timing and units treated, the ATT estimator that averages \( \hat{\Delta}_{it} \) over the post-treatment period is

\[ \hat{\Delta}_{1} = \frac{1}{T_2} \sum_{T_1+1}^{T} \hat{\Delta}_{it} \]

where \( T_2 = T - T_1 \) is the number of post-treatment time periods., where \( \hat{\Delta}_{it} = y^{(1)}_{it} - y^{(0)}_{it} \). While the data are available for \( y^{(0)}_{it} \) but not for \( y^{(1)}_{it} \), we need to estimate this quantity. The methods differ according to the model used for this estimation. Table C.1 below shows the models and differences.

To summarize, the ADID method is more flexible than DID because it controls for slope in addition to intercept differences in pre-treatment periods. This also means that it only requires that the treated unit’s trend is parallel to a slope-adjusted trend of the control unit in pre-treatment periods. Compared to Synthetic Control Methods that require no intercept and weights sum to one, ADID require equal weights, but the weights can sum to any value.

**Double Machine Learning**

Traditional machine learning methods are good at predictions, but not good at causal relationships because they can produce biased estimates due to regularization and overfitting. Double machine learning (Chernozhukov et al., 2018a) is a recently developed framework that combines machine learning methods and causal inference in that it deals with regularization bias through orthogonalization and deals with overfitting bias through sample splitting. Double machine learning enables in obtaining approximately unbiased and \( \sqrt{N} \)-consistent estimation under high dimensional complex covariates and confounding variables. Researchers will find the Python and R packages of DoubleML (Bach et al., 2021, 2022) valuable in the implementation of the double/debiased machine learning framework of Chernozhukov et al. (2018a).
Table C.1: A summary of formulae for different methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Model for $y_{1t}$</th>
<th>Formula for $\hat{y}_{1t}$</th>
<th>Formula for ATT</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADID</td>
<td>$y_{1t} = \delta_1 + \delta_2 y_{co,t} + \text{error}$; $t = 1, \ldots, T$</td>
<td>$\hat{y}_{1t, ADID} = \hat{\delta}_1 + \hat{\delta}<em>2 y</em>{co,t}$</td>
<td>$\frac{1}{T_2} \sum_{T_1+1}^{T} (y_{1t} - \hat{y}_{1t, ADID})$</td>
<td>Compared HCW, if $\beta_j$s are equal, yields ADID</td>
</tr>
<tr>
<td>DID</td>
<td>$y_{1t} = \delta_1 + \bar{y}_{co,t} + \text{error}$; $t = 1, \ldots, T$</td>
<td>$\hat{y}_{1t, DID} = \hat{\delta}_1 + \hat{\delta}<em>2 y</em>{co,t}$</td>
<td>$\frac{1}{T_2} \sum_{T_1+1}^{T} (y_{1t} - \hat{y}_{1t, DID})$</td>
<td>Compared HCW, if $\beta_j$s are equal and positive, and sum to 1, yields DID</td>
</tr>
<tr>
<td>SC</td>
<td>$y_{1t} = z_t' \beta + \text{error}; \beta_{SC}$ minimizes error sum of squares</td>
<td>$\hat{y}<em>{1t, sc} = z_t' \hat{\beta}</em>{sc}$</td>
<td>$\frac{1}{T_2} \sum_{T_1+1}^{T} (y_{1t} - \hat{y}_{1t, sc})$</td>
<td>Compared HCW, if $\beta_j$s are all positive and sum to 1 with no intercept, yields SC</td>
</tr>
<tr>
<td>MSC</td>
<td>Same as for SC with the conditions, $\beta_1 = 0; \sum_{j=2}^{N} \beta_j = 1; \beta_j \geq 0$, $z_{1t} = z_{t+1} \cdots z_{T}$</td>
<td>$\hat{y}<em>{1t, MSC} = z_t' \hat{\beta}</em>{MSC}$</td>
<td>$\frac{1}{T_2} \sum_{T_1+1}^{T} (y_{1t} - \hat{y}_{1t, MSC})$</td>
<td>Compared HCW, if $\beta_j$s are all positive, yields MSC</td>
</tr>
<tr>
<td>HCW</td>
<td>$y_{1t} = z_t' \beta + \text{error}; t = 1, \ldots, T$; $\beta_{OLS}$ is the least squares estimate of $\beta$</td>
<td>$\hat{y}<em>{1t, HCW} = z_t' \hat{\beta}</em>{OLS}$</td>
<td>$\frac{1}{T_2} \sum_{T_1+1}^{T} (y_{1t} - \hat{y}_{1t, HCW})$</td>
<td>This is the base case against which comparisons are made.</td>
</tr>
</tbody>
</table>
Useful Resources


Useful Resources

of stacked discounts: Identifying revenue and cost effects of discount

Frisch, R. and F. V. Waugh (1933). “Partial time regressions as com-
pared with individual trends”. *Econometrica: Journal of the Econo-
metric Society*: 387–401.

Gelman et al. (2009). “Statistical modeling, causal inference, and social

of Economic Literature*. 57(3): 535–574.

drops a product category? investigating the consequences of ending


“Disentangling the effects of ad tone on voter turnout and candidate
243.

A large-scale exploration of non-experimental approaches to adver-

“A comparison of approaches to advertising measurement: Evidence
from big field experiments at Facebook”. *Marketing Science*. 38(2):
193–225.

Prentice-Hall.


IBM (n.d.[a]). “What is computer vision?” URL: https://www.ibm.com/topics/computer-vision.

Useful Resources


Lim, J. H., R. Rishika, R. Janakiraman, and P. Kannan (2020). “Competitive effects of front-of-package nutrition labeling adoption on


Full text available at: http://dx.doi.org/10.1561/1700000080


References


References


Full text available at: http://dx.doi.org/10.1561/1700000080


References


Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases. 8: 65–70.

Journal of Educational Psychology. 51(6): 309.


DOI: 10.1177/00222437231210267.


Full text available at: http://dx.doi.org/10.1561/1700000080