bEhaVioral Insights anD Effective eNergy policy acTions

Project No. 957117

Project acronym: EVIDENT

Project title:

Behavioral Insights and Effective Energy Policy Actions

Deliverable 4.1

Analytical Qualitative and Quantitative Tools Requirements (Econometric Models)

Programme: H2020-LC-SC3-EE-2020-1

Start date of project: December 01, 2020

Duration: 36 months

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957117

Document Control Page

Deliverable Name	Analysis of best practices
Deliverable Number	D4.1
Work Package	WP4
Associated Task	T4.1
Covered Period	M01 – M09
Due Date	M09 – 31/08/2021
Completion Date	M09 – 30/08/2021
Submission Date	M09 – 31/08/2021
Deliverable Lead Partner	Democritus University of Thrace (DUTH)
Deliverable Author(s)	Ioannis Pragidis (DUTH), Paris Karypidis (DUTH), Georgios Geronikolaou (DUTH), Mitropoulos Fotios (DUTH), Vaso Kotsirou (DUTH), Tilemachos Efthimiadis (JRC)
Version	V3.0

Dissemination Level		
PU	Public	**X**
CO	Confidential to a group specified by the consortium (including the Commission Services)	

Document History

Version	Date	Change History	Author(s)	Organisation
0.1	June, 2021	Initial version	Ioannis Pragidis	DUTH
0.2	July, 2021	Updated version	Ioannis Pragidis	DUTH
0.3	August, 2021	Final version	Ioannis Pragidis	DUTH

Internal Review History

Name	Institution	Date
Ioannis Neokosmidis	Bi2S	August 20, 2021
Peter Rosenberg	CW	August 23, 2021

Quality Manager Revision

Internal Review History

Name	Institution	Date
Dimostheni Ioannidis	CERTH	August 25, 2021

Table of Contents

List of Figures

List of Tables

Acronyms

Acronym	Explanation
CW	Check Watt
DiD	Difference in Differences
FD	First Differencing
IV	Instrumental Variable
QTE	Quantile Treatment Effect
OLS	Ordinary Least Squares
PPC	Public Power Corporation
RCT	Randomised Control Trials
RD	Regression Discontinuity
2SLS	Two-Stage Least Squares
WP	Work Package

Executive Summary

Many empirical questions in economics and other social sciences depend on the causal effects of programs or policies. In the last two decades, much research has been done on the statistical and econometric analysis of such causal effects. In this deliverable, we attempt to present state-of-the-art statistical and econometric methods that are relevant for analysing data from experimental and quasi-experimental cases. This is not a technical review, meaning that most technical details and derivation proofs are mostly being avoided. We would like to think of this review more as a researcher's companion, providing with the necessary tools for applied research and with the appropriate references for a more thorough analysis for interested readers. The central topics presented in this deliverable are that of evaluating the effect of exposure of a set of units to treatment on some outcome, and of analysing data stemming from surveys and other quasi-experiments. The review starts from interpreting causal statements as a comparison of so-called potential outcomes and then continues with reviewing methods based on panel data and other specialised methods such as quantile regressions.

1. Purpose and Overall Structure of the Deliverable

1.1 Purpose of the Deliverable

This deliverable aims to provide the baseline and the theoretical foundation for the econometric analysis to be conducted in the EVIDENT project. Based on the experiments' design, this deliverable offers an overview of the most relevant analytical methods. This task includes the review of the available econometric methods for panel and cross-section data in evaluating policy interventions and the specifications of the chosen models using research developments in the econometric and statistical literature. It is crucial to exactly specify the models, based on behavioural economics theory, intuition, and common sense along with their asymptotic distributions to motivate our regression models for inferring casual relationships. D2.1 will provide the necessary input for the exact identification of the econometric methods to be implemented. This information will be provided during the time of implementation of D2.1.

1.2 Relation with other Deliverables and Tasks

Work Package (WP) 4 and more specifically D4.1, is about statistics and econometric analytics and lies at the heart of the EVIDENT project. The key objective of WP4 is to design and implement the econometric analysis for evaluating the policy measures developed and implemented in WP2 and WP3 respectively and produce the relevant reports and coding. D4.1 is the first deliverable of WP4 and will review, categorise, analyse and propose the econometric models to be used for inferencing the output of the implemented policy interventions. D4.1 is part of T4.1, that is about the exploration of analytical qualitative and quantitative tools requirements. It receives input from D2.1 "Specifications of field studies and surveys". Its output will be received as input from D4.2 "Econometric analysis and robustness tests".

1.3 Structure of the Document

This deliverable is structured as follows:

- Section 2 – Introduction, causal inference, internal and external validity
- Section 3 – Analysis of classical randomised experiments
- Section 4 – Panel data
- Section 5 – Other types of regression methods

2. Introduction

Experimental randomised and quasi-experimental studies have gained pace over the last few years and became important in the program evaluation literature in the context of economics and development economics (Duflo et al., 2007). The reason for this expansion is that digitisation has decreased the cost of their use while their main advantage is that randomisation, sets the ground for causal inference which is the "Holy Grail" of empirical research. Although experimental randomised experiments (RCTs) avoid many of the challenges of observational studies for causal inference, still many statistical issues arise in evaluating the effect of a treatment. Even in its simplest form, with observable homogenous units receiving a binary treatment, there are many statistical issues of how to conduct inference. Furthermore, behavioural analysis and policy recommendations, for example in energy consumption, have also advanced from the latest developments in experimental design.

There is a proliferation of analytical methods and each time a researcher should choose according to the design specifics of the different experimental and quasi-experimental studies, the data availability and most importantly, according to the research question of interest. For example, research interest may be about the average treatment effect of an intervention, the effect on the total distribution of the output variable, or about the effect's heterogeneity on different samples. To accommodate these issues, and following the related literature, the analysis in this deliverable is based on three main parts. The first is about randomised-based methods of inference, the second is about regression-based methods of inference and the third is about panel data methods and other types of regression.

Our review begins by presenting the analysis of randomised experiments by Fisher's (1935) exact p-value and Neyman's (1923, 1990) repeated sampling approach. These methods that are based crucially on the randomisation assumption, which eliminates issues related to endogeneity and self-selection biases, are simple to implement and provide useful insights about the treatment effect. In the same vein, Rubin's (1975) interpretation of causal statements as potential outcome provides an attractive analytical framework that allows for general heterogeneity in the effects of the treatment and definition of parameters without reference to particular statistical models.

Next, we present regression-based methods for estimating causal effects that on the contrary of Fisher's and Neyman analysis, are based on different sampling perspectives of the outcome for an infinite super-population of units. The basic framework of analysis is a linear regression model with no covariates using OLS estimators for inference. Regression methods also allow for a straightforward incorporation of covariates that increases the statistical power. A useful extension of the regression-based methods is the use of the propensity score as a mean of partitioning the sample into different small strata and then using regression methods within each stratum to adjust for differences in covariates and thus achieving adjusted standard errors and valid confidence intervals. Rosenbaum and Rubin (1983), showed that propensity scores maintain the unconfoundedness assumption and thus estimators' unbiasedness. Similarly, matching methods, proposed by Imbens and Rubin (2015), are used for estimating average effects and other causal estimands, for example, the difference in the median or other quantiles by treatment status, or differences in variances.

Instrumental Variables (IVs) methods, are also presented, since they provide a consistent estimate for average treatment effects when the unconfoundedness assumption fails (its exact definition is presented in section 3.1 below). In cases where a covariate that is relevant for estimating the treatment effect is unobservable, we use an IV that is found to be related with the treatment assignment however is unrelated with any other covariates. IVs can be used for both estimating constant and heterogenous

effects. However, as noted in the related literature, it is a quite difficult task to identify a valuable IV, and in most cases it simple does not exist.

Next, we provide an analysis regarding panel data methods. Panel data consist of repeated observations on the same cross unit, for example, a consumer. Repeated observations in time, control for unobserved factors or missing data and provide a rich information set. Unfortunately, panel data are not readily available in most experimental studies, as they presuppose existence of available data before the implementation of an experiment. Although panel data, have a wide range of methods, we mainly focus here, for reasons that will be apparent later, on fixed effects, first differencing, and difference in differences method. Under fixed effects, we assume that unobserved factors are common across units and thus can be eliminated by demeaning observations. An alternative to fixed effects method is first differencing (FD) and using first differences we can eliminate the unobserved fixed effect by taking first differences. This is an easy-to-use method and is often being use in applied research. Finally, the difference-in-differences (DiD) approach, relies on the presence of additional data before and after the treatment. The usefulness of the DiD method is based on the key assumption of the parallel trends, which states that the outcome variable of interest in the treatment group, would have followed the same time trend as the control group if no treatment had taken place. This means, that although observable and unobservable factors may affect the level of the outcome, this difference must be constant over time. The difference in differences approach has been used in labour economics and for estimating the effects of different types of taxes (Tsoutsoura, 2015).

In the final part of this deliverable, we present three different types of approaches in the realm of regression methods, that can be used for estimating the treatment effect: the regression discontinuity design, the quantile regression and the discrete choice models (logit and probit models). First, regression discontinuity design, has a long tradition in statistics (Cook et al., 2002) and recently has been used in economics for estimating bouncing effects around a threshold. Discontinuity designs are used in cases where the treatment assignment is a deterministic function of an observable variable crating a natural experiment. For example, assume that there is an income threshold for a consumer to be included in a social housing electricity tariff. Then, we can further assume that consumers that lie on the neighbourhood of this threshold (on its right and left), will be similar in every other aspect. This creates two distinct groups with the only in between difference to be the different energy tariff they confront with. Under this framework we can estimate for example, the price elasticity of retail electricity demand.

Quantile regressions are extremely important in cases where the research interest doesn't lie exclusively on the average treatment, rather it is important to estimate treatment effect's entire distribution. Quantile regression can answer research questions of how a treatment effect varies conditional on specific covariates. For example, how the impact effect of a behavioural nudge varies conditional on a consumer's initial energy consumption level. Other researchers have used quantile regressions for answering for example, how inequality varies conditional on other covariates like education and experience (see, e.g., Buckinsky, 1994).

Finally, the framework of using discrete regression models for estimating the effect on binary dependent variables is presented. Until now, we implicitly assumed that the dependent variable of interest takes on continues values. However, it could be the case that the dependent variable of interest takes on only discrete values. For example, the output variable of interest could be whether a consumer aggress upon receiving a behavioural nudge, or as it will be shown below, what is the probability of a consumer to be included in the treatment group conditional on other covariates like place of residence, house size and initial level of energy consumption. The output variable could then take values of zero and one (think of the propensity score), with zero indicating that the consumer is in the control group and one indicating

that the consumer is in the treatment group. Discrete regression models are usually being estimated using Probit or Logit models.

2.1 Definition of causal effects

In this introductory section, we also present the basic framework for causal inferences. It is important to highlight it, since the statistical and econometric methods described below are relevant to this. causal inferences are widely used in everyday life. For example, a person might expect that a training program might increase his future labour income. Thus, through causal inference, we try to identify the independent, actual effect of a phenomenon, intervention, or action. For causal inference, the most challenging part is to identify the actual effect and to provide credible evidence on causal effects. In doing so, we need to have access to rich datasets that include all possible variables that might be relevant to the analysis.

However, in most cases, especially in observational studies, this is a demanding or infeasible task and rarely an estimation of an effect can be characterised as causal. Also, it should be noted that correlation does not imply causality. Nevertheless, well-designed randomised experiments can be used for causal analysis. For example, in cases where an experiment can be designed without the assignment mechanism to depend on characteristics of the units, or on other observables and unobservables (Rosenbaum, 2002) then the estimated relationships can be accounted as causal inferences.

2.2 Internal validity

In this sub-section, short definitions about internal and external validity of an experimental study are provided. An analysis has internal validity, if the observed covariance between a treatment and an outcome variable reflects "*a causal relationship …… in which the variables were manipulated*" (Cook et al., 2002). The word "manipulation" here refers to the existence of a treatment that ultimately has an impact on the outcome variable. Essentially, well-executed randomised experiments by definition have internal validity, since its design focuses on specific testable hypotheses. In other words, internal validity refers to the statistical power of an experiment with results being considered as having statistical significance. Overall, internal validity refers to the ability of a study to estimate causal effects within the study population.

2.3 External validity

The second aspect of the validity of experimental studies is external validity. "*External validity concerns inferences about the extent to which a causal relationship holds over variation in persons, settings, treatments and outcomes*" (Cook, 2002). Thus, external validity is about generalizing causal inferences, drawn for a particular population and setting, to others, where these alternative settings could involve different populations, different outcomes, or different contexts. In other words, it's about whether an idea that takes hold in a small group can do the same in a much larger one. The most important concern for external validity is Deaton's (2010) view that "RCTs, like nonexperimental results, cannot automatically be extrapolated outside the context in which they were obtained".

External validity is of major importance for policy makers, since the optimum would be to carry out a randomised experiment in different settings and still have the same results. In this direction, the researcher should account for differences in the distribution of characteristics across settings and units and adjust for these differences in unit-level characteristics (by reweighting the units, Hotz et al., 2005). Consequently, major concerns for external validity are sampling bias, market equilibrium effects,

spillovers, site selection bias, and political reactions (Banerjee et al., 2017). To provide an intuition, sample bias is apparent in cases where the sample is not representative of the population. The market equilibrium effect is related to changes in the market conditions as a result of a large-scale intervention. For example, a small experiment is in many cases consistent with a partial equilibrium analysis: all relative market prices can be assumed to stay constant which by contrast, in a large experiment do not usually stay constant. Furthermore, many treatments have spillovers on neighboring units, which implies that those units are not ideal control groups. Finally, selection bias is concerned with cases where organisations or individuals who agree to participate in an early experiment may be different from the rest of the population (Heckman, 1992).

3. The analysis of classical randomised experiments

3.1 Introduction

The EVIDENT project aims to provide insights about the effectiveness of behavioural-based interventions on residential energy conservation. The impact evaluation of such interventions is a well-investigated problem and it is best performed through natural field experiments, designed as classical randomised experiments that could provide estimations for causal relations. Under this framework, different types of treatments will be received by participants (control and treatment group), while the differences between the response of the treatment and the control group account for the causal effect of the treatment. Thus, the object of interest is a comparison between the two outcomes for the same unit (customer or participant) when exposed and when not exposed to the treatment. A main limitation of this approach is that only a single outcome can be observed because the unit can be exposed to only one level of treatment. This means that the fundamental problem of causal inference is the missing data.

Rubin (1975) used the notion of potential outcomes, each corresponding to one of the levels of the treatment. Each outcome is *a priori* observable and it depends on the treatment assignment, however, *a posteriori* only one outcome is observable. Both the notion of potential outcome and the assignment mechanism, the process for determining which units receives the treatment and which unit does not, are central to causal inference (Holland, 1986).

The analysis reviewed in this deliverable is relevant to randomised experiments, where the assignment mechanism is both known and controlled by the researcher (Cochran, 1965). Furthermore, the methods proposed in this section assume three basic restrictions on assignment mechanisms (Imbens and Rubin, 2015):

1. Individualistic assignment. A unit's assignment probability of a treatment effect id independent of potential outcomes for other units.
2. Probabilistic assignment. The probability of assignment for each treatment and each unit is nonzero.
3. Unconfounded assignment. The assignment probability is independent of the potential outcome.

Based on the above restrictions and the overall design of the EVIDENT project, we consider only one class of assignment mechanism, namely the classical randomised experiment. In this type of experiment, the assignment mechanism fulfils all three restrictions, individualistic, probabilistic, and unconfoundedness. Under these restrictions, estimates have statistical power and finite sample inferences are possible.

Classical randomised experiments are classified in four different types, namely Bernoulli trials, completely randomised experiments, stratified randomised experiments (randomised blocks), and paired randomised experiments. All four types have different sets of assignment vectors. For reasons related to the specifics of the field experiments in the EVIDENT project, we focus on the completely randomised experiments and stratified randomised experiments.

Completely randomised experiments

In this design type, a fixed number of individuals is assigned to receive the active treatment. In its simplest form, the sample is equally and randomly divided between the treatment and the control group $N_t = \frac{N}{2} = N_c$. The assignment mechanism is:

$$\Pr(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1)) = \begin{cases} \binom{N}{N_t}^{-1}, & \text{for all W such that } \sum_{i=1}^{N} W_i = N_t \\ 0 & otherwise \end{cases}$$

, where $\boldsymbol{W}$ is the assignment vector, taking the value 1 of the individual is assigned to the treatment group and 0 otherwise. $\boldsymbol{X}, \boldsymbol{Y}(0)$, and $\boldsymbol{Y}(1)$ denote the covariate column vector, the observable output in the control group, and the observable output in the treatment group, respectively.

Stratified randomised experiments

In this type of experiments, units are first partitioned into blocks or strata using predefined covariates so that the units in each block are similar with respect to these covariates. Thus, in the cases in which covariates are available, stratified experiments are preferred over completely randomised experiments since they provide greater statistical power (Athey and Imbens, 2017). Then, in each block, a completely randomised experiment is being conducted with assignments independent across blocks. As a result, a stratified randomised experiment with J blocks is a classical randomised experiment with an assignment mechanism satisfying:

$$\Pr(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1)) = \begin{cases} \prod_{j=1}^{J} \binom{N(j)}{N_t(j)}^{-1}, & \text{for all W such that } \sum_{i=1}^{N} W_i = N_t \\ 0 & otherwise \end{cases}$$

If the covariates correspond to substantive information about the units, in the sense that are predictive of the potential outcomes, randomizing within strata will lead to more precise inferences by eliminating the possibility that all or most units of a certain type, as defined by the blocks, are assigned to the same level of treatment (Imbens and Rubin, 2015).

3.2 Estimands and inferences

This section presents the methods used for analysing the results from the completely and stratified randomised experiments reviewed above. There exist two main classes of methodologies. The first one is about methodologies that are based on the randomisation attribute of the assignment mechanism, with the most prominents being Fisher's exact p-values (Fisher, 1935) and Neyman's repeated sampling approach (Neyman, 1923 and 1990). These statistical methods are justified by randomisation, in contrast to the more traditional econometrics methods. Randomisation methods assume that the subject's potential outcome is fixed, and considers the assignment of subjects to treatments as random. To this

end, the most important assumption for ensuring causal results is unconfoundedness, which states that the assignment mechanism is independent of the potential outcomes. More formally we have:

$$W_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big)$$

Under the unconfoundedness assumption, estimands are unbiased.

In contrast, econometrics, that is sampling-based approach considers the treatment assignments to be fixed, while the outcomes are random. Thus, through randomisation a researcher avoids making assumptions about the theoretical distributions of the dependent variables as well as for the covariates. Almost always, these assumptions are difficult to verify. However, both approaches have their pros and cons as it will be showed below. The selection of the method to be used is mostly related to the sample available and the research question.

3.3 Randomised-based estimands and hypotheses

3.3.1 Fisher's exact p-values

Fisher introduced the *sharp null hypothesis* of no effect of the active versus the control treatment. This means that at least a unit is being affected by the active treatment. It is important to note that the null hypothesis is different from the more general hypothesis of the average effect. This hypothesis is weaker than the null hypothesis since it may not be able to reject the null hypothesis of no average effect in cases that are positive effects for some units as soon as these effects are counterbalanced by negative effects in other units. Fisher's exact p-values methods is now being used increasingly in applied research in economics, covering a wide range of topics (Plackett and Burman, 1946, Casari and Cason, 2009, Heckman et al, 1997, Kaplan and Wolf, 2017, Gigerenzer, 2009, Heckman and Vytlacil, 2007, and Bohnet and Frey, 1999)

As discussed earlier, the main problem for causal inference is the missing values of the potential outcomes. A researcher observes only the realised outcome that corresponds to the one out of two different treatment assignments (in cases with a binary treatment). Under Fisher's null hypothesis and under sharp null hypotheses, the other potential outcome is "known" for each unit in the sample and for each assignment mechanism. Actually, is being inferred through the sharp null hypothesis and is a truly nonparametric procedure in the sense that does not rely on other model's assumptions and unknown parameters.

For any test statistic, $\boldsymbol{T}$, that is a function of the stochastic assignment vector, $\boldsymbol{W}$**;** the observed outcomes, $\boldsymbol{Y^{obs}}$, and any pre=treatment variables, $\boldsymbol{X}$**,** the sharp null hypothesis allows to infer the distribution of $\boldsymbol{T}$ which is generated through the stochastic nature of the assignment mechanism. Under the classical randomisation experiment and the three restrictions reported above, the assignment mechanism is known to the researcher and no further assumptions are needed for inferring the distribution of $\boldsymbol{T}$ as is the case in the model-based methods. Using the distribution**,** we can compare the actually observed value of the chosen statistic, $\boldsymbol{T^{obs}}$, against the distribution of $\boldsymbol{T}$**,** under the null hypothesis. Unlikely values of the observed distribution could be taken as evidence against the null hypothesis of no treatment effect.

The first step in the analysis includes the selection of the sharp null hypothesis. In most empirical applications researchers choose the sharp null hypothesis of no effect, however, this always should follow

from the substantive question of interest. The selection of any other form for a null hypothesis is a straightforward extension of the no effect case. The second step is about the choice of the test statistic and should be chosen to have statistical power against a scientifically interesting alternative hypothesis. Following Imbens and Rubin (2015) we define the statistic as a real-valued function $T(\boldsymbol{W}, \boldsymbol{Y^{obs}}, \boldsymbol{X})$ of the vector of assignments, $\boldsymbol{W}$, the vector of observed outcomes, $\boldsymbol{Y^{obs}}$, and the potential matrix of pre-treatment variables, $\boldsymbol{X}$.

Fisher's exact p-value is easy to compute. First, the researcher obtains an estimation of the chosen statistic, $\boldsymbol{T^{obs}}$. Next, he assumes that the outcomes are fixed between treatment and control groups and reassigns the treatments and recalculates the statistic $\boldsymbol{T}$. This procedure is repeated for all possible combinations of the assignment mechanism and the exact p-value is calculated as the number of cases that the statistic $\boldsymbol{T}$ would be more extreme that our observed value of $\boldsymbol{T}$. For example, in the table below we present the estimated $\boldsymbol{T}$ from a hypothetical exercise that includes 7 different assignment mechanisms $\boldsymbol{W} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{i=7}} \boldsymbol{W_i}$.

Table 1: Assignment vectors and hypothetical T

Assignment mechanism	$\boldsymbol{T}$
$\boldsymbol{W_1}$	$\boldsymbol{T^{obs}}$: 4
$\boldsymbol{W_2}$	2
$\boldsymbol{W_3}$	1
$\boldsymbol{W_4}$	3
$\boldsymbol{W_5}$	4
$\boldsymbol{W_6}$	1
$\boldsymbol{W_7}$	0

Results in Table 1 above show that only $\boldsymbol{W_5}$ has value equal or greater than $\boldsymbol{W_1}$(both have value of 4). Thus, Fisher's exact p-value is 2/7=0,28 which translates as not rejecting the null hypothesis.

These calculations could be done exactly in cases where the samples are small. In general, with $\boldsymbol{N_t}$, units assigned to the treatment group and $\boldsymbol{N_c}$ units assigned to the control group, the number of distinct values of the assignment vector is $\binom{N_t+N_c}{N_t}$, which can grow very quickly and it may be infeasible to calculate the test statistic for each distinct value of the assignment vector. In this case, there is the option for the researcher to calculate the statistic for only a randomly chosen subset of possible assignment vectors. For each draw from the total set the probability of being drawn is $1 \Big/ \binom{N_t+N_c}{N_t}$. The researcher calculates the statistic from the first draw and repeats this process $K-1$ times, in each instance drawing a new vector of assignments and calculating its time the statistic $\boldsymbol{T}$. Then, the p-value for our statistic is approximated by the fraction of these K statistics that are as extreme as, or more extreme than the observed value of the statistic, $\boldsymbol{T^{obs}}$.

Next, we discuss the choice of the statistic. As mentioned before, the choice is based mainly on the nature of data and the research question under investigation. The common statistic used is the absolute value of the difference in average outcomes by treatment status:

$$T^{dif} = \left|\bar{Y}_t^{obs} - \bar{Y}_c^{obs}\right| = \left|\frac{\sum_{i:w=1} Y_i^{obs}}{N_t} - \frac{\sum_{i:w=0} Y_i^{obs}}{N_c}\right| \quad (1)$$

The test statistic in eq. 1 is mostly used when the frequency distributions of $Y_i(0)$ and $Y_i(1)$ (these are the values that correspond to the control and active treatment outcomes respectively) have few outliers. An obvious alternative is to transform in logarithms the outcomes before comparing average differences between treatment levels. This is mostly appealing when the frequency distributions of outcomes have outliers, units with zero values, or the treatment effect is more likely to be multiplicative than additive.

$$T^{log} = \left|\frac{\sum_{i:w=1} \ln\left(Y_i^{obs}\right)}{N_t} - \frac{\sum_{i:w=0} \ln\left(Y_i^{obs}\right)}{N_c}\right| \quad (2)$$

If outliers are a concern or if the treatment effect may have a multiplicative impact as before the researcher may use statistics based on trimmed means or other estimates of location. On choice is the absolute value of differences in medians in the two different treatment assignments:

$$T^{median} = \left|med_t\left(Y_i^{obs}\right) - med_c\left(Y_i^{obs}\right)\right| \quad (3)$$

, where $med_t\left(Y_i^{obs}\right)$ and $med_c\left(Y_i^{obs}\right)$ are the observed sample medians of the subsamples with $W_i = 0\ and\ W_i = 1$ respectively. Other test statistics based on robust estimates of location include the average in each subsample after trimming the lower and upper 5% or 25% of the two subsamples.

$$T^{quant} = \left|q_{\delta,t}\left(Y_i^{obs}\right) - q_{\delta,c}\left(Y_i^{obs}\right)\right| \quad (4)$$

, for $\delta \in (0,1)$ indicating the quantiles of the empirical distribution of Y_i^{obs} for each distinct treatment assignment.

Finally, an important class of test statistics involves transforming the outcomes to ranks before considering differences by treatment status. This choice is mostly attractive in cases where outcomes have a distribution with a substantial number of outliers.

$$\tilde{R}_i = \tilde{R}_i\left(Y_1^{obs} \dots\dots\dots\dots Y_N^{obs}\right) = \sum_{j=1}^{N} \mathbf{1}_{Y_j^{obs} < Y_i^{obs}} \quad (5)$$

Unlike the simple difference in means, or the difference in logarithms, rank-based statistics do not have a direct interpretation as estimated causal effects, however, they can lead to more powerful tests due to their insensitivity to thick-tailed or skewed distributions.

3.3.2 Neyman's repeated sampling approach

A second method for estimating treatment effects that has similarities with Fisher's exact p-value is Neyman's repeated sampling approach (Neyman 1923, 1990, Winship and Morgan, 1999, Rubin, 2005, Rubin, 2008, Levitt and List, 2009, Rosenbaum and Rosenbaum, 2010). As previously, the distribution of the estimated statistics, for example the average treatment effect, depends on the randomisation distribution with all potential outcomes regarded as fixed. Neyman's approach concentrates on estimating both unbiased point estimates and interval estimators for the causal effects based on an unbiased estimator for the sampling variance under repeated sampling.

Neyman's population average treatment effect is given by:

$$\tau = \frac{1}{N}\sum_{i=1}^{N}\big(Y_i(1) - Y_i(0)\big) = \bar{Y}(1) - \bar{Y}(0) \qquad (6)$$

, where $\bar{Y}(1)$ and $\bar{Y}(0)$ are the averages of the potential control and treated outcomes respectively. Because of the randomisation a natural estimator would be:

$$\hat{\tau} = \bar{Y}_t^{obs} - \bar{Y}_c^{obs} \qquad (7)$$

, where

$$\bar{Y}_t^{obs} = \frac{1}{N_t}\sum_{i:W_i=1} Y_i^{obs} \text{ , and } \bar{Y}_c^{obs} = \frac{1}{N_c}\sum_{i:W_i=0} Y_i^{obs} \qquad (8)$$

For a proof regarding the unbiasedness of the estimator $\hat{\tau}$ the reader is referred to Imbens and Rubin (2015).

Next, we display Neyman's unbiased estimator for the sampling variance. This estimator is unbiased under the assumption of a constant additive treatment effect:

$$\mathbb{V}_{neyman} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t} \qquad (9)$$

, where s_c^2 and s_t^2 are the population variances of $Y_i(0)$ and $Y_i(1)$ respectively and are estimated as:

$$s_c^2 = \frac{1}{N_c-1}\sum_{i:W_i=0}\left(Y_i(0) - \bar{Y}_c^{obs}\right)^2 = \frac{1}{N_c-1}\sum_{i:W_i=0}\left(Y_i^{obs} - \bar{Y}_c^{obs}\right)^2 \quad (10)$$

$$s_t^2 = \frac{1}{N_t-1}\sum_{i:W_i=1}\left(Y_i(1) - \bar{Y}_c^{obs}\right)^2 = \frac{1}{N_t-1}\sum_{i:W_i=0}\left(Y_i^{obs} - \bar{Y}_t^{obs}\right)^2 \quad (11)$$

Until now, we assume that the sample is the population of interest, however it could be the case that the sample is a part of an infinite super-population. $\mathbb{V}_{neyman}$ remains unbiased estimator for the sampling variance of $\hat{\tau}$ of the infinite super-population average treatment effect.

To construct confidence intervals for deriving interval estimators, we use $\mathbb{V}_{neyman} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}$ from eq. 9 and base the interval on a normal approximation to the randomisation distribution of $\hat{\tau}$. For example, if we would like to construct a 90% confidence interval, we use the 5th and 95th percentile of the standard normal distribution, -1.645 and 1.645, to calculate a nominal central 90% confidence interval as:

$$CI^{0.90}(\tau) = (\hat{\tau} - 1.645\sqrt{\mathbb{V}_{neyman}}, \hat{\tau} + 1.645\sqrt{\mathbb{V}_{neyman}}) \quad (12)$$

The sampling variance can also be used carry out tests of hypotheses regarding the average treatment effect. Assume for example that the average treatment effect is zero against the alternative hypothesis that the average effect differs from zero:

$$H_0: \frac{1}{N}\sum_i (Y_i(1) - Y(0)) = 0$$

$$H_a: \frac{1}{N}\sum_i (Y_i(1) - Y(0)) \neq 0$$

The resulting t-statistic is:

$$\mathrm{t} = \frac{\bar{Y}_t^{obs} - \bar{Y}_c^{obs}}{\sqrt{S_Y^2/N_t + S_Y^2/N_c}} \approx \frac{\bar{Y}_t^{obs} - \bar{Y}_c^{obs}}{\sqrt{\sigma^2/N_t + \sigma^2/N_c}} \quad (13)$$

In situations that are discrete covariates, the research should proceed by partitioning the sample into subsamples defined by the value of the covariate and then conduct the analysis separately using these subsamples. The average of these estimators weighted by the sample size derive the overall average treatment effect and should remain unbiased.

3.4 Regression methods estimands and hypotheses

Regression methods could also be used for estimating causal effects as they offer more interpretable results and provide a straightforward way to incorporate a large number of covariates resulting in more precise causal inferences, than the exact methods, in the case when the included covariates have explanatory power. Nevertheless, Freedman (2008a) points out that randomisation does not justify ordinary least squares (OLS) main assumptions and their use often restricts the set of questions that can be addressed.

With respect to the previous methods that were based on randomisation, regression-based methods rely on different sampling perspectives of the outcome for an infinite super-population of units. This means that we consider our sample as a random sample drawn from an infinite population. These assumptions are in most of the times hard to be evaluated in practice as we have no information about the joint distribution in the outcome and we only observe outcomes and covariates in our sample.

We begin our discussion with the linear regression with no covariates and we next include covariates. It is important first to interpret the estimator of the causal effect of the treatment and then to discuss approaches to inference. The linear regression function for the observed outcome, Y_i^{ols}, is:

$$Y_i^{ols} = \alpha + \tau \cdot W_i + \varepsilon_i \qquad (14)$$

, where the unobserved residual ε_i captures unobserved covariates or misspecification errors and W_i as before is the treatment indicator. Coefficient τ indicates the causal effect of the treatment. In the OLS approach, the error term ε_i is independent of, or at least uncorrelated with, the treatment indicator, τ. However, this assumption is difficult to be evaluated in practice. For estimating the coefficients τ and α, the following equation is minimised:

$$(\hat{\tau}_{ols}, \hat{a}_{ols}) = \arg\min_{\tau,\alpha} \sum_{i=1}^{N} \left(Y_i^{obs} - \alpha - \tau \cdot W_i\right)^2 \qquad (15)$$

Having the following solutions:

$$\hat{\tau}_{ols} = \frac{\sum_{i=1}^{N}(W_i - \bar{W}) \cdot \left(Y_i^{obs} - \bar{Y}^{obs}\right)}{\sum_{i=1}^{N}(W_i - \bar{W})^2} = \bar{Y}_t^{obs} - \bar{Y}_c^{obs}, \qquad (16)$$

and

$$\hat{\alpha}_{ols} = \bar{Y}^{obs} - \hat{\tau}_{ols} \cdot \bar{W}$$

, where

$$\bar{Y}^{obs} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{obs} \text{ , and } \bar{W} = \frac{1}{N}\sum_{i=1}^{N} W_i = \frac{N_t}{N}$$

Imbens and Rubin (2015), show that the OLS estimator $\hat{\tau}_{ols}$, is identical to the difference in average outcomes by treatment status:

$$\hat{\tau}_{ols} = Y_t^{obs} - Y_c^{obs} = \hat{t}^{diff} \quad (17)$$

, where, as before

$$\bar{Y}_t^{obs} = \frac{1}{N_t}\sum_{i:W_i=1} Y_i^{obs} \text{ , and } \bar{Y}_c^{obs} = \frac{1}{N_c}\sum_{i:W_i=0} Y_i^{obs}$$

$\hat{\tau}_{ols}$ is the estimate of the causal impact of the treatment. It is important to note that, although OLS methods usually produce only an association measure between two variables when used in observational studies, in the case of randomised experiments, the $\hat{\tau}_{ols}$ is transformed to an estimation of the causal impact as a result of being identical to $Y_t^{obs} - Y_c^{obs}$.

Next, the analysis includes the inference approach of the OLS method. Inference is based on the variance of the residuals in eq. 16 and could be estimated as:

$$\hat{\sigma}_{y/w}^2 = \frac{1}{N-2}\sum_{i=1}^{N} \hat{\varepsilon}_i^2 = \frac{1}{N-2}\sum_{i=1}^{N} (Y_i^{obs} - \hat{Y}^{obs})^2 \quad (18)$$

Consequently, the variance in eq.18 can be rewritten as

$$\hat{\sigma}_{y/w}^2 = \frac{1}{N-2}\left(\sum_{i:W=0}^{N} \left(Y_i^{obs} - \hat{Y}_i^{obs}\right)^2 + \sum_{i:W=1}^{N} \left(Y_i^{obs} - \hat{Y}_i^{obs}\right)^2\right) \quad (19)$$

Accordingly, the estimator for the sampling variance of $\hat{\tau}_{ols}$ is

$$\widehat{\mathbb{V}}^{homosk} = \frac{\hat{\sigma}^2_{y/w}}{\sum_{i=1}^{N}(W_i - \bar{W})^2} \quad (20)$$

, which is equal to $\mathbb{V}_{neyman} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}$ from eq. 9 above. Thus, in the case of no additional explanatory or pre-treatment variables the estimated variance for the OLS approach is identical to the Neyman's sampling variance.

However, it is common in practice that higher precision in estimates can be achieved by including covariates that are related to the outcome. This increase in precision is translated to a lower estimated variance. With additional covariates the regression function is specified as:

$$Y_i^{ols} = \alpha + \tau \cdot W_i + X_i\beta + \varepsilon_i \quad (21)$$

, where X_i is a row vector of covariates. In the context of EVIDENT, this could be the house size or a household's main heating source. As above, the regression coefficients are estimated through OLS:

$$\left(\hat{\tau}_{ols}, \hat{a}_{ols}, \hat{\beta}_{ols}\right) = \arg\min_{\tau,\alpha} \sum_{i=1}^{N}\left(Y_i^{obs} - \alpha - \tau \cdot W_i - X_i\beta\right)^2 \quad (22)$$

Imbens and Rubin (2015) find that irrespective of whether the specified and subsequently estimated regression is truly linear in the covariates in the population, $\hat{\tau}_{ols}$, is still unbiased in large samples. The key insight is that by randomising the treatment assignment, the population correlation between the treatment indicator and the covariates becomes zero. Therefore, the inclusion of the additional covariates matters for the sampling variance of the estimators which is now estimated as:

$$N * \widehat{\mathbb{V}}^{homosk} = \frac{\hat{\sigma}^2_{Y\cdot W,X}}{\rho(1-\rho)} \quad (23)$$

, where ρ is the probability limit of the ratio of the number of treated units to the number of total units, $\rho = plim\left({}^{N_t}/_{N}\right)$. It is clear from eq. 23 that, if the covariates explain much of the variation in the potential outcomes, then the conditional variance $\hat{\sigma}^2_{Y\cdot W,X}$ will be much smaller than the variance in eq. 20, which will lead in an increase in the precision. In practice, the sampling variance for the average treatment effect can be estimated using standard OLS as:

$$\widehat{\mathbb{V}}^{homosk} = \frac{1}{N(N-1-\dim(X_i))} \frac{\sum_{i=1}^{N}\left(Y_i^{obs}-\hat{\alpha}-\hat{\tau}\cdot W_i-X_i\hat{\beta}\right)^2}{\bar{W}(1-\bar{W})} \quad (24)$$

Thus, linear regression models could be used for both estimating average treatment effects and for the presence of treatment effects or testing hypotheses concerning the heterogeneity in the treatment effect. As in the exact models method, analysed in section 4.3, randomisation has an important role as it validates the consistency of the OLS estimators that do not have to rely on the validity of the regression model as is the case in observational studies. However, these attractive features of the regression methods are only present in large samples.

3.4.1 Regression methods using covariates

Covariates could also be useful in estimating average treatment effects in cases where the treatment assignment, w, and the outcomes, (y_0, y_1), are allowed to be correlated. This may be relevant for EVIDENT's use cases 3 and 4, in which participants will be self-selected (self-selection bias). Rosenbaum and Rubin (1983), introduced the ignorability of treatment assumption and used the observed covariates to assume ignorability in a conditional mean independence sense. The assumption is the following:

Assumption 1: $(a)\ E(y_0|x,w) = E(y_0|x); and\ (b)\ E(y_1|x,w) = E(y_1|x)$ (25)

Eq. 25, indicates that even though (y_0, y_1) and w might be correlated, they are uncorrelated once we partial out x. We further decompose the counterfactual outcomes into their means and a stochastic part with zero mean:

$$y_0 = \mu_0 + v_0 \quad , \quad E(v_0) = 0$$
$$y_1 = \mu_1 + v_1 \quad , \quad E(v_1) = 0$$

Under assumption 1 and assuming in addition, that:

$$E(v_1|x) = E(v_0|x)$$

, then standard parametric regression methods can be used to estimate average treatment effects that are given as follows:

$$E(y|w,x) = \mu_0 + aw + g_0(x) \quad (26)$$

, where a, is the average treatment effect and $g_0(x) = E(v_0|x)$. $g_0(x)$ can be either a linear or a non-linear function of, x, and is an example of a control function and when is added in a regression as in eq. 26, it controls for possible self-selection bias.

This result can be achieved even without assumption 1 in eq.25, which means the effect of treatment is the same for everyone in the population. Now, the estimated equation is:

$$E(y|w,x) = \mu_0 + aw + g_0(x) + w[g_1(x) - g_0(x)] \quad (27)$$

, where a, is the average treatment effect and $g_0(x) = E(v_0|x)$ and $g_1(x) = E(v_1|x)$. In practice and in empirical application, eq.27 takes the following form:

$$E(y|w,x) = \gamma + aw + x\beta_0 + w[x - \psi]\delta \quad (28)$$

, where β_0 and δ are vectors of unknown parameters and $\psi \equiv E(x)$. Subtracting the mean from x, ensures that the average treatment effect is a. Thus, eq. 28 can be further simplified to:

$$E(y|w,x) = \gamma + aw + x\beta_0 + w[x - \bar{x}]\delta \quad (29)$$

This type of regression estimator, as in eq. 29, can be also applied in regression discontinuity designs. In this case, the treatment assignment, w, is typically a step function of a covariate, $w = f(s)$, in which $w = 1[s \leq s_0]$, where s_0, is a known threshold. Two remarks should be highlighted; the first is that assumption 1 must hold, while the second is that the g function should be continuous in x. If g is a discontinuous function of x, then the changes in y due to a change in x will be indistinguishable.

Finally, it should be noted, that not all control variables are good variables. This means, that the inclusion of a covariate or a set of covariates may introduce problems in interpreting results, since a comparison of the dependent variable on other covariates does not have a causal interpretation. Usually, good controls are variables that can be thought of as having been fixed at the time the regressor of interest was determined.

3.4.2 Regression methods using propensity scores

Propensity scores are usually regarded as alternatives to regression methods, however, in this deliverable as well as in the EVIDENT project, propensity scores will be interpreted as regressions. The idea behind the propensity score is to partition the sample into different strata based on this score. The main objective

towards this is to maintain the assumption of unconfoundedness, which is an important assumption for estimands to be unbiased. However, this is a difficult assumption to verify in practice (with the exception of most randomised experiments). Rosenbaum and Rubin (1983), showed that propensity score can contribute in making more tractable the unconfoundedness assumption.

Propensity score is defined as the probability of a unit to be assigned in the treatment group on different values of a covariate:

$$e(x) = pr(W_i = 1|X = x_i)$$

and thus, unconfoundedness assumption now turns to:

$$W_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big)|\, e(X_i) \qquad (30)$$

Eq. 30, denotes that the assignment mechanism is orthogonal to the potential outcomes condition on the propensity score which suffices to control for $e(x)$ rather than X to remove biases associated with a non-random assignment mechanism. To apply the propensity score method, we first need to estimate $e(x)$ usually through a logistic regression function and then to sort the observations according to their propensity scores:

$$\hat{e}(X_{i1}) \leq \hat{e}(X_{i2}) \leq \hat{e}(X_{i3}) \dots\dots\dots\dots \hat{e}(X_{ij})$$

Then, define B_{ij}, for $i = 1, \dots\dots, N$ and $j = 1, \dots\dots\dots, J$ as the indicators:

$$B_{ij} = \begin{cases} 1 \\ 0 \end{cases} if\ c_{j-1} \leq e(X_i) \leq c_j$$

and

$$B_{ij} = 1 - \sum_{j=1}^{J-1} B_{ij}$$

Finally, we split the sample into J evenly size strata using the sorted propensity score and, in each stratum j, estimate the average treatment effect $\tau_j = E[Y_i(1) - Y_i(0)|B_{ij} = 1]$ as:

$$\hat{\tau}_j = \bar{Y}_{j1} - \bar{Y}_{j0}$$

, where

$$\bar{Y}_{jw} = \frac{1}{N_{jw}} \sum_{i:W_i=w}^{J-1} B_{ij} \times Y_i$$

and

$$N_{jw} = \sum_{i:W_i=w}^{N} B_{ij}$$

If strata are large and differences within strata in the propensity scores are relatively small, then the propensity score within strata is constant and treatment effects estimators are unbiased, as if were generated by a completely randomised experiment. OLS-type regressions could then be used to estimate the effect as was the case in subsection 4.4.1. The average treatment effect is then estimated as the weighted average of the within-stratum estimates:

$$\hat{\mathrm{T}}_{block} = \sum_{j=1}^{J} \hat{\mathrm{T}}_j \cdot \left(\frac{N_{j0} + N_{j1}}{N} \right)$$

As a rule of thumb, Cochran (1968), showed that if the sample is be partitioned in five equalised strata, then the remaining bias is less than 5%. However, it should be noted that this should be decided case by case and should depend on the sample size and the joint distribution of the data. Finally, for inference we estimate the variance of the estimator which is calculated conditional on the strata indicators based on the random assignment assumption within the strata j. The variance within strata is estimated as follows:

$$\hat{V}_j = \hat{V}_{j0} + \hat{V}_{j1}$$

, where

$$\hat{V}_{jw} = \frac{S_{jw}^2}{N_{jw}},$$

, where

$$S_{jw}^2 = \frac{1}{N_{jw}} \sum_{i:B_{ij}=1, W_i=w} \left(Y_i - \bar{Y}_{jw} \right)^2$$

The overall assumption is then estimated as:

$$\widehat{\mathbb{V}}_{(\hat{\tau}_{block})} = \sum_{j=1}^{J} \left(\hat{V}_{0j} + \hat{V}_{1j}\right) \cdot \left(\frac{N_{j0} + N_{j1}}{N}\right)^2$$

In a different approach, the inverse of the propensity score could be used as weight to the treated and control samples for estimating the expectation of the unconditional response to the treatment (inverse-propensity weighting). Under this specification, the average treatment effect is estimated as:

$$\tau_{PATE} = E\left[\frac{W_i \cdot Y_i}{e(X_i)} - \frac{(1-W_i)\cdot Y_i}{1-e(X_i)}\right] \quad (31)$$

An obvious estimate for eq. 31 is:

$$\hat{\tau}_{weight} = \frac{1}{N} \times \sum_{i=1}^{N}\left[\frac{W_i \cdot Y_i}{e(X_i)} - \frac{(1-W_i)\cdot Y_i}{1-e(X_i)}\right] \quad (32)$$

Note that, in eq. 32, $e(X_i)$ is not known. However, we can use an estimator of $e(X_i)$ from a logistic regression function as it was the case in the previous method (Hirano, Imbens, and Ridder, 2003). Then, given the estimated propensity score we estimate the average treatment effect as follows:

$$\hat{\tau}_{ipw} = \sum_{i=1}^{N} \frac{W_i \cdot Y_i}{\hat{e}(X_i)} \Big/ \sum_{i=1}^{N} \frac{W_i}{\hat{e}(X_i)} - \sum_{i=1}^{N} \frac{(1-W_i)\cdot Y_i}{1-\hat{e}(X_i)} \Big/ \sum_{i=1}^{N} \frac{W_i}{1-\hat{e}(X_i)} \quad (33)$$

Imbens and Rubin (2009) highlight the fact that, when the covariate distributions are substantially different for the two treatment groups, concerns may arise for the use of inverse propensity weights (IPW). This results to propensity score getting closer to zero or one and these extreme values, in turn, cause identification issues. This means that alternative parametric models, such as Probit or Logit may result to very different estimates.

3.4.3 Instrumental Variables and Two-Stage Least Squares

3.4.3.1 Constant effects

IVs have been widely used in econometrics and statistics (Sargan, 1958, Baum et al., 2003, Miguel et al., 2004, and Phillips and Hansen, 1990) and have been found to consistently estimate the parameters. In experiments, they are mostly used for estimating average treatment effects when we suspect failure of the unconfoundedness assumption, for solving the problem of missing or unknown control variables, and for estimating heterogenous treatment effects. In this section, we present the IV method along with a sample estimates and inference and its use with heterogenous potential outcomes.

If a covariate that is relevant for estimating the treatment effect is unobservable, then we may use IV methods. To this end, we need an instrument, for example, z_i, to predict treatment assignment, w_i, but uncorrelated with any other determinants of the dependent variable. That is, the instrument to be redundant in a certain conditional expectation and is called an exclusion restriction. z_i should have a clear effect on the treatment.

IV estimands are estimated under a two-step procedure. In the first step, we regress the treatment on the instrument and possible covariates. This is for ensuring that the instrument does affect the treatment, as in a different case the use of the instrument would be irrelevant. In the second step, we regress potential outcome on the instrument and covariates also used in the first step. The mathematical representation of the two steps is formulated as:

$$w_i = X'_i\pi_{10} + \pi_{11}Z_\iota + \varepsilon_{1i} \quad (33a)$$

$$Y_i = X'_i\pi_{20} + \pi_{21}Z_\iota + \varepsilon_{2i} \quad (33b)$$

, where Y_i is the potential output and ε_{ji} are random errors that are uncorrelated with the IV. The parameter π_{11}, from above, indicates the first-step of the IV on the treatment effect.

Now, assume that the casual relationship is given be the following equation:

$$Y_i = \alpha' X_i + \rho w_i + \eta_i \quad (34)$$

, where η_i is a compound error term that includes the unobservable covariate and an error term. If we substitute the first-step equation 33a into the casual relationship we get equation 33b:

$$Y_i = X'_i\pi_{20} + \pi_{21}Z_i + \varepsilon_{2i}$$

After some algebraic manipulations, that are beyond the scope of this deliverable (for more details the reader is being referred to Angrist and Pischke (2008), the second-step equation, 33b can be written as:

$$Y_i = X'_i\pi_{20} + \rho\widehat{w}_i + [\eta_i + \rho(w_i - \widehat{w}_i)] \quad (35)$$

Using the Two-Stage Least Squares method (2SLS), we can estimate the treatment effect in two steps. In the first step, using OLS we estimate the $\widehat{w}_i$ from eq. 33a. and in the second step we estimate eq. 35. The estimated parameter ρ is the average treatment effect. However, in practice we don't usually construct 2SLS estimates in two-steps since the resulting standard errors would be wrong. This is done using

specialised software routines such as STATA. The extension to the multi-instrument case is straightforward.

3.4.3.2 Heterogenous effects

We turn our analysis from the constant to the heterogenous treatment effects using the IVs method. This type of analysis is important for both internal and external validity. In order for a policy intervention to be effective at a larger scale, its findings should have predictive value in different contexts. For example, in the PPC use case, the treatment group receiving the information material regarding peer-comparison energy consumption data will be lottery-drafted. The findings from this analysis will not be directly comparable to the observable impact when the treatment expands to additional PPC's customers, since this expansion will be based on customer's own willingness to receive the information material, which will cause selection bias. IVs methods could be used to alleviate this selection bias and elicit any heterogenous effects among the customers that will contribute in better estimating the real impact of the intervention at scale.

For estimating heterogenous effects, Angrist and Pischke (2008) state that an IV is needed for initiating a causal chain, where the instrument, Z_i, affects the treatment status, W_i, which is the variable of interest, which in turns affect outcomes, Y_i. In this respect, they impose three assumptions. The first is that the instrument is as good as randomly assigned and it is independent of potential outcomes and potential treatment assignments:

$$[\{Y_i(d,z), \forall d, z\}, D_{1i}, D_{0i}] \coprod z_i$$

As before, D_{1i}, D_{0i} indicate whether a unit belongs to the treatment or the control group respectively. This independence assumption captures the causal effect of the instrument, Z_i on the assignment mechanism, D_{1i}, D_{0i}.

The second assumption is that the potential outcome, $Y_i(d,z)$, is merely only a function of the assignment mechanism, D_{1i}, D_{0i}. In our PPC example, this means that although a lottery-draft will affect the assignment mechanism, however, this will not affect consumer's energy consumption. The latter should be only affected by the treatment assignment. As it was shown in the case of the constant effect above this assumption is named the exclusion restriction. The exclusion restriction would fail if for example consumers that do not receive the informational material were affected in some way other than through a receipt of the informational material.

The third and final assumption for the heterogenous IV model, is that while the instrument may have no effect on some people, all of those who are affected are affected in the same way. This is named the monotonicity assumption. In the PPC example, those who will be eligible for receiving the informational material through the lottery, will receive it. Without monotonicity, IVs estimators are not guaranteed to estimate a weighted average of the underlying individual causal effects.

These three assumptions postulate that an instrument which is as good as randomly assigned, affects the outcome through a single known channel, has a first-stage, and affects the causal channel of interest only in one direction, can be used to estimate the average causal effect on the affected group.

3.4.4 Matching methods

Another popular method for estimating average treatment effects that could potentially be used in all EVIDENT's use cases is matching. Matching estimators can be considered as the most intuitive estimators for causal effects. In applied research, several variations of the matching estimators are used. The first is a simple case where we match each treated unit to a single control unit, with exactly the same values of the covariates, using each control unit at most once as a control.

We mainly focus on average effects and pair matching, although the methods already extend to estimating other causal estimands, for example, the difference in the median or other quantiles by treatment status, or differences in variances. The natural estimator for the average treatment effect for the treated units is simply the average difference within the pairs, and one can estimate the sampling variance by the sample variance of the within-pair differences divided by the number of pairs. Several other extensions include choosing a unit for match with replacement, that is, units that are used more than once, and cases where multiple matches are used. The latter accounts for cases that several units from the control group are matched to a unit from the treatment group.

The procedure for implementing the matching method for empirical analysis includes three steps. The first step is about defining the distance to be used as the matching indicator (i.e., how should covariates be transformed or scaled in order to be used in matching), while the second step is about the distance metric itself. In the literature, several distance metrics have been proposed such as the Mahalanobis or the Euclidean metric. These metrics are used to match treatment units to the their "closest" control counterpart. Finally, the third step is about estimating the treatment effect.

Regarding the first step, it is important for the researcher to choose how covariates will be transformed or scaled. The issue of the choice of metric is compounded by the presence of multiple covariates, each of which can be continuous, discrete, or a simple indicator variable. It is rather important in cases where covariates have no natural scale, and therefore one should use a metric that is invariant to their scale. Hence, after a transformation is chosen (e.g., logarithm versus level) for a covariate, researchers typically should normalise all covariates to a common variance before matching.

In the second step, Imbens and Rubin (2015) propose distance metrics of the form:

$$d_v(x, x') = (x' V^{-1} x)^{1/2}$$

, where V, is the distance, which can be estimated using the Mahalanobis metric:

$$V_m = \frac{1}{2}\left(\frac{1}{N_c}\sum_{i:w_i=0}(X_i - \bar{X}_c)^T(X_i - \bar{X}_c) + \frac{1}{N_t}\sum_{i:w_i=1}(X_i - \bar{X}_t)^T(X_i - \bar{X}_t)\right.$$

As it is apparent from the equation above, the metric is based on correlations across covariates. It is important to note that matches based on the Mahalanobis metric are invariant to affine transformations of the covariates. A second choice of metric is the Euclidean:

$$V_E = diag(V_M)$$

, which is the diagonal matrix with variances on the diagonal ignoring the covariances.

Ideally, when considering alternative distance metrics in the pursuit of estimating treatment effects for treated units, the intermediate goal is to obtain a metric that creates matched pairs such that the expected control outcomes at the covariate values are identical, or at least very similar. In cases the covariates are nearly uncorrelated the choice of the distance metric has little effect on the estimated results.

Next, we estimate an unbiased estimator for the average treatment effect for the case a one-to-one match between units in the treatment and the control group.

$$\hat{t}_t^{match} = \frac{1}{N_t}\sum_{i:W_i} \hat{t}_t^{match} = \frac{1}{N_t}\sum_{i:W_i} Y_i^{obs} - Y_{m_c^i}^{obs} = \frac{1}{N_t}\sum_{i:W_i} (Y_i(1) - Y_{m_c^i}(0))$$

, where m_c^i is the set of control indices containing the matches for the treated unit i. Accordingly, the estimator for the sample variance is:

$$\hat{\mathbb{V}}_{(\hat{t}_t^{match})} = \frac{1}{N_t}\sum_{i:W_i} (Y_i^{obs} - Y_{m_c^i}^{obs} - \hat{t}_t^{pair})^2$$

Next, we present the case of matching with replacement. This means that we allow for a control unit to be used more than once as a matching unit. As before, the choice of the best matching unit is relevant to the proximity of a control unit to the treatment unit. Replacement has the advantage that reduces the bias of the matching indicator, as now, the best matching control units are always available to be chosen more than once.

However, there is also one disadvantage by using control units repeatedly as this increases the sampling variance of the estimator. The point here is that replacement forces us to use fewer control units than before. Moreover, this may also lead to difficulties in estimating the sample variance of the estimator.

Empirically, the optimal match in now:

$$m_c^i = argmin_{i' \in C} \| X_1 - X_{1'} \|$$

The average treatment effect for the treated is estimated as:

$$\hat{t}_t^{repl} = \frac{1}{N_t}\sum_{i:W_i} Y_i^{obs} - Y_{m_c^i}^{obs} = \frac{1}{N_t}\sum_{i:W_i}(Y_i(1) - Y_{m_c^i}(0))$$

A very important variable now is the number of times a control unit is used a matching unit which is given by:

$$L(i) = \sum_{j=1}^{N} \mathbf{1}_{j \in M_i^c}$$

For control unit i; $L(i) = 0$ for all treatment units. The simple matching estimator of the sample average treatment effect of the treated is estimated as:

$$\hat{t}_t^{repl} = \frac{1}{N_t}\sum_{i=1}\left(W_i Y_i^{obs} - (1 - W_i)L_i Y_i^{obs}\right)$$

$$= \frac{1}{N_t}\sum_{i=1}^{N}(W_i Y_i(1) - (1 - W_i)L_i Y_i(0))$$

This formula shows that the matching estimator is a weighted average of treated and control outcomes within the full sample. For the treated units the weights are all $\frac{1}{N_t}$, and for the control units the weights sum to one, but vary, with the value of the weight reflecting each control units' relative value as a comparison unit for the treated units.

In cases that M matches are used for each treatment units the average treatment effect estimator is given by:

$$\hat{t}_t^{repl} = \frac{1}{N_t}\sum_{i=1}^{N}(Y_i(1) - \frac{1}{M}\sum_{j \in M^c(i)} Y_j\,(0))$$

and the corresponding sampling variance is:

$$\widehat{\mathbb{V}}\left(\hat{t}_t^{match,\ M}\right) = \frac{1}{N_t}(\sigma_t^2 + \frac{\sigma_c^2}{M})$$

, with σ_t^2, and σ_c^2 being the variances of the treatment and control group respectively. We can use in their places the sample counterparts and still get unbiased results. Overall, matching can be used for estimating the average treatment effect for the sample and for the treated. We can select between different matching metrics and matching methods such as with and without replacement. All these methods lead to robust estimates.

4. Panel data

4.1 Introduction

The previous sections presented methods key to causal inference, based on cross section data, where at a given point in time, units are selected at random from the population. However, panel data, which consist of repeated observations on the same cross section of individuals, consumers, firms, states, countries, etc, could also be used to deal with unobserved confounders. Please note, that our main goal is to estimate causal effects by controlling observed confounding factors and in the case of potential outcome we can only observe one state of treatment for a unit. Thus, all methods are about imputing missing data. Panel data having a time and a cross section dimension may be used to control for unobserved but fixed omitted variables by requiring the counterfactual trend behaviour of treatment and control groups to be the same. We also have the option to control for lagged dependent variables.

Panel data can be used under certain assumptions to obtain consistent estimators in the presence of omitted variables. To be more specific, suppose a linear model with the unobserved variable, c, entering additively along with the other observable covariates, x_j:

$$E(y|\boldsymbol{x}, c) = a_0 + \boldsymbol{x}\boldsymbol{a_1} + c$$

The main assumption of interest is now whether c is correlated with each x_j. If it is uncorrelated then c is another unobserved factor affecting y that is not systematically related to the observable explanatory variables. On the other hand, if $Cov(x_j, c) \neq 0$, putting the unobserved factor into the error term may cause serious issues.

Before proceeding in the analysis of using data for estimating the treatment effect, we should briefly present two key conditions for OLS that need to hold for consistently estimating the parameters of interest. First, is the orthogonality condition:

$$E(\Delta x' \Delta u) = 0$$

The orthogonality condition postulates that the observed covariates should be orthogonal to error terms, that is to unobserved factors. The second assumption is the rank condition:

$$rank\ E(\Delta x' \Delta x) = K$$

The rank condition ensures that there are no exact linear relationships between the covariates. x is a full column rank meaning that the columns of x are linearly independent and that there are at least K observations. A final comment is about asymptotic analysis which is useful for providing a reasonable approximation to the finite sample properties of estimators and statistics. For example, we are confident that $N \to \infty$ asymptotics are more appropriate than $T \to \infty$ asymptotics. That is, if N (the cross sections) is sufficiently large relative to T, and we can assume rough independence in the cross section, then our

asymptotic analysis should provide suitable approximations (Wooldridge, 2010). In cases where N is of the same order as T, additional assumptions about the nature of the time series dependence are needed. In the EVIDENT project, we have cases with $N \to \infty$ and T held fixed as is the case of PPC, and $T \to \infty$ with N held fixed as in the case of CW where hourly data are available.

4.2 Fixed effects

The fixed effects method has been widely used in economics (Ashenfelter, and Rouse, 1998, Freeman, 1984, and Bai, 2009). For presenting the fixed method in panel data, we again consider the linear unobserved effects model for T periods:

$$E(Y_{it}|A_i, X_{it}, t, D_{it})$$

The observed Y_{it} is either Y_{0it} or Y_{1it} potential outcome from the control or the treatment group respectively, A_i is the unobserved factor, for example, could indicate how environmentally sensitive is a consumer, t is a dummy variable for time (hour, day, week, month, year, etc) and D_{it} indicates a unit's treatment status in time t.

The sample counterpart of the expectations equation for the control group, $E(Y_{0it}|A_i, X_{it}, t)$ is given by:

$$Y_{0it} = a + t + A'_i\gamma + X_{it}\delta + \varepsilon_{it} \qquad (36)$$

The first fixed effects assumption is strict exogeneity of the explanatory variables conditional on A_i:

$$E(\varepsilon_{it}|X_{it}, A_i) = 0, \qquad t = 1,2, \ldots\ldots, T$$

Next, we assume that the unobserved factor A_i is fixed in time and thus it appears in the equation without the t subscript, and that the causal effect, for example, of the information material, is additive and constant, ρ . Then the equation for the treatment is given by:

$$E(Y_{1it}|A_i, X_{it}, t, D_{it}) = E(Y_{0it}|A_i, X_{it}, t) + \rho$$

This implies

$$(Y_{1it}|A_i, X_{it}, t, D_{it}) = a + t + \rho D_{it} + A'_i\gamma + X_{it}\delta + \varepsilon_{it} \qquad (37)$$

, where ρ is the casual effect of interest. The sample counterpart of eq. 37 is:

$$Y_{it} = a_i + t + \rho D_{it} + X_{it}\delta + \varepsilon_{it} \quad (38)$$

, where $a_i = a + A'_i\gamma$ and now eq. 38 is called the fixed-effects model. Imagine that $N = 500$, according to eq. 38 we have to estimate at least 500 parameters that are coefficients on dummies for each individual while the time effects are coefficients on time dummies. These dummies for each individual are related to the unobserved idiosyncratic characteristic.

The idea for estimating ρ, under the assumption of strict exogeneity is to transform eq. 38 to eliminate the unobserved effect a_i. When more than one time periods are available, as is the case in panel data, there are several transformations that accomplish this purpose. We will present the within transformation also called the fixed effects transformation. It is called within estimator because it uses the time variation within each cross section. The within transformation is obtained by first averaging eq. 38 over $t = 1,2, \ldots\ldots, T$ to get the cross-section estimation:

$$\bar{Y}_i = a_i + \bar{t} + \rho\bar{D}_i + \bar{X}_i\delta + \bar{\varepsilon}_i \quad (39)$$

, where $\bar{Y}_i = T^{-1}\sum_{t=1}^{T} Y_{it}$, $\bar{D} = T^{-1}\sum_{t=1}^{T} D_i$, $\bar{X}_i = T^{-1}\sum_{t=1}^{T} X_{it}$, and $\bar{\varepsilon}_i = T^{-1}\sum_{t=1}^{T} \varepsilon_{it}$. Then, subtracting eq. 39 from eq. 38 for each t gives the within transformation equation:

$$Y_{it} - \bar{Y}_i = t - \bar{t} + \rho(D_{it} - \bar{D}_i) + \delta(X_{it} - \bar{X}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i) \quad (40)$$

The time demeaning of the original eq. 38 has removed the individual specific effect a_i. And can be estimated with OLS. For the fixed effect estimator to be efficient a further assumption is needed:

$$E\left(\varepsilon_i\varepsilon_{it}^{'}|X_i, a_i\right) = \sigma_\varepsilon^2 I_T$$

, where σ_ε^2 is the variance of the error term. This assumption ensures that the idiosyncratic errors, ε_i, have a constant variance across t and are serially uncorrelated. Under this assumption, fixed effects estimators are efficient leading to simple computation of standard errors and t-statistics.

Also, note that we assumed time-invariant omitted variables for implementing the fixed effects method. However, in many cases this may not be a plausible assumption. Thus, it may be convenient to alter our estimation strategy described in eq. 38 above and control for past values of the dependent variable. Then, the equation to be estimated would be:

$$Y_{it} = a_i + t + \theta Y_{it-h} + \rho D_{it} + X_{it}\delta + \varepsilon_{it}$$

Y_{it-h} could be a vector that includes values form a multi-period in the past. One potential pitfall from using past values of the dependent variable is that the necessary conditions for estimating consistent estimates of the treatment effect, β, are now different. This can be seen if we take first differences of the equation above:

$$\Delta Y_{it} = \theta \Delta Y_{it-1} + \Delta t + \rho \Delta D_{it} + \delta \Delta X_{it} + \Delta \varepsilon_{it}$$

Nickell (1981) showed that the lagged dependent variable, ΔY_{it-1} is now correlated to the $\Delta \varepsilon_{it}$, and OLS do not consistently estimate the regression parameters. A solution to this issue would be to use IVs. For example, Y_{it-2} could be used as an instrument for the ΔY_{it-1}. However, we still need to make the strong assumption that both Y_{it-2} and $\Delta \varepsilon_{it}$ are uncorrelated which in many cases is unplausible. As a rule of thumb, we can estimate both equations an if both get similar results then our results are consistent.

As stated above, panel data and more specifically fixed effects will be relevant in the EVIDENT project. For example, using hourly data from CW it can be estimated whether COVID-19 has any impact on demand responses. During lockdowns, many people work from home and this on the one hand increases overall consumption, and on the other hand people are able to better schedule their energy consuming tasks and switch consumption to off-peak hours. Whether a consumer reacts by switching its consumption or not may be related to behavioural inattention. Fixed effects are used in the preliminary analysis for indicating any consumption patterns on a daily, day of the week, and a monthly basis. Although this analysis will be presented in more details in D4.2, we include below three indicative graphs.

Figure 1 presents the hour of the day pattern and was estimated using fixed effects.

Figure 1: Hour of day consumption - Fixed effects

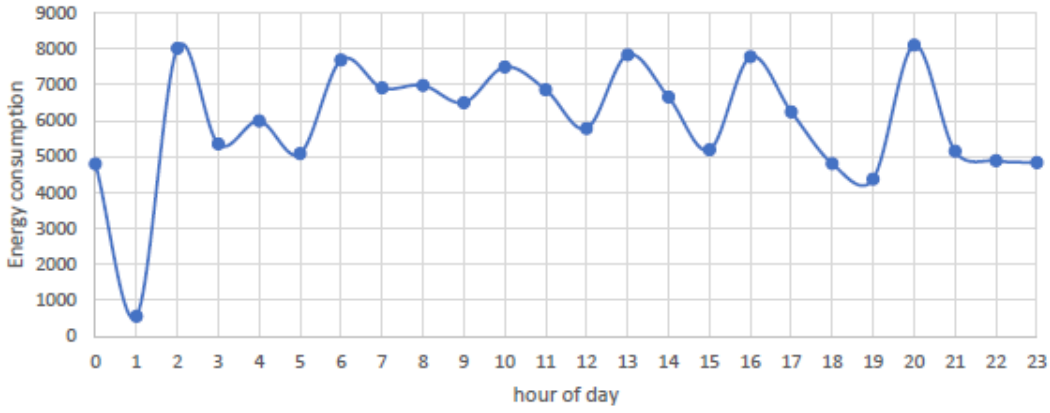


Figure 2 presents the day of the week consumption using fixed effects. Number 1 in the horizontal axis denotes “Monday”, number 2 denotes “Tuesday”, number 3 denotes “Wednesday” and so forth.

Figure 2: Day of the week consumption - Fixed effects

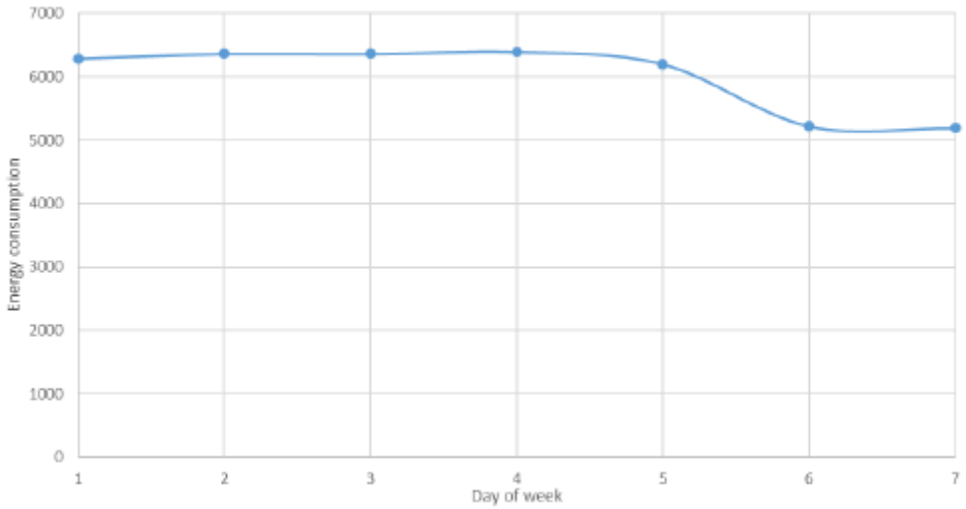
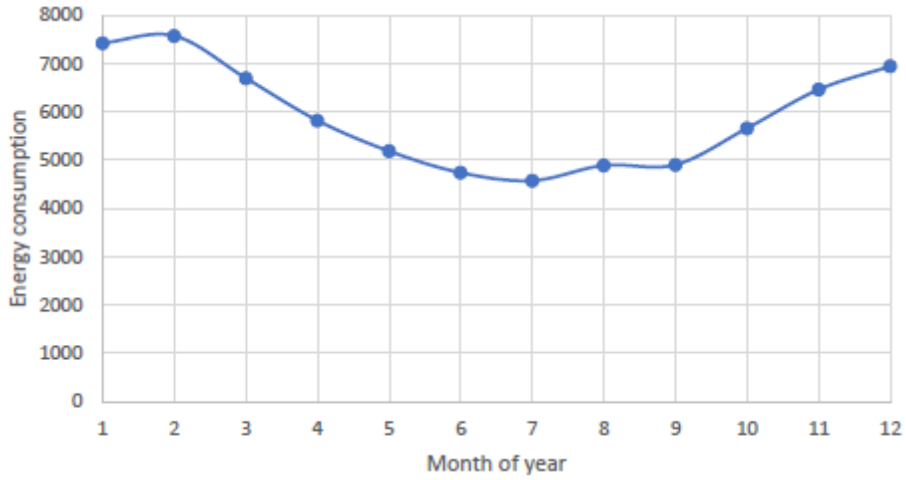


Figure 3 presents the month of year consumption using fixed effects. Number 1 in the horizontal axis denotes “January”, number 2 denotes “February”, number 3 denotes “March” and so forth.

Figure 3: Month of year consumption - Fixed effects



4.3 First differencing

An alternative to deviations from means method is FD. Using FD, we can eliminate the unobserved fixed effect A_i. With differencing the impact effect is estimated as the parameter ρ in the equation below:

$$\Delta Y_{it} = \Delta t + \rho \Delta D_{it} + \delta \Delta X_{it} + \Delta \varepsilon_{it} \quad (41)$$

, where $\Delta Y_{it} = Y_{it} - Y_{it-1}$, $\Delta D_{it} = D_{it} - D_{it-1}$, $\Delta X_{it} = X_{it} - X_{it-1}$, and $\Delta \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{it-1}$. In differencing we lose one observation for each cross section, the first time period and we have overall $T-1$ observations rather than T as in the case of the fixed effects methods. FD can also be estimating using OLS. One reason to prefer first differencing to the fixed effect estimation is that FD is easier to implement. From a statistical perspective, a fixed effect estimator is asymptotically more efficient in the class of estimators under the exogeneity assumption.

Furthermore, remember that a key assumption for the fixed effect estimator to be efficient is $E(\varepsilon_i \varepsilon_{it}' | X_i, a_i) = \sigma_\varepsilon^2 I_T$. It assumes homoskedasticity and no serial correlation in ε_{it}, However, assuming that error terms are not serially correlated may be a too strong assumption. An alternative assumption is that the first differences of the idiosyncratic errors are serially uncorrelated and have constant variance (Wooldridge, 2010):

$$E\left(\varepsilon_i \varepsilon_{it}^{'} | X_i, a_i\right) = \sigma_\varepsilon^2 I_{T-1}$$

Under this type of homoscedastic assumption, it can be shown that the FD estimator is most efficient in the class of estimators using the strict exogeneity assumption.

4.4 Difference-in-differences

Another method that is based on a combination of before-after and treatment-control group comparisons, and is related to the fixed effects method, is the DiD method (Donald and Lang, 2007, Abadie, 2005, Athey and Imbens, 2006, Grima and Görg, 2007, Conley and Taber, 2011, and Goodman-Bacon, 2021). DiD is mostly used when treatment assignment is non-random, and when regressors of interest vary only at a more aggregate level such as state or any other type of cohort. For example, what is the impact of a new government regulation, or an introduction of a new law on energy consumption? DiD in method has been used in many applications in economics with labour economists being the first to apply it (Card and Krueger, 2003, Angrist and Krueger, 1999, Bertrand et al., 2004, and Wing and Bello-Gomez, 2018).

Since not all potential outcomes are observable, only a signle outcome from one state for each unit (Y_{0ist}, Y_{1ist}) is observed. The DiD method assumes an additive structure for potential outcomes in the no-treatment group for each group:

$$E(Y_{0ist} | s, t) = \gamma_s + \lambda_t$$

, where, the numbers 0, and 1 denote treatment, s denotes the group, i and t, denote the unit and time respectively. The above expectation implies that the outcome in the control group, is determined by the sum of a time-invariant group effect (γ_s), and a time effect that is common across all groups (λ_t).

Regarding the outcome in the treatment group for each group, we expect that will be determined as in the case of the control group plus an impact effect (βD_{st}) and its expectation will be given as:

$$E(Y_{1ist} | s, t) = \gamma_s + \lambda_t + \beta D_{st}$$

Now, the population DiD estimands is given by:

$$E(Y_{ist}|s = control, t = after) - E(Y_{ist}|s = control, t = before)$$

$$- E(Y_{ist}|s = treatment, t = after) - E(Y_{ist}|s = treatment, t = before) = \beta$$

Given the group and the time the sample estimable counterpart of the above is:

$$Y_{ist} = \gamma_s + \lambda_t + \beta D_{st} + \varepsilon_{ist} \quad (42)$$

, where, $Y_{ist} = E(Y_{1ist} - Y_{0ist}|s, t)$. The expected values of Y_{ist} is estimated by the sample average $\bar{Y}_{is} = T^{-1}\sum_{t=1}^{T} Y_{ist}$. If data from multiple time periods are available, then can be used to estimate the common trends assumption.

We can use regression to estimate eq. 42 and would take the following form:

$$Y_{ist} = a + \gamma S + \lambda d_t + \beta(S \cdot d_t) + \varepsilon_{ist} \quad (43)$$

, where S is a dummy for the treatment group a unit belongs to, and d_t is a time-dummy that switches on for observations that obtained after the intervention was initiated, and the parameter β of the interaction term $S \cdot d_t$, indicates the impact effect of the treatment. This regression formulation in eq. 43 offers a useful way for estimating both treatment effects and standard errors for performing inference analysis.

The DiD method can be extended in a more general framework than the 2x2 case reported here. If several treatments and control groups are available, the analysis remain the same, however, now the γ_s indicates fixed effects for each different group, while λ_t stands for time effects for each control group.

The usefulness of the DiD method is based on the key assumption of the parallel trends, which states that the outcome variable of interest in the treatment group, would have followed the same time trend as the control group if no treatment had taken place. This means that, although observable and unobservable factors may affect the level of the outcome however this difference must be constant over time. As the potential outcome for each group is only observed in one state (treatment or control), the assumption is fundamentally untestable. Year shocks can have a negative effect on DiD models, since state-and time-specific random effects generate a clustering problem that affects statistical inference. This problem is often called serial correlation and is often apparent in time series data. Using data from previous periods, we can test whether the assumption holds or not and correct any inconsistency induces by random shocks.

5. Other types of regression

5.1 Regression discontinuity designs

In many cases, the treatment assignment could be a function of a variable, x_i. Imagine there is a threshold and when the variable takes on values below or above it, then a treatment is being assigned to a group of units. These types of treatment effects can be exploited effectively using regression discontinuity designs (RD). More formally, sharp regression discontinuity is used in cases when treatment assignment is a deterministic function of a variable, x_i (Imbens and Lemieux, 2008, Cook, 2008, Lee and Lemieux, 2010, Cook and Wong, 2008, Calonico et al., 2014, Skovron and TiTiunik, 2015, Kolesár, and Rothe, 2018 and Gelman and Imbens, 2019):

$$D_i = \begin{cases} 1 & if\ x_i \geq x_0 \\ 0 & if\ x_i < x_0 \end{cases}$$

, where x_0 is a known threshold. Thus, once the value of x_i is known, the treatment assignment to a unit is also known. To formalise the RD method, we use the following model from Angrist and Pischke (2008):

$$E[Y_{0i}|x_i] = a + \beta x_i$$

$$Y_{1i} = Y_{0i} + \rho$$

That leads to the following regression:

$$Y_i = a + \beta x_i + \rho D_i + \eta_i \quad (44)$$

ρ, is the causal effect of interest. Several assumptions hold regarding the derivation of eq. 44 and for estimating consistent parameters. First, it is assumed that the effect is additive to the control group. Second, the trend relation of the outcome variable with the variable x_i, $E[Y_{0i}|x_i]$ should be smooth however the relation could be linear as in the case of eq. 44 or nonlinear. In the nonlinear case, suppose that:

$$E[Y_{0i}|x_i] = f(x_i)$$

which means that the outcome variable is a smooth function of x_i. Now, the RD function to be estimated is:

$$Y_i = f(x_i) + \rho D_i + \eta_i \quad (45)$$

where

$$D_i = \begin{cases} 1 & if\ x_i \geq x_0 \\ 0 & if\ x_i < x_0 \end{cases}$$

The key assumption here is that $f(x_i)$ is still a continuous function in the neighbourhood of x_0. For modelling $f(x_i)$, Angrist and Pischke (2008), propose a $p^{th} - order$ polynomial and thus the RD estimated can be estimated from the regression below:

$$Y_i = a + \beta_1 x_i + \beta_2 x_i^2 + \cdots \ldots \ldots \ldots + \beta_\rho x_i^\rho + \rho D_i + \eta_i \quad (46)$$

Alternatively, we can estimate a different specification that is based on a normalisation of x_i, by differencing x_i from its threshold value, x_o:

$$\tilde{x}_i \equiv x_i - x_0$$

Now, the regression to be estimated takes the following form:

$$Y_i = a + \beta_1 \tilde{x}_i + \beta_2 \tilde{x}_i^2 + \cdots \ldots \ldots \ldots + \beta_\rho \tilde{x}_i^\rho + \rho D_i + \beta_{\rho+1} D_i \tilde{x}_i + \cdots \ldots \ldots \ldots \ldots \ldots + \beta_{\rho+\rho} D_i \tilde{x}_i^\rho + \eta_i \quad (47)$$

The model in eq. 47 has the advantage that it imposes no restrictions on the underlying conditional mean functions. It is important to note, that the validity of either eq. 46 or 46 estimates, is based on whether polynomial models adequately describe $E[Y_{0i}|x_i]$. If this not the case, then the estimated parameters of the treatment effect might be estimation of an unaccounted-for nonlinearity mean function.

In the EVIDENT project, RD methods could be used for estimating the price elasticity of energy consumption. Supposing there is a cut-off value in energy consumption (measured in KWh) above which the electricity price increases. Then, around these cut-off values, RD estimations could estimate whether there are any bouncing or discontinuity effects. The presence of such effects could be used as an indicator for electricity's price elasticity.

5.2 Quantile regression

In most cases, applied economists are concerned about mean effects. If for example, a consumer receives a behavioural nudge, we implicitly assume that the treatment effect will describe the entire distribution. However, the effect might be different for consumers being in the upper quantile than for consumers being in the lower quantile, in terms for prior energy consumption. Policy-makers and economists have been especially concerned with potential treatment effects across the distribution rather than the average itself. Furthermore, quantile regression is often preferred to average regression to reduce susceptibility to outliers.

We referred to quantile estimates in sections 4.3.1 and 4.4.2, however, in this section we will present more formally estimators and assumptions that are based on. More importantly, we will be interested for conditional and unconditional quantile treatment effects (QTE), assuming that the treatment is either exogenous or endogenous (Hao et al., 2007, Koenker and Bassett, 1978, Abadie, Angrist and Imbens, 2002, Machado and Mata, 2005, Chernozhukov and Hansen 2005, Firpo, 2007, and Frölich, and Melly, 2008).

The estimation of conditional QTEs with exogenous treatment, requires the assumption of a linear model for potential outcomes and covariates being conditional on some confounders:

$$Y_i^d = X_i\beta^\tau + D\delta^\tau + \varepsilon_i$$

, D is a binary treatment effect, X_i is the covariates and β^τ and δ^τ are the unknown parameters of the model. More specifically, δ^τ representes the conditional quantile treatment effects at quantile τ. Assuming the treatment to be exogenous means that it is independent of the potential outcomes and the covariates. Under these assumptions, the unknown coefficients can be estimated by the classical quantile regression estimator proposed be Koenker and Bassett (1978), defined by:

$$(\hat{\beta}^\tau, \hat{\delta}^\tau) = argmin \sum \rho_t(Y_i^d - X_i\beta^\tau + D\delta^\tau) \quad (48)$$

, where $\rho_t(u) = \tau \max(u, 0) + (1-\tau)\max(-u, 0)$ is a quantile loss function. In this case, u is the error of a single data point and the $\max$ function returns the largest value in the parentheses. This means that if the error is positive, then the check function multiples the error by τ, and if the error is negative, then the check function multiplies the error by $(1-\tau)$.

In many applications, the treatment assignment, D, is endogenously self-selected. As was described in subsection 4.4.3, for the IVs, in such cases we need to use an IV identification strategy to recover the true effects. Assuming that an IV, Z, exists and further assuming that:

$$0 < \Pr(Z = 1|X) < 1$$
$$E(D_1|X) \neq E(D_0|X)$$
$$\Pr(D_1 \geq D_0|X) = 1$$

The assumptions above, in addition to a conditional independence assumption on the IV, indicate that the treatment effect can be identified only for the compliers group. That is, for participants that comply with their treatment assignment. The estimated coefficients for the quantile treatment effect are given by the following:

$$(\hat{\beta}^\tau, \hat{\delta}^\tau) = argmin \sum W_i \cdot \rho_t(Y_i^d - X_i\beta^\tau + D\delta^\tau)$$

, where W_i is a weight:

$$W_i = 1 - \frac{D_1(1-Z_i)}{1-\Pr(Z=1|X_i)} - \frac{(1-D_i)Z_i}{\Pr(Z=1|X_i)} \quad (49)$$

From the above remark, it is obvious that a preliminary estimator for $\Pr(Z = 1|X_i)$ is needed as a first step. $\Pr(Z = 1|X_i)$ is a propensity score as it was discussed in subsection 4.4.2. A Logit or a Probit model can be used to estimate the propensity score. These estimation methods will be discussed in the next subsection.

Unconditional quantile treatment effects that provide estimation for the population of interest can be considered rather than confining the results as a function of the observables. Furthermore, we have an advantage using unconditional QTEs, since we can include covariates that are independent from the treatment in the estimation, without changing the limit of the estimated QTEs. Finally, using unconditional QTEs we don't need to make any parametric restrictions. We need to have in mind that the interpretation of the unconditional effects is slightly different than before. For instance, if we are interested in consumers being in the low quantile of energy consumption, the unconditional QTE will summarise the effect with a relatively low absolute energy consumption. The bivariate quantile regressor estimator for the unconditional QTE with endogenous treatment is given by the following:

$$(\hat{\beta}^{\tau}, \hat{\delta}^{\tau}) = argmin \sum W_i \cdot \rho_t(Y_i^d - X_i\beta^{\tau} + D\delta^{\tau})$$

, where as before, W_i is a weight:

$$W_i = \frac{z_i - \Pr(Z = 1|X_i)}{\Pr(Z = 1|X_i)\{1 - \Pr(Z = 1|X_i)\}}(2D_i - 1) \quad (50)$$

The notation is the same as before. Note that the weight equation is now different than in eq. 49.

Finally, we consider the case where the treatment is exogenous conditional on X. We additionally assume that:

$$0 < \Pr(D = 1|X) < 1$$

The bivariate quantile regressor estimator for the unconditional QTE with exogenous treatment is given by the following:

$$(\hat{\beta}^{\tau}, \hat{\delta}^{\tau}) = argmin \sum W_i \cdot \rho_t(Y_i^d - X_i\beta^{\tau} + D\delta^{\tau})$$

, where as before, W_i is a weight:

$$W_i = \frac{D_i}{\Pr(D = 1|X_i)} + \frac{1 - D_i}{1 - \Pr(D = 1|X_i)} \quad (51)$$

The weight eq. 51 is traditional propensity-score weighting estimator, also known as inverse probability weighting, and a preliminary estimator for $\Pr(D = 1|X_i)$ is needed to estimate.

5.3 Discrete regression models

In this subsection, we discuss models that are used to estimate the effect on binary dependent variables. Many nonlinear econometric models are intended to explain limited depended variables. For instance, assume that the variable to be explained is discrete and takes on a finite number of values, for example, in the case of the propensity score. As previously stated, the propensity score indicates the probability of a unit to receive the treatment, $D_i = 1$, thus the dependent variable takes on only two values, zero and one.

A simple linear regression of y on x is not appropriate, since among other things, the implied model of the conditional mean assumes tight restrictions on the residuals of the model. Furthermore, the fitted value of the dependent variable, y, from a simple linear regression is not restricted to lie between zero and one. Instead, we use binary response models, probit and logit of the form (Wooldridge, 2010, Newey, 1985, Greene, et. al., 1993, Chen and Khan, 2003, and Newey, 1987):

$$P(y = 1|x) = G(x\beta) \equiv p(x)$$

, where $G(z) = z$, which takes on values in the open unit interval $< 0G(z) < 1$, is the identity function. Note, that the response probabilities cannot be between zero and one for all values of x. Using binary response models, we estimate the effect of the exogenous variables (covariates) on the response probability $P(y = 1|x)$.

Next, we provide a formal representation of the most commonly used binary models in economics, the probit and the logit models that differ only on the form of the identity function, $G(z)$. In the probit model, the identity function, $G(z)$, is as follows:

$$G(z) \equiv \Phi(z) \equiv \int_{\infty}^{z} \varphi(v)d(v) \qquad (52)$$

, where $\Phi(z)$ is the standard normal density function:

$$\Phi(z) = (2\pi)^{-1/2}\exp\left(\frac{-z^2}{2}\right)$$

In the logit model the identity function, $G(z)$, is as follows:

$$G(z) = \Lambda(z) \equiv \exp(z)/[1 + \exp(z)] \qquad (53)$$

Given both probit and logit specifications, we can estimate the parameters of this model using the method of conditional maximum likelihood. The likelihood function, with N independent, identically distributed observations, is given by:

$$l_i = \sum_{i=1}^{N} y_i \log[G(x\beta)] + (1 - y_i)\log\,[1 - G(x\beta)] \quad (54)$$

From the general maximum likelihood results, we know that estimators are consistent and asymptotically normal. The sample variance is given by the following form:

$$\hat{var} = \left\{ \sum_{i=1}^{N} \frac{[g(x\beta)^2 x'x}{G(x\beta)[1 - G(x\beta)]} \right\}^{-1}$$

, where

$$g(x) \equiv \frac{dG}{dx}(x)$$

The identity function, $G(z)$, does not have to be a cumulative distribution function, rather it can take any form as long as it takes values between zero and one for all real numbers z. For successfully applying the Probit and the Logit model, it is important to know how to interpret the estimated parameters β_i, on both continuous and discrete explanatory variables. If the explanatory variables take on continuous values the effect is given by:

$$\frac{\partial p(x)}{\partial x_i} = g(x\beta)\beta_i$$

Therefore, the sign and the size of the partial effect is given by the sign and the size of the estimated parameter β_i. Also, the relative effects do not depend on the explanatory variables, x, as seen below:

$$\frac{\frac{\partial p(x)}{\partial x_i}}{\frac{\partial p(x)}{\partial x_j}} = \frac{\beta_i}{\beta_j}$$

If the explanatory variable of interest is binary, then the partial effect from changing the explanatory variable from zero to one is given by:

$$G(\beta_1 + \beta_2 x_2 + \cdots \ldots \ldots + \beta_{k-1} x_{k-1} + \beta_\kappa) - (\beta_1 + \beta_2 x_2 + \cdots \ldots \ldots + \beta_{k-1} x_{k-1})$$

Knowing the sign and the size of β_κ is enough to determine the impact of the binary dependent variable on the explanatory variable.

6. Conclusion

The ultimate goal for econometric and statistical analysis in experimental and quasi-experimental studies is first to achieve internal validity and then to provide estimates that also have external validity. As it was shown at the beginning of this document, external validity is rather important for policy making since it validates results for larger scale interventions. The methods discussed include statistical methods for analysing data from randomised experiments, with a primary focus on randomisation- and model-based methods. We started our analysis with classic methods developed by Fisher and Neyman, up to recent work in panel data and quantile regressions and the convergence of the statistical and econometric literature, with the Rubin potential outcomes framework. We stressed the importance of estimating heterogeneous effects and the effect on the distribution of the dependent variable rather than the effects on its mean. Methods were presented in a simple format, abstracting from rather technical issues, and focusing, where possible, on the choice of the most suitable case for each case.

A general takeaway would be, researchers in analysing experimental data prefer stratified samples as it may make analysis more informative and can construct test statistics that may have more power than statistics that do not depend on strata. Similarly, by estimating average treatment effects within strata, and then averaging up the estimates appropriately, the results will be more precise if the covariates, that used to form the strata, are sufficiently strongly correlated with the potential outcomes. Finally, if randomisation was compromised, adjusting for covariance differences, through strata, may remove biases. As a final note, research in practice shows that in most cases, carefully applied to coherent causal questions, regression and 2 stage least squares almost always make sense.

References

[1] Abadie, A., Angrist, J. and Imbens, G., 2002. IVs estimate of the effect of subsidized training on the quantiles of trainee earnings. Econometrica, 70(1), pp.91-117.
[2] Abadie, A., 2005. Semiparametric difference-in-differences estimators. The Review of Economic Studies, 72(1), pp.1-19.
[3] Angrist, J.D. and Krueger, A.B., 1999. Empirical strategies in labour economics. In Handbook of labour economics (Vol. 3, pp. 1277-1366). Elsevier.
[4] Angrist, J.D. and Pischke, J.S., 2008. Mostly harmless econometrics. Princeton university press.
[5] Ashenfelter, O. and Rouse, C., 1998. Income, schooling, and ability: Evidence from a new sample of identical twins. The Quarterly Journal of Economics, 113(1), pp.253-284.
[6] Athey, S. and Imbens, G.W., 2006. Identification and inference in nonlinear difference-in-differences models. Econometrica, 74(2), pp.431-497.
[7] Athey, S. and Imbens, G.W., 2017. The econometrics of randomised experiments. In Handbook of economic field experiments (Vol. 1, pp. 73-140). North-Holland.
[8] Bai, J., 2009. Panel data models with interactive fixed effects. Econometrica, 77(4), pp.1229-1279.
[9] Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M. and Walton, M., 2017. From proof of concept to scalable policies: Challenges and solutions, with an application. Journal of Economic Perspectives, 31(4), pp.73-102.
[10] Baum, C.F., Schaffer, M.E. and Stillman, S., 2003. IVs and GMM: Estimation and testing. The Stata Journal, 3(1), pp.1-31.
[11] Bertrand, M., Duflo, E. and Mullainathan, S., 2004. How much should we trust differences-in-differences estimates? The Quarterly journal of economics, 119(1), pp.249-275.
[12] Bohnet, I. and Frey, B.S., 1999. The sound of silence in prisoner's dilemma and dictator games. Journal of economic behavior & organization, 38(1), pp.43-57.
[13] Buchinsky, M., 1994. Changes in the US wage structure 1963-1987: Application of quantile regression. Econometrica: Journal of the Econometric Society, pp.405-458.
[14] Calonico, S., Cattaneo, M.D. and Titiunik, R., 2014. Robust nonparametric confidence intervals for regression-discontinuity designs. Econometrica, 82(6), pp.2295-2326.
[15] Card, D. and Krueger, A.B., 1993. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania.
[16] Casari, M. and Cason, T.N., 2009. The strategy method lowers measured trustworthy behavior. Economics Letters, 103(3), pp.157-159.
[17] Chernozhukov, V. and Hansen, C., 2005. An IV model of quantile treatment effects. Econometrica, 73(1), pp.245-261.
[18] Chen, S. and Khan, S., 2003. Rates of convergence for estimating regression coefficients in heteroskedastic discrete response models. Journal of Econometrics, 117(2), pp.245-278.
[19] Conley, T.G. and Taber, C.R., 2011. Inference with “difference in differences” with a small number of policy changes. The Review of Economics and Statistics, 93(1), pp.113-125.
[20] Donald, S.G. and Lang, K., 2007. Inference with difference-in-differences and other panel data. The review of Economics and Statistics, 89(2), pp.221-233.
[21] Duflo, E., Glennerster, R. and Kremer, M., 2007. Using randomization in development economics research: A toolkit. Handbook of development economics, 4, pp.3895-3962.
[22] Gigerenzer, G., 2004. Mindless statistics. The Journal of Socio-Economics, 33(5), pp.587-606.

[23]Girma, S. and Görg, H., 2007. Evaluating the foreign ownership wage premium using a difference-in-differences matching approach. Journal of International Economics, 72(1), pp.97-112.
[24]Cochran, W.G. and Chambers, S.P., 1965. The planning of observational studies of human populations. Journal of the Royal Statistical Society. Series A (General), 128(2), pp.234-266.
[25]Cook, T.D., Campbell, D.T. and Shadish, W., 2002. Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.
[26]Cook, T.D., 2008. "Waiting for life to arrive": a history of the regression-discontinuity design in psychology, statistics and economics. Journal of Econometrics, 142(2), pp.636-654.
[27]Cook, T.D. and Wong, V.C., 2008. Empirical tests of the validity of the regression discontinuity design. Annales d'Economie et de Statistique, pp.127-150.
[28]Deaton, A., 2010. Instruments, randomization, and learning about development. Journal of economic literature, 48(2), pp.424-55.
[29]Freedman, D.A., 2008. On regression adjustments to experimental data. Advances in Applied Mathematics, 40(2), pp.180-193.
[30]Firpo, S., 2007. Efficient semiparametric estimation of quantile treatment effects. Econometrica, 75(1), pp.259-276.
[31]Fisher, R. A. (1935), Design of Experiments, Oliver and Boyd.
[32]Freeman, R.B., 1984. Longitudinal analyses of the effects of trade unions. Journal of Labour Economics, 2(1), pp.1-26.
[33]Frölich, M. and Melly, B., 2008. Unconditional quantile treatment effects under endogeneity.
[34]Gelman, A. and Imbens, G., 2019. Why high-order polynomials should not be used in regression discontinuity designs. Journal of Business & Economic Statistics, 37(3), pp.447-456.
[35]Greene, W.H., Greene, L.K. and Seaks, T.G., 1995. Estimating the functional form of the independent variables in probit models. Applied Economics, 27(2), pp.193-196.
[36]Goodman-Bacon, A., 2021. Difference-in-differences with variation in treatment timing. Journal of Econometrics.
[37]Hao, L., Naiman, D.Q. and Naiman, D.Q., 2007. Quantile regression (No. 149). Sage.
[38]Heckman, J., 1992. Randomization and Social program. Evaluating welfare and training programs, pp.201-230.
[39]Heckman, J.J., Ichimura, H. and Todd, P.E., 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. The review of economic studies, 64(4), pp.605-654.
[40]Heckman, J.J. and Vytlacil, E.J., 2007. Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. Handbook of econometrics, 6, pp.4779-4874.
[41]Hirano, K., Imbens, G.W. and Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica, 71(4), pp.1161-1189.
[42]Holland, P.W., 1986. Statistics and causal inference. Journal of the American statistical Association, 81(396), pp.945-960.
[43]Hotz, V.J., Imbens, G.W. and Mortimer, J.H., 2005. Predicting the efficacy of future training programs using past experiences at other locations. Journal of Econometrics, 125(1-2), pp.241-270.
[44]Imbens, G.W. and Lemieux, T., 2008. Regression discontinuity designs: A guide to practice. Journal of Econometrics, 142(2), pp.615-635.
[45]Imbens, G.W. and Rubin, D.B., 2015. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

[46]Kaplan, A. and Wolf, J.C., 2017. The conduct of inquiry: Methodology for behavioral science. Routledge.
[47]Koenker, R. and Bassett Jr, G., 1978. Regression quantiles. Econometrica: journal of the Econometric Society, pp.33-50.
[48]Kolesár, M. and Rothe, C., 2018. Inference in regression discontinuity designs with a discrete running variable. American Economic Review, 108(8), pp.2277-2304.
[49]Lee, D.S. and Lemieux, T., 2010. Regression discontinuity designs in economics. Journal of economic literature, 48(2), pp.281-355.
[50]Levitt, S.D. and List, J.A., 2009. Field experiments in economics: The past, the present, and the future. European Economic Review, 53(1), pp.1-18.
[51]Machado, J.A. and Mata, J., 2005. Counterfactual decomposition of changes in wage distributions using quantile regression. Journal of applied Econometrics, 20(4), pp.445-465.
[52]Miguel, E., Satyanath, S. and Sergenti, E., 2004. Economic shocks and civil conflict: An IVs approach. Journal of political Economy, 112(4), pp.725-753.
[53]Neyman, J. (1923, 1990), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," translated in Statistical Science, (with discussion), Vol. 5(4): 465–480, 1990.
[54]Newey, W.K., 1985. Maximum likelihood specification testing and conditional moment tests. Econometrica: Journal of the Econometric Society, pp.1047-1070.
[55]Newey, W.K., 1987. Efficient estimation of limited dependent variable models with endogenous explanatory variables. Journal of econometrics, 36(3), pp.231-250.
[56]Nickell, S., 1981. Biases in dynamic models with fixed effects. Econometrica: Journal of the econometric society, pp.1417-1426.
[57]Phillips, P.C. and Hansen, B.E., 1990. Statistical inference in IVs regression with I (1) processes. The Review of Economic Studies, 57(1), pp.99-125.
[58]Plackett, R.L. and Burman, J.P., 1946. The design of optimum multifactorial experiments. Biometrika, 33(4), pp.305-325.
[59]Rosenbaum, P.R. and Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), pp.41-55.
[60]Rosenbaum, P.R., 2002. Overt bias in observational studies. In Observational studies (pp. 71-104). Springer, New York, NY.
[61]Rosenbaum, P.R., Rosenbaum, P.R. and Briskman, 2010. Design of observational studies (Vol. 10). New York: Springer.
[62]Rubin, D.B., 1975. Bayesian inference for causality: The importance of randomization. In The Proceedings of the social statistics section of the American Statistical Association (Vol. 233, p. 239). Alexandria, VA: American Statistical Association.
[63]Rubin, Donald B. "Causal inference using potential outcomes: Design, modeling, decisions." Journal of the American Statistical Association 100, no. 469 (2005): 322-331.
[64]Rubin, D.B., 2008. For objective causal inference, design trumps analysis. The Annals of Applied Statistics, 2(3), pp.808-840.
[65]Sargan, J.D., 1958. The estimation of economic relationships using IVs. Econometrica: Journal of the Econometric Society, pp.393-415.
[66]Skovron, C. and Titiunik, R., 2015. A practical guide to regression discontinuity designs in political science. American Journal of Political Science, 2015, pp.1-36.

[67]Tsoutsoura, M., 2015. The effect of succession taxes on family firm investment: Evidence from a natural experiment. The Journal of Finance, 70(2), pp.649-688.

[68]Wing, C., Simon, K. and Bello-Gomez, R.A., 2018. Designing difference in difference studies: best practices for public health policy research. Annual review of public health, 39.

[69]Winship, C. and Morgan, S.L., 1999. The estimation of causal effects from observational data. Annual review of sociology, 25(1), pp.659-706.

[70]Wooldridge, J.M., 2010. Econometric analysis of cross section and panel data. MIT press.