ERUDITE

ÉQUIPE DE RECHERCHE SUR L'UTILISATION DES DONNÉES INDIVIDUELLES EN LIEN AVEC LA THÉORIE ÉCONOMIQUE

Sous la co-tutelle de :
UNIVERSITÉ GUSTAVE EIFFEL
UPEC · UNIVERSITÉ PARIS-EST CRÉTEIL

Series of ERUDITE Working Papers

N° 01-2021

Title

Measuring the effect of health events in the labour market

Author

Emmanuel Duguet

Measuring the effect of health events in the labour market

Emmanuel Duguet[1,2]

[1]Université Paris Est Créteil, ERUDITE EA 437
[2]FR TEPP CNRS 2042

January 2021

Abstract

This paper presents a survey of the estimation methods available to measure the effect of a health events on an outcome variable. We review the most widespread methodology that the profession uses in order to evaluate the impact of health events. We consider both matching and differencing methods, with a focus on the panel data regressions which can be used in order to asses the effect of health on labour market outcomes.

JEL: : C18, C21, C25, J24, I10.
Keywords: labour, health, matching, differencing, treatment effects, panel data.

Introduction

Human capital fully relies on health (Grossman, 1972). Therefore it is important to evaluate the consequences of health variations on the labour market performances. For a long period of time, many labour market studies have ignored health issues so that health was the missing variable, a source of “unobservable" individual variations in the labour market variables. Recently more data have been made available. First, one can find rather easily survey data, with many variables on health and some on labour. Most of them are declarative. Second, more recently, administrative data have been made available to the researchers. They are not declarative and thus offer a better measurement of some diseases. They also have a systematic time dimension because they are collected by the administration for its own purpose (e.g., reimbursement of health expenditures.). This opens the way to panel data analysis. This chapter presents the most common methods used for evaluating the effect of a health event (illness, accident) on outcome variables, like participation in the labour market.

1 The Rubin approach

We wish to estimate the effect of a health event on a labour market outcome, using the Rubin causal model framework (Rubin, 1974). In statistical terms, the health event will in fact be called a “treatment". Since our treatment is dichotomous, the outcome can take on two values $Y(0)$ when no health event happened, and $Y(1)$ when it did. We observe the following value:

$$Y = (1-W) \times Y(0) + W \times Y(1) = \begin{cases} \text{Y(0)} & \text{if W=0} \\ \text{Y(1)} & \text{if W=1} \end{cases}$$

Therefore, we have an observational problem. We only observe one of the two potential outcomes at a time. From the data, we would like to infer one of the following quantities:

- $\text{E}(Y(1) - Y(0))$, the average treatment effect (ATE);
- $\text{E}(Y(1) - Y(0)|W = 1)$, the average effect of the treatment on the treated (ATT). It is the treatment effect evaluation, since it refers to the population who was actually treated. It is also the most commonly used.
- $\text{E}(Y(1) - Y(0)|W = 0)$, the average effect of the treatment on the not-treated (ATN). This effect represents a prospective evaluation, since it tries to evaluate what would have happened to the population who did not get the treatment.

There are many ways to estimate these effects. We will consider the most widespread methods in this chapter.

Originally, the goal was to estimate the effect of a treatment on a given population. Ideally, we would like to have experimental data. The people would be given the treatment at random. The people in the treatment group would be compared to the people in the placebo group or control group. The randomness of the selection is here to ensure that, on a large enough sample, the distribution of both observable an unobservable characteristics are similar in the two samples. There would remain two sources of differences: treatment and outcomes.[1]

[1] On the method adapted to randomized experiments, see the part II of Imbens and Rubin (2015).

Unfortunately, it is not always possible to collect experimental data in the social sciences. Consider health events, like accidents or chronic illnesses. The social science researcher can obviously not deliberately cause a car accident to someone or inoculate a disease to a given worker. There is no other choice than to work with observational data. More and more, health insurance data includes decades of information about large samples of individuals. We will observe a health event and labour market outcomes at the same time.

Unfortunately, we cannot compare the people with and without accidents directly, at least because of differences in the *confounding variables*. These variables are related both to the health events and to the labour market outcomes. Consider gender: women and men do not have the same illnesses and do not have the same participation in the labour market when they are not ill. By comparing two data sets with and without health events, the outcome variables may differ because these data sets have different proportions of women, and not only because of differences in health status. Two other typical confounding variables are the age and the education level. Some disease, like cancers, will occur more often in an older population and the education level is also associated with different behaviours towards risks. Therefore, when we compare the treated and the control groups, the differences may stem from both the confounding variables and the treatment.

A related issue is unobservable heterogeneity, which is considered when panel data are available. Non observable characteristics, like genetics or the childhood living conditions, may be related to the occurrence of some diseases or accidents and, at the same time, influence the outcome in the labour market. If the sample really is random, these characteristics should be balanced in the treated and control groups. Otherwise, bias reducing techniques should be used.

Consider the following mean difference, which has an observable counterpart:

$$\underbrace{\mathrm{E}(Y(1)|W=1)-\mathrm{E}(Y(0)|W=0)}_{\text{Average outcome difference}} = \underbrace{\mathrm{E}(Y(1)|W=1)-\mathrm{E}(Y(0)|W=1)}_{\text{ATT}} + \underbrace{\mathrm{E}(Y(0)|W=1)-\mathrm{E}(Y(0)|W=0)}_{\text{Selection bias}}$$

The left hand gives the mean outcome difference in the two groups. The right-hand gives the sum of two terms. Firstly, what we are looking for, the ATT, and, secondly, an additional term which is the mean outcome difference in the absence of treatment $Y(0)$ between the treated group $W=1$ and the control group $W=0$. This term comes from a group selection difference (treated or not) and is termed a selection bias, because it impeaches to use the mean difference as a valid ATT estimator from observational data. With experimental data, we would not have this problem because, by definition, the distribution of the outcomes would be the same in the two groups, so that the selection bias cancels. We could use the outcomes of the not-treated to estimate what would have happened to the treated if they had not been treated. Statistically, it corresponds to the assumption:

$$Y(0) \perp\!\!\!\perp W.$$

This assumption may be too strong for observational data, we need a weaker assumption. When observational data are available, it is still possible to estimate the effect of the treatment on the treated (ATT) with the following conditional independence assumption:

$$\text{CIA-0: } Y(0) \perp\!\!\!\perp W|X \tag{1}$$

This assumption means that, given X, the people in the two groups would reach the same outcome in the absence of treatment $Y(0)$. Consider participation in the labour market so that X includes age, gender and the education level. People with the same age, gender and education should reach comparable outcomes on average in the labour market in the absence of treatment (here, in the absence of disease). When this assumption holds, it is possible to use the data about the people in the control group for inferring what would have happened to the treated in the absence of treatment. We do this because it is not possible to observe the outcome of the treated when they are not treated. The credibility of the CIA-0 depends on the data available. The basic ingredient of our estimators will be the conditional mean outcome difference:

$$\begin{aligned} c(x) &= \mathrm{E}(Y(1)|W=1, X=x) - \mathrm{E}(Y(0)|W=0, X=x) \\ &= \mathrm{E}(Y|W=1, X=x) - \mathrm{E}(Y|W=0, X=x)\,. \end{aligned}$$

The second line in the previous expression states that $c(x)$ has an empirical counterpart, since Y is always observable. Now, consider the estimation of the ATT. We let:

$$c_1(x) = \mathrm{E}(Y(1) - Y(0)|W=1, X=x)$$

and we can rewrite the ATT:

$$c_1 = \mathrm{E}(c_1(X)) = \int \mathrm{E}(c_1(x))\,\mathrm{dP}_{X|W=1}(x).$$

where P denotes the cdf of $X|W=1$. When X is continuous we get:

$$\int c_1(x)\mathrm{p}(x|W=1)\mathrm{d}x$$

where p is the density of $X|W=1$, and when X is discrete, we get:

$$\sum_x c_1(x)\Pr(x|W=1),$$

where $\Pr(x|W=1)$ is the probability that $X=x$ among the treated. This simply means that one can compute the ATT in two steps. First, we compute the average outcome differences $c(x)$ for each group of individuals $X=x$. Second, we average the averages according to the distribution of X *in the treated group* $W=1$. We can write:

$$\begin{aligned} c_1 &= \mathrm{E}(c_1(X)) \\ &= \mathrm{E}(\mathrm{E}(Y(1)|X, W=1) - \mathrm{E}(Y(0)|X, W=1))\,, \end{aligned}$$

assuming that CIA-0 holds:

$$\mathrm{E}(Y(0)|X, W=1) = \mathrm{E}(Y(0)|X, W=0) \Rightarrow c_1(x) = c(x), \forall x$$

and we get the following expression, which has an empirical counterpart:

$$c_1 = \mathrm{E}(c(X))\,.$$

Overall, the CIA-0 assumption opens the possibility to use some outcomes of the control group in order to estimate the average outcome of the treated if they had not been treated. This will justify the following method: set a value of $X=x$, compute the average of $Y(1)$ in the treated group with $X=x$, the average of $Y(0)$ in the control group

with the same value $X = x$ and take their difference $c(x)$. Then aggregate these differences according to the distribution of X *in the treatment group* in order to get the ATT.

This method implies the following additional assumption: there must be both treated and controls for *all* the values of $X = x$. This is a common support hypothesis. Notice that it can fail if some people are excluded from the treatment group for specific values of X.

The estimation of the ATN is similar to the ATT. One just need to define the absence of treatment as the treatment and proceed as before. The ATN is defined by:

$$c_0 = \mathrm{E}(c_0(X))$$

with

$$c_0(x) = \mathrm{E}\left(Y(1) - Y(0)|W = 0, X = x\right)$$

The identifying assumption becomes:

$$\text{CIA-1: } Y(1) \perp\!\!\!\perp W|X \tag{2}$$

therefore, we will be able to use the data of the treated to estimate the outcome the controls would have had if they had been treated. Under CIA-1, we get the conditional effect $c_0(x) = c(x), \forall x$, so that:

$$c_0 = \mathrm{E}\left(c_0(X)\right) = \int c(x)\mathrm{dP}_{X|W=0}(x).$$

We also use $c(x)$ but with a different weighting distribution than in the ATT case. Also notice that we implicitly assume that we can find treated individuals with the same values of X than the not-treated, all along the distribution of $X|W = 0$. Last, we can estimate the Average Treatment Effect (ATE), c_2, by assuming:

$$\text{CIA-2: } Y(0), Y(1) \perp\!\!\!\perp W|X \tag{3}$$

It is the most demanding assumption and one may well question which of the three effects is the most relevant in order to avoid useless assumptions. The ATE uses the marginal distribution of X:

$$c_2 = \mathrm{E}\left(c(X)\right) = \int c(x)\mathrm{dP}_X(x).$$

2 Cross-section estimation

2.1 Matching

We will first consider multidimensional matching methods. Health and labour often include categorical variables and few continuous variables, so that we will consider exact matching and calliper matching. When these methods are not applicable, propensity score matching may be welcome.

Exact matching. Consider the estimation of the treatment effects. All the individuals in the data set have an identification variable. The levels of this variable allow to form index sets of treated and controls. $\mathbf{I}^W$ will denote the index set in the treatment group W, and I^W the number of elements in $\mathbf{I}^W$. By convention, we will let $W = 0$ denote the control group, $W = 1$ denote the treated group and $W = 2$ denote all the individuals in the data set ($W = 0$ and $W = 1$). Therefore we have $\mathbf{I}^2 = \mathbf{I}^0 \cup \mathbf{I}^1$ and $I^2 = I^0 + I^1$, because an individual cannot be in the treated group and in the control group at the same time. Each index set defines a treatment effect: the ATT relies on the distribution of X in $\mathbf{I}^1$, the ATN in $\mathbf{I}^0$ and the ATE in $\mathbf{I}^2$.

We take the estimation of the ATT as an example. The derivations of the other estimators are similar and we will just give the associated results. The treated are identified by their index $i \in \mathbf{I}^1$. For each value of i, we have twins in the control group, who share the same value of the matching variables. With exact matching, we can define the twins of the treated i by a matching function giving directly the index set of the twins $\mathbf{M}$. Let X_i be the matching variables for the treated i, the index set of the twins is denoted $\mathbf{M}(X_i)$ or, shortly $\mathbf{M}(i)$:

$$\mathbf{M}(i) = \{j \in \mathbf{I}^0 : X_j = X_i\},\ i \in \mathbf{I}^1$$

Similarly, we can apply this matching function to a whole set of individuals, so that $\mathbf{M}(I^1)$ is the twin set of all the treated. By convention, M will denote the number of elements in $\mathbf{M}$. The estimator of the ATT is defined by:

$$\hat{c}_1 = \frac{1}{I^1} \sum_{i \in \mathbf{I}^1} \left(Y_i - \frac{1}{M(i)} \sum_{j \in \mathbf{M}(i)} Y_j \right).$$

With a perfect matching, the values of X define groups of treated and controls and we can decompose the previous difference according to the values of X. With categorical data, there will be a finite number of groups. In order to use the CIA-0 condition, we will partition the treated set $\mathbf{I}^1$ according to the values of X, into G^1 groups. Matching is straightforward for categorical data. Each realisation of X can be interpreted as a sorting key that defines groups of data. Consider gender (2 levels, 1 and 2) and education (3 levels, 1, 2 and 3), we can define 6 groups from this data $\mathbf{G}^1 = \{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3)\}$. We can index the groups by $g \in \mathbf{G}^1$ and let $G^1 = 6$ denotes the number of groups. Inside each group g there are I^1_g treated and $M(\mathbf{I}^1_g)$ controls. We let:

$$\mathbf{I}^1 = \bigcup_{g \in \mathbf{G}^1} \mathbf{I}^1_g \text{ and } \mathbf{I}^1_g \cap \mathbf{I}^1_{g'} = \emptyset \ \ \forall g \neq g'$$

so that:

$$\hat{c}_1 = \frac{1}{I^1} \sum_{g \in \mathbf{G}^1} \sum_{i \in \mathbf{I}^1_g} \left(Y_i - \frac{1}{M(i)} \sum_{j \in \mathbf{M}(i)} Y_j \right)$$

The first term of the difference is proportional to the mean of the outcome variable in the treatment group g denoted ($\bar{Y}^1_g$):

$$\sum_{i \in \mathbf{I}^1_g} Y_i = I^1_g \times \frac{1}{I^1_g} \sum_{i \in \mathbf{I}^1_g} Y_i = I^1_g \times \bar{Y}^1_g.$$

For the second term, consider that inside each set $\mathbf{I}_g^1$ the control group $\mathbf{M}(i)$ does not depend on i because all the treated i have the same value of X and, therefore, exactly the same twins. We can set $\mathbf{M}(i) = \mathbf{M}(\mathbf{I}_g^1)\ \forall i \in \mathbf{I}_g^1$ and the second term in the sum is simply the mean of the outcome variable in the control group g, denoted $\bar{Y}_g^0$:

$$\frac{1}{M(i)} \sum_{j \in \mathbf{M}(i)} Y_j = \frac{1}{M(\mathbf{I}_g^1)} \sum_{j \in \mathbf{M}(\mathbf{I}_g^1)} Y_j = \bar{Y}_g^0$$

which implies that:

$$\sum_{g \in \mathbf{G}^1} \frac{1}{M(i)} \sum_{j \in \mathbf{M}(i)} Y_j = I_g^1 \times \bar{Y}_g^0$$

and we estimate the ATT by:

$$\begin{aligned} \hat{c}_1 &= \sum_{g \in \mathbf{G}^1} \frac{I_g^1}{I^1} \left(\bar{Y}_g^1 - \bar{Y}_g^0 \right) \\ &= \sum_{g \in \mathbf{G}^1} \omega_g^1 \left(\bar{Y}_g^1 - \bar{Y}_g^0 \right) \end{aligned}$$

with $\omega_g^1 = I_g^1 / I^1$. Notice that ω_g^1 is simply the proportion of treated that lies in the $\mathbf{I}_g^1$ set, that is the empirical distribution of $X|W = 1$. It is easily checked that $0 < \omega_g^1 < 1$ and $\sum_g \omega_g^1 = 1$.

This presentation strongly simplifies the computation of the variance for the following reasons (Barnay et al., 2019). First, by construction, there is no individual in common between different $\mathbf{I}_g^1$ sets and this property extends to the twins because the matching is perfect. Second, by definition, there is no intersection between the treated and the control sets (one individual is treated or not). This implies that:

$$\begin{aligned} \mathrm{V}(\hat{c}_1) &= \sum_{g \in \mathbf{G}^1} \mathrm{V}\left(\omega_g^1 \left(\bar{Y}_g^1 - \bar{Y}_g^0 \right) \right) \\ &= \sum_{g \in \mathbf{G}^1} (\omega_g^1)^2 \left(\mathrm{V}\left(\bar{Y}_g^1 \right) + \mathrm{V}\left(\bar{Y}_g^0 \right) \right) \end{aligned}$$

The same principles can be used for the two other effects. The ATN is identified under the conditional independence assumptions CIA-1 (2) and the ATE is identified under CIA-2 (3). Notice that the support of X differs for each estimator: $X|W = 1$ for the ATT, $X|W = 0$ for the ATN and X for the ATE. We get the following estimates of the three effects:

$$\hat{c}_W = \sum_{g \in \mathbf{G}^W} \omega_g^W \left(\bar{Y}_g^1 - \bar{Y}_g^0 \right),\ W = 0, 1, 2$$

with $\omega_g^W = I_g^W / I^W$. The variance equals:

$$\mathrm{V}(\hat{c}_W) = \sum_{g \in \mathbf{G}^W} (\omega_g^W)^2 \left(\mathrm{V}\left(\bar{Y}_g^1 \right) + \mathrm{V}\left(\bar{Y}_g^0 \right) \right),$$

and we use the following unbiased estimators ($W = 0, 1$) for the variances of the means:

$$\hat{\mathrm{V}}\left(\bar{Y}_g^W\right) = \frac{1}{(I_g^W - 1)\, I_g^W} \sum_{i \in \mathbf{I}_g^W} \left(Y_i - \bar{Y}_g^W\right)^2.$$

There are two sources of difference between these estimators:

1. The support of the matching variables, which defines the matching rates. Each estimator potentially relies on a different distribution of the confounding variables X. One may check directly the variations in this distribution when comparing the effects. Notice, that we must find enough individuals in the counterfactual group for the computation to make sense.

2. The weights ω_g^W used in the computation, for a given support. These weights can vary strongly when X includes determinants of the treatment variable, because the treatment is not allocated at random between the individuals.

Continuous variables. The CIA assumptions may hold for continuous variables too and perfect matching may be impossible because there are too few observations in the data set. This will obviously raise the problem of the matching rate. For the ATT, it is the proportion of treated who *can* be matched. In order to fix this problem, we need to adapt the matching method. Among them, we could use coarsened exact matching, calliper matching, nearest neighbour matching or propensity score matching. In all these cases, we should care about the matching rate because it indicates the fulfilment degree of the common support condition. One important decision to take is whether we use a twin only once or several times (i.e. for several treated). Obviously, drawing will replacement will always make more twins available, increase the matching rate and reduces bias. But the variance will be bigger and more difficult to compute. Drawing without replacement makes the computations easier but if the twins are scarce, some treated cannot be matched, and the ones who are matched may belong to a non representative set. The trade-off is clearly related to the size of the data set. With a very large reservoir of controls, drawing without replacement will not be a problem because all the treated can be matched with several twins. But if the data set is smaller, drawing with replacement may be the only relevant option. In this case, matching everyone with all the available twins may be the only way to avoid biases. Also notice that the availability of panel data may easier matching for the following reason: one may match the treated with both the people who are never treated and with the ones who will be treated at a later date. This increases the number of available matches.

Several adaptations are possible: coarsened exact matching, calliper matching or nearest neighbour matching. Coarsened exact matching simply transforms continuous variables into categorical ones by taking intervals. Exact matching is applied afterwards to the new categorical variables. Calliper matching is more interesting because it sets a maximum distance for the matching. The difference between the two methods is important. Consider a matching on age. With a coarsened exact matching we would define intervals like 16-29, 30-49 and 50+. This implies that someone aged 30 will be matched with someone aged 49, 19 years of age difference, but not with someone age 29 while there is only one year of age difference. With calliper matching one sets the difference explicitly like, say, 3 years. Someone aged 30 will be matched with people in the 27-33 range, which looks much better than with intervals.

Calliper matching. When matching variables include continuous variables, it is not possible to perform exact matching anymore. But one can set a tolerance margin for the continuous variable. This margin is called a *calliper*. For example, if we study the effect of cancer on employment participation, we may tolerate a difference of, say, 3 years between the treated and the controls. It could be justified by the fact that, on the one hand, the incidence probability of cancer does not vary much in three years, as well as the participation rate in the labour market, provided that the workers are far enough from the legal retirement age. More generally, explicit methods should be preferred to automatic methods, because they constrain the researcher to justify the distance allowed between the treated and the controls. A good survey of the literature should allow to determine a reasonable calliper for the main continuous variables. In what follows, we take the example of one continuous variable, but the extension to several variables is straightforward. Let X denote the matching variables. We will separate the discrete variables D from the continuous variables C, so that $X = (D, C)$. Considering the discrete variables first, we get a matching reservoir $\mathbf{M}(i)$:

$$\mathbf{M}(i) \triangleq \{j \in \mathbf{I}^0 : X_j = X_i\},\ i \in \mathbf{I}^1$$

the important point is that this reservoir does not depend on i. All the individuals $i \in \mathbf{I}_g$ share the same reservoir $\mathbf{M}(i)$ because they share the same value of the discrete matching variables. Introducing continuous variables will restrict the part of the reservoir that we can use and the matching will become individual. We denote it $\mathbf{R}(i)$:

$$\mathbf{R}(i) \triangleq \mathbf{M}(i) \cap \{j \in \mathbf{I}^0 : |C_j - C_i| \leq r\}.$$

where $r > 0$ is the calliper. It is simply the largest difference tolerated for a matching on the continuous variable. By convention $R(i)$ denotes the number of twins for the treated i.

The introduction of a calliper creates the following complication, when matching is done with replacement.[2] For the same value of the discrete variables the twins' sets are not disjoint anymore. In other words, if the calliper equals three years, the treated with less than three years difference will have similar twins and this will create a correlation between their performance differences. This must be accounted for both in the mean and in the variance formulas.

In order to simplify the exposition of the ATT, we use the following matching dummies, $r_{ij} = 1$ when j can be matched with i:

$$\forall j \in \mathbf{M}(i),\ r_{ij} = \begin{cases} 1 & \text{if } j \in \mathbf{R}(i) \\ 0 & \text{otherwise} \end{cases}$$

With this definition, the number of twins for the treated i equals:

$$R(i) \triangleq \sum_{j \in \mathbf{M}(i)} r_{ij}$$

and their mean outcome variable can be written indifferently:[3]

$$\bar{Y}_i^0 \triangleq \frac{\sum_{j \in \mathbf{R}(i)} Y_j}{R(i)} = \frac{\sum_{j \in \mathbf{M}(i)} r_{ij} Y_j}{\sum_{j \in \mathbf{M}(i)} r_{ij}}.$$

[2] We favour matching with replacement because it gives the highest matching rates.

[3] In the absence of continuous variables, all the dummies $r_{ij} = 1$ and we get $\bar{Y}_0(i) = \bar{Y}_0,\ \forall i$, the mean in the control group.

The ATT can be estimated by:

$$\hat{c}_1 = \frac{1}{I^1} \sum_{g \in \mathbf{G}^1} \sum_{i \in \mathbf{I}_g^1} \left(Y_i - \bar{Y}_i^0 \right)$$

Here we should notice that, for $i \in \mathbf{I}_g^1$, we have $\mathbf{M}(i) = \mathbf{I}_g^0$ so that we can rewrite the ATT in the following way:

$$\hat{c}_1 = \sum_{g \in \mathbf{G}^1} \omega_g^1 \left(\bar{Y}_g^1 - \tilde{Y}_g^0 \right)$$

with:

$$\tilde{Y}_g^0 = \frac{1}{I_g^0} \sum_{j \in \mathbf{I}_g^0} \bar{r}_j Y_j$$

with the counterfactual weights:

$$\bar{r}_j \triangleq \frac{1}{I_g^1} \sum_{i \in \mathbf{I}_g^1} \left(\frac{r_{ij}}{\frac{1}{I_g^0} \sum_{j \in \mathbf{I}_g^0} r_{ij}} \right).$$

Consider some special cases. When one control j can be matched with all the treated, we get $r_{ij} = 1 \ \forall i$ and $\bar{r}_j = 1$ so that we get the exact matching case $\tilde{Y}_g^0 = \bar{Y}_g^0$. When the control j cannot be matched with any of the treated i, we get $r_{ij} = 0$ and $\bar{r}_j = 0$ so that Y_j is excluded from the computation of the counterfactual. In the standard case, the control j can be matched with a part of the treated, and its contribution to the counterfactual will be increasing with its number of matches (i.e. the number of i such that $r_{ij} = 1$).

More generally, the effects can be written:

$$\hat{c}_W = \sum_{g \in \mathbf{G}^W} \omega_g^W \left(\bar{Y}_g^1 - \tilde{Y}_g^0 \right), \ W = 0, 1, 2$$

and their variance (Duguet and Le Clainche, 2020):

$$\mathrm{V}(\hat{c}_W) = \sum_{g \in \mathbf{G}^W} (\omega_g^W)^2 \left(\mathrm{V}\left(\bar{Y}_g^1 \right) + \mathrm{V}\left(\tilde{Y}_g^0 \right) \right)$$

but now, the variance of the counterfactual must account for the fact the same controls are used several times. Let us assume that the variance of the matchable twins' outcome inside the group g is $(\sigma_g^0)^2$, we get:

$$\mathrm{V}\left(\tilde{Y}_g^0 \right) = (\sigma_g^0)^2 \frac{\sum_{j \in \mathbf{I}_g^0} \bar{r}_j^2}{(I_g^0)^2}.$$

and one can use the following estimator:[4]

$$(\hat{\sigma}_g^0)^2 = \frac{1}{\sum_{j \in \mathbf{I}_g^0} r_{gj}} \sum_{j \in \mathbf{I}_g^0} \left(r_{gj} Y_j - \frac{\sum_{j \in \mathbf{I}_g^0} r_{gj} Y_j}{\sum_{j \in \mathbf{I}_g^0} r_{gj}} \right)^2$$

where $r_{gj} = 1$ if there exists $i \in \mathbf{I}_g^1$ such that $r_{ij} = 1$. In short, we keep the only controls that can be matched with at least once in $\mathbf{I}_g^1$ in order to compute the outcome variance.

[4] We do not use the correction for the degrees of freedom so that $\hat{o}_g^0 = 0$ when there is only one twin in some subgroups g, as expected.

Propensity score matching. Exact or even calliper matching may not be possible when there are many continuous variables. In this situation, it is possible to reduce the dimension of the matching problem by using the propensity score method. The propensity score, $e(X)$, is the probability to get the treatment conditional of the observable variables X:

$$e(X) = \Pr(W = 1|X).$$

Rosenbaum and Rubin (1983) showed that the following property holds for evaluating the ATE:

$$(Y_0, Y_1) \perp\!\!\!\perp W|X \text{ and } 0 < e(X) < 1,\ \forall X$$
$$\Rightarrow (Y_0, Y_1) \perp\!\!\!\perp W|e(X) \text{ and } 0 < e(X) < 1,\ \forall X$$

For the other effects, we use:

$$Y_k \perp\!\!\!\perp W|X \text{ and } 0 < e(X) < 1,\ \forall X$$
$$\Rightarrow Y_k \perp\!\!\!\perp W|e(X) \text{ and } 0 < e(X) < 1,\ \forall X$$

with $k = 0$ for the ATT and $k = 1$ for the ATN.

These results open the possibility to match on the propensity score $e(X)$, a real number, rather than directly on X. The intuition is the following: if two individuals have the same probability to be treated, and one is treated while the other is not, then the treatment may be allocated at random among them. This reduces the matching to one dimension. Several methods may then be applied in order to estimate the causal effects.[5]

The reader may however be conscious that matching on the propensity score is inefficient compared to matching on X (Frölich, 2007), so that this method should be used when direct matching is not feasible.

In practise, we do not observe $e(X)$ but an estimate $\hat{e}(X)$. It may be obtained by a Logit or a Probit model. This will have consequences for the regression methods presented later. Let W^* be a latent variable determining the (random) allocation to treatment:[6]

$$W_i^* = X_i\theta + u_i,\ i \in \mathbf{I}^2$$

where u is the disturbance of the model with cdf F_u and X includes a constant term.[7] We observe the following treatment variable:

$$W_i = \begin{cases} 1 & \text{if } W_i^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

which is Bernoulli distributed, with probability:

$$\begin{aligned} \Pr[W_i = 1|X_i] &= \Pr[W_i^* > 0|X_i] \\ &= \Pr[u_i > -X_i\theta] \\ &= 1 - F_u[-X_i\theta] \\ &= e(X_i;\theta). \end{aligned}$$

[5] For a presentation of propensity score methods, see Caliendo and Kopeinig (2008) and Austin (2011).

[6] This section requires some knowledge in limited dependent variables and maximum likelihood estimation. See Wooldridge (2002), chapter 15.

[7] Without loss of generality because matching is always perfect on the constant term.

In a Probit model, this reduces to $e(X_i;\theta) = \Phi(X_i\theta)$ where $F_u(u) = \Phi(u)$ is the cdf of the standard normal distribution; in a Logit model, we get $e(X_i;\theta) = \Lambda(X_i\theta)$, where $F_u(u) = 1/(1+\exp(-u))$ is the cdf of the logistic distribution. The estimation proceeds by maximum likelihood. The log-likelihood for one observation i is:

$$\ell_i(W_i|X_i;\theta) = W_i \ln(e(X_i;\theta)) + (1-W_i)\ln(1-e(X_i;\theta))$$

and the maximum likelihood estimator is defined as:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i\in\mathbf{I}^2} \ell_i(W_i|X_i;\theta).$$

We get the estimated propensity score by:

$$\hat{e}(X_i) = e(X_i;\hat{\theta}).$$

From there, we can use a calliper matching, a kernel prediction of the counterfactual or a regression including the propensity score.[8]

Common support and balancing. The propensity score should verify $0 < e(X) < 1$. This means that no individual is excluded from the treatment, or imposed the treatment. This implies that, for all values of X, there may be some randomness in the allocation of the treatment among individuals. We have a distribution of the treatment probability in the interval $\mathbf{E}_1 = [e_1^-, e_1^+]$ and, similarly for the not-treated $\mathbf{E}_0 = [e_0^-, e_0^+]$. Generally, the distribution of the treated will be shifted to the right because the treated tend to have higher probabilities to be treated. If we wish to estimate the ATT, we will need to find matches for the probabilities in $\mathbf{E}_1$. Ideally, this requires $\mathbf{E}_1 \subseteq \mathbf{E}_0$. When this property fails, we can still compute a matching rate and proceed to the estimation when it remains close to 100%. Obviously, all the effects can be estimated when $\mathbf{E}_0 = \mathbf{E}_1$. However, nothing guarantees that this condition holds and it should be examined carefully. Notice that taking the common support automatically, $\mathbf{E}_0 \cap \mathbf{E}_1$, may cause a bias when it removes an important part of the data points. When the matching possibility are reduced, it is recommended to turn to methods with replacement, so as to use each match several times.

The condition $Y_k \perp\!\!\!\perp W|e(X)$ does not imply the condition $Y_k \perp\!\!\!\perp W|X$, which remains the motivation the method. For this reason, authors have suggested to check for covariate balancing inside subclasses of the propensity score (Imbens and Rubin (2015), chap. 14). From a matching perspective, one should focus on the continuous variables. Indeed, with replacement, one can always match the value of a categorical variable if there is a least one individual in the matching reservoir. Since one can match both on the categorical variables and the propensity score at the same time, the only difference should come from the continuous variables.[9] Rosenbaum and Rubin (1985) propose to compute the following *normalized difference* for each stratum defined by the propensity score:[10]

$$\hat{\Delta} = 100 \times \frac{\bar{z}_1 - \bar{z}_0}{\sqrt{(s_1^2 + s_0^2)/2}}$$

[8] It is possible to match both on the propensity score and other variables. It may be useful for some categorical variables, like gender.

[9] Notice that one can include the categorical variables in the propensity score estimation, and match on both the propensity score and the exact value of categorical variables in a later step.

[10] In practice, one may choose quantiles of the estimated propensity score to insure equal size strata. The choice of the quantile should be guided by the number of observation in each stratum. On the use of the normalized differences, see the chapters 15 and 16 of (Imbens and Rubin, 2015)

where $\bar{z}_W = (1/I^W)\sum_{i\in\mathbf{I}^W} z_i$, $s_W^2 = 1/(I^W-1)\sum_{i\in\mathbf{I}^W}(z_i-\bar{z}_W)^2$ and $W = 0,1$. This statistic is different from the Student statistic, since it measures the deviation of means expressed in standard errors. The Student statistics for testing the equality of the means is:

$$\hat{T} = \frac{\bar{z}_1 - \bar{z}_0}{\sqrt{s_1^2/I^1 + s_0^2/I^0}},$$

and it will increase with the sample size when the means are different, while $\hat{\Delta}$ may not. More generally, other differences may be examined, involving variances or a measure of the distributions overlap, as indicated in Imbens and Rubin (2015).

Dehejia and Wahba (2002) propose an algorithm that relates the specification of the propensity score to its capacity to ensure the balancing of the observables X. After estimation, the propensity score intervals are made and the Student tests are performed. If the tests pass for all the strata, the algorithm stops. Otherwise, if a variable does not pass the test, the authors suggest to add higher order and interaction terms in the specification of the propensity score and run the test again.

Calliper matching on the propensity score. Let $\hat{e}_i = \hat{e}(X_i)$, then the matching set for the individual i on the population $W \in \{0,1\}$:

$$\mathbf{M}_i^W \triangleq \{j \in \mathbf{I}^W : |\hat{e}_j - \hat{e}_i| \leq r\},\ i \in \mathbf{I}^{1-W}$$

where $r > 0$ is the calliper. The associated counterfactual for an individual i will be:

$$\bar{Y}_i^W = \frac{1}{M_i^W}\sum_{j\in\mathbf{M}_i^W} Y_j.$$

where M_i^W denotes the number of observations in $\mathbf{M}_i^W$. The use of a calliper creates an overlap between the matching sets of the individuals. The ATT will be estimated by:

$$\hat{c}_1 = \frac{1}{I^1}\sum_{i\in\mathbf{I}^1}\left(Y_i - \bar{Y}_i^0\right)$$

the ATN by:

$$\hat{c}_0 = \frac{1}{I^0}\sum_{i\in\mathbf{I}^0}\left(\bar{Y}_i^1 - Y_i\right)$$

and the ATE by:

$$\hat{c}_2 = \frac{1}{I^2}\sum_{i\in I^2}(1-W_i)\left(\bar{Y}_i^1 - Y_i\right) + W_i\left(Y_i - \bar{Y}_i^0\right) = \frac{I^0}{I^2}\hat{c}_0 + \frac{I^1}{I^2}\hat{c}_1. \tag{4}$$

With $I^2 = I^0 + I^1$, the ATE is obtained as the weighted mean of the two other treatment effect estimates. The variances are more complicated to compute, so that the bootstrap method may be used (Efron and Tibshirani, 1994).[11] The sampling should be done separately for the treated and the controls. Then, one should run the two estimation steps, in order to account for the variability in the estimated propensity scores. Notice, however, that the standard bootstrap method may fail for the nearest neighbour estimators with a fixed number of matches (Abadie and Imbens, 2008).

[11] This estimator is similar to a kernel matching with a uniform kernel, as shown in the next section.

Kernel estimation. Consider the estimation of the ATT. The main issue is the estimation of the counterfactual:

$$\mathrm{E}(Y^0|e, W = 1).$$

Following Heckman et al. (1997), we will treat this issue as a nonparametric prediction problem. More precisely, we will consider that that Y^0 is a function of the propensity score $e(X)$ instead of X:

$$Y^0 = m(e) + u$$

where m is an unknown function, $u \perp\!\!\!\perp W|e$ and $\mathrm{E}(u|e) = 0$. We get:

$$\mathrm{E}(Y^0|e, W = 1) = \mathrm{E}(Y^0|e, W = 0).$$

In order to estimate this quantity, we will first restrict ourselves to the not-treated sample ($W = 0$) and perform a non parametric regression of Y on the estimated propensity score $\hat{e}$. The choice of a non-parametric regression is motivated by the will to avoid restrictions on the function m.

In a second step, we will compute the predictions from the propensity scores in the treated sample ($W = 1$). This is possible when the support condition holds for the estimated propensity scores. For each treated i, we compute the prediction:

$$\hat{Y}_i^0 = \hat{m}(\hat{e}_i)$$

and compute the ATT in the following way:

$$\hat{c}_1 = \frac{1}{I^1} \sum_{i \in \mathbf{I}^1} \left(Y_i - \hat{Y}_i^0 \right).$$

In order to compute $\hat{m}$, we use the Nadaraya-Watson estimator (Nadaraya (1964), Watson (1964)).[12]

$$\hat{m}(e_i) = \sum_{j \in \mathbf{I}^0} \omega_j(e_i) Y_j$$

where $\omega_j(e_i)$ is the weight of the (not-treated) match j in the computation of the local average at e_i. This weight of j should be maximal when the match is perfect, $e_i = e_j$, and decrease with the distance from the match $|e_i - e_j|$. A relevant choice is kernel weighting. It is defined by:

$$\omega_j(e_i) = \frac{K(z_{ij})}{\sum_{j \in \mathbf{I}^0} K(z_{ij})}$$

where $z_{ij} = (e_i - e_j)/h$, K is the kernel chosen and h the bandwidth parameter. This parameter is a smoothing parameter for the estimation of the m function. In the standard cases, the kernel function is positive, symmetric around 0 and integrates to 1. It reaches its maximum in 0, that is when $e_i = e_j$. The bandwidth parameter h is determined by cross validation on the not-treated sample. Let $\hat{h}$ be the estimated bandwidth. It is obtained through cross-validation (Frölich, 2004).[13]

$$\hat{h} = \arg\min_h \sum_{j \in I^0} \left(Y_j - \hat{m}_{-j}(\hat{e}_j) \right)^2$$

[12] The part that follows requires some notions of nonparametric estimation, see Pagan and Ullah (1999).

[13] The method was originally introduced by Clark (1975).

where $\hat{m}_{-j}$ is the kernel estimation obtained by omitting the j-th observation. Finally, the effect of the treatment on the treated is estimated by:

$$\hat{c}_1 = \frac{1}{I^1} \sum_{i \in \mathbf{I}^1} \left(Y_i - \sum_{j \in \mathbf{I}^0} \hat{\omega}_j(\hat{e}_i) Y_j \right)$$

with

$$\hat{\omega}_j(\hat{e}_i) = \frac{K(\hat{z}_{ij})}{\sum_{j \in \mathbf{I}^0} K(\hat{z}_{ij})}, \hat{z}_{ij} = \frac{\hat{e}_i - \hat{e}_j}{\hat{h}}.$$

The main determinant of the weight is the propensity score difference. The weight is strictly decreasing with this difference and may be 0 if the difference is too strong. A widespread kernel that use all the observations is the Gaussian kernel:

$$K(z) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \ z \in \mathbb{R}.$$

There is always a match with this kernel, but it can be far. Another interesting special case is the uniform kernel:

$$K(z) = \begin{cases} \frac{1}{2h} & \text{if } |z| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

With the uniform kernel, the prediction reduces to a local mean of Y_j values for the $\hat{e}_j$ that falls in the interval $[\hat{e}_i - h, \hat{e}_i + h]$. In this case, the bandwidth parameter can be interpreted as a calliper on the propensity score ($h = r$).

The variance of the Nadaraya-Watson estimator may be estimated by the bootstrap. In this situation, we keep the estimate obtained for the original sample, $\hat{c}_1$ and use the bootstrap repetitions in order to estimate the variance. The B draws are done separately in the treated and not-treated samples. The resulting estimation for the bootstrap repetition $b = 1, ..., B$ is denoted $\hat{c}_{1b}$ and the variance is estimated by:[14]

$$\hat{\mathrm{V}}(\hat{c}_1) = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{c}_{1b} - \hat{c}_1)^2.$$

Notice that the standard bootstrap method works well the for the kernel estimators but may fail for other estimators. Abadie and Imbens (2008) showed that the bootstrap may fail for the nearest neighbour estimators with a fixed number of matches.

The extension to the ATN is straightforward since the we can define a treatment equal to the absence of treatment. The ATE will clearly require two non parametric regressions instead of one. A first regression will estimate $\mathrm{E}(Y^1|W = 0)$ and another one $\mathrm{E}(Y^0|W = 1)$, with different bandwidth parameters. Then, we should apply (4) in order to get the estimate $\hat{c}_2$, and perform bootstrap repetitions for estimating the variance of the corresponding estimator.

2.2 Regression

Regression on the confounding variables. The first type of regression simply uses the matching variables as regressors (Rubin (1973), Lee (2005)). Consider the following model:

$$Y^W = \alpha_W + X\beta_W + u_W, \ W = 0, 1$$

[14] On the choice of the number of bootstrap repetitions, see Andrews and Buchinsky (2000).

They represent the two potential outcomes for the same individual depending on s-he is treated ($W = 1$) or not ($W = 0$). We observe:

$$\begin{aligned} Y &= (1 - W) \times Y^0 + W \times Y^1 \\ &= \alpha_0 + X\beta_0 + W(\alpha_1 - \alpha_0) + (WX)(\beta_1 - \beta_0) + u \end{aligned}$$

with $u = (1 - W)u_0 + Wu_1$. We obtain a model with the cross products WX, which we rewrite:

$$Y = \alpha + X\beta + W\gamma + WX\delta + u \tag{5}$$

with $\alpha = \alpha_0$, $\beta = \beta_0$, $\gamma = \alpha_1 - \alpha_0$ and $\delta = \beta_1 - \beta_0$. Several specific cases are of interest:

1. If the treatment is allocated at random and data are collected about Y only, we can write the model without X:

   $$Y = \alpha + W\gamma + u,$$

   and the OLS estimation of γ will simply give the difference of the outcome means in the treated and control samples: $\hat{\gamma} = \bar{Y}_1 - \bar{Y}_0$. This is also one of the estimators used with experimental data.

2. The variables X have the same effect in the treatment and the control groups, $\beta_0 = \beta_1$. Although the assumption deserves to be tested, this model is often used in applications, without the test. We get:

   $$Y = \alpha + X\beta + W\gamma + u,$$

   we obtain a model in which the treatment dummy W is included as a standard additional regressor.

In these two cases however, we get a constant treatment effect in the following sense. In the first model, we get the following ATT:

$$c_1 = \mathrm{E}(y_1 - y_0|W = 1) = \gamma + \mathrm{E}(u_1 - u_0|W = 1).$$

Further assuming $\mathrm{E}(u_1) = 0$ and a mean independence assumption about u_0, namely $\mathrm{E}(u_0|W = 0) = \mathrm{E}(u_0|W = 1)$, we get $c_1 = c$. Now consider the ATN:

$$c_0 = \mathrm{E}(y_1 - y_0|W = 0) = \gamma + \mathrm{E}(u_1 - u_0|W = 0)$$

and assuming $\mathrm{E}(u_0) = 0$ and $\mathrm{E}(u_1|W = 0) = \mathrm{E}(u_1|W = 1)$, we get $c_0 = c$. The reader can check that we get the same result for the ATE by simply assuming $\mathrm{E}(u_1) = \mathrm{E}(u_0)$. In the three cases, the OLS will provide an estimate of γ and this method of estimation does not allow for different values of the average treatment effects in the treated and the control groups. Adding the explanatory variables X does not change this property since their effect is the same whatever W is.

The only way to allow for different estimations of the three treatment effects is to include the cross products WX in the regression. In model (5) we get, under the relevant mean independence assumptions:

$$\begin{aligned} c_0 &= \gamma + \mathrm{E}(X|W = 0)\delta, \\ c_1 &= \gamma + \mathrm{E}(X|W = 1)\delta, \\ c_2 &= \gamma + \mathrm{E}(X)\delta. \end{aligned}$$

Different conditional distributions for X will give different evaluations. Letting $(\hat{\gamma},\hat{\delta})$ be the OLS estimator of (γ,δ) and $\bar{X}_W$ be the vector of sample means for the samples defined by $W = 0,1,2$, we can use the following estimators:

$$\begin{aligned}
\text{ATN: } \hat{c}_0 &= \hat{\gamma} + \bar{X}_0\hat{\delta},\\
\text{ATT: } \hat{c}_1 &= \hat{\gamma} + \bar{X}_1\hat{\delta},\\
\text{ATE: } \hat{c}_2 &= \hat{\gamma} + \bar{X}_2\hat{\delta}.
\end{aligned}$$

with

$$\bar{X}_W = \frac{1}{I^W}\sum_{i\in I^W} X_i.$$

But this way to present the problem does not lead to the simplest expression of the variance of these estimators, since $\bar{X}_W$ and $(\hat{\gamma},\hat{\delta})$ are correlated. For getting the variances directly from the OLS output, it is better to *center* the X variables before to take the cross products. Here it is important to notice that the mean used in the centering *depends on the effect* we wish to estimate. For the effect $\hat{c}_W$, we use $\bar{X}_W$, with $W = 0,1,2$. We rewrite the model (5) under the equivalent formulation:

$$y = \alpha + X\beta + W\gamma_W + W(X - \bar{X}_W)\delta + u$$

with $\gamma_W = \gamma + \bar{X}_W\delta$. Notice that the definition of γ_W changes with the mean used in the centring. The estimates will now be obtained with:

$$c_W = \gamma_W + \mathrm{E}(X - \bar{X}_W|W)\delta$$

which can be estimated by:

$$\begin{aligned}
\hat{c}_W &= \hat{\gamma}_W + \frac{1}{I^W}\sum_{i\in I^W}(X_i - \bar{X}_W)\\
&= \hat{\gamma}_W.
\end{aligned}$$

and we can use the OLS variance of $\hat{c}_W$. Here, one may remark that the structure of the disturbance $u = (1-W)u_0 + Wu_1$ suggests a heteroskedasticity correction since the error term differs in the treated and control groups (see White (1980)).

Regression on the propensity score. When there are many variables, it may be easier to perform a regression on the propensity score. The implementation is simple: in a first step, one estimates the propensity score $\hat{e}(X)$ and, in a second step, one uses $\hat{e}(X)$ instead of X. All the previous methods can be used. The model is (Rosenbaum and Rubin, 1985):

$$Y_W = \alpha_W + \beta_W e(X) + u_W,\ W = 0,1$$

and we would like to estimate:

$$y = \alpha + e(X)\beta + W\gamma_W + W(e(X) - \overline{e(X)}_W)\delta + u$$

in order to get $\hat{c}_W = \hat{\gamma}_W$. Notice that $\overline{e(X)}_W$ is the mean of the propensity scores in group W. Unfortunately, we cannot observe $e(X)$ and we use an estimate of it instead,

denoted $\hat{e}(X)$. This is called a *generated regressor* in the literature. The regression becomes:

$$y = d_0 + d_{1W} W + d_2 \hat{e}(X) + d_3 W(\hat{e}(X) - \overline{\hat{e}(X)}_W) + v.$$

The OLS estimation $\hat{d}_{1W}$ will still provide a consistent estimator of c_W, but the OLS standard errors are wrong. One important point to notice is that the usual heteroskedasticity corrections are not a good approach to this problem. A specific correction must be made.[15] Our presentation follows the general case of Murphy and Topel (2002). Separating the standard regressors (Z_1) from the generated ones (Z_2), we rewrite the model under the form:

$$y_i = Z_{1i}\delta_1 + Z_{2i}\delta_2 + v,\ i \in \mathbf{I}_2.$$

with

$$\begin{aligned}
Z_{1i} &= (1, W_i), \\
Z_{2i} &= f(X_i;\hat{\theta}) \triangleq (\hat{e}_i, W_i(\hat{e}_i - \overline{\hat{e}}_{iW})), \\
\hat{e}_i &= e(X_i;\hat{\theta}), \\
\delta_1 &= (d_0, d_{1W})',\ \delta_2 = (d_2, d_3)', \\
v_i &= u_i + f(X_i;\theta) - f(X_i;\hat{\theta}).
\end{aligned}$$

Letting:

$$Z = (Z_1, Z_2) \text{ and } \delta = \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}$$

the OLS estimator is given by:

$$\hat{\delta} = \left(Z'Z\right)^{-1} Z'y$$

and its asymptotic variance is deduced from:

$$\begin{aligned}
\text{avar}\left(\sqrt{I_2}(\hat{\delta} - \delta\right) &= \sigma^2 Q_0^{-1} \\
&\quad + Q_0^{-1}\left(Q_1 J_1^{-1} Q_1' - Q_1 J_1^{-1} Q_2' - Q_2 J_1^{-1} Q_1'\right) Q_0^{-1}
\end{aligned}$$

with the empirical counterparts:

$$\begin{aligned}
\hat{\sigma^2} &= \frac{1}{I_2} \sum_{i \in \mathbf{I}_2} \hat{v}_i^2, \\
\hat{v}_i &= y_i - Z_i \hat{\delta}, \\
\hat{Q}_0 &= \frac{1}{I_2} \sum_{i \in \mathbf{I}_2} Z_i' Z_i, \\
\hat{Q}_1 &= \frac{1}{I_2} \sum_{i \in \mathbf{I}_2} Z_i' \frac{\partial f}{\partial \theta}(X_i;\hat{\theta})\hat{\delta}_2, \\
\hat{Q}_2 &= \frac{1}{I_2} \sum_{i \in \mathbf{I}_2} Z_i' \hat{v}_i \frac{\partial \ell_i}{\partial \theta}(X_i;\hat{\theta}), \\
\hat{J}_1 &= -\frac{1}{I_2} \sum_{i \in \mathbf{I}_2} \frac{\partial^2 \ell_i}{\partial \theta \partial \theta'}(X_i;\hat{\theta}).
\end{aligned}$$

[15] This issue was originally addressed by Pagan (1984). For a more detailed treatment, see the sections 6.1 and 12.4 of Wooldridge (2002).

The first term $\sigma^2 Q_0^{-1}$ corresponds to the usual OLS variance (under homoskedasticity). The other terms represent the bias of the OLS variance. It can have any sign, so that the OLS estimation can overestimate or underestimate the true standard errors.

3 Panel data estimation

3.1 Data structure

Treatment group. We consider a panel data with maximum of T years and I_2 individuals. The panel is unbalanced, so that the individuals can enter or exit the panel at any time $t = 1, \ldots, T$. We assume that the data is missing at random. Compared to the cross-section case, the treatment will now be identified by its date of appearance, so that the sample of treated varies over time. In order to simplify the presentation, we assume that each individual i is present over the period $[t_i^-, t_i^+]$ and is treated at date t_i. When the individual is never treated, we use the convention of an infinite treatment date $t_i = \{+\infty\}$, so that the individual cannot be treated in the panel time span.[16]

The time dimension modifies the definition of the treatment and control groups since the status of an individual varies over time. Consider a health event, like an accident. The control group includes the individuals who did not have an accident *yet*. After that date, we may consider that the individual has moved into the treatment group. In most studies, we will not allow for individuals to get back to the control group after the health event, because it may have long term effect. An accident can be benign, but it can also cause a permanent disability. A similar definition is used for chronic diseases, since they are long lasting. Interestingly, this convention allows for measuring treatment effects after several periods of time. In this approach, we distinguish the effect at the date of the treatment, from the effects several periods after. This allows for seeing whether the health consequences are on the short run or on a longer term. Overall, the control group includes the workers who did not had the treatment yet. Notice that this definition is the only one that is compatible with cross-section data, because we do not observe the future values of the treatment in this case. With a cross section, there should be a part of the not-treated that will be treated at a later date because the future value of the treatment are not available at the time of the sample construction. When the treatment has long lasting consequences, we may well consider that t_i is the starting date of the treatment rather than the treatment date. In this case, we will use:[17]

$$W_{it} = \begin{cases} 0 & \text{if } t < t_i \\ 1 & \text{if } t \geq t_i \end{cases} \tag{6}$$

This raises the question of which effects can be estimated. One advantage of panel data is to allow for examining how the intensity of the effect varies over time. In this presentation, we will focus on the ATT a periods after the treatment:

$$c_1(a) = \mathrm{E}\left(Y_t(1) - Y_t(0) | W_{it} = 1, t = t_i + a\right), \ a \geq 0$$

where t_i is the treatment date. Here we insist on the dependence of the effect on the time from treatment a. Notice that, on yearly data, the date of the treatment is not considered as fully informative, because the health event can happen at any time between January 1st and December 31st. This is why the first difference often studied is

[16] One can allow randomly missing observations in $[t_i^-, t_i^+]$ without altering what follows.

[17] With this definition, notice that the convention $t_i = \{+\infty\}$ implies $W_{it} = 0 \ \forall t$.

not for $a = 0$ but for $a = 1$. Since we observe $Y_t(1)$ for the treated after the treatment, we will need to estimate a counterfactual representing the outcome the treated would have had if they had not been treated. Matching and regression can be used to fill this purpose.

3.2 Difference in differences

A popular estimator on panel data is the Difference-in-differences (henceforth, DiD) estimator.[18] We can derive it from the following panel data model:

$$Y_{i,t}(W_{i,t}) = c_0 + c_1(t - t_i)W_{i,t} + \alpha_i + \beta_t + \varepsilon_{i,t} \tag{7}$$

where c_0 is the constant term, α is an individual fixed effect, also called "correlated effect" in the literature, β a time effect representing a flexible time trend and ε the idiosyncratic error term, uncorrelated with the other elements of the model. Without loss of generality, we can set $\mathrm{E}(\alpha) = \mathrm{E}(\beta) = \mathrm{E}(\varepsilon) = 0$. The treatment dummy W is defined by (6).

In this model, it is readily checked that the ATT after a periods is given by $c_1(a)$, with the convention $c_1(a) = 0,\ \forall a < 0$:[19]

$$\mathrm{E}\left(Y_t(1) - Y_t(0) | W_{it} = 1, t = t_i + a\right) = c_1(a),\ a \geq 0$$

There are several ways to estimate this model. A popular one is the Difference-in-Differences estimator. This method consist in eliminating (c_0, α, β) by differencing, and to attenuate ε by averaging. First, consider the treatment date of one treated individual i and compute the outcome variation b periods before and a periods after this date.[20] We get:

$$\begin{aligned} D_i(a,b) &= Y_{i,t_i+a} - Y_{i,t_i-b} \\ &= c_1(a) + \beta_{t_i+a} - \beta_{t_i-b} + \varepsilon_{i,t_i+a} - \varepsilon_{i,t_i-b} \end{aligned}$$

Second, take one not-treated that is present on the same dates and compute the before-after outcome difference. Here, we should define what a not-treated is in a dynamic context. One can adopt a static definition: anyone who is treated at any date. With this convention, the not-treated belong to the set:

$$\mathbf{M}^S(i) = \left\{ j \in I^2 : (t_j = +\infty) \cap (t_j^- \leq t_i - b) \cap (t_j^+ \geq t_i + a) \right\}, \tag{8}$$

but one can also choose to adopt a *dynamic matching* convention by taking all the individuals in $\mathbf{I}^2$ who did not had a treatment up to $t_i + a$:

$$\mathbf{M}^D(i) = \left\{ j \in I^2 : (t_j > t_i + a) \cap (t_j^- \leq t_i - b) \cap (t_j^+ \geq t_i + a) \right\}. \tag{9}$$

Clearly, we have $M^S(i) \subseteq M^D(i)$ since the condition $(t_j > t_i + a)$ is always fulfilled for $t_j = lbrace + \infty\}$. Once the set is chosen, we take the difference:

$$\begin{aligned} D_j(a,b) &= Y_{j,t_i+a} - Y_{j,t_i-b} \\ &= \beta_{t_i+a} - \beta_{t_i-b} + \varepsilon_{j,t_i+a} - \varepsilon_{j,t_i-b}, \end{aligned}$$

[18] On this estimator, see Lechner (2010).

[19] The last condition means that there is no treatment effect before the treatment date. This convention is compatible with the convention $t_i = +\infty$ for the not-treated.

[20] Standard values are $a = b = 1$.

therefore the difference in differences equals:

$$D_i(a,b) - D_j(a,b) = c_1(a) + \varepsilon_{i,t_i+a} - \varepsilon_{i,t_i-b} - (\varepsilon_{j,t_i+a} - \varepsilon_{j,t_i-b})$$

and we get an unbiased estimate of the ATT:

$$\mathrm{E}(D_i(a,b) - D_j(a,b)) = c_1(a).$$

In practice, we compute the mean of all the double differences, so that the estimator is:

$$\hat{c}_1(a) = \frac{1}{I^1} \sum_{i \in \mathbf{I}^1} \left((Y_{i,t_i+a} - Y_{i,t_i-b}) - \frac{1}{M(i)} \sum_{j \in \mathbf{M}(i)} (Y_{j,t_i+a} - Y_{j,t_i-b}) \right). \tag{10}$$

where $\mathbf{M}(i)$ is either the static (8) or the dynamic (9) matching set. Even though the previous notations insist on the dependence of the ATT on a, it obviously depends on the reference period b. There are often robustness checks on this parameter. In the context of program evaluations Chabé-Ferret (2015) even recommends the use of symmetric DiD methods (with $a = b$).

The computation of the variance can be made by several methods. When the differences are uncorrelated, one can use the methods adapted to cross-sectional data. This assumption is generally valid for the terms in $Y_{i,t_i+a} - Y_{i,t_i-b}$ when the individuals have independent outcomes. For the terms in $Y_{j,t_i+a} - Y_{j,t_i-b}$ it may also be the case if each twin is used only once. But, in the general case, the computations involve correlated components because the same twins are used at different dates, and we can reasonably expect that their time series are autocorrelated. With the time dimension, belonging to a matching group does not exclude a twin from another matching group. A bootstrap method may fix this issue. In this case, we draw the individuals separately in the treated and the control groups, and keep their entire time series each time.

3.3 Matching

If we use the time dimension for differencing and estimate longer run effects, we may well match as before. Consider the evaluation of the ATT. What we need is an estimate of the outcome of the treated if they had not been treated. Indeed, we do observe what the treated did before to be treated thanks to the time dimension of the panel.

Then, one may think that a simple assumption like $Y_{t+a}(0) \perp\!\!\!\perp W|X$ may be enough. However, this would miss the possibility to eliminate the fixed effect from the outcome equation. For this reason, we take a before-after difference $Y_{t+a}(0) - Y_{t-b}(0)$ and this explains the form of the conditional independence assumption below. Another point will be on the matching variables. In order to avoid endogeneity issues, we may match on lagged and time constant variables. The lag used will typically be $t_i - b$ but one may well use a summary of past values for matching.

Exact matching. Considering an additive model, we need to take a difference in order to eliminate the fixed effect. Then, we will look for a similar difference for the not-treated, in order to compute a DiD estimator.[21] Therefore, in order to estimate the $c_1(a)$, we will assume:

$$\mathrm{CIA-0P}: Y_{t+a}(0) - Y_{t-b}(0) \perp\!\!\!\perp W_t | X_{t-b}, \text{ for some } b \geq 1$$

[21] Notice that the not-treated difference is based on the *treated* dates.

Under this assumption, we get:

$$\mathrm{E}\left(Y_{t+a}(0) - Y_{t-b}(0) | W_t = 1, X_{t-b}\right) = \mathrm{E}\left(Y_{t+a}(0) - Y_{t-b}(0) | W_t = 0, X_{t-b}\right)$$

and we can use the outcome variation of the not-treated in order to evaluate the outcome variation of the treated if they had not been treated. The matching set know depends on observable variables. They typically include the variables that do not vary over time, like gender, the period of birth or the education level, denoted X_i^1, and lagged time varying variables $X_{i,t-b}^2$. For the treated i, we get the dynamic matching set:[22]

$$\mathbf{M}^D(i) = \left\{ j \in \mathbf{I}_2 : (t_j > t_i + a) \cap (X_j^1 = X_i^1) \cap (X_{j,t_i-b}^2 = X_{i,t_i-b}^2) \right\}$$

and the estimators are given by (10).

Calliper matching. Here, we just need to add constraints on the continuous variables. It may vary over time or not. Let C_i^1 a continuous variables stable over time, with calliper r_1 and $C_{i,t}^2$ its time-varying equivalent with calliper r_2. The matching set becomes:

$$\mathbf{M}(i)^D = \left\{ j \in \mathbf{I}_2 : (t_j > t_i + a) \cap (X_j^1 = X_i^1) \cap (X_{j,t_i-b}^2 = X_{i,t_i-b}^2) \right.$$
$$\left. \cap (|C_j^1 - C_i^1| \leq r_1) \cap (|C_{j,t_i-b}^2 - C_{i,t_i-b}^2| \leq r_2) \right\}$$

and the same formula as before (10) applies for the ATT estimation.

Propensity score matching. The most difficult issue in this case is, in fact, the estimation of the propensity score itself. The use of panel data opens the possibility to account for the existence of individual heterogeneity in the probability to get the treatment. This is relevant when one thinks that the list of explanatory variables does not include all the determinants of the treatment allocation. It will happens if time constant variables have been used to allocate the treatment and if these variables are not observable by the econometricians. In practice however, researchers rarely estimate a propensity score with an individual effect. Here are some practices and suggestions:

- Pooled regressions. One just run a regression like in the cross-section case, with adding a time varying constant term.
- Separate regressions. Each treatment year is used to estimate a binary variable model separately. All the regression coefficients vary with each year.
- Panel data estimation. Ideally, one would like to estimate a model like:[23]

  $$e(W_{it}; \theta) = \mathrm{P}(\theta_t^0 + \theta_i^1 + X_{i,t-b}^2 \theta_2),$$

  where θ_t^0 is a time effect and θ_i^1 is an individual fixed effect. After the estimation, we would use:

  $$e(W_{it}; \hat{\theta}) = \mathrm{P}(\hat{\theta}_t^0 + \hat{\theta}_i^1 + X_{i,t-b}^2 \hat{\theta}_2).$$

[22] We favour the dynamic version of the matching set, with replacement, in order to keep the highest number of matches. Otherwise, including many matching variables may make the matching impossible for many treated and bias the estimation.

[23] The effect of the time constant variables X_i^1 is included in the θ_i^1 term.

Unfortunately, the estimation of such models is complicated by several problems (Hsiao, 1996). Also notice that we need an estimate of the individual effect $\hat{\theta}_i^1$ in order to estimate the propensity score. Contributions for solving this problem include Chamberlain (1980), Greene (2004), Fernández-Val (2009) and Stammann et al. (2016).[24]

3.4 Regression

Other methods than matching may be used in order to estimate the ATT. Regression methods naturally extend to panel data and allow for controlling unobserved heterogeneity. Notice that, when the model is fully saturated, regression may be even be equivalent to matching (Angrist and Pischke, 2008).[25]

Data structure. The first thing to do is to select which dates of the treated are to be included in the panel. Indeed, in most situations all the effects cannot be estimated due to the lack of data. The number of treated available tends to decrease when we get away from the treatment date. Therefore, we set an estimation windows k, which represents the longest tractable period available for the estimation:

$$c_1(a) = \mathrm{E}\left(Y_{t+a}(1) - Y_{t+a}(0) | W_t = 1\right) \text{ with } 0 \leq a \leq k$$

Once that k is selected, one should drop all the observations of the treated for which $t > t_i + k$, otherwise this case could go in the reference group and bias the estimation. Another possibility is to build an aggregate dummy for all the cases $t > t_i + k$ and compute an average effect over this interval.[26] We will define $k+1$ treatment dummies, where each dummy corresponds to a lag in the treatment effect:

$$W_{i,t}(a) = \begin{cases} 1 & \text{if } t = t_i + a \\ 0 & \text{otherwise} \end{cases}$$
$$a = 0, \ldots, k$$

The first dummy is $W_{i,t}(0)$ and should be thought off as an incomplete measurement of the effect, unless the treatment is measured at the beginning of period and the outcome at the end of period. When there are not enough observations for estimating a separate effect per period, one may regroup them into intervals.

Regression on the confounding variables. We include the explanatory variables X_{it}. The model is:

$$Y_{i,t} = \beta_0 + X_{i,t}\beta_1 + \sum_{a=0}^{k} W_{i,t}(a)\left(\gamma_0^r(a) + (X_{i,t} - \bar{X}_r)\gamma_1(a)\right) + u_{i,t}, \tag{11}$$

[24] The last contribution is associated with the R package “bife", which provides estimation solutions.

[25] When X includes only categorical data, this is equivalent to run a regression with all the dummies and their cross products (including for more than two variables). In practice, we may include a selection of the cross products to approximate a fully saturated model.

[26] This raises the symmetric issue of whether we should keep all the other values of the treated or only $t = t_i - b$. We assume that we keep all the values prior to the treatment, when the support of the treatment dates distribution t_i covers all the periods available in the panel. Otherwise, a different convention may be used.

with $\gamma_0^r(a) = \gamma_1(a) = 0, \forall a < 0$. $\bar{X}_r$ is the mean of the reference population (treated, not-treated, total) and $u_{i,t}$ the error term with a two-way fixed effects structure. We let:

$$u_{i,t} = \alpha_i + \beta_t + \varepsilon_{i,t},\ \mathrm{E}(\alpha) = \mathrm{E}(\beta) = \mathrm{E}(\varepsilon) = 0.$$

where α and β are the individual and the time constant fixed effects, potentially correlated with W and X, and ε is the idiosyncratic error term, uncorrelated with the other components of the model. Notice that the values of the coefficients $\gamma_0^r(a)$ depend on the mean used for the centring of the explanatory variables ($\bar{X}_r$).

For the ATT, we condition the effect on the treated population, so that we use the reference mean $\bar{X}_1$. On panel data, this mean must be taken for a specific reference year since the data values vary over time. This condition is fulfilled when the explanatory variables are constant over time. However, even in this case, the unbalanced nature of the panel may alter the mean for different estimation lags a since the set of treated individuals varies over time. By convention, one may take the structure in $t_i - 1$ since it is the first period before the treatment. This convention can be used when the variables vary over time. We obtain:

$$\mathrm{E}\left(Y_{t+a}(1) - Y_{t+a}(0) | W_{i,t} = 1, X_{i,t} = \bar{X}_1\right) = \gamma_0^1(a),$$

since $\bar{X}_r = \bar{X}_1$. The model (11) is a classic fixed effect model, so that we can use the within estimator.[27] It consists in applying the within transformation to all the time varying variables and run the OLS method. The within transformation of a variable $Y_{i,t}$, denoted $\tilde{Y}_{i,t}$, is simply its deviation from the individual mean $Y_{i,\bullet}$:

$$\tilde{Y}_{i,t} = Y_{i,t} - Y_{i,\bullet}, \text{ with } Y_{i,\bullet} = \frac{1}{T_i} \sum_{t \in \mathbf{T}_i} Y_{i,t},\ \forall (i,t).$$

Notice that the within transformation of a time constant variable is zero. In order to account for the time effects, we include time dummies in the regression.[28] Let:

$$d_{i,\tau} = \begin{cases} 1 & \text{if } t = \tau \\ 0 & \text{otherwise} \end{cases}$$
$$\tau = 1, \ldots, T$$

We also define the treatment dummies:

$$W_{i,t}^a = \begin{cases} 1 & \text{if } t = t_i + a \\ 0 & \text{otherwise} \end{cases}$$
$$a = 0, \ldots, k$$

so that the regression (11) becomes:

$$Y_{i,t} = \beta_0 + X_{i,t}\beta_1 + \sum_{a=0}^{k} W_{i,t}^a \gamma_0^r(k) + \sum_{a=0}^{k} W_{i,t}^k (X_{i,t} - \bar{X}_r)\gamma_1(a) + \sum_{\tau=2}^{T} d_{i,\tau}\beta_\tau + \alpha_i + \varepsilon_{i,t}.$$

[27] For an presentation of panel data models, see Balestra (1996a), Balestra (1996b) and the chapter 10 in Wooldridge (2002). The within estimator is also known as the "fixed effect" or LSDV (Least-Squares Dummy Variables) estimator in the literature.

[28] One of the time dummies must be excluded from the regression in order to avoid a perfect collinearity case. We choose to drop the first one by convention, without loss of generality.

Here, we should notice that the time constant parts of the model will be cancelled by the within transformation. The confounding variables are written:

$$X_{i,t} = (X_i^1, X_{i,t}^2)$$

so that only $X_{i,t}^2$ will remain after the within transformation. The individual effects α_i will also be cancelled but not the cross products of X_i^1 by the treatment dummies, because the treatment dummies vary over time. Overall, we get the following regression, to which we apply OLS:

$$\tilde{Y}_{i,t} = \tilde{X}_{i,t}^2 \beta_1 + \sum_{a=0}^{k} \tilde{W}_{i,t}^a \gamma_0^r(k) + \sum_{a=0}^{k} \tilde{S}_{i,t}^{a,r} \gamma_1(k) + \sum_{\tau=2}^{T} \tilde{d}_{i,\tau} \beta_\tau + \tilde{\varepsilon}_{i,t}$$

with $S_{i,t}^{a,r} = W_{i,t}^a (X_{i,t} - \bar{X}_r)$. The covariance matrix of the within estimator already allows for the autocorrelation involved by the within transformation itself, but we saw that the treatment model suggests that the variance could differ between the treated and the controls. The simplest solution is to account for autocorrelation and heteroskedasticity of an unknown form, following Arellano (1987).[29]

When the variables are adequately centred before to take the cross products and before to take the within transformations, the effects will be given by the estimations of the $\gamma_0^r(a)$ coefficients. If the centring is performed according to the mean of treated group, we will get the ATTs at different lags. A centring according to the not-treated group $\bar{X}_0$ will give the ATNs and the ATEs will be obtained when the centring is done according to the whole population ($\bar{X}_2$).

Regression on the propensity score. Once the propensity score has been estimated, it is possible to use it instead of the matching variables in a regression similar to the previous one. The model will just include time and fixed effects, compared to the cross section case. The covariance matrix could be obtained like in Murphy and Topel (2002) provided that the propensity score is obtained as an M-estimator with continuous second order derivatives. It may be more complex to derive than in the cross-section case. The use of the bootstrap is possible, but it will be more time consuming on panel data.

References

Abadie, A. and G. W. Imbens (2008): "On the failure of the bootstrap for matching estimators," Econometrica, 76, 1537–1557. 13, 15

Andrews, D. W. K. and M. Buchinsky (2000): "A three-step method for choosing the number of bootstrap repetitions," Econometrica, 68, 23–51. 15

Angrist, J. D. and J.-. Pischke (2008): Mostly harmless econometrics: An empiricist's companion, Princeton University Press. 23

Arellano, M. (1987): "Computing robust standard errors for within-groups estimators," Oxford bulletin of Economics and Statistics, 49, 431–434. 25

Austin, P. C. (2011): "An introduction to propensity score methods for reducing the effects of confounding in observational studies," Multivariate behavioral research, 46, 399–424. 11

[29] The method is implemented under R in the plm package, with the vcovHC function, and under SAS, with the hccme=4 option of the "proc panel".

BALESTRA, P. (1996a): "Fixed effect models and fixed coefficient models," in The Econometrics of Panel Data, ed. by L. Mátyás and P. Sevestre, Springer, 34–49. 24

——— (1996b): "Introduction to linear models for panel data," in The Econometrics of Panel Data, ed. by L. Mátyás and P. Sevestre, Springer, 25–33. 24

BARNAY, T., E. DUGUET, AND C. LE CLAINCHE (2019): "The effects of breast cancer on individual labour market outcomes: an evaluation from an administrative panel in France," Annals of Economics and Statistics, 103–126. 7

CALIENDO, M. AND S. KOPEINIG (2008): "Some practical guidance for the implementation of propensity score matching," Journal of Economic Surveys, 22, 31–72. 11

CHABÉ-FERRET, S. (2015): "Analysis of the bias of matching and difference-in-difference under alternative earnings and selection processes," Journal of Econometrics, 185, 110–123. 21

CHAMBERLAIN, G. (1980): "Analysis of Covariance with Qualitative Data," The Review of Economic Studies, 47, 225–238. 23

CLARK, R. M. (1975): "A calibration curve for radiocarbon dates," Antiquity, 49, 251–266. 14

DEHEJIA, R. H. AND S. WAHBA (2002): "Propensity score-matching methods for non-experimental causal studies," Review of Economics and Statistics, 84, 151–161. 13

DUGUET, E. AND C. LE CLAINCHE (2020): "The socioeconomic and gender impacts of health events on employment transitions in France: a panel data study," Journal of Human Capital. 10

EFRON, B. AND R. J. TIBSHIRANI (1994): An introduction to the bootstrap, CRC press. 13

FERNÁNDEZ-VAL, I. (2009): "Fixed effects estimation of structural parameters and marginal effects in panel probit models," Journal of Econometrics, 150, 71–85. 23

FRÖLICH, M. (2004): "Finite-sample properties of propensity-score matching and weighting estimators," Review of Economics and Statistics, 86, 77–90. 14

——— (2007): "On the inefficiency of propensity score matching," AStA Advances in Statistical Analysis, 91, 279–290. 11

GREENE, W. (2004): "The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects," The Econometrics Journal, 7, 98–119. 23

GROSSMAN, M. (1972): "On the concept of health capital and the demand for health," Journal of Political Economy, 80, 223–255. 2

HECKMAN, J. J., H. ICHIMURA, AND P. E. TODD (1997): "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme," The review of Economic Studies, 64, 605–654. 14

HSIAO, C. (1996): "Logit and probit models," in The Econometrics of Panel Data, ed. by L. Mátyás and P. Sevestre, Springer, 410–428. 23

IMBENS, G. W. AND D. B. RUBIN (2015): Causal inference in statistics, social, and biomedical sciences, Cambridge University Press. 2, 12, 13

LECHNER, M. (2010): "The estimation of causal effects by difference-in-difference methods," Foundations and Trends in Econometrics, 4, 165–224. 20

LEE, M.-J. (2005): Micro-econometrics for policy, program, and treatment effects, Oxford University Press. 15

MURPHY, K. M. AND R. H. TOPEL (2002): "Estimation and inference in two-step econometric models," Journal of Business & Economic Statistics, 20, 88–97. 18, 25

NADARAYA, E. A. (1964): "On estimating regression," Theory of Probability & Its Applications, 9, 141–142. 14

PAGAN, A. (1984): "Econometric issues in the analysis of regressions with generated regressors," International Economic Review, 221–247. 18

PAGAN, A. AND A. ULLAH (1999): Nonparametric econometrics, Cambridge university press. 14

ROSENBAUM, P. R. AND D. B. RUBIN (1983): "The central role of the propensity score in observational studies for causal effects," Biometrika, 70, 41–55. 11

——— (1985): "The bias due to incomplete matching," Biometrics, 103–116. 12, 17

RUBIN, D. B. (1973): "The use of matched sampling and regression adjustment to remove bias in observational studies," Biometrics, 185–203. 15

——— (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies," Journal of Educational Psychology, 66, 688–701. 2

STAMMANN, A., F. HEISS, AND D. MCFADDEN (2016): "Estimating fixed effects logit models with large panel data," Conference Paper, beiträge zur Jahrestagung des Vereins für Socialpolitik 2016: Demographischer Wandel - Session: Microeconometrics, No. G01-V3. 23

WATSON, G. S. (1964): "Smooth regression analysis," Sankhyā: The Indian Journal of Statistics, Series A, 359–372. 14

WHITE, H. (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," Econometrica, 817–838. 17

WOOLDRIDGE, J. M. (2002): Econometric analysis of cross section and panel data, MIT Press. 11, 18, 24