Causal Inference: Controlling for bias in observational studies using propensity score methods

Submitted in partial fulfilment of the requirements for
the degree

Magister Scientiae

by

Mxolisi Msibi
Department of Statistics

Faculty of Natural and Agricultural Sciences

University of Pretoria

August 2020

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

This is entirely, dedicated to my lovely daughter Ntewàa Babalwa Msibi.

- a significant pillar of all my strength -

Acknowledgements

I wish to express my deepest gratitude, first and foremost, to my supervisor, Dr Lizelle Fletcher, whose genius and availability guided me through this rough terrain. Her guidance, in clearing-up and or simplifying most of the tough concepts pertaining to this specific literature, is second to none. It was her voice that always echoed those most needed encouraging words that ensured that I remain inspired, with all three of my eyes firmly on the ball. She is also responsible for the professionalism dripping on this body of work, of which I hopefully hope to maintain and improve upon going forward, and the will to do the right thing even when the going was tough. I honestly believe that without her persistent help, and belief that we'll make it through when I felt hopeless, which was at most times, the goal of this project would not have been realized. Thank you Doc, words can't explain how grateful I am to have had this journey with you. Further, a special mention also deservedly goes to Miss Fransonet Reyneke, thanks for sharing your research data and insight with us. I'll forever be grateful to you for showing such a kind gesture.

Acknowledgements are also due to all the authors, whose work is what our research is based on for sampling, referencing, and consolidation so that this dissertation, as a whole, was created. Thanks a lot for paving the way on this tough but interesting topic. From that seminal Rosenbaum and Rubin 1983 paper to the more recent Greifer in relation to the cobalt package, I pay all my respects, in acknowledgments. The true belief that something is real and achievable comes from seeing that others have already walked the path and that you can follow in their footprint as big as they are (for a minor) to fill. I went through a great length of journal searching, looking, and definitely reading looking at all the ideas on how I can compile all of the train of thought acquired once done reading all these papers through my research process. Another special mention here is Jason A. Roy, Ph.D. of the University of Pennsylvania whose A Crash Course in Causality: Inferring Causal Effects from Observational Data course, offered through Coursera, was a real eye-opener for me in the midst of this research process. I managed to complete that course and thus gained better traction, due to the simplified knowledge acquired henceforth, in my overall research endeavor. Thanks for a well-planned and instructed course, you really simplified all that I read through the various articles that I went through as part of my bibliography to understand causal inference.

I wish to also appreciate the Statistics Department at the University of Pretoria, all the staff, the students I met and made friends with, and the wonderful lecturers that taught me some more mathematical statistics during my tenure here. The endless list have names such as Dr. Kanfer, Dr. Human, Dr. Millard, Dr. Paul van Staden, Dr. Human, and the international lecturers who also came through for module lecturing and talks, *i.e.* Prof Chen, Prof Wilson, and Dr. Jacobs. I really learnt a lot from all of your interesting topics. Not forgetting Dr. Inger Fabris-Rotelli, who seem to have been the de facto to go to person every time I had issues, with almost everything, thank you for always being available Doc.

This work also had the luxury of being critically-analyzed, in a technical sense, by 'my trusted proof-reading crew', consisting of Philani Mbhele, Nkululeko Majozi, Teboho Nkopane, and brother Thokozane 'Twoboy' Msibi, whose contribution will always be highly appreciated. I value each and every one of your contributions very highly and were all certainly not in vein. You better believe me that without any of your lost valuable sleep time, turned my correction-nights, plus the invaluable

moral support this daunting project may have taken another twist. At last, I feel relieved that I will not longer be confusing you guys and randomly throwing 'causal-inference' related pile of junk your way to read for spellchecks corrections, or at least for the foreseeable couple of months. Not forgetting to thank my fellow masters, Wanda Ndamse and Sphiwe Skhosana, for their help, where they could, when I had some LyX and LaTeX related struggles.

Lastly, I wish to forward my gratitude to acknowledge the support and great love of my family; my parents Mr. Mvelase Msibi and Mrs Mumsey Msibi; all my countless siblings plus all direct or indirect relatives, if I had too much margin I was going to list all of you by name. Not forgetting Sis Aggie Matlhadisha, Stephen Windell and everyone from the Insight Actuaries family, where it all began, your help and support from the beginning and during the course of this journey is appreciated. All the friends that kept the fire in my eyes burning though their constant motivations, during this tough three and a half years, I sincerely appreciate and love all of you. You all kept me going on, even when it felt impossible and I wanted to give-up, and believe me when I say this work would not have come to fruition without all your kind words of encouragement.

There must also have been some super-forces, or higher power, behind all the strength to keep going at this project even in the midst of the hard times I have been on lately at both professional and personal level. So it's best I acknowledge that as well.

Declaration

I, Mxolisi Msibi, declare that this mini-dissertation, which I hereby submit for the degree Master of Science in Mathematical Statistics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Name:

Signed:

Date:

Abstract

Adjusting for baseline pre-intervention characteristics between treatment groups, through the use of propensity score matching methods, is an important step that enables researchers to do causal inference with confidence. This is critical, largely, due to the fact that practical treatment allocation scenarios are non-randomized in nature, with various inherent biases that are inevitable, and therefore requiring such experimental manipulations. These propensity score matching methods are the available tools to be used as control mechanisms, for such intrinsic system biases in causal studies, without the benefits of randomization (Lane, To, Kyna , & Robin, 2012). Certain assumptions need to be verifiable or met, before one may embark on a propensity score matching causal effects journey, using the Rubin causal model (Holland, 1986), of which the main ones are conditional independence (unconfoundedness) and common support (positivity). In particular, with this dissertation we are concerned with elaborating the applications of these matching methods, for a 'strong-ignorability' case (Rosenbaum & Rubin, 1983), *i.e.* when both the overlap and unconfoundedness properties are valid. We will take a journey from explaining different experimental designs and how the treatment effect is estimated, closing with a practical example based on two cohorts of enrolled introductory statistics students prior and post-clickers intervention, at a public South African university, and the relevant causal conclusions thereof.

Keywords: treatment, conditional independence, propensity score, counterfactual, confounder, common support

Table of Contents


treatment
causal
effect
bias
difference
control
response
allocation
selection
common
outcome
scores
observed
average
matched
treated
randomized
variables
exposed
support
model
method
conditional
propensity
due
binary
covariates
subject
score
note
regression
test
probability
estimated
logistic
subjects
data
analysis
matching

List of Tables

List of Figures

1

Introduction

In instances where interventions are to be measured for their successes or failures propensity score adjustment techniques are often applied. This dissertation is concerned with the analysis of causality, which is the effect a given intervention may have, positive or negative, on some outcome. The famous 'correlation is no implication of causation' anecdote is attributed to G. A. Barnard (Holland, 1985 and 1986). In both papers, Holland talks up the philosophy behind the concept of cause and effect and how mathematics helps to identify distinguishing factors between models for causation and those for association. Since these kinds of experimental studies puts more emphasis in differentiating cause and effect from simple associations (Mariani & Pêgo-Fernandes, 2014). Good experimental design processes will allow us to mix these two worlds into one hence making it easier to reason, with a degree of confidence, about the causal inference made.

Causal models are useful in a variety of practical fields, such as social, economic, and political sciences, more specifically epidemiology in the medical sciences, and other natural sciences, in determining the effect that interventions have on responses. In their 1983 seminal article, Rosenbaum & Rubin refer to randomized trials being the "gold standard" method for experimental design, even though they are not always feasible, practically, in real-life scenario implementations due to data sets often being observational in nature. It is interesting to further note that there may also be ethical implications when randomized trials are applied on human subjects. This therefore, necessitates the need for different experimental designs such that researchers can make meaningful conclusions through causal inference. These are quasi-experimental designs, whose sole purpose is balancing observational data by controlling for some of the intrinsic biases within them so that they may eventually mimic random trials, which are discussed in Chapter 2. Study biases, common with observational data, are those of selection bias which implies that subjects may self-select to the treatment and information bias or observation bias which relates to inaccuracies in its collection, measurement, and interpretation, which will compromise both the internal validity and generalizability of the applied model (Kukull

& Ganguli, 2012). Such biases, especially in selection, tend to cause some of the baseline or pre-intervention characteristics to confound, meaning some may directly impact both treatment allocation and the outcome (Starks, Diehr & Curtis, 2009).

A common consensus with causal inference, for observational studies, is that the Rubin Causal model (Holland, 1985) mostly referred to as 'the potential-outcomes framework' model is the technique of choice. Caused by a deeply-rooted problem related to making causal inference with 'missing data', due to not being able to observe both outcomes from the same subject (Imai, 2011 and Holland, 1986) which is coined 'the fundamental problem of causal inference' by Holland, (1986). Economists often call this 'the fundamental problem of program evaluation' in their literature. In any way scientists are in agreement or admit to the fundamental nature of the requirement of the counterfactual[1] or non-observable outcome, thus requiring control mechanisms such as propensity score modeling, to address this issue. Therefore, in order to estimate, for each subject, the impact on the response with or without the intervention a counterfactual outcome is needed. This is typically done through comparing subjects from the active treatment to those with similar features from the control group. Therefore, quasi-experiments are seen as subsequent control methodologies when analyzing causality with observational data thus correcting for biases and the fundamental problem of causal inference.

Layout. In Chapter 2 the literature review introduce the reader to, the different kinds of, experiment-designs, *i.e.* randomized control trials (RCTs) and quasi-experimental designs (QEDs). The latter is, typically, used for real-life observational studies in order to control for the mentioned biases that are innate within its data form, and hence help imitate random trial behavior. Comparisons of these designs, in relation to their advantages and disadvantages, are done together with the mathematical definitions of the efficacy and or effectiveness of interventions.

Propensity scores are defined and then introduced, before being thoroughly discussed in the following Chapter 3, as the measures needed, to be estimated, so that matching techniques can be applied. Chapter 3 continue expanding on the theory behind adjusting with propensity scores, together with the introduction of causal assumptions or properties required, the mathematics behind them, the different kinds of matching methods, and finally the typical steps navigated when matching is applied.

Application. Propensity score matching applications, in a practical scenario through the mentioned steps of Chapter 3, is the sole purpose of Chapter 4. For this dissertation, two cohorts of an introductory statistics course enrollees at the University of Pretoria

[1]Counterfactual refers to what could have been on subjects allocated treatment or intervention but never partake

(prior and post a clickers intervention) is utilized. The university implemented a classroom response system, referred to as clickers, device intervention to improve student interaction, amongst them and with their instructors, ease of accessing assignments and collaboration at home or in class, and or sharing course-related knowledge online, for this module. With the ultimate aim being to ensure students score higher marks and eventually better odds at success or progressing. Further, the data collection mechanisms of these devices allow lecturers to consolidate from the real-time information analysis and pinpoint where the challenge is for the rest of the class. Of utmost importance, to note, here is that these students were not enrolled at random to either group and so we have an observational study.

Our cohorts are the class of 2014, or the control group, and that of 2017, which we refer to as the treatment group from here on, exclusively. Their respective sample sizes are 1,625 in the control group and 1,486 in the treatment group. Propensity score adjustment techniques are applied, three of them in addition to first starting with an unadjusted case, for causal inference analysis. These methods are greedy matching with a caliper, optimal matching, and inverse probability of treatment weights. All these were consistent, in terms of their individual outcome model estimates, with some slight differences on the estimated causal risk difference of which was much bigger for the greedy method, between the two groups for this data.

2

Literature review

For experimental-design studies, estimating the causal effect requires a researcher to hypothesize the success of some intervention on treated subjects (Rosenbaum & Rubin, 1983 and Dehejia & Wahba, 2002). Experimental designs are, therefore, devised as means to somehow simplify, and observe these effects on the outcome (response) variable whenever thorough examination was conducted. This is usually done through thorough manipulation of the independent variable(s), via the intervention, and examining their effects on the response (dependent) variable. The objective is that of correcting or controlling for any imbalances between the treatment groups, *i.e.* exposed and control, whenever randomization was not applicable.

Rosenbaum & Rubin (1983) warn that researchers need to be wary of selection bias and confounding variables[2] when working with observational data. This awareness being of crucial importance since selection bias and confounders are an inevitability of such a process. The coin-toss nature of randomized experiments will self-correct for such problems by giving each subject (or observation unit) an equal chance at allocation (Austin, 2011). This in turn allows for covariates in both the treatment groupings to be evenly distributed, hence achieving the balance required for analysis. Therefore, the ability to distinguish between randomized control and non-randomized experiments is an important skill for a mathematical statistician and or data scientist to master. These skills aid the relevant researcher when identifying the experimental-paradigm nature of given data, and hence swiftly finding the right analytical tool that fit the specified problem design.

Rosenbaum, & Rubin (1983); Dehejia & Wahba (2002) and Thavaneswaran & Lix (2008) argue that a randomized control trial (RCT) is an ideal exercise for data mining, more so for experimental design problems. Not all data will, however, adhere to its required set-up in practical scenarios. For instance, as explained above, selection bias will violate the independence assumption of the response in observational data scenarios, due to the

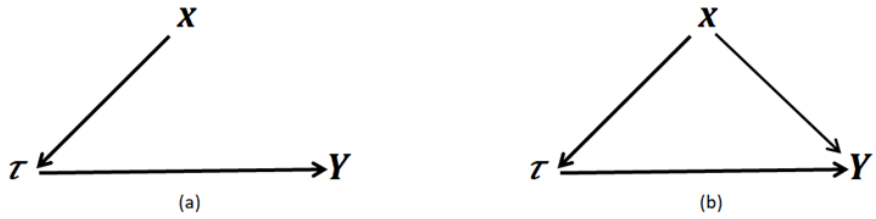
[2]Confounders are those variables that the researcher failed to control for, compromising the internal validity of an experiment

inevitability of confounders or confounding variables existing. Confounder variables tend to affect both the treatment and the outcome, refer to Figure 1, and the basic experimental design requirement is to control for them. Controlling for confounders ensure that a quasi-experimental design (QED) process, which will almost behave randomly, is achieved (Heinrich, Maffioli, & Vázquez, 2010).

Causality using graphs

Directed acyclic graphs (DAGs) are typically used to depict relations between covariates, treatment status and the outcome variables. Their innate character is that they are set of vertices (or nodes) linked through pointed arrows (or a set of edges) for causal direction (Imbens, 2019 and Scheines, 1997). For example Figure 1 shows a pair of DAGs that indicate two possible causal scenarios, *i.e.* confounded and non-confounded cases.

Graphic representation of a confounder variable, on the right. The ideal required situation is the one on the left, i.e. (a) Ideal case where covariate $\boldsymbol{X}$ impacts only the treatment $\boldsymbol{\tau}$ (b) Unwanted case where $\boldsymbol{X}$ is a confounder impacting both the treatment $\boldsymbol{\tau}$ and outcome $\boldsymbol{Y}$



(Source: Own elaboration)

Figure 1 Directed acyclic graphs (confounder & non-confounder)

Following the notation from Scheines, (1997), Figure 1 can be interpreted as follows:

Given the sets of three vertices and two edges respectively $\{X, \tau, Y\}$ and $\{X \rightarrow \tau, \tau \rightarrow Y\}$, the diagrammatical nature of our DAG representation is $\{X \rightarrow \tau \rightarrow Y\}$, where the treatment status is given by τ, the confounder X, and the outcome Y, respectively.
Similarly, for the same set of three vertices $\{X, \tau, Y\}$ and three edges $\{X \rightarrow \tau, X \rightarrow Y, \tau \rightarrow Y\}$, the DAG may be represented as follows $\{Y \leftarrow X \rightarrow \tau \rightarrow Y\}$. In this case X is an 'observed confounder', meaning that causal inference is inadequate unless this confounded effect is controlled for (Imbens, 2019).

The concepts of observational and randomized experiments will be discussed in more detail, Section 2.1. Specifically, Subsection 2.1.1 and 2.1.2 explains the theory behind randomized control trials (RCTs) and quasi-experimental designs (QEDs) experimental design paradigms, together with their advantages and disadvantages, as adapted from Thavaneswaran and Lix (2008). The efficacy or impact of an intervention or a treatment will be expanded on in Section 2.3, more so the math behind the process, as referred to the article by Heinrich, Maffioli & Vázquez (2010). As stated above, the main idea, behind these techniques is that of controlling for confounder variables through minimizing both overt, and to some extent covert, bias in non-randomized trials or observational studies. Therefore in Section 2.4 we introduce the technique of propensity score matching that helps control for confounding in observational studies.

2.1 Experimental designs

It is accepted that randomized control trials (or RCTs) are the "gold standard" methodology in estimating causal effects on the response (outcome) variable (Rosenbaum & Rubin 1983 and Austin 2011). Such experiments allow for a random mechanism to drive the allocation (or assignment) of subjects to treatment groups (Heinrich et al., 2010). Furthermore, Austin (2011) emphasizes that random treatment allocations should eliminate the possibility of baseline features confounding since all participants have an equal chance of being selected into either treatment group. In empirical quasi-experimental designs, however, subjects have options to self-select and normal standard research procedure is controlling for selection bias 'prior' to estimating the treatment effect (Rosenbaum & Rubin, 1983). And so, comparing systematic differences between treatment groups through techniques like propensity score matching (PSM) may be useful (Lane et al., 2012).

Possible biases with non-randomized designs

As explained above, non-random selection of subjects may lead to variables that confounds and selection bias. You may balance for this by using a propensity score matching algorithm. This is because segmenting on these scores will ensure that comparisons are on an equitable basis, “apples with apples”. The implication is that these comparisons will be done between groups exhibiting similar characteristics, except for the treatment. Austin (2011) and Rosenbaum & Rubin (1983) allude to ”a steady growth”, witnessed over the years, in interest from researchers for quasi-experimental design data analysis techniques. But despite this increase in the popularity of quasi-experimental designs, Rosenbaum and Rubin (1983) and Austin (2011) note that researchers using these models still face unavoidable complexities when interpreting response causal effect.

2.1.1 Randomized control trials

Intra-individual variation is a possibility with experiment designs leading to differences on the baseline covariates (Stuart, Bradshaw & Leaf, 2015). A most likely scenario with random trials is that each participant or subject will get an equal chance to be treated (Rosenbaum & Rubin, 1983). Meaning that, we thus have inherent advantages such as:

- subjects being matched equally on all characteristics (covariates),
- inference on the causal effect can easily be made, and
- that the balancing on both the observed and unobserved features is conditional.

Random allocation ensures that one is exposed to the treatment independent of both their observed and non-observed covariates (Heinrich, Maffioli & Vázquez, 2010). This result in an even match on all of the characteristic features (covariates) and therefore causal inference can be made with confidence. Maintaining an ideal situation where robustness of the inference gets better with sample size increases. For a binary treatment experiment, this implies independence of the responses against the treatment status (z),

$$(Y_1, Y_0) \perp\!\!\!\perp z \,\big|\, \boldsymbol{X} \tag{1}$$

where (Y_1, Y_0) are the respective responses or outcomes for treatment and control groups,

z denotes the treatment group or status

$\perp\!\!\!\perp$ is the independence symbol emphasizing *strong-ignorability*

$\boldsymbol{X}$ denotes the subjects characteristic features, and

$z \,\big|\, \boldsymbol{X}$ indicates that the treatment was allocated conditioning on the observed covariates

Heinrich, Maffioli & Vázquez, (2010) explain that, proportion-wise, Equation (1) suggests that each subject encompasses of evenly distributed features, or covariates, between both treatment groups. Thus implying that the two treatment groups are equal on average, mathematically written as,

$$\mathbb{E}(Y_0 \Big| treated) = \mathbb{E}(Y_0 \Big| control) \Rightarrow \mathbb{E}(Y_0 \Big| z = 1) = \mathbb{E}(Y_0 \Big| z = 0) \tag{2}$$

and interchanging the counterfactual, in other words what would have happened to the treatment group without administering treatment, on the left-hand side of Equation (2) with the observable on the right-hand side allow for an easier estimation of τ_{ATT}, the average treatment effect on the treated (ATT). The counterfactual on the left hand side of Equation (2), above, is the factual response that a subject is non-treated even though it was allocated the treatment (Holland, 1986 and Imai, 2011).

With such randomness in the allocation an advantage of a zero selection bias is ensured, meaning that the treatment effect will just be the difference in expected responses between the treatment and control groups (Heinrich, Maffioli & Vázquez, 2010). And so, as a consequence for random trials, an ordinary least-squares (OLS) regression on the treatment allocation (z) variable and some constant term (α) will be adequate when estimating the efficacy of the intervention program (Heinrich, Maffioli & Vázquez, 2010). That is,

$$Y = \alpha + \beta z + \varepsilon \tag{3}$$

where β - represents the efficacy (or impact) of the treatment program, and ε - are the error (residual) terms.

2.1.2 Quasi-experimental designs

In practice, however, allocation to treatments will most likely be of non-random nature (Heinrich, Maffioli & Vázquez, 2010). The non-randomness causes selection bias and thus compromises the control group, making it weaker to the treated group, Heinrich, et. al. (2010). Listed below, in bullet form, are some of the benefits with quasi-experimental designs (QEDs):

- they control for confounding and extraneous variables , these are any variables that you were not intentionally studying in your experiment or test,
- they tend to require less resources than randomized control trials (RCTs) would, Schweizer, Braun & Milstone (2016),
- QEDs are also used when adjusting for the estimate of the treatment effect in non-random cases, and that

- they are pragmatic, or tends to represents real-world (or practical) problems.

Equation (1), above, introduced the mathematical notation for denoting the expected responses (outcomes) related to subjects in either treatment group. That is, the response Y for each treatment exposure level is given as:

$$Y = \begin{cases} Y_1, & if \quad exposed \\ Y_0, & otherwise \end{cases} \tag{4}$$

The indicator for the observed subjects, in this binary case, is z. The observed response is thus given by

$$Y = zY_1 + (1 - z)Y_0 \tag{5}$$

For a subject in the treatment group it will be

$$Y = 1 \cdot Y_1 + (1 - 1) \cdot Y_0 = 1 \cdot Y_1 + 0 \cdot Y_0 = Y_1$$

and similarly, for the control group subjects

$$Y = 0 \cdot Y_1 + (1 - 0) \cdot Y_0 = 0 \cdot Y_1 + 1 \cdot Y_0 = Y_0.$$

As noted in Section 2.1.1, with randomized trials the responses together with the influential characteristics will be independent of the treatment exposure (Heinrich, Maffioli & Vázquez, 2010). Without random allocations, however, a relationship may exist between the treatment and these influential features of the responses, Y_0, Y_1, and the simple mean treatment effect difference between the groups, as in the randomized case, won't suffice (Heinrich, Maffioli & Vázquez, 2010). This is due to the fact that exposed subjects may be distinct to those in the control group, in spite of the treatment effect, therefore, rendering the simple mean response difference useless in evaluating the causal effect, Heinrich, Maffioli & Vázquez, (2010).

In Section 2.3 below, the effectiveness or impact of the treatment is examined and one of the solutions, propensity scores with focus on propensity score matching (PSM), to help curb the above-mentioned correlation problems is suggested in Section 2.4. The PSM method, which is also the main focus of application for this dissertation will help minimize, and possibly eliminate, the above-mentioned possible biases. It will be discussed in further detail in Section 3.2, expanding from the introduction to propensity scores in subsection 3.1.2.

2.2 Disadvantages

Both randomized control trials and quasi-experimental designs seem to have shortfalls in some aspects. Some of the problems that one may encounter when applying either of the

two designs are discussed below.

2.2.1 Disadvantages with RCTs

Possible problems with randomized trials are (Stuart, Bradshaw & Leaf, 2015):

(1) cumbersomeness and or the costliness in terms of execution,

(2) feasibility may not always be accomplished, due to one or more reasons and or concerns, since the results can often not be easily replicated on humans due to ethical concerns, such concerns range, amongst others, from subject consenting to the study and or parental proxy for minors; due to the possibility of subjects risking lives with clinical trials, and

(3) that at most the design may end up not being generalizable or easily replicated on human subjects.

Other disadvantages of randomized trials are, if studies were conducted in a natural habitat for subjects, such as in a school or prison, imposition of control on the extraneous variable[3] will be near impossible. This difficulty is attributed to complications in getting cooperation from study subjects due to their perceived feeling pertaining to the intervention being administered, and the measurement of intra-individual human characteristics such as personal feelings and or emotional well-being. Some of these are due to the fact that, sometimes, subjects will need to consent for treatment in randomized trials, and that in particular, as noted by Stuart, Bradshaw & Leaf (2015), subjects who agreed to participate at the start exhibit distinct characteristics to those that will jump on the bandwagon, later-on, once the efficacy of the program is evident.

2.2.2 Disadvantages with QEDs

Despite its increasing use, QEDs have a number of glaring shortcomings. Most notably, some adapted from Schweizer, Braun & Milstone (2016), these drawbacks includes:

(1) the inevitability of selection bias, as stated in Section 2.1.2, due to lack of random selection, which leads to

(2) the struggles in making an assumption on causal inference between the treatment and response,

(3) not having control on the influence extraneous variables have on the response, and that

[3] An extraneous variable represents any covariate that is not the independent factor under investigation for outcome effect (McLeod, 2019).

(4) the reasoning on causal inference is weaker compared to randomized experiments.

As a result, Rosenbaum & Rubin (1983) note that QED studies will, unlike those with randomized control trials, require that the probability of treatment allocation be estimated since it is unknown. This can be explained, in a Bayesian sense, to imply that the 'posterior' probabilities are estimates of the probabilities of exposure conditioned on the baseline covariates. The concept(s) relating to response efficacy on non-random (observational) studies is broken down further in Section 2.3, below.

2.3 Defining the influence of the treatment

In this part we detail the effectiveness of an intervention or treatment on the expected response, *i.e.* the potential outcome, as adapted from Heinrich, Maffioli & Vázquez (2010). The treatment effect is defined as the difference on responses between exposed subjects and those in the control group, mathematically given as below for any subject i:

$$\delta_i = Y_{1i} - Y_{0i} \tag{6}$$

for $i = 1, 2, , \ldots, n$, where $n = n_\tau + n_C$ is the sum of all available subjects, where

n_τ counts the subjects that were exposed to the treatment

n_C the count of subjects that remained for control.

In general, Heinrich et al. (2010) refer to the following as measures related to all non-observable parameters (or parameters based on counterfactual responses):

The first measure is the expected (average) treatment effect (ATE), and we estimate it from Equation (6) as;

$$\tau_{ATE} = \mathbb{E}(\delta) = \mathbb{E}(Y_1 - Y_0) \tag{7}$$

A second measure is that of the average treatment effect on the treated (or ATT);

$$\tau_{ATT} = \mathbb{E}(Y_1 - Y_0 \Big| z = 1) \tag{8}$$

We get a third measure, the average treatment effect on the unexposed (ATC) as;

$$\tau_{ATC} = \mathbb{E}(Y_1 - Y_0 \Big| z = 0) \tag{9}$$

Now, using the fact that the average of a difference is the difference in averages, the ATT in Equation (8) can be rewritten as:

$$\tau_{ATT} = \mathbb{E}(Y_1 \Big| z = 1) - \mathbb{E}(Y_0 \Big| z = 1), \tag{10}$$

where $\mathbb{E}(Y_0|z=1)$ is the 'non-observed' (counterfactual) average response for treated subjects that were unexposed. This is a direct consequence of the "Fundamental Problem of Causal Inference" (Holland, 1986). Drichoutis, Nayga, Jr. & Lazaridis (2009) advise against trying to estimate the counterfactual using $\mathbb{E}(Y_0 \Big| z=0)$, since this is what give rise to self-selection bias. For the 'observed' response of untreated subjects Y_0, it is given by $\mathbb{E}(Y_0 \Big| z=0)$. Thus, we have that:

$$\Delta = \mathbb{E}(Y_1 \Big| z=1) - \mathbb{E}(Y_0 \Big| z=0), \tag{11}$$

where Δ is the average difference between the outcomes of treated and untreated subjects. Manipulating the right-hand side of Equation (11) by adding and subtracting $\mathbb{E}(Y_0|z=1)$, will give us the following difference between Δ and the ATT:

$$\begin{aligned}
\Delta &= \mathbb{E}(Y_1 \Big| z=1) - \mathbb{E}(Y_0 \Big| z=0) + \mathbb{E}(Y_0 \Big| z=1) - \mathbb{E}(Y_0 \Big| z=1) \\
&= \mathbb{E}(Y_1 \Big| z=1) - \mathbb{E}(Y_0 \Big| z=1) + \mathbb{E}(Y_0 \Big| z=1) - \mathbb{E}(Y_0 \Big| z=0 \\
&= \mathbb{E}(Y_1 - Y_0 \Big| z=1) + \mathbb{E}(Y_0 \Big| z=1) - \mathbb{E}(Y_0 \Big| z=0) \\
&= \tau_{ATT} + \{\mathbb{E}(Y_0 \Big| z=1) - \mathbb{E}(Y_0 \Big| z=0)\}
\end{aligned}$$

i.e.

$$\Delta = \tau_{ATT} + \{\mathbb{E}(Y_0 \Big| z=1) - \mathbb{E}(Y_0 \Big| z=0)\} \tag{12}$$

On the right hand side (RHS) of Equation (12), $\{\mathbb{E}(Y_0 \Big| z=1) - \mathbb{E}(Y_0 \Big| z=0)\}$ is the 'selection bias'. This 'selection bias' is given as the difference between the counterfactual for exposed or treated subjects and the observed response (Heinrich, Maffioli & Vázquez, 2010). Or shortly, it represents differences between the two groups in the scenario when the program was not administered (Zaga Szenker, 2015). We therefore re-write Equation (11) as follows:

$$\Delta = \tau_{ATT} + \quad selection \quad bias \tag{13}$$

Some important notes, pertaining to the above '*selection bias*' term, are that a case of selection bias approaching zero will imply that the estimate for the ATT is just the difference between the mean of the observed responses between the two groups (Heinrich et al., 2010). However, it is further argued by Heinrich et al. (2010) that consequences of a non-zero selection bias, as is usually the norm in practice, are a biased estimator for the ATT from the difference in Equation (11), that is $\hat{\tau}_{ATT} = \mathbb{E}(Y \Big| z=1) - \mathbb{E}(Y \Big| z=0)$. Therefore, the real purpose of a propensity scores related study is to control for a selection bias equal to zero in order to correctly estimate the causal parameter of interest.

2.4 Propensity score analysis

The probability that a subject gets allocated to the treatment, or intervention under investigation, given a set of pre-exposure characteristics is referred to as the propensity score (Rosenbaum & Rubin, 1983). Below we explore some specifics related to the applications of propensity-scores based methods for causal inference. We cover why we need propensity scoring based techniques, the intrinsic mechanics with one of them *i.e.* matching, and the ideal scenario for their application in practice, bearing in mind possible biases with observational data experiments and the fundamental problem with causality, as discussed above.

2.4.1 Need for propensity scores

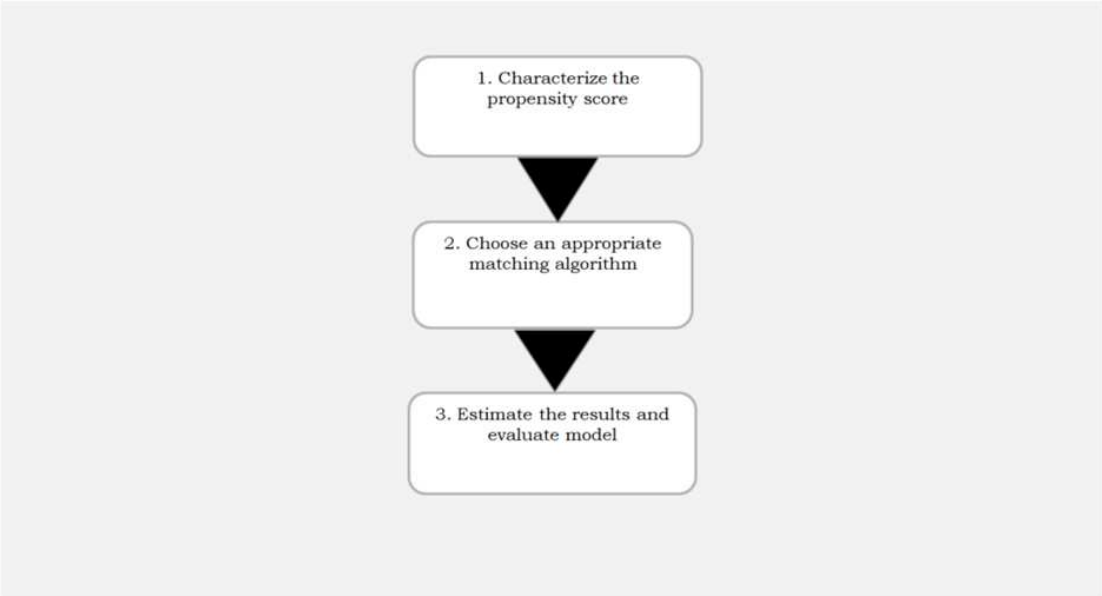
Whenever randomized allocation processes are not possible, which is often the case in practice, such as it is with social, education, socio-political, economic, biostatistics, epidemiological and or other medical science data analysis projects, matching via propensity scores is used in order to control for possible confounding and treatment effect biases (Lane, To, Kyna & Robin, 2012). This is because such studies are, in their nature, response-based. Thus implying that analysis projects of this nature are concerned with determining the effect of some sort of intervention, or treatment, such as how a change in political regime may have influenced the economy of a country, or the impact some medication had on patients in curing or controlling for a certain ailment. Heinrich et al. (2010), give an example of studying the efficacy of introducing incentives for teachers on improving pupils' pass rate (performance) and another of studying an impact of a Honduran youth skill training programme on employment prospects of undereducated children from poor backgrounds.

Matching mechanics

Figure 2, below, is a diagram depicting basic generic parts, as explained in the article by Heinrich et al. (2010), required in the practical implementation of propensity score matching algorithms. The next chapter expands more, in detail, on these phases of propensity score matching and will expanding the third part pre and post with internal diagnostic steps of covariate balance and hidden-bias analysis.

(1) The first step is obtaining the propensity score, through a classification model of choice, checking for confounding variables and estimating subjects' propensity scores.

(2) The second step is choosing the matching algorithm to be used. This requires precise considerations on key parameters and the bias-variance trade-off, *i.e.* the trade-off in the efficiency and bias of the estimator of interest.



(Source: Own elaboration)

Figure 2 Propensity scores analysis (typical implementation phases)

Selecting a matching algorithm is purely dependent on the bias and variance compromise for that specific dataset (Heinrich et al., 2010). For instance, using only the nearest neighbour method will tend to cause higher variance, since only the control subjects that most resemble (or are identical to) the treated ones are likely to be matched in the construction of the counterfactual, thus causing a definite reduction in the bias hence excluding a number of control group items, Heinrich et al. (2010). An application of multiple nearest neighbour algorithms, however, will evidently allow for an exploitation of more control subjects by the estimator hence increasing both efficiency, due to lower variance, and bias (Heinrich et. al., 2010; Caliendo & Kopeinig, 2005).

It is important to note that Heinrich et al. (2010) argue further that when matching against a lot of neighbours, the expected increase in the efficiency of the study will still come with a high bias cost due to possible mismatching.

(3) The final, and third step, is that of result estimation and then the evaluation of the impact of the intervention or treatment via the propensity scores model. Heinrich, Maffioli & Vázquez, (2010) present the evaluation of the intervention's impact, done through computing the average treatment effect on the responses between exposed and control subjects, based on the chosen matching algorithm.

2.4.2 Uses of propensity scores

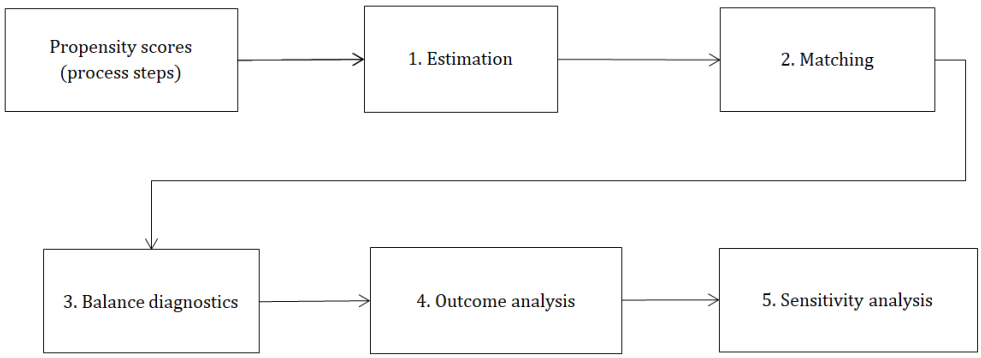
The theory of propensity scores, its estimation and that of the propensity score matching (PSM) method, with applications, will be discussed in more detail in sections 3.1 and 3.2 of this dissertation. Rosenbaum & Rubin (1983) mention that the main objective of propensity score matching is to emulate, or mimic, random trials also known as the true experimental design, when the main assumptions of conditional independence (unconfoundedness) and common support are satisfied. Unconfoundedness is an assumption that none of the observed or baseline factors are impacting both outcome and treatment allocation (de Luna & Lundin, 2009). Whenever these two main assumptions are met, the process is referred to be 'strongly ignorable' (Rosenbaum & Rubin, 1983).

Strong ignorability and 'strongly ignorable' are interchangeable and so are the assumptions conditional independence and common support with unconfoundedness and positivity (or overlap) respectively. We'll discuss these two matching assumptions, which validate strong ignorability when both met, in Section 3.2.1. Specifically, the propensity score or the treatment allocation probability is estimated using classification tools such a probit or logit (logistic regression) model, bagging and or boosted regression trees (CART), *i.e.* random forests (Thavaneswaran & Lix, 2008). Note that, logistic regression is said to be the most propitious one for dichotomous treatment allocation cases (Austin, 2011) and therefore it is the method usually applied.

3

Theoretical background

In this chapter, the theory related to the applications of non-randomized control trials, particularly propensity score matching (PSM), is presented. Different matching tools or methods for matching using propensity scores are discussed, as well as the logit regression model, typically used as the technique of choice to estimate these scores in the first place. Despite growth in popularity, Rosenbaum & Rubin (1983) note that PSM is still a generally underutilized method in practice. The focus of this part of the dissertation is discussing the steps in formulating matching via propensity scores for a binary treatment study, see Figure 3. In order for us to be concise, in explaining the process in the applications of propensity score matching, our research follows these steps with each for the subsequent subsections of interest such that the process is easier for the reader to follow. Figure 3 is a depiction of these steps, and the relevant sections where they will be detailed is 3.1 - 3.5.



(Source : Harris, H & Horst, S.J.. (2016). Slight adaptations)

Figure 3 Steps in propensity score analysis

3.1 Estimation

As mentioned in Section 2.4.2, logistic regression is one of the classification tools that is applied when estimating propensity scores. The logistic regression algorithm, its theory, and how it connects to propensity scores are detailed below.

3.1.1 Logistic regression

Logistic regression is an algorithm that gets utilized for binary classification experiments, for example our dichotomous treatment problem here, where one group of subjects are exposed and another group is the control. It is thus referred, along with discriminant analysis, to be the method of choice for categorical response problems whenever one seeks the relationship between the predictors and the response probability (Hair, Jr., William, Black & Babin, 2010, p. 315). Logistic regression is a statistical algorithm that gives the output as the log odds of an event occurring, in our case the likelihood that a subject gets allocated the treatment.

Binary logistic regression

Accordingly, Hair, Jr. et al., p. 316, explains that logistic regression aims to address two kinds of specific research questions:

- pointing at impacts of the covariates on the response variable, and
- classify subjects into the class they belong.

For a binary classification problem, the algorithm separates the data space into two distinct groups, to allot the data points or subjects, linearly through some boundary (Joglekar, 2015). The logit learning algorithm and the subsequent input data will determine the form of such a boundary. Hair, Jr. et al. (2010), p. 315, define a logistic regression model, in its general form, as an algorithm whose sole purpose is to estimate the relationship between a single nonmetric (binary) response and a mixture of numeric and categorical predictor variables.

Comparing logistic regression to linear regression

Linear regression often fails to deal with outliers in binary response experiments since it can misclassify treatment class subjects as non-exposed (Agrawal, 2017 and Joglekar, 2015). Further, since our target variable Y is binary, we define, for some baseline covariate set $\boldsymbol{X}$, $p = P(Y = 1 \,|\, \boldsymbol{X})$ as the likelihood that a subject gets the allocated treatment. Assuming that this probability can be estimated via linear regression, we will have that

$$p = P(Y = 1 \,|\, \boldsymbol{X}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon, \tag{14}$$

Note that the range of the above linear model is infinite and consequently violates the probability property that a true probability ranges from zero to one, *i.e.* $0 < p < 1$, as per the expected predictions (Hair, Jr. et al. 2010, p. 320). The general consensus, in order to avoid the above mentioned problems especially the one relating to outlier cases, is that one may take log of the odds to construct the sigmoid or logistic function (Hair, Jr. et al.2010, p. 320). In essence, this is accomplished through the following steps:

Odds

The odds of allocation to treatment is

$$odds = \left(\frac{p}{1-p}\right) \Rightarrow odds \cdot (1-p) = p$$

$$\Rightarrow odds - p \cdot odds = p$$

$$\Rightarrow odds = p + p \cdot odds$$

$$\Rightarrow odds = p \cdot (1 + odds)$$

$$\Rightarrow p = \frac{odds}{1 + odds}$$

Log-odds

Given the logistic regression model as

$$ln\left(\frac{p}{1-p}\right) = \boldsymbol{X}\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \boldsymbol{\varepsilon}. \tag{15}$$

$ln\left(\frac{p}{1-p}\right)$ is the logit function or the log-odds function of interest. Therefore, the above linear predictors are mapped into the log-odds of the response being 'classified 1' or allocated the treatment (Agrawal, 2017).

Sigmoid function

Taking the inverse of the log-odds function gives the, S-shaped, sigmoid or logistic function. This function always maps the linear combination of predictors to some true response probability, $p \in (0, 1)$. This is due to the logistic function being asymptotic at these two extreme points (Agrawal, 2017). We then take exponents on both sides of Equation (15) and it evaluates to

$$p = \frac{e^{\boldsymbol{X}\boldsymbol{\beta}}}{(1 + e^{\boldsymbol{X}\boldsymbol{\beta}})}, \tag{16}$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$, where $\boldsymbol{X}$ is the design matrix of dimension size np, and$\boldsymbol{\beta}$ is a column vector of size p. That is,

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & . & . & . & . & x_{1p} \\ x_{21} & x_{22} & & & & & x_{2p} \\ . & & . & & & & . \\ . & & & . & & & . \\ . & & & & . & & . \\ . & & & & & . & . \\ & & & & & . & \\ x_{n1} & & . & . & . & . & x_{np} \end{pmatrix} \text{ and } \boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_p]^T.$$

Letting $\eta = X\boldsymbol{\beta}$ and multiplying the numerator and denominator of Equation 16 by $e^{-\eta}$ gives the defined sigmoid curve, for treatment allocation probabilities in Figure 4, as $S(\eta) = \frac{1}{(1+e^{-\eta})}$.

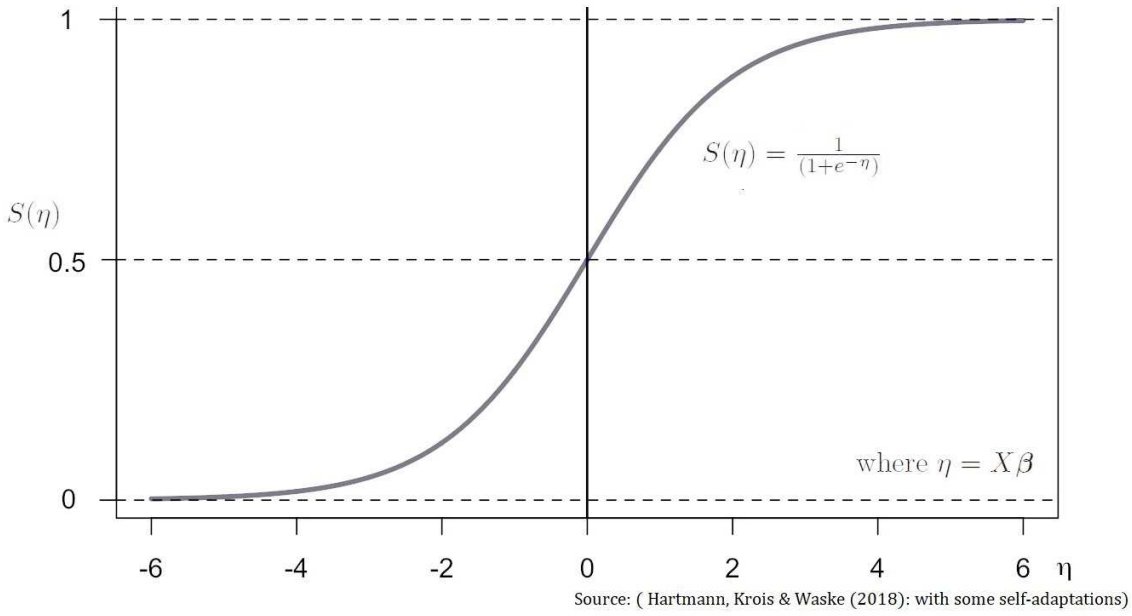


Figure 4 Sigmoid function maps the probability of treatment into interval (0,1)

Logistic regression formulation and explanation

Consider the probability that a data unit or subject is classified into the treatment class, denoted $P(Y = 1)$. The following three steps will explain the logic behind the logistic regression algorithm;

Step 1

Evaluate the boundary function, *i.e.* the log-odds function,

$$log\left(\frac{p}{1-p}\right) = \boldsymbol{X}\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \boldsymbol{\varepsilon},$$

Step 2

The odds ratio, which gives the probability of success over failure, is evaluated through $OR(X) = e^{X\boldsymbol{\beta}}$. Noting that, $X\boldsymbol{\beta} = ln(OR(X))$.

Step 3

The probabilities are then obtained by,

$$\begin{array}{rcl} P(Y=1) & = & \frac{OR(X)}{(1+OR(X))} \\ \Rightarrow P(Y=1) & = & \frac{e^{X\boldsymbol{\beta}}}{(1+e^{X\boldsymbol{\beta}})} \\ \Rightarrow P(Y=1) & = & \frac{e^{X\boldsymbol{\beta}}}{(1+e^{X\beta})} \cdot \left(\frac{e^{-X\boldsymbol{\beta}}}{e^{-X\boldsymbol{\beta}}}\right) \\ \therefore P(Y=1) & = & \frac{1}{(1+e^{-X\boldsymbol{\beta}})} \end{array} \tag{17}$$

These probabilities are the propensity scores or ,in our case, the treatment allocation probabilities.

Estimating logistic regression coefficients

Maximum likelihood estimation (MLE), instead of ordinary least-squares (OLS) estimation, is used when one estimates logistic regression coefficients (Hair, Jr. et al. 2010, p. 322). With the MLE method the advantage is a set of coefficients that maximizes the probability of the event being observed from our data (Agrawal, 2017). Given a binary data experiment, where the probability of being allocated treatment is p, that is;

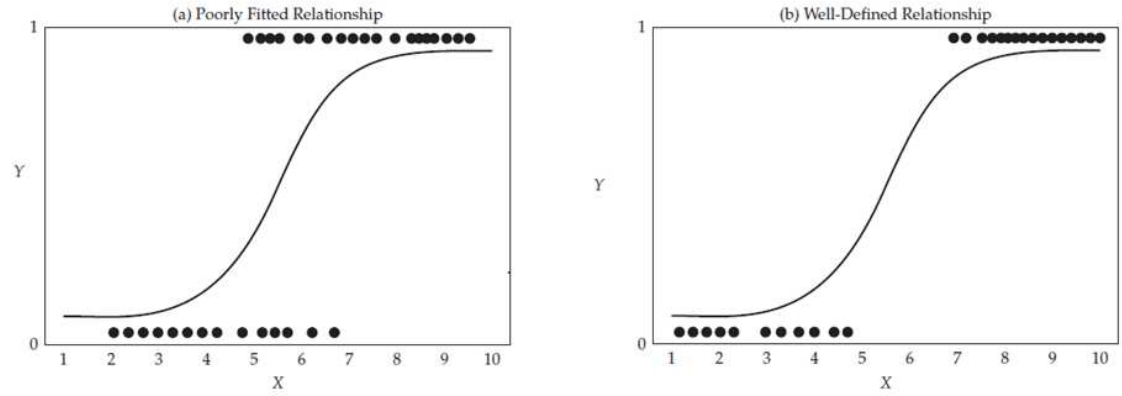
$$P(Y=1\Big|\boldsymbol{X}) = \begin{cases} p, & if \quad exposed \\ 1-p, & otherwise \end{cases}$$

the likelihood function is defined as:

$$\mathfrak{L}(\boldsymbol{\beta}, y) = \prod_{i=1}^{N} \left(\frac{p_i}{1-p_i}\right)^{y_i} \cdot (1-p_i). \tag{18}$$

Taking the log of the likelihood in Equation (18) and differentiating with respect to the parameter estimate of interest then setting this derivative to zero provide optimal parameter estimates. One may also accomplish this through iterative optimization methods such as Newton-Raphson when evaluating this maximum of the log-likelihood (Agrawal, 2017).

Logistic regression diagnostics



(Source: Hair et al. (2010) pg. 320)

Figure 5 Logistic regression comparisons of bad (left) and good (right) fits

Figure 3 (Hair, Jr. et al. 2010, p. 320) depicts graphical representations of good vs bad binary logit fits. These two types of fits are part (a) the case when the logistic curve was not a perfect fit to the data, whereas part (b) represents a well-defined curve. Deviance measures are utilized to measure the goodness-of-fit of a logistic regression model (Hair, Jr. et al. 2010 p. 323). Such measures are useful in monitoring for overlaps and or perfect curvature for bad and good fit logistic regression. Hair, Jr. et al. (2010), p. 323 discuss how a poorly fit model will tend to have higher value for the deviance metric, meaning that the researcher may still have room to improve the model. With logistic regression we'll use two types of deviance metrics, *i.e.* the null deviance, and model deviance.

Model accuracy is typically measured through the use of confusion-matrices or a "classification matrix" for binary cases (Hair, Jr. et al. 2010 p. 324). A confusion matrix is a typical two-by-two tabulation, also called a contingency table, for all possible combinations of actuals against predicted values. Table 1, below, depicts the confusion matrix. It is used when calculating model performance measures, also known as evaluation metrics, such as Precision, Recall, and Accuracy, Specificity and Sensitivity (Matsumoto & Del-Moral-Hernandes, 2013).

Table 1: Confusion matrix or contingency table (actual v. predicted)

		Predicted $\hat{y}$	
		0	1
Actual y	0	TN	FP
	1	FN	TP

where TP= true positive classification of subjects; TN= true negatives; FP= false postives; and FN= false negatives

Deviance measures

The deviance is defined as the likelihood difference between some base (or null) model and the logit fit, and is comparable to a multivariate modeling F test (Hair, Jr. et al. 2010, p. 323). The perfect model will have a likelihood equaling a unit and thus the deviance will be given as

$$deviance = -2ln\{\mathfrak{L}(\boldsymbol{\beta})\} = -2LL. \tag{19}$$

As a consequence, the deviance for a perfect fit is the minimum when evaluating the null and other proposed models, Hair, Jr. et al. (2010), p. 323. Further, the authors Hair, Jr. et al., have noted that this comparison test is an iterative stepwise, parameter addition, process where conclusions of improvements of the fits are made if the null model deviance is greater than that of the model. In effect, a model with the highest hit ratio (or accuracy) will have a model deviance very closer to zero to confirm its perfect fit status.

Pseudo -$\boldsymbol{R}^2$

In Hair, Jr. et al., (2010), pp. 323, three measures for pseudo-R^2 were listed as other available diagnostic tools for logistic regression fits. And these are:

- the Cox and Snell R^2,
- the Nagelkerke R^2, and
- a pseudo R^2 measure based on the reduction in the $-2LL$ value".

The authors further note that, comparatively, these three pseudo-R^2 measures are interpreted similarly to the R^2 of OLS regression.

Accuracy

For logistic regression the prediction accuracy is evaluated using confusion matrices (Hair, Jr. et al. 2010, p. 324). Model predictive accuracy is measured through taking the sum of correctly classified subjects over the entire subject list (Hair, Jr., et al., 2010, p. 332 and Matsumoto & Del-Moral-Hernandes, 2013), giving a so called "hit ratio" metric.

The summarized outcomes from a logistic regression model are coded, cf. Table 1, as follows:

- TP and TN are the true positives and negatives respectively, similarly
- FN and FP are false negatives and positives.

The accuracy of the model or the 'hit ratio' (Hair, Jr., et al. 2010, p. 335) where $\hat{y}$ and y are the predicted and actual responses respectively, is given as follows:

$$hit \quad ratio = \left(\frac{TN+TP}{TN+TP+FN+FP}\right) = \left(\frac{TN+TP}{all \quad subjects}\right). \tag{20}$$

Several other measures of interest with classification exercises are precision, recall, specificity, and Jaccard index. Receiver characteristic curves (ROC), together with the area under the curve (AUC), may also be evaluated for model accuracy and used as diagnostics. These measures are explained below.

Precision

The proportion of correctly classified positive predictions is called the precision (Lipton, Elkan, & Naryanaswamy, 2014; Saxena, 2018; Matsumoto & Del-Moral-Hernandes, 2013).

$$precision = \left(\frac{TP}{TP+FP}\right) = \left(\frac{TP}{predicted \quad positives}\right). \tag{21}$$

It is usually a good measure to apply in those instances where misclassifying positive classes induce a higher cost. Since a higher precision indicates that the fit correctly/ 'precisely' identifies the positive class with a higher percantage or degree of confidence the most times, *i.e.* subjects allocated treatment in our case. So a good fit, in terms of higher precision, implies that false negative predictions are forced down or compromised.

Recall or sensitivity

Lipton, et al. (2014); Saxena, (2018) and Matsumoto & Del-Moral-Hernandes (2013) note that when one was interested in the percentage of positives that got perfectly classified out of the actual positives, *i.e.* recall (or sensitivity) measure, the formulation is given as:

$$recall = \left(\frac{TP}{TP+FN}\right) = \left(\frac{TP}{actual \quad positives}\right). \tag{22}$$

Specificity

For correct classification of the negative class, the typical proportional measure is the specificity (Matsumoto & Del-Moral-Hernandes, 2013).

$$specificity = \left(\frac{1-FP}{TN+FP}\right) = \left(\frac{1-FP}{actual \quad negatives}\right). \tag{23}$$

F_1 score

In classification exercises the F_1 score, also called the harmonic mean of the *precision* and *recall* measures, is used for balance between these two diagnostic measures (Saxena, 2018). It comes in handy in cases where the treatment indicator, or target variable, exhibits some imbalance with majority of subjects being from the control group. The F_1 score is thus given as:

$$F_1 \quad score \quad = 2 \cdot \left(\frac{precision \cdot recall}{precision + recall} \right). \tag{24}$$

Jaccard index

Authors (Lipton, et al. 2014) defines the Jaccard index as:

$$Jaccard = \left(\frac{TP}{TP + FN + FP} \right). \tag{25}$$

This index is used to measure similarity between the groups or classes, if they are non-separable. For instance an index closer to 1 or 100% will indicate some great deal of overlap between the treatment classes. So, a Jaccard index of zero is indication that the treatment groups are exclusive, whereas when it equals 0.5 the interventions gets split half, *i.e.* 50% each.

3.1.2 Propensity scores

Rosenbaum and Rubin (1983) present the propensity score as the conditional probability of exposure for an individual based on their characteristics or "baseline covariates". This means that instead of comparing subjects in higher dimensions the "propensity scores summarize all of the covariates into one scalar: the probability of being treated." (Stuart, 2010). Such a balancing score will range from 0 to 1, and is calculated mainly by using logistic regression. Propensity scores are useful in allowing one to match subjects on a single number or scalar, *i.e.* the propensity score, and control for confounder-covariates in regression analysis problems.

Suppose that you have n subjects (or individual units), let z denotes the treatment condition, and Y be the subsequent response, such that $z_i = 1$ for a subject allocated the treatment and its response is Y_{1i} (Pan & Bai, 2015 and Rosenbaum & Rubin, 1983). And for the control group, $z_i = 0$ for subject i resulting in the subsequent response Y_{0i}. The propensity score p_i of treatment exposure ($\forall$ subject i), conditional on or given a vector $\boldsymbol{X}_i$ of observed or pre-treatment features, is thus given as:

$$p_i = P(z = 1 \,\big|\, \boldsymbol{X}_i) \tag{26}$$

The propensity score p_i for subject i, can be estimated from logistic regression using the log-odds of the given binary treatment condition (Pan & Bai, 2015) as follows:

$$ln\left(\frac{p_i}{1-p_i}\right) = ln\left(\frac{P(z=1\Big|\boldsymbol{X}_i)}{1-P(z=1\Big|\boldsymbol{X}_i)}\right) = \boldsymbol{\beta}^T\boldsymbol{X}_i, \tag{27}$$

where

$$\begin{aligned} p_i &= \beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k+\cdots+\beta_px_p = \boldsymbol{X\beta} \\ &= \frac{e^{\boldsymbol{X}\beta}}{1+e^{\boldsymbol{X}\beta}} = \frac{e^{\boldsymbol{X}\beta}}{1+e^{\boldsymbol{X}\beta}}\times\frac{e^{-\boldsymbol{X}\beta}}{e^{-\boldsymbol{X}\beta}} = \frac{e^{\boldsymbol{X}\beta}}{1+e^{-\boldsymbol{X}\beta}} \end{aligned}$$

and therefore:

$$p_i = \left(\frac{e^{\boldsymbol{X}\beta}}{1+e^{-\boldsymbol{X}\beta}}\right), \tag{28}$$

with

β_0 representing the intercept term,

β_k is the regression coefficient, for $k = 1, 2, ..., p$,

x_k depicts the random treatment feature variables, and

$\boldsymbol{X}_i$ is the observed value of the covariates (feature variables).

Note that $\left\{1-P(z=1\Big|\boldsymbol{X}_i)\right\}$, above, represents the probability that the observed ith subject was not exposed to this treatment, in other words it remained within the control group.

3.2 Matching

Once the propensity scores have been estimated, for the two groups, different adjustment and or matching methods can be chosen and applied (Rosenbaum & Rubin, 1983). The following are listed, in the literature, as some of the available methods in the applications of propensity scores in practice (Lane, To, Kyna & Robin, 2012; Rosenbaum & Rubin, 1983; Austin, 2011; Beal & Kupzyk, 2014; Pan & Bai, 2015):

(1) matching,

(2) stratification,

(3) regression or covariate adjustment, and

(4) the inverse propensity score weighting also referred to as "the inverse probability of treatment weighting (IPW)".

From Section 3.1.2 above, using logistic regression the propensity score of Equation (26) reduces to:

$$p_i = P(z = 1 \Big| \boldsymbol{X}_i) = \left(\frac{e^{\boldsymbol{X}_i \beta_i}}{1 + e^{-\boldsymbol{X}_i \beta_i}} \right). \tag{28}$$

Rosenbaum & Rubin (1983) define a balancing score as any function $b(\boldsymbol{X})$ such that $\boldsymbol{X} \perp\!\!\!\perp z \Big| b(\boldsymbol{X})$, conditional on $b(\boldsymbol{X})$, and that the propensity score $p(\boldsymbol{X})$ is some balancing score where the distribution of $(\boldsymbol{X})$ is uncorrelated from that of the binary treatment allocation (z). They prove in Theorem 2 that the propensity score is the finest balancing score that gives the probability of allocating or enrolling subjects to a treatment program. Note that the proof for Theorem 2, that the propensity score is the finest balancing score, is given in Appendix 1.

In the following section we look at propensity scores, their underlying techniques in practice, together with the relevant causal assumptions, especially those numbered (1) and (2). Recall that, whenever these two conditions (*i.e.* (1) and (2)) are simultaneously met 'strong ignorability' is confirmed.

3.2.1 Propensity score matching

Observational data requires that researchers imitate a conditional randomized trial, of which implies applications of propensity score matching, Beal & Kupzyk (2014). Propensity score matches are done through the selection of subgroups from the control subjects and then matching them, on their scores, to those treated. Thavaneswaran & Lix (2008) state that propensity score analysis is concerned with computing unbiased estimates for the impact of treatment based on the assumption that subjects with similar base character exists in both treatment groups. Encompassed in the steps before matching is done are:

(1) determination of how many controls are matched per treatment,

(2) deciding on what is the appropriate algorithm to be used, and

(3) establishing acceptable similarity levels for the estimated propensity scores.

All the above matching steps are possible with propensity score matching whenever the two properties of unconfoundedness (conditional independence) of the covariates and overlaps (existence of some common region of support and or positivity) in the distributions of the scores from each treatment group are satisfied. After the propensity scores are estimated, the following matching methods, listed in literature as commonly available in practice, may be implemented for appropriate adjustments (cf. Figure 6):

(1) nearest neighbour (NN) (or greedy) matching,

(2) radius(or caliper) matching,

(3) many-to-many (N:N) matching,

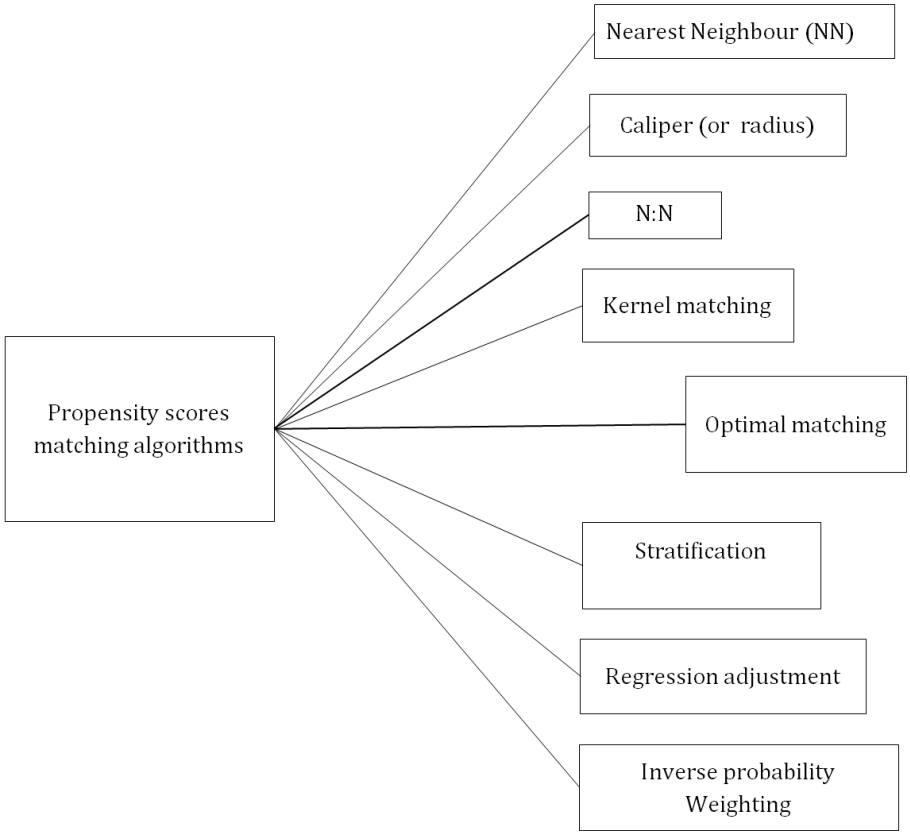
(4) stratified (or interval) matching,

(5) optimal matching,

(6) kernel matching,

(7) regression adjustment, and

(8) inverse probability of treatment weighting (IPW).



NN - nearest neighbor; N:N - N:N matching algorithm; PS - Propensity Scores

(Source: Thavaneswaran & Lix (2008), self-adapted illustration)

Figure 6 Propensity score matching methods

Causal assumptions

There are three main causal inference assumptions, mostly mentioned in literature, for observational studies or quasi-experimental design (QED) projects. In order to match on propensity scores, one needs to understand the twin causal properties of the common support condition and conditional independence or unconfoundedness assumption, the 1st

and 2nd assumptions below, as explained in Rosenbaum & Rubin, (1983). We list the three causal assumptions as:

(1) the conditional independence assumption,

(2) the common support or 'overlap' condition, and

(3) the single unit treatment value assumption.

The third property has implications that subjects won't interfere once the treatment is allocated and that treatment exhibits a singular form (Ning, Ghosal & Thomas, 2019). As explained by Rosenbaum & Rubin (1983), for any given unit or subject pair i and j there should be no conflict over resources, in other words distinct individual outcomes $Y_{\tau i}$ and $Y_{\tau j}$, for treatment τ, exist. This is due to the fact that subjects can only be exposed to either version of this treatment and their outcomes are exclusive meaning one subject being exposed won't affect another's outcome (Gordon, Zettelmeyer, Bhargava & Chapsky, 2018). Ning, Ghosal & Thomas, (2019) further states about the conditional independence assumption that it means subjects have an equal treatment allocation chance, as with a coin toss, whilst those of the common support (or 'overlap') condition will be that all subjects have a positive exposure probability.

Rosenbaum & Rubin (1983) point out that whenever the common support and conditional independence assumptions are met strict-ignorability or strong-ignorability of matching is achieved. The implications of being 'strongly ignorable' are that the potential responses will not be affected by the treatment status for any given baseline covariate vector ($\boldsymbol{X}$). Further stated by Rosenbaum & Rubin is that 'strong ignorability' will stay true, when both these causal assumptions are valid, unless the estimate of interest for the researcher (investigator) was 'only' the average treatment effect on the treated (ATT). Therefore, our discussion below is on the causal assumptions with focus on the 'strong ignorability' case being true, the main requirement in applications of propensity score matching (PSM).

Assumption (1): The conditional independence assumption

With this assumption subjects are allocated to the treatment based on observable features from the sample regardless of the treatment status, meaning that allocation is conditioned on these characteristics or features. Such conditioning will therefore lead to potential outcomes that are independent of treatment allocation (Gordon, et al. 2018). This implies that in the absence of an intervention the potential response (outcome) will be independent of such status (Heinrich, Maffioli & Vázquez, 2010). This method is given mathematically as follows (taken from Equation (1), and where the $\perp\!\!\!\perp$ is for "strong ignorability");

$$(Y_1, Y_0) \perp\!\!\!\perp z \Big| \boldsymbol{X} \tag{29}$$

In other words, there exist a set $\boldsymbol{X}$ of features, observable at baseline or prior the intervention, to the researcher such that controlling for them leads to potential outcomes that are independent of treatment exposure (Rosenbaum & Rubin, 1983 and Heinrich, Maffioli & Vázquez, 2010). The proof for conditional independence or the unconfoundedness property, as adapted from Grilli & Rampichini (2011), is given below.

To prove Equation (29) it will be sufficient to show that

$$P\{z = 1 \Big| Y_1, Y_0, p(\boldsymbol{X})\} = P\left\{z = 1 \Big| p(\boldsymbol{X})\right\} = p(\boldsymbol{X}).$$

Implying that the independence of the responses (Y_1, Y_0) and z, the treatment allocation, are conditional on the propensity scores $p(\boldsymbol{X})$

$$\begin{aligned}
P(z = 1 \Big| Y_1, Y_0, p(\boldsymbol{X})) &= \mathbb{E}[z = 1 \Big| Y_1, Y_0, p(\boldsymbol{X})] = P(z = 1 \Big| p(\boldsymbol{X})) \\
&= \mathbb{E}\{\mathbb{E}[z = 1 \Big| Y_1, Y_0, p(\boldsymbol{X}), \boldsymbol{X}] \Big| Y_1, Y_0, p(\boldsymbol{X})\} \\
&= \mathbb{E}\{\mathbb{E}[z = 1 \Big| Y_1, Y_0, \boldsymbol{X}] \Big| Y_1, Y_0, p(\boldsymbol{X})\} \\
&= \mathbb{E}\{\mathbb{E}[z = 1 \Big| \boldsymbol{X}] \Big| Y_1, Y_0, p(\boldsymbol{X})\} \\
&= \mathbb{E}[p(\boldsymbol{X}) Y_1, Y_0, p(\boldsymbol{X})] = p(\boldsymbol{X}),
\end{aligned}$$

where the last equality implies the unconfoundedness property. Similarly, we can evaluate that,

$$\begin{aligned}
P(z = 1 \Big| p(\boldsymbol{X})) &= \mathbb{E}[z = 1 \Big| p(\boldsymbol{X})] \\
&= \mathbb{E}\{\mathbb{E}[z = 1 \Big| \boldsymbol{X}] \Big| p(\boldsymbol{X})\} \\
&= \mathbb{E}[p(\boldsymbol{X}) \Big| p(\boldsymbol{X})] = p(\boldsymbol{X})
\end{aligned}$$

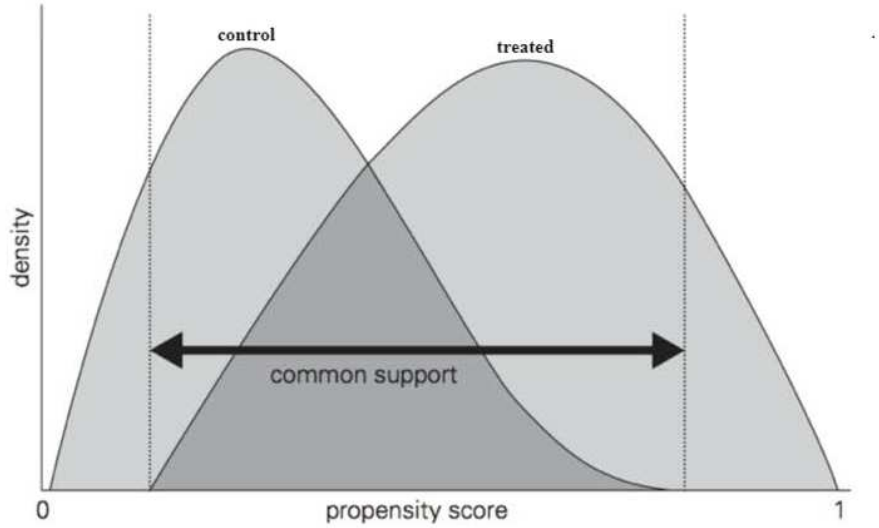
This renders the proof complete.

Equation (29) above is simply the mathematical notation for the idea expressed in the previous paragraphs that the potential outcomes are independent of treatment status given the baseline covariates, also referred to 'conditional on observables' by Gordon, Zettelmeyer, Bhargava & Chapsky (2018). In other words, after controlling for $\boldsymbol{X}$, the treatment allocation will effectively be "as good as random" (Heinrich, Maffioli & Vázquez, 2010).

This is the 'unconfoundedness' property, also known as the selection on observables property in literature. The conditional independence assumption is of crucial importance in correctly identifying the impact of the treatment on the response since it ensures that differences in treatment groups will be accounted for to reduce selection bias, thus allowing that subjects in the control group be used in constructing the counterfactual for those in the treatment group.

Assumption (2): The common support ('overlaps') condition

Examining possible overlaps in the distributions of propensity scores between the exposure and control groups is a critical step since the average treatment effect on the treated (ATT) is only valid within the common support region (Caliendo & Kopeinig, 2005). That is, valid estimates for a treatment effect are possible only if all exposed subjects are matched successfully to those in the control group (Gertler, Martinez, Premand, Rawlings & Vermeersch, 2010). In order for that to happen, all subjects must have positive treatment allocation probabilities (Ning, Ghosal & Thomas, 2019).



(Source: Zaga Szenker, 2015)

Figure 7 The common support region

It may happen that scores differ for exposed subjects to those in the control group, see Figure 7 above, especially at the extreme ends or tails (Gertler et al., 2010). The consequence is that at these tails, closer to one or zero, exposed subjects will have no control match at the higher end, and similarly for control subjects at the lower end. That is, subjects at the tails of these distributions are outside the common support region, at times necessitating the need for extrapolation. Figure 7 is a typical illustration of the lack of overlaps (the problem of the common support) where propensity score distributions between treatment groups are not smooth (Gertler et al., 2010).

There are various methods to deal with, or investigate, this issue that are suggested by Caliendo & Kopeinig (2005), such as visualization of the propensity score distributions as done in Figure 7. Others may include a comparison of the minimum and maximum propensity scores of both groups and also estimation of the groupings' density distributions. This requirement is said to rule out the perfect predictability phenomenon ensuring

that subjects with similar features have a chance of being allocated treatment. In mathematical notation this condition is presented as follows:

$$0 < P(z = 1 \Big| \boldsymbol{X}) < 1. \tag{30}$$

Consequences (of the common support) are that subjects will have positive probabilities of being allocated treatment and hence it is also referred to as the positivity assumption in literature (Ning, Ghosal & Thomas, 2019). This means that a subject will have a probability that it is exposed somewhere in the (0,1) interval. Probability theory tells us then that the probability of non-exposure must also lie within this same interval. This formula is interpreted, using the common support requirement defined above due to these possible overlaps in matching, as:

$$P(z = 0 \Big| \boldsymbol{X}_i) = 1 - Pr(z = 1 \Big| \boldsymbol{X}_i). \tag{31}$$

Equation (31) is the probability that a subject was not allocated the said treatment. As stated above, treatment allocations are 'strongly ignorable' whenever both the conditional independence assumption and the common support condition are satisfied. On that consequence, we then rewrite Equation (31), if the conditional independence assumption is met, as follows (Rosenbaum & Rubin, 1983):

$$P(z = 1 \Big| Y_{1i} Y_{0i}, \boldsymbol{X}_i) = P(z = 1 \Big| \boldsymbol{X}_i) = p(\boldsymbol{X}_i) = p_i, \tag{32}$$

i.e. a consequential direct application of the unconfoundedness proof from Equation (29).

Now, under 'strong ignorability' of treatment allocation, the average causal effect can be estimated by conditioning on the propensity score p_i instead of the pre-treatment characteristics $\boldsymbol{X}_i$ (Rosenbaum & Rubin, 1983). This, according to Rosenbaum & Rubin, simplifies a multidimensional problem to a one dimensional or 'scalar' when matching. This means that the responses from the study are independent of the treatment allocation conditional on the covariates and also that treatment allocation is a true probability, *i.e.* it is between zero and one (Rosenbaum & Rubin, 1983 and Beal & Kupzyk, 2014). In short, treatment allocation is ignored and assumed to be random. Matching can thus be done, only, through the one-dimensional propensity score instead of computing multivariate distances of $\boldsymbol{X}_i$.

If the parameter of interest was the average causal effect on those subjects exposed to treatment (ATE) then one may relax the assumptions of conditional independence and common support (Rosenbaum & Rubin, 1983 and Heinrich, Maffioli & Vázquez, 2010). From Equation (8), another parameter of interest is the average causal effect of the treated participants (ATT), and was thus defined as follows:

$$\tau_{ATT} = \mathbb{E}(Y_1\text{-}Y_0 \,\big|\, z = 1) = \mathbb{E}(Y_1 \,\big|\, z = 1) - \mathbb{E}(Y_0 \,\big|\, z = 1) \tag{33}$$

The non-observable outcome or the counterfactual for treated subjects is represented by the second term, $\mathbb{E}(Y_0 \,\big|\, z = 1)$, on the right hand side of Equation (33), implying that the average difference in responses over the common support is the estimate of the average causal effect (Caliendo & Kopeinig, 2005).

Benefits with matching

The three main advantages with matching according to Rosenbaum & Rubin (1983) and Posner & Ash (2012) are:

(1) Researchers or investigators get to work with better characterized matched data, thus leading to a simple representation.

(2) Covariate balance in matched samples leads to lower variability of the average treatment effect than it would have been for randomized trials (RCTs), and therefore matched samples will tend to lower the estimate for the average treatment effect due to similarity in covariate distributions after matching.

(3) Outcome-based models are, typically, resilient to any violations of the underlying assumptions. Posner & Ash (2012) credit this to the robustness of model-based approaches especially in cases where deviations from model assumptions were visible.

3.2.2 Matching algorithms

Logistic regression is the method of choice in the calculation of propensity scores. Strong ignorability must still remain valid (Rosenbaum & Rubin, 1983; Heinrich et al., 2010 and Caliendo & Kopeinig, 2005). Various authors give the following primary factors, to take into account, when one is selecting a method of choice when matching exposed (treated) subjects to those in the control or comparison group.

- Will the matching be done with or without replacement?
- What is the neighbourhood or closeness of the match?
- Are the study subjects to be weighted for the analysis?
- Will subjects be counted from the control group and then matched to exposed ones?

Various matching algorithms are available to match control subjects to those treated, embedded within different statistical software packages. Matching algorithms are explained in more detail, within their respective subsections, below. First, we expand on the implementation steps. The following are the key steps in the implementation of propensity score matching methods, to be discussed in the next section (Stuart, 2010):

Design

- Step 1: Define the 'closeness' or the distance measures between matches
- Step 2: Implement the matching using the defined closeness measures
- Step 3: Assess the matched results for quality (steps 1 - 2 may be repeated to improve matching quality)

Analysis

- Step 4: Estimate the treatment effect off the high quality matches from above.

Below is a brief explanation of the above steps:

Defining the '*closeness*' requires that we do variable selection for covariate inclusion in the matching algorithm, followed by calculating their distance measure. The 'strong ignorability' assumption, as discussed above, implies that all differences between the control and treatment groups were observed whilst adjusting for selected covariates (Rosenbaum & Rubin, 1983).

Rosenbaum and Rubin (1983) define distance as a similarity measure between any two subjects from either the exposed or control group. Four distance measures are used to determine similarity between two subjects:

(1) the exact,

(2) Mahalanobis,

(3) the propensity score, and lastly

(4) the linear propensity score distance measures.

We depict the math notation of these distance measures, denoted D_{ij}, between any subject pair i and j below.

(1) The exact distance measure

For subjects i and j, we get that

$$D_{ij} = \begin{cases} 0 & if\ X_i = X_j \\ \infty & otherwise \end{cases}, \tag{34}$$

where X_i are the covariates for subject i, and X_j for subject j.

(2) The Mahalanobis distance measure

$$D_{ij} = (X_i - X_j)' \sum^{-1} (X_i - X_j) \tag{35}$$

where $\sum$ is the variance-covariance matrix, X_i and X_j represents the covariates for subjects i and j. This distance is scaled through this variance-covariance matrix for each subject (Posner & Ash, 2012). For instance, for a case with a variance of X_i double that of X_j, Mahalanobis requires that an equidistant subject is twice as far. An analogy from Posner & Ash (2012) is that one can think of rough (or 'rocky') and smooth terrain, in orthogonal directions, south to north then east to west, respectively. The authors note that an hour drive from either direction is not necessarily similar, as a subject driving west-side will definitely go the farthest than one driving up north, making Mahalanobis distance somehow comparable to the time it takes one to get somewhere. Posner & Ash (2012) also state some degree of equivalence between this measure and the Euclidean distance, as a consequence of standardizing by the variance-covariance matrix $\sum$. For continuous key features of interest, the Mahalanobis distance measure is defined as follows (Stuart, 2010):

$$D_{ij} = \begin{cases} (X_i - X_j)' \sum^{-1} (X_i - X_j), & if \left| logit(p_i) - logit(p_j) \right| \leq c \\ \infty & otherwise \end{cases}$$

where c denotes the tolerance matching range. Although calculation of this metric works best for continuous features of X, the covariates will have to be converted to binary form, *i.e.* dummy coded, if they are categorical or factor variables (Stuart, 2010). Stuart further observes that the variance-covariance matrix of X will be dependent on what interest the researcher harbours, meaning the form of $\sum$ is based on what effects are being investigated, whether it is the average treatment effect (ATE) or the average treatment effect on the treated (ATT). We note that, for ATT, the variance-covariance matrix $\sum$ will be from the full control group, whilst for ATE; $\sum$ is from the full treatment exposure group.

(3) The propensity score distance

$$D_{ij} = \left| p_i - p_j \right| \tag{36}$$

where p_i and p_j denotes the respective propensity scores for subjects i and j.

(4) The linear propensity score distance

$$D_{ij} = \left| logit(p_i) - logit(p_j) \right| \tag{37}$$

This distance measure is viewed as another way to effectively reduce selection bias

(Rosenbaum & Rubin, 1983). One major drawback mentioned in Stuart (2010) about Mahalanobis and the exact distance measures is their lack of robustness within high dimensional feature spaces.

After the propensity scores are calculated, using some classifier tool like logistic regression, the next step then is to use them to match the exposed (treated) subjects to those in the control group. The matching techniques in Figure 6 will be discussed within the subsequent individual sub-sections below. Note that 'strong ignorability' is still a requirement.

Nearest neighbour ('greedy') matching

The rationale behind this matching method is that the propensity scores of control group subjects will be compared to those that received the treatment and the closest ones will be matched (Heinrich, Maffioli & Vázquez, 2010). In a greedy matching exercise, randomly selected treated subjects are matched to a control subject whose propensity score is closer (Austin, 2011). That is done through minimizing the difference in the propensity scores between the two groups. The steps that are typically taken in a greedy matching exercise are (Posner & Ash, 2012 and Austin, 2011):

- each subject from the smallest group is matched with a partner closest in propensity score from the other group,
- the matched pair then gets extrapolated from the data,
- repeat this process until the matching group is exhausted.

Caliendo & Kopeinig, (2005) note that this is the easiest method and it is applied in variants of matching either with or without replacements. Allowing for these variants in matching will enable us to control for the bias-variance trade-off, Caliendo & Kopeinig (2005) and Heinrich, Maffioli & Vázquez (2010). A matching performed allowing for repeats, with replacement, will have lower bias thus producing high quality matches.

If the data has a different distribution of control vis-à-vis the treatment group, bad matches are inevitable; thus allowing for replacement, when matching, is a solution. Another technique to address the bias-variance trade-off is referred to as 'oversampling', when the researcher applies a number of nearest neighbour methods (Caliendo & Kopeinig, 2005). They explain that this involves making choices on weighting the pairs (i, j) to match to the treated subjects and the count of these created pairs. With ordered control and treatment groups we ensure that the first exposed subject is in comparison to a first control subject with a similar propensity score (Thavaneswaran & Lix, 2008). Mathematically that is written as:

$$C_{p_i} = min \left| p_i - p_j \right| \tag{38}$$

where C_{p_i} is the class of control subject j matched to treated subject i,

p_i is the estimated propensity score for exposed subject i , and

p_j is the estimated propensity score for control subject j.

Caliper ('radius') matching

To avoid the potential risk of mismatches due to the nearest-neighbour algorithm, one may use a caliper, also referred to as the tolerance on the maximum propensity score, method (Heinrich, Maffioli & Vázquez, 2010 and Caliendo & Kopeinig, 2005). With this algorithm, each treated subject is matched with that from the control group on a predefined or pre-determined interval for all propensity scores (Thavaneswaran & Lix, 2008). Heinrich, Maffioli & Vázquez (2010) argue that this has the consequence of now matching around some radius-neighbourhood, therefore, mapping a group of subjects instead of only the nearest neighbour. This method allows wide varieties of subjects to be used in comparisons within the radius or caliper, ensuring that those adjacent enough to be mapped are considered.

In a many-to-one caliper matching scenario, the estimator can be written as follows (assuming subjects are matched with replacement):

$$\mathbb{E}(\Delta Y) = \frac{1}{n_\tau} \sum\nolimits_{i=1} (Y_{1i} - \bar{Y}_{1j(i)}) \tag{39}$$

where $\bar{Y}_{1j(i)}$) is the average response for all control subjects matched with treated subject i, Y_{1i} is the response for subject i, and for a subject sample of size n (combining the exposure and control groups), $n = n_\tau + n_C$, such that n_τ is the number of subjects in the exposed group, and n_C is the number of subjects in the control group.

For this matching method a predetermined caliper range is defined, usually within one-quarter of the standard error of the estimated score (Thavaneswaran & Lix, 2008). This ensures that values that fall outside this tolerance level are extrapolated, *i.e.* enforcing the common-support directly. This range is given mathematically by:

$$\left| p_i - p_j \right| < r \tag{40}$$

where p_i represent the estimated propensity score for the exposed subject i,

p_j is the estimated propensity score for the control subject j, and

r is the predetermined tolerance range of the values (or the caliper).

N:N matching

In this method, a randomized order of the first n subjects from both treatment groups, the exposed and control groups are matched (Thavaneswaran & Lix, 2008). The closest propensity scores will be applied as criteria for such matching.

Kernel matching

The above-mentioned algorithms, greedy, N:N, and caliper, typically match on few individuals from the control group to create the counterfactual or imaginary response (Caliendo & Kopeinig, 2005). They, together with Li (2012), further state this as the reason why non-parametric matching estimators such as kernel-matching (KM) that utilize all control group subjects, weighted on their averages, in constructing the counterfactual response are looked at. Matching with a kernel is applied to an inversely proportional weighted average of the control subjects in matching to a treated subject (Thavaneswaran & Lix, 2008).

The main advantage with these methods is that more information will most likely be examined thus lowering the variance, whereas the disadvantage may be that of possible poor matches on some subjects with an unintended high bias consequence (Caliendo & Kopeinig, 2005). This typically give rise to a classic 'bias-variance tradeoff' statistical exercise for the researcher. A typical kernel estimator, for the potential outcome $\mathbb{E}(Y_{0i} \Big| p_i, z = 0)$ for the ATT is thus given, adapted from Li (2012), as

$$\mathbb{E}\{Y_{1i} - Y_{0i} \Big| p_i; z = 1\} = \frac{1}{n_\tau} \sum_{i=1\{z_i=0\}}^{n_C} \left\{ Y_{1i} - \sum_{j=1\{z_j=0\}}^{n_C} Y_{0j} \cdot W_k(p_j) \right\}, \tag{41}$$

where the weights $W_k(p_j)$ are given by

$$W_k(p_j) = \frac{K\left\{\frac{p_j - p_i}{h_n}\right\}}{\sum_{k=1\{z_k=0\}}^{n_C} \cdot K\left\{\frac{p_k - p_i}{h_n}\right\}}$$

with $p_i = p(\boldsymbol{X}_i)$ is the treatment allocation probability, *a.k.a.* the propensity score for subject i, similarly for p_k and p_j, K represent the defined kernel function, and h is the smoothing parameter (or bandwidth). Typically, Li (2012) further states, a researcher will have to have to postulate a Gaussian kernel, together with an appropriate bandwidth smoothing parameter, in order to estimate the causal effects.

Optimal matching

When minimizing the total within-pair difference of the propensity score, the optimal matching technique is preferred to greedy matching (Austin, 2011). This traditionally complex approach is made possible lately due to improvements in computation power

(Olmos & Govindasamy, 2015). With this approach the average propensity score and the distance within the matched pairs are evaluated and the match group with the lowest average is chosen. In order to minimize the total propensity score difference for any given data, matched groups of treatment and control subjects are to be identified, and mathematically defined as

$$\Delta = \sum_{\mathcal{S}=1}^{S} \omega\{\left|A_{\mathcal{S}}\right|, \left|B_{\mathcal{S}}\right|\} \cdot \delta(A_{\mathcal{S}}, B_{\mathcal{S}}), \tag{42}$$

where ω is the weighting for the number of subjects, or size, in the grouping, $\left|A_{\mathcal{S}}\right|$ or $\left|B_{\mathcal{S}}\right|$ is the size or number of elements within the stratum in groups A or B, and δ is the distance between elements in the stratum. Note that, this approach ensures a total propensity score distance of an optimal minimum value between matched subjects (Roy, 2018 and Olmos & Govindasamy, 2015).

3.2.3 Stratification

This method segments the common-support, or the $(0, 1)$ interval, of the propensity scores into separate strata or sub-intervals (Caliendo & Kopeinig, 2005). It is also referred to as "interval matching" or "sub-classification" (Rosenbaum & Rubin, 1983). These intervals are segmented through the use of a range of values as one strategy, consisting of subjects with similar average propensity scores from the treatment or control group (Thavaneswaran & Lix, 2008).

Stratification matching works as follows:

(1) check if the propensity score is balanced per strata; if not

(2) split the strata into smaller subjects.

The rationale behind stratification is that the data sample will be ordered, or ranked on their propensity scores and then subdivided into equivalent subsets or layers; typically divided into five strata, usually in quantiles of size 0.2 (Beal & Kupzyk, 2014; Austin, 2011 and Posner & Ash, 2012). Beal & Kupzyk, 2014 argue that this equivalence in classes when matching helps in eliminating the possibility of covariate-bias.

Analyses of the propensity scores will be done on each of these different layers or equal-sized quantiles (Austin, 2011). This will allow us to infer causality with confidence due to the propensity score distributions of the treatment groups overlapping per strata (Posner & Ash, 2012). Austin further notes a correction of at least 90% on possible covariate-bias when the researcher is stratifying based on continuous confounder variables. In order to obtain the average treatment effect, for instance, the average difference measure of the response outcomes will be calculated for each stratum weighted by the distribution of

the exposed subject across all the strata. However, there are possible complications in making decisions with this method as it is said to be time-consuming, as an exercise, since matching decisions are to be distinct across all strata levels. Further drawbacks, taken from (Posner & Ash, 2012), are listed below:

(1) Possible extrapolation of large amounts of data in cases where strata is skewed towards one group. For example when one of the groups consist a small number of observations in particular (therefore directly impacting the power and accuracy measures of the analysis).

(2) Random selection may cause problems with replicability, since it means there is a possibility that two investigators may reach differing conclusions due to the analysis being based on distinct random samples.

3.2.4 Regression adjustment

Being exposed to one treatment over the other may cause inter-individual variation towards the expected response. For regression or covariate adjustment to be deemed appropriate the common support condition (CSC), or substantial overlaps between the groups, case must be clearly verifiable (Thavaneswaran & Lix, 2008). This prompted researchers to add the estimated propensity scores into a regression model so as to balance out for such variability (Rosenbaum & Rubin, 1983). In that sense, the expected average response difference between the treatment and control groups will be the required causal effect of the treatment variable when each subject have an adjusted propensity score (Beal & Kupzyk, 2014). What gets to be evaluated here, between the two groups are the differences between the respective means of their propensity scores, the ratio of their variances, and the ratio of their respective covariates' residuals (Thavaneswaran & Lix, 2008). In case when the conditional independence or unconfounded property holds, Hirano & Imbens (2001) formulate the estimated treatment effect by

$$\begin{aligned} \tau_{ATT} \quad (\boldsymbol{x}) = & \quad \mathbb{E}\{Y_1\text{-}Y_0 \,\big|\, X = \boldsymbol{x}\} \\ = & \quad \mathbb{E}\{Y_1 \,\big|\, X = \boldsymbol{x}, z = 1\} - \mathbb{E}\{Y_0 \,\big|\, X = \boldsymbol{x}, z = 0\} \\ = & \quad \mathbb{E}\{Y \,\big|\, X = \boldsymbol{x}, z = 1\} - \mathbb{E}\{Y \,\big|\, X = \boldsymbol{x}, z = 0\} \end{aligned}$$

And estimating the average treatment effect is done through the following equality,

$$\tau_{ATT} = \mathbb{E}\{\tau_{ATT}(\boldsymbol{x})\},$$

giving the final estimate needed for separate exposure and control groups as:

$$\hat{\tau}_{ATT} = (\bar{Y}_\tau - \bar{Y}_C) - \beta(\bar{X}_\tau - \bar{X}_C), \tag{43}$$

where β is the weighting parameter for the impact of treatment, $(\bar{Y}_\tau = \bar{Y}_1)$,$(\bar{Y}_C = \bar{Y}_0)$ are the average responses for the treatment groups, respectively, and $\bar{X}_\tau$,$\bar{X}_C$ are the average covariate differences, when comparing the two groups.

Unlike in the above methods, where the propensity score is applied to the separate groups prior to matching, the covariate adjustment method requires that they be applied in the final stage of analysis. This method consequently use the actual propensity scores to do the matching, instead of the estimated propensity scores (Thavaneswaran & Lix, 2008).

3.2.5 Inverse probability weighting

This method relies on weighting based on the inverse of estimated propensity scores, or the inverse probability of treatment allocation weighting (Rosenbaum & Rubin, 1983; Austin, 2011; Pan & Bai, 2015 and Thavaneswaran & Lix, 2008). It is part of the methods that are not commonly utilized, in practice, compared to those mentioned above. Weighting requires that subjects be weighed for better representations of population dynamics on either treatment group (Thavaneswaran & Lix, 2008). Subjects will get scaled by their treatment allocation probabilities based on the actual treatment group to which they belong (Austin, 2011). Using propensity scores when removing possible confounder influence (Kuang, et al. 2020).

Pseudo-populations, or 'synthetic samples', of weighted copies per subjects, where treatment allocation will not depend on the covariates $\boldsymbol{X}$, results from this technique, Austin (2011). Under-represented subjects will be weighed higher on treatment allocation whilst those over-represented will get scaled lower such that treatment populations are comparable. This ensures that subjects are equally likely to be treated, creating a design that mimics that of randomized control trials (RCTs), as opposed to if it were just the original population. We define, the weight of an exposed subject i as:

$$w_i = \frac{1}{\hat{p}_i} \tag{44}$$

i.e. as the inverse of the propensity score for that ith subject (or the inverse probability of being allocated the treatment) and the weights for each control subject i are:

$$w_i = \frac{1}{1 - \hat{p}_i} \tag{45}$$

Note that, the pseudo-populations created from these weights will ensure that all the data are utilized, via down and up weighting, instead of discarding unmatched subjects (Roy, 2018). Estimations are done on the balanced weighted pseudo-populations or 'synthetic samples' using the 'now' unconfounded relationship between the treatment z and the potential outcome Y.

Some drawbacks are that a propensity score closer to zero will cause these weights to approach infinity and thus cease to exist (Thavaneswaran & Lix, 2008). Below, we depict how the treatment effect is, typically in some ways, estimated with the inverse probability of treatment allocation weighting (IPW) method. We also introduce the concept of marginal structural models (MSMs) and how they are applied in conjunction with this estimation technique. Further in-depth reads pertaining to inverse probability of treatment allocation weighting and the regression (covariate) adjustment techniques can be found in Rosenbaum & Rubin (1983); Austin (2011) and Pan & Bai (2015).

Rationale for the IPW technique

The logic of the inverse probability weighting method is shown here, adapted from Austin & Stuart (2015) and Barter (2017). Recalling, from sub-section 2.1.2, that the observed outcome or response is given as $Y = z \cdot Y_1 + (1-z) \cdot Y_0$, recalling Equation (5), and in the presence of observed confounders we estimate the causal effect using the IPW as:

$$\hat{\tau}_{ipw} = \frac{1}{n_\tau}\sum\nolimits_{z=1}\frac{Y_1}{\hat{p}_i} - \frac{1}{n_\tau}\sum\nolimits_{z=0}\frac{Y_0}{(1-\hat{p}_i)} = \frac{1}{n}\sum\nolimits_{z=1}\frac{zY_1}{\hat{p}_i} - \frac{1}{n}\sum\nolimits_{z=0}\frac{(1-z)Y_0}{(1-\hat{p}_i)},$$

where $\hat{p}_i$ is the estimated propensity score and $z \in \{0, 1\}$ is the treatment status.

To show that the above quantity will remain unbiased in nature, it suffices that we prove the following (Barter, 2017):

$$\mathbb{E}\left\{\tfrac{zY_1}{\hat{p}_i}\right\} = \mathbb{E}[Y_1] \quad and \quad \mathbb{E}\left\{\tfrac{(1-z)Y_1}{1-\hat{p}_i}\right\} = \mathbb{E}[Y_0]$$

That is

$$\begin{aligned}
\mathbb{E}\left\{\tfrac{zY_1}{\hat{p}_i}\right\} &= \mathbb{E}\left\{\mathbb{E}\left\{\tfrac{zY_1}{\hat{p}_i}\middle|\boldsymbol{X}_i\right\}\right\} \\
&= \mathbb{E}\left\{\mathbb{E}\left\{\tfrac{zY_1}{\hat{p}_i}\middle|\boldsymbol{X}_i\right\}\right\} \\
&= \mathbb{E}\left\{\frac{\mathbb{E}[z\middle|\boldsymbol{X}_i]\cdot\mathbb{E}[Y_1\middle|\boldsymbol{X}_i]}{\hat{p}_i}\right\} \\
&= \mathbb{E}\{\mathbb{E}[Y_1\middle|\boldsymbol{X}_i]\} = \mathbb{E}[Y_1]
\end{aligned}$$

$\therefore$ the estimator for Y_1 is consistent.

Similarly, we work out consistency checks for the Y_0 estimand as

$$\begin{aligned}
\mathbb{E}\left\{\frac{(1-z)Y_0}{1-\hat{p}_i}\right\} &= \mathbb{E}\left\{\mathbb{E}\left\{\frac{(1-z)Y_0}{1-\hat{p}_i}\,\middle|\,\boldsymbol{X}_i\right\}\right\} \\
&= \mathbb{E}\left\{\mathbb{E}\left\{\frac{(1-z)\cdot Y_0}{1-\hat{p}_i}\,\middle|\,\boldsymbol{X}_i\right\}\right\} \\
&= \mathbb{E}\left\{\frac{\mathbb{E}\{(1-z)\,|\,\boldsymbol{X}_i\}\cdot\mathbb{E}\{Y_0\,|\,\boldsymbol{X}_i\}}{1-\hat{p}_i}\right\} \\
&= \mathbb{E}\left\{\frac{(\mathbb{E}[1\,|\,\boldsymbol{X}_i]-\mathbb{E}[z\,|\,\boldsymbol{X}_i])\cdot\mathbb{E}[Y_0\,|\,\boldsymbol{X}_i]}{1-\hat{p}_i}\right\} \\
&= \mathbb{E}\left\{\mathbb{E}[Y_0\,|\,\boldsymbol{X}_i]\right\} = \mathbb{E}[Y_0],
\end{aligned}$$

which confirms that there is consistency in the weighting estimator for Y_0.

Estimating the IPW treatment effect

Under the overlap condition, or positivity, and exchangeability assumptions, the impact of a treatment or an intervention is estimated by

$$\mathbb{E}(Y_1) = \frac{\sum_{i=1}^{n} I(z=1)\frac{Y_i}{p_i}}{\sum_{i=1}^{n} I(z=1)\frac{1}{p_i}}, \tag{46}$$

where p_i is the propensity score for subject i, the numerator $\sum_{i=1}^{n} I(z=1)\frac{Y_i}{p_i}$ is the sum of the outcome in the treated pseudo-population, and the denominator $\sum_{i=1}^{n} I(z=1)\frac{1}{p_i}$ is the number of subjects in the treated pseudo-population. The indicator function I for the treated subjects is given as

$$I(z=1) = \begin{cases} 1 & if \quad treated \\ 0 & otherwise \end{cases}.$$

Similarly,

$$\mathbb{E}(Y_1) = \frac{\sum_{i=1}^{n} I(z=0)\frac{Y_i}{p_i}}{\sum_{i=1}^{n} I(z=0)\frac{1}{p_i}}, \tag{47}$$

where $\sum_{i=1}^{n} I(z=0)\frac{Y_i}{p_i}$ sums the control outcome, and the denominator $\sum_{i=1}^{n} I(z=0)\frac{1}{p_i}$ is the number of subjects, respectively, in the control pseudo-population. The control indicator function I will thus be given by

$$I(z=0) = \begin{cases} 1 & if \quad control \\ 0 & otherwise \end{cases}.$$

Marginal structural models (an introduction)

Another common practice is that of controlling for confounder variables through adjusting via marginal structural models or MSMs (Chiba, Azuma & Okumura, 2009). These are

particular causal models, which at most utilize two step least squares regressions ($2-SLS$), which are different to traditional regression models. Some in-depth discussions pertaining $2-SLS$ are covered in Angrist & Imbens (1995). MSMs are concerned with modelling the mean of the potential outcomes (Angrist & Imbens, 1995). Such models are '*marginal*' for being based on the population average rather than conditional on confounders (Roy, 2018). Their '*structural*' nature is derived from the fact that they are used in modelling potential outcomes instead of observed outcomes. Further, for more complex causal models, MSMs will help in allowing for inverse probability of treatment weights to be used as an estimation method. We detail different kinds of such MSMs below, taken from Roy (2018).

Linear MSM

Let the binary treatment allocation remain as is, *i.e.* $z = 0$ for a subject in control and $z = 1$ for a subject allocated the treatment, respectively. We define the linear MSM as

$$\mathbb{E}(Y_z) = \varphi_0 + \varphi_1 z \tag{48}$$

Now, $\mathbb{E}(Y_0) = \varphi_0$ and $\mathbb{E}(Y_1) = \varphi_0 + \varphi_1$ from Equation (37) and, therefore,

$$\mathbb{E}(Y_1) - \mathbb{E}(Y_0) = \varphi_0 + \varphi_1 - \varphi_0 = \varphi_1$$

Hence, we conclude that φ_1 is the average causal effect, typically used with continuous outcome experiments.

Logistic MSM

In the case of binary outcome experiments, with the treatment allocation z still defined as above, we will apply logit transformation as follows

$$logit\{\mathbb{E}(Y_z)\} = \varphi_0 + \varphi_1 z, \tag{49}$$

where $\mathbb{E}(Y_z) = P(Y_z = 1)$. We get the causal odds-ratio e^{φ_1}, after logit exponentiation, that is

$$\varphi_1 = \frac{odds \quad (Y_1 = 1)}{odds \quad (Y_0 = 1)}.$$

The above numerator and denominator represents the outcome odds, where $odds \quad (Y_1 = 1\ ) = \frac{P(Y_1=1)}{1-P(Y_1=1)}$, and the $odds \quad (Y_0 = 1) \ = \frac{P(Y_0=1)}{1-P(Y_0=1)}$.

Effect modification MSM

The complexity of marginal structural models led scholars, commonly in the fields of epidemiological sciences, to investigate requirements of estimating with effect modifiers (Chiba, Azuma & Okumura, 2009). Typically, the logistic MSM is a technique utilized to test for bias in the effect modifier estimate (Chiba, Azuma & Okumura, 2009). Suppose that a modifier variable M exists for the treatment effect, then the marginal structural model including the effect modifier is defined as:

$$\mathbb{E}(Y \Big| M) = \varphi_0 + \varphi_1 z + \varphi_3 M + \varphi_4 z M \tag{50}$$

where $z \in \{0, 1\}$ still represents our binary treatment status and $\mathbb{E}(Y_z \Big| M)$ is the expected (potential) outcome conditioned on this modifier variable. The potential outcome conditional on M will have a mean given by

$$\mathbb{E}(Y_1 \Big| M) - \mathbb{E}(Y_0 \Big| M) = \varphi_1 + \varphi_4 M,$$

which is the average causal effect of interest.

General MSM

One can use the following general formulation to define an MSM:

$$g\{\mathbb{E}(Y_z \Big| M)\} = h(z, M; \varphi), \tag{51}$$

where $g(\cdot)$ is the link function, like in generalized linear models, except that here the potential or expected outcomes are used instead of the observed (Roy, 2018). It should be kept in mind that $g(\cdot)$ is a function that represents treatment z and modifier M in parametric forms and will typically be linear and/ or additive in nature. Such potential outcomes will not necessarily be the same as those from observed data, for instance in Equation (51), the left hand side involves potential outcomes instead of observed.

Use structural models for IPW estimation

Marginal structural models look a lot like generalized linear models, for example $\mathbb{E}(Y_{zi}) = g^{-1}(\varphi_0 + \varphi_1 z)$, but due to covariate confounding, Roy (2018) argued that this model is not an equivalent of the regression model $\mathbb{E}(Y_i \Big| z) = g^{-1}(\varphi_0 + \varphi_1 z_i)$. However, with the application of the inverse probability of treatment weights (IPW) method the resulting pseudo-populations are free from confounding when the assumption of strong ignorability holds across all treatment allocation levels (Austin & Stuart, 2015). This implies that one can directly infer causality from the above equation. Estimation of the MSM parameters

is done through finding solutions for the estimating equations from the observed data of the pseudo-population or the synthetic sample.

$$\sum_{i=1}^{n} \frac{\partial \mu_i^T}{\partial \varphi} M_i^{-1} w_i \{Y - \mu(\varphi)\} = 0, \tag{52}$$

where w_i is the weight for subject i given as $w_i = \frac{1}{z_i \cdot P(z=1|\boldsymbol{X}_i) + (1-z_i) \cdot P(z=0|\boldsymbol{X}_i)}$, and z_i is still the binary treatment allocation, z_1 for case when treatment is allocated to subjects and z_0 for those that remained for control.

3.2.6 Limitations to propensity scores

Some of the limitations associated with propensity score matching are:

- large samples are usually required for propensity score matching;
- the matching algorithm only controls for selection bias based on observed variables (*i.e.* only on the measured) due to lack of randomization, and so hidden-bias from the omitted variables may still cause problems;
- the more confounding variables you try to match on, the harder the matching becomes;
- a deviation from ordinary least-squares (OLS) assumptions is that propensity score matching assumes independence to be conditional on the baseline characteristics;
- useful cases or subjects may be excluded at both extreme ends of the propensity scores;
- in trying to find the perfect matching, some subjects may be excluded due to incomplete matching, *i.e.* no match found with the same propensity score;
- some confounders may not be matched but still have an effect on your results.

3.3 Balance diagnostics

The requirement with observational studies is that the propensity scores will be estimated, through the given data, since they will be unknown (Austin, 2011 and Xie, Brand & Jann, 2012). Now, recalling that the propensity score is the realest (finest) balancing score, Rosenbaum & Rubin (1983), this means that treatment allocation won't have an impact on the propensity score distribution and the subject's baseline covariates (Austin, 2011). Since the true balancing score is the propensity score (Rosenbaum & Rubin, 1983), it is expected that the distribution be similar in the strata of subjects in either treatment group. Appropriate methods in this evaluation stage involve, amongst others:

- similarity in baseline covariate distribution between the two groups especially for analogous propensity scores;
- with propensity score matching, adequacy of the model involves comparing the exposed control group to the matched sample;
- with the inverse probability of treatment weighting, model adequacy requires comparing the exposed control group to the sample matched on treatment's inverse probability weights;
- for stratification, the adequacy involves comparing the exposed control group to the sample matched within strata of the propensity scores.

For categorical variables, the distributions between exposed and control subjects are evaluated whereas the means or medians are of interest with continuous variables, Austin (2011).

Diagnostics

The process of examining the adequacy of your propensity score model, also known as 'matching diagnostics' and or diagnosing the quality of matches, is a critical step for any researcher to embark on (Stuart, 2010). It is referred to as "perhaps the most important step" in the process by Stuart and is vital in evaluating the validity or quality of the matching model. Generally, investigators (or researchers) are advised to reject any matching techniques whose response samples are lacking in terms of stability (or balance). Several balance metrics, or diagnostic methods of matching algorithms, are discussed below, divided into numerical and graphical diagnostics subsections, as detailed in the Stuart (2010) article.

3.3.1 Numerical balance diagnostics

Stuart (2010) states that this is the most frequently used balance metric.

For continuous variables:

The t-test mean difference is typically used. This metric is given in mathematical definition as follows (Stuart, 2010);

$$SMD = \frac{\bar{X}_\tau - \bar{X_C}}{\sigma_\tau}, \tag{53}$$

where SMD is the standardized mean difference, $\bar{X}_\tau$, $\bar{X}_C$ are sample covariate means for the exposed and control groups, and σ_τ is the pooled treatment standard deviation, the bias is given as $\mu_\tau - \mu_c \sim \bar{X}_\tau - \bar{X}_C$. Rosenbaum and Rubin (1985) refer to this as the standardized-bias or the standardized difference of the means. The rule of thumb

is that the absolute standardized-bias must be less than 5%, with the percentage bias reduction constrained at greater than 80%, else balance is compromised. We can also rewrite Equation (53) as follows:

$$SMD = \frac{\mu_\tau - \mu_C}{\sqrt{\frac{s_\tau - s_C}{2}}}. \tag{54}$$

For binary variables:

The standardized mean difference, where a norm set-up is dummy-coding multi-level categories (Austin, 2011), is defined by

$$SMD = \frac{\hat{\pi}_\tau - \hat{\pi}_c}{\sqrt{\frac{\hat{\pi}_\tau(1-\hat{\pi}_\tau)+\hat{\pi}_C(1-\hat{\pi}_C)}{2}}} \tag{55}$$

where $\hat{\pi}_\tau$ and $\hat{\pi}_C$ are sample proportions in the intervention and control groups, respectively.

The standardized bias:

This measure is calculated (with the rule of thumb that it be less than 5%.) as:

$$SB = \frac{Bias}{\sqrt{\frac{\nu_\tau + \nu_c}{2}}} \times 100 \tag{56}$$

where $Bias$ represents the bias measure.

The percent bias reduction:

Noting that, the rule of thumb here is for a statistic that is greater than 80%.

$$PBR = \frac{Bias_{before} - Bias_{after}}{Bias_{before}} \times 100 \tag{57}$$

where $Bias_{before}$ and $Bias_{after}$ are the bias measures before and after matching, respectively.

3.3.2 Graphical balance diagnostics

The wearisome nature of diagnosing the quality of matches using numerical metrics, especially with high dimensional data, is what necessitates the usefulness of plots in assessing feature balances (Stuart, 2010). There following are plot diagnostic methods Stuart (2010), applicable in practice:

(1) Propensity score distribution checks, from the original then matched group (similar to when checking for the common support, see Figure 7);

(2) QQ-plots, whose examination is by comparing each of the treatment groups per quantiles compared in each group (for similar empirical distributions of the treatments, the QQ-plot will produce a 45-degree straight line);

(3) the standardized mean difference plots, plotted to view how subject covariates improve their balance after matching (here the researcher investigates, specifically the increase or drop of the standard difference in means);

(4) histograms for the propensity score for each characteristic (*for instance, SPSS software enable one to produce population pyramid*s).

3.4 Outcome analysis

Once balance diagnostics related to confounder variables are done, and assuming they were done properly and to the contentment of the researcher, the next step in the process is response analysis (Olmos & Govindasamy, 2015). Since covariate confounder control is now intact, causal inference can be made assuredly, and with confidence (Baek, Park, Won, Park & Kim, 2015). Covariate balance will have implication that, Baek, et al. 2015, the matched set will be ripe for analysis of response difference between the treatment groups. Therefore, the matching results are of critical importance in outcome analysis methods and will always be taken into account.

Typical hypothesis

A typical hypothesis being tested with outcome analysis is that of the treatment having no effect on the outcomes, in other words the impact of the intervention is as good as if it were non-administered. This null hypothesis is tested against the alternative that an apparent treatment effect is evident. Our approach here will be concerned with estimating the treatment effect and its confidence interval, hence giving the required test statistics.

$$H_0\text{: } \textit{there is no treatment effect}$$

$$H_a\text{: } \textit{the treatment had an effect}$$

Olmos & Govindasamy (2015) detail how the near neighbour and Mahalanobis distance metrics can be used with techniques such as $ANCOVA$, linear regression models and or even matched t-tests for outcome modelling. It is also important to bear in mind, Baek, Park, Won, Park & Kim (2015), that these matched pairs may still exhibits intra-individual covariate differences even though between-group distributions seem to depict some degree of consistency.

3.4.1 Randomization tests

When running randomization tests, also known as permutation and or exact tests, the following are the steps that researchers must take (Roy, 2018);

- first compute test statistics from observed data, then
- assume that the null hypothesis of the treatment not having an impact is true, now
- permute the treatment allocations within pairs at random and recalculate the test statistic, and
- iterate the above steps a number of times checking for changes and or consistency in your observed test-statistic

The test statistic should thus lie around the mean of the distributions of these iterated computations in order to conclude that the treatment effect made a difference. Note that, the above step may easily be accomplished, straight forward, by using the McNemar test statistic instead of iterative randomization (Fay, 2016). Due to the fact that the McNemar tests are a straightforward equivalent of the randomization tests, Roy (2018).

3.4.2 McNemar test

In general, a McNemar's test is applied whenever the significance of the relationship between response and treatment group variables needs pinpointing (Baek, Park, Won, Park & Kim, 2015). Thus given a contingency table, also known as a confusion matrix, for binary response data we will test for treatment and control outcomes, Fay (2016). And the McNemar test statistic, of which is approximately "*chi-squared*", is thus evaluated in the following manner:

$$\chi^2 = \frac{(b-c)^2}{b+c}, \tag{58}$$

The argument, from Fay, 2016, is that 'only' the discordant pairs b and c gets used with this technique, as the other two will not help in identifying the impact of the treatment. McNemar test will be distributed $chi-square$, as mentioned above, with freedom degree 1, ${\chi_1}^2$. And for large sums of c and b such a closer approximation, as stated in Fay (2016), requires continuity correction and will thus be given below as:

$$\chi^2 = \frac{(\left|b-c\right| - 1)^2}{b+c}, \tag{59}$$

In the case of the exact binomial test, when sample size is small, the exact two-tailed $p-value$ is calculated as follows, under the null hypothesis of the odds-ratio being 1 (Fay, 2016),

$$p_{val} = 2\sum_{k=b}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) 0.5^k (1-0.5)^{n-k}, \tag{60}$$

where n is the sum $(b + c)$ of the discordant pairs. Note that Equation (58) - (60) are based on the McNemar confusion matrix, Table 2 depicted below.

Table 2: McNemar's confusion matrix
(Source: Fay, 2016)

		treatment τ	
		0	1
control	0	a	b
	1	c	d

3.4.3 The paired t-test

In case of continuous outcome data, a paired t-test is an equivalent to the McNemar test (Baek, Park, Won, Park & Kim, 2015). This test can either be used or the Wilcoxon signed-rank test to test for inter-group differences (Baek, et al., 2015). Further note that, for binary response and or survival event data other listed outcome models exists, available for use, and they are detailed below.

3.4.4 Conditional logistic regression

For matched binary outcome data, or binary response case, it is as good a strategy to apply conditional logistic regression outcome models (Roy, 2018). This method is detailed more, or further, by Kuo, Duan, & Grady, in their 2018 article. Here the authors, Kuo, Duan & Grady (2018), infer the use of conditional logit on the "sparse data problem", of which is a phenomenon associated with loose-matching or matching when you have few confounders.

3.4.5 Stratified Cox model

In survival analysis, or when dealing with time-to-event outcome data, a prudent strategy is that of stratified log-rank test and or of a stratified Cox proportional hazard model (Baek, Park, Won, Park, & Kim, 2015). This is due to the case that the baseline hazard will be stratified on the matched sets (Roy, 2018). Though, at individual covariate level, a worry is that of possible pairwise differences despite collective distribution similarity per group (Baek, et al., 2015).

3.4.6 Generalized estimating equations (GEE)

This technique allows for the group specifier variable to be matched. In case of dichotomous outcomes, like our binary treatment case, the causal risk difference, causal risk ratio, or

causal odds-ratio (depending on your link function) will have to be estimated (Roy, 2018). Also note that for a binary response case, the conditional logistic regression model above may be as good enough as the GEE for logistic regression as stated by Baek, et al. (2015) (Since this method allow for the use of odds-ratios as treatment measures for further requirement).

3.5 Sensitivity analysis

We discuss the concept of sensitivity analysis through diagnostics of heterogeneity, the possibility that multiple studies may not show the same effect, and then the common support failure, these two are the Rosenbaum and Lechner bounds, respectively (Lechner, 2000 and Rosenbaum, 2002). This is to check for hidden-bias, especially in cases where there were unobserved (omitted or 'missed') confounders. Glynn, Schneeweiss & Stürmer (2006) noted that matching methods will, unlike randomized experiments, most likely balance confounders that were only observed. This necessitates the use/ need of sensitivity analysis methods to some extent.

In Caliendo & Kopeinig (2005), the Rubin and Lechner bounds are discussed as means for performing sensitivity analysis. Stuart (2010) states that analyzing the sensitivity of unobserved variables is performed to eliminate the treatment effect observed, in cases where correlations between the observed and the hypothetical unobserved covariate features are to be monitored. Marco Caliendo and Sabine Kopeinig separated their section on sensitivity analysis into the following two parts,

(1) the unobserved heterogeneity case, and

(2) the failure of the common support,

i.e. the Rubin and Lechner bounds, respectively. These two cases are discussed below.

3.5.1 The unobserved heterogeneity case

For the Rubin bounds, the assumptions of "strong ignorability" as given by Equation (26) must be solidly satisfied (valid), *i.e.* the probability of allocation for treatment is conditionally independent and within the common support. We rewrite Equation (26) from Caliendo & Kopeinig (2005) as:

$$p_i = p(\boldsymbol{X_i}) = P(z = 1 \Big| \boldsymbol{X}_i) = F(\beta \boldsymbol{X_i} + \gamma \boldsymbol{u}_i), \tag{61}$$

where $\boldsymbol{X}_i$ are the observed confounders (or characteristics) for subject i, and letting $p(\boldsymbol{X}_i) = p_i$ so that our notation above get simplified, β is the impact of $\boldsymbol{X}_i$ on the

treatment allocation, $\boldsymbol{u}_i$ represents the unobserved confounder variable, and γ is the impact that the unobserved confounder $\boldsymbol{u}_i$ have on treatment allocation.

Further note that, a hidden-bias free study will have $\gamma = 0$. Thus the treatment allocation probability will be determined through the features $\boldsymbol{X}_i$ of each subject only. However, as noted in Caliendo & Kopeinig (2005), a process exhibiting hidden-bias will tend to produce distinct propensity scores even for a similar pair subjects, on all baseline characteristics. In other words, two subjects will have different treatment allocation probabilities even though they are an exact identical pair.

Under the assumption that function F in Equation (61) is logistic, the odds–ratio of a given matched pair i and j will be:

$$\frac{\frac{p_i}{(1-p_i)}}{\frac{p_j}{(1-p_j)}} = \frac{p_i(1-p_j)}{p_j(1-p_i)} = \frac{e^{(\beta \boldsymbol{X}_j + \gamma \boldsymbol{u}_j)}}{e^{(\beta \boldsymbol{X}_i + \gamma \boldsymbol{u}_i)}}, \tag{62}$$

where $\frac{p_i}{(1-p_i)}$ and $\frac{p_j}{(1-p_j)}$ are the odds-ratios of treatment exposure for subjects i and j, respectively. We also note that the implications of Equation (62) are that the odds-ratio will be bounded in nature, Rosenbaum (2002). And that matched subjects will be those allocated this treatment:

$$\frac{1}{e^{\gamma}} \leq \frac{p_i(1-p_j)}{p_j(1-p_i)} \leq e^{\gamma} \tag{63}$$

i.e.

$$\Rightarrow \frac{1}{\Gamma} \leq \frac{p_i(1-p_j)}{p_j(1-p_i)} \leq \Gamma,$$

where $\Gamma = e^{\gamma}$ for some $\gamma > 0$.

This means that the propensity score, or the treatment allocation probability, is similar only when $\Gamma = 1$ or in a hidden-bias free scenario. Effectively, this will imply the perfect case of randomized-trials design (Li, 2011). And that for $\Gamma = 2$ subjects will have odds of treatment that differ by a factor of 2, regardless of their covariates being similar (Rosenbaum, 2002 and Caliendo & Kopeinig, 2005). Nicely put in Li (2011) as, sufficient adjustment of pre-treatment covariates implies that the presence of a hidden or unobserved confounder will be twice as likely in the intervention group as in control. Thus, application of this Γ measure is best used in the monitoring of any strays from a hidden-bias free study (Rosenbaum, 2002).

3.5.2 The failure of the common support case

For the case pertaining to the failure of common support, recall from Section 3.2.1 Assumption 2 that subjects outside the common support region will be extrapolated. This is due to impossibilities in matching extreme values from either treatment group. As a consequence, the researcher has options of letting go of causal inference or hypothesis and

impose structural imputation with common support failure (Xie, Brand & Jann, 2012). But extrapolation or removal of subjects outside this region tends to break down the process by allowing for the phenomenon referred to as "ignoring the common support" (Lechner, 2000 and Caliendo & Kopeinig, 2005). The authors argue further that this is most likely due to heterogeneity of the causal effect.

Graphical solution

Figure 7 (Section 3.2.1) is a typical graphical depiction of the common support region, bounded by dotted vertical lines, between the distributions of the propensity scores for both treatment groups (Zaga Szenker, 2015). Whenever a large proportion of subjects are removed in the analysis, the problem or failure of the common support arises (Zaga Szenker, 2015). Although validity of the common support condition assures us that the groups will be compatible, problems may still arise at the extremes. This happens especially when exposed subjects exhibit higher propensity scores whilst control group subjects have lower propensities. This means that we will have no matches at the tails between control and exposed subjects as a consequence, in other words the is no overlap.

Analytical solution

Lechner, 2000, devised a non-parametric robustness check for when causal effects were estimated in midst of common support failure, or when some subjects exist outside of the common support region. Working out the math behind this approach requires that information be gathered from these outlying subjects (Lechner, 2000 and Caliendo & Kopeinig, 2005). In both papers, a new notation was devised and we introduce it for the purpose of our discussion below.

We'll define Ω as the subset of the treatment exposure space, where the treatment status was defined as $z \in \{0,1\}$, from a binary treatment perspective as is in Equation (1) above, and the pre-treatment characteristics set is given by X. Now, we let $\Omega^{\tau_{ATT}}$ be defined by $\{(z = 1) \times X\}$ such that it be the subset of the treatment exposure that only contains subjects enrolled to the said treatment, and then define

$$W^{\tau_{ATT}} = \begin{cases} 1, & if \quad observed \quad subject \quad \in \quad \Omega^{\tau_{ATT}} \\ 0, & otherwise \end{cases} \tag{64}$$

Further, we set $W^{\tau_{ATT}*} = 1$ when observations are within the region of common support and $\widetilde{\Omega}^{\tau_{ATT}}$ will be the complement subset, *i.e.* the subset where such an effect is not present. And now, let $P(W^{\tau_{ATT}*} = 1 \,\big|\, W = 1)$ denote the share of subjects that are within the common support region, relative to the total number, and λ_0^1 depicts the

mean of the treatment response Y_1, the expected outcome, for subjects outside the support region. The average treatment effect on the treated τ_{ATT} (ATT) of interest exists when the second assumption is applied (Lechner, 2000). Therefore, the following are identifiable (Caliendo & Kopeinig, 2005);

$$P(W^{\tau_{ATT}*} = 1 \Big| W^{\tau_{ATT}} = 1)$$

and

$$\lambda_0^1 = \mathbb{E}(Y^1 \Big| W^{\tau_{ATT}*} = 1 \Big| W^{\tau_{ATT}} = 1)$$

In addition, using the fact that the potential outcome of the control subjects Y_0 is bounded implies that

$$P(\underline{Y} \leq Y_0 \leq \overline{Y} \Big| W = 0 \Big| W^{\tau_{ATT}} = 1) = 1.$$

Now, from all these assumptions, the Lechner bounds for the average treatment effect on the treated (ATT) $\tau_{ATT}(\Omega^{\tau_{ATT}}) \in [\underline{\tau}_{ATT}(\Omega^{\tau_{ATT}}), \overline{\tau}_{ATT}(\Omega^{\tau_{ATT}})]$ can be re-written as, Caliendo & Kopeinig (2005);

$$\underline{\tau}_{ATT}(\Omega^{\tau_{ATT}}) = \tau_{ATT}(\Omega^{\tau_{ATT}}) \cdot P(W^{\tau_{ATT}*} = 1 \Big| W^{\tau_{ATT}} = 1)$$

$$+(\lambda_0^1 - \overline{Y}) \cdot [P(W^{\tau_{ATT}*} = 1 | W^{\tau_{ATT}} = 1)] \tag{65}$$

$$\overline{\tau}_{ATT}(\Omega^{\tau_{ATT}}) = \tau_{ATT}(\Omega^{\tau_{ATT}}) \cdot [P(W^{\tau_{ATT}*} = 1 \Big| W^{\tau_{ATT}} = 1)]$$

$$+(\lambda_0^1 - \underline{Y}) \cdot [P(W^{\tau_{ATT}*} = 1 | W^{\tau_{ATT}} = 1)] \tag{66}$$

$\Omega^{\tau_{ATT}}$ is a representation of the effect of ignoring subjects outside the common support region, *i.e.* those with neither a perfect match, Lechner, (2000). Lechner further warns researchers that pleading ignorance of the common support problem, through exclusively making estimations of τ_{ATT} using only subjects inside the region, would be detrimental. Readers keen on in-depth understanding of these common-support-failure scenarios may consult the following articles, Lechner (2000) and Caliendo & Kopeinig (2005). Note that, in this dissertation we will only focus on the heterogeneity case in the application of sensitivity analysis.

4

Applications

Success rates for first year introductory level modules are generally low, especially in statistics courses (Reyneke, Fletcher, & Harding, 2018). The authors argue that this problem is of concern among universities across South Africa, even at the global stage, and the University of Pretoria is not immune. They describe various interventions that have been introduced by the university in order to try curbing these persistent low success rates. Of specific interest with this chapter, in particular, is determining whether the introduction of classroom response systems or clickers, for audience response and interaction in a blend of a traditional and inverted classroom set-up, was successful in improving the average final examination mark of participating students.

The study comprises two cohorts of students enrolled for a first year introductory level module, in the first semesters of 2014 and 2017 respectively. The 2014 cohort represents the class without clickers use and the 2017 cohort is the intervention group, or the group that got to utilize the said classroom response system. Respective cohort sizes were 1,625 in 2014 and 1,486 for the clickers class. In particular, our analysis is based on the investigation of whether propensity scores adjustment of the data, in order to balance pre-intervention covariates between the two cohorts, depict evidence of differences in the outcomes compared to simply utilizing these two groups in their raw form.

Flipped classrooms are those organized in such a manner that usual class activities get flipped around with those that students would traditionally do at home and vice versa (Lage, Platt, & Treglia, 2000; Brame, 2013; Cabi, 2018; Crouch & Mazur, 2001; Du, 2011; Nouri, 2016 and Zainuddin & Halili, 2015). Scholars report success for flipped classes, in literature, compared to those where content is still relayed in the traditional fashion (Zainuddin & Halili, 2015). These pedagogical models are proving to be beneficial, especially for underperforming students, in improving performance and confidence in tackling and engaging with the subject matter and or other students, hence strengthening the case for proponents of such classrooms in institutions of higher learning and even at basic education level (Nouri, 2016). Much better success is reported, especially when these classes

are blended and taken with aid of classroom response systems or clickers, as instruction support tools (Bojinova & Oigara, 2011 and Heid & Boshoff, 2011). This is because these devices allows for improved, continuous, and active student engagements, Siau, Sheng & Nah (2006), amongst students themselves, between them and their instructor, and in turn with the subsequent subject matter. In particular, the ability to poll the class solution replies right on-the-spot remains an attractive feature, for it allows instructors to focus on explaining the misunderstood concepts that seem to be troubling the class right there and then or as soon as possible (Bojinova & Oigara, 2011).

Study participants

Our study participants are cohorts of enrolled students for a first semester introductory first year statistics module. Mostly, these students are from the commerce faculty, known as Economic and Management Sciences (EMS), and were enrolled for the academic years 2014 and 2017. The 2014 cohort did not get to use the classroom response devices (clickers), in their time, meaning that 2017 enrollees are part of the exposed or intervention group. As mentioned above, a total of 1,625 students are in the 2014 cohort and 1,486 were enrolled in 2017.

Analysis

In order to estimate the differences in the average final mark between the given cohorts, outcome modeling, together with independent two-sample t-tests, will be conducted across all prior and post adjustment methods. Specifically, the mean difference in the outcome is estimated after first isolating the bias, that is natural phenomena with observational data, by controlling participants pre-treatment characteristics using propensity score methods (Kabunga, 2014). This is because student enrollment to either cohort was likely non-randomized, *i.e.* could have been through self-selection and or due to relevant faculty requirements. In particular, three propensity score matching methods were used, *i.e.* greedy with a caliper, optimal, and inverse probability of treatment weights, to match the two cohorts. The results are summarized in Table 6 in Section 4.4. Overall, we will follow the process steps from Chapter 3 when conducting this analysis.

Baseline covariates

Table 3 below displays the variables that were assessed, at baseline or pre-intervention, for use in propensity score application.

Table 3 Variables used for propensity score estimation

Variable (*name*)	**Description**	**Type**	***Sample data***
Study Outcome:			
Exam_mark	Module examination score.	Quantitative	0 - 100
Intervention:			
Clicker exposure	Treatment status or class.	Binary	1 (*2017*) 0 (*2014*)
Baseline covariates:			
Repeats	Indicator for module repeaters.	Binary	1 (*Yes*) 0 (*No*)
Gender_Desc	Student gender.	Binary	1 (*Female*) 0 (*Male*)
Race	Ethnicity group or race.	Categorical	*African, Coloured, Indian, and White*
Authority+	Matriculation authority.	Categorical	*The 9 Provinces, CMB, FRC,IEB, and NA*
Faculty+	Student enrolment faculty.	Categorical	*EMS, Humanities Law, NAS, and other*
Home_Language_Instruction	Indicator for home language instruction.	Binary	1 (*Yes*) 0 (*No*)
Math_Grade12	Matric mathematics mark.	Quantitative	20-99
Years prior	Years prior first attempt.	Quantitative (discrete)	1-18

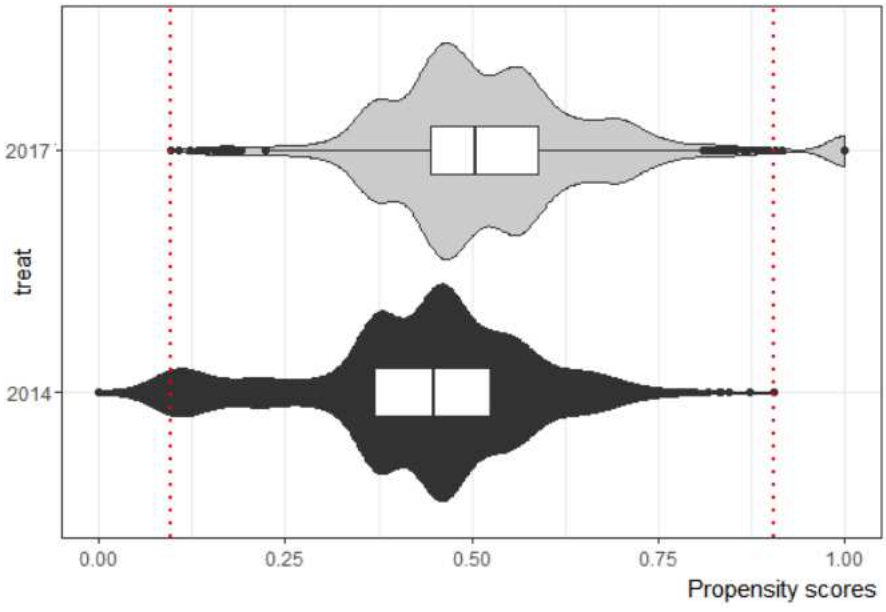
+The abbreviations, pertaining to the respective factor levels of the covariates examination authority and faculty, in Table 3 are as follows: Student matriculation authorities are EC (Eastern Cape), FS (Free State), GP (Gauteng), LI (Limpopo), MP (Mpumalanga), NC (Northern Cape), NW (North West), and finally WC (Western Cape). We also have CMB (Cambridge), IEB (Independent Education Board), NA (unknown matric authority), and lastly FRC represent students whose matric was done in a Foreign country. Further, the typical faculties are shortened as EMS (Economic and Management Sciences), and NAS (Natural and Agricultural Sciences).

4.1 Propensity score estimation

In order to estimate student's probability of being exposed to clickers, *i.e.* their propensity scores, we fit a logistic regression classification model of the following form (based on the observed pre-treatment characteristics in Table 3 above);

$$\begin{aligned} log(p_i) = log\left(\frac{1}{1+e^{-\mathbf{x}_i\beta_i}}\right) \quad = \quad & \beta_0 + Repeats\cdot\beta_1 + Gender\cdot\beta_2 + Race\cdot\beta_3 \\ & + Authority\cdot\beta_4 + Faculty\cdot\beta_5 + HomeLanguage\cdot\beta_6 \\ & + Grade12Math\cdot\beta_7 + YearsPrior\cdot\beta_8 \end{aligned}$$

This results in the parameter estimates of Table A1. And the subsequent estimation equation, based on significance of the parameter results, is also part of the Appendix section. After successfully fitting the logistic regression classifier, in order to estimate the propensity scores, one needs to check if the estimated scores validate strong ignorability. This property requires that both the 'ignorability' and 'positivity' causal assumptions are met for the combined cohorts. Visualizing the estimated scores as in Figure 8 and 9 below, which depict the common support or overlaps assumption per treatment group, is one way this can be done. Figure 8 below gives a panoramic view of the common support region per exposure group.



º(treat) is the cohort or treatment class

Figure 8 Graph of the common support

The ignorability assumption

Unconfoundedness is an assumption met through the fact that the potential outcomes do not depend on the cohort the students are in but is only conditional on the observed baseline covariates.

The positivity assumption

The common support condition plots in Figures 8 and 9 show existence of positive probabilities of enrolling students, whenever the baseline characteristics depict some degree of similarity across the treatment groups, due to substantial overlaps. And, for our data space, such a region is given by the interval (0.097, 0.904), *i.e.* the red dotted vertical lines in the plots, due to there being no treated (2017) and control (2014) student enrollees below and above it respectively. Therefore we conclude that there exists great deal of overlap in propensity scores in the probability interval (0, 1) thus satisfying the common support condition. Note that Figures 8 and 9 are two sides of the same coin and allow for an inspection of the common support condition from differing angles.

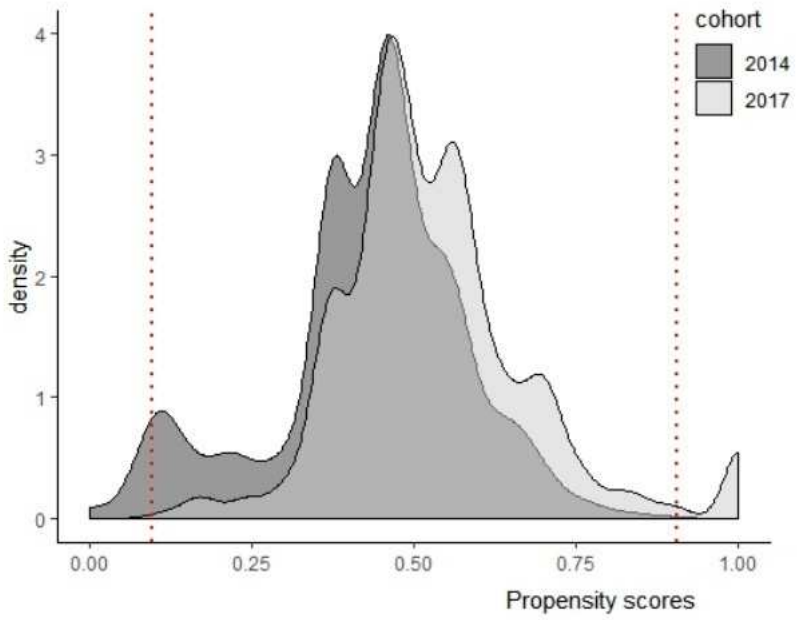


Figure 9 Common support region (traditional view)

Extreme value discrepancies, in the propensity scores outside the interval of common support, between the enrollment classes are visible in both Figures 8 and 9. These plots characterize the plausibility of the common support or positivity assumption in a sense that only a few students, visible at both ends of the interval, may get extrapolated through the matching techniques. In this way, post-adjustment data will tend to have similar

distributions to the prior case. Since the validity of both the conditional independence and positivity assumptions satisfies strong-ignorability, the next step in the analysis is the matching step.

4.2 Matching using estimated propensity scores

Displayed in Table 4 below is the breakdown or splits, in terms of matched participants from all three matching techniques for the 3,111 students (1,625 in the 2014 control cohort and 1,486 in the 2017 treatment group).

Table 4 Post-matching group counts for the three methods

Matching method	*Treated matched*	*Control matched*	*Treated unmatched*	*Control unmatched*
Greedy with caliper	1,430	1,430	56	195
Optimal	1,486	1,486	0	139
IPW	3061.7	3081.4	0	0

From Table 4 we see that only 251 and 139 participants get dropped, respectively from both the greedy with a caliper and optimal matching methods, which is 8.07 and 4.47% of the combined cohorts. These large post-adjustments samples support the substantial pre-adjustment overlaps. In other words there is homogeneity in the characteristics of enrolled students across the two cohorts. Note that the pseudo-populations or synthetic-samples, obtained from the inverse probability weighting technique, created synthetic samples of about 3,062 (from 1,486) treatment and 3,081 (from 1,625) control subjects each.

4.2.1 Greedy matching (with a caliper)

Jitter plots are visual aids to inspect post-matching propensity score distributions of the data, with a nice feature of drilling-down on units. We applied a greedy matching technique, on the 2014 to 2017 cohorts, by matching at a ratio of 1:1 with a caliper of 0.2 times the estimated propensity score. Using package *MatchIt* (Ho, Imai, King & Stuart, 2011), the following post-adjustment jitter plot was produced. The 195 control and 56 treated subjects that remained unmatched are visible at the most lower and upper section of this

jitter plot, in Figure 10. Figure 10 below is the jitter plot visualization for the greedy with a caliper matching technique.

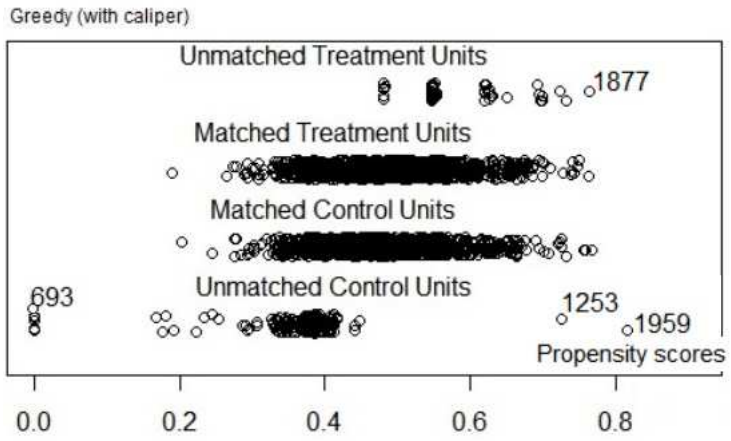


Figure 10 Greedy match jitter plot

4.2.2 Optimal matching

The optimal minimum distance is taken into account with this matching technique. Post to pre-intervention cohorts matching is done through the application of the *optmatch* package (Hansen & Klopfer, 2006). The 139 unmatched subjects were all from the 2014 (control) enrolment group. In Figure 11 these are consolidated in a jitter plot, and the 139 unmatched control subjects are at the bottom.

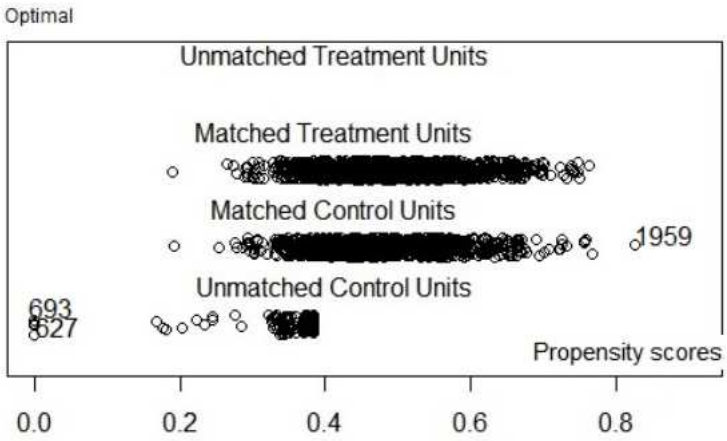


Figure 11 Optimal match('jitter' plot)

Drilling-down one can tell that student number 1959 got matched successfully with the optimal matching method, whilst the nearest-neighbour (greedy) with a caliper method could not match this 2014 student to an enrollee, refer to Figures 10 & 11.

4.2.3 Inverse probability of treatment weighting

For the inverse probability of treatment weighting technique, package *ipw* (van der Wal & Geskus, 2011) was utilized to create pseudo-populations or synthetic samples. Post-adjustment figures evaluate to sample sizes of 3,061.70 (3,062) and 3,081.42 (3,081) subjects in the treated and control groups respectively, with the maximum and minimum weights equaling 14.5 and 1.0 in that order.

Visualizing inverse probability weighting weights

Figure 12 enable us to view the calculated weights from the inverse probability of treatment weighting method, in lieu of the above jitter plots. Weights are typically plotted so that they help in visual identification of abnormal weights, if any exists. These are weights that are either extremely large or too small. Abnormally large-scaled weights tend to be consequences of low treatment allocation probability (Thavaneswaran & Lix, 2008) and therefore may require truncation. Figure 12 depicts the graphical representation of pre- and post-truncation (at 1th and 99th percentiles) distribution of our inverse probability weights.

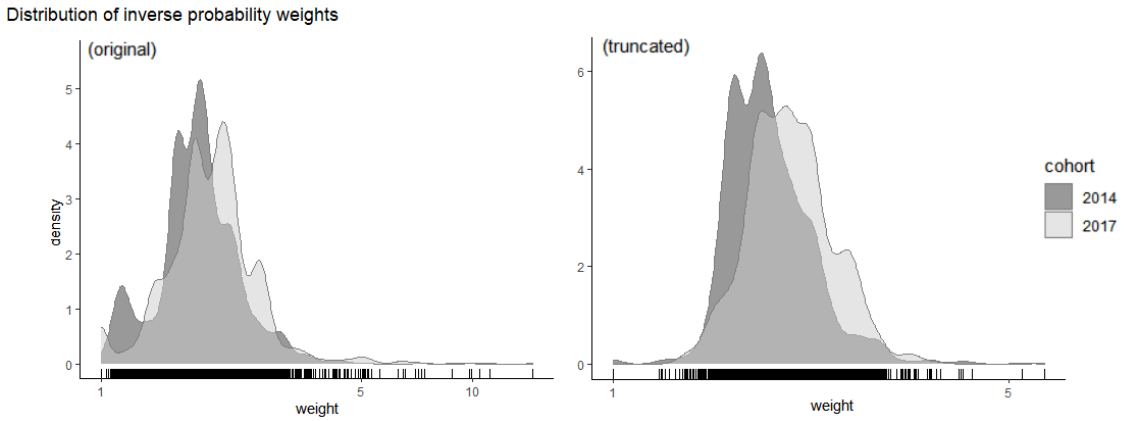


Figure 12 Weight distribution plots (pre- and post-truncation)

Clearly, there is no evidence of extremely large weights in Figure 12. Before truncation, the original minimum and maximum weight were given by 1 and 14.503 (peaking at mean 1.98), respectively. These were fairly reasonable weights, but still needed to be truncated, as a rule of thumb, for better analysis of the data in case skewed weights were observed. Post truncation weights, on the other hand, were respectively 1 and 5.792 (with mean 2.00).

4.3 Evaluating covariate balance

In performing balance diagnosis, methods from Zhang, Kim, Lonjon & Zhu (2019) are followed and then consolidated using relevant R software packages. We separate this section into numeric (Zhang et al., 2019) and graphical (Love, 2019) diagnostics for the inspection of pre- and post-adjustment covariate balance.

4.3.1 Numeric balance diagnostics

For all the three matching methods we have summarized descriptive statistics and frequencies (with their respective standardized mean differences), including the unadjusted case. These are combined in Table 5 below, from *tableone* (Yoshida, 2019) outputs.

Table 5 Covariate balance comparison (pre- and post-matching)

	Pre - match					Post - match											
	unadjusted					Greedy matching (with 'caliper')					Optimal matching					Inverse probability weights matching				
	*Standardized mean difference (SMD)															*Standardized mean difference (SMD)*				
Student baseline characteristics		**2014**		**2017**			**2014**		**2017**			**2014**		**2017**			**2014**		**2017**		
	*Size (**n**)*	***1,625***		***1,486***		***SMD****	***1,430***		***1,430***		***SMD***	***1,486***		***1,486***		***SMD***	*3,081.4*		*3,061.7*		***SMD***
Repeater = 1 (%)	98	(6.00)	131	(8.80)	*0.106*	84	(6.60)	95	(7.50)	*0.034*	95	(6.60)	131	(8.80)	*0.083*	248.9	(8.10)	253.9	(8.30)	*0.008*
Gender = Male (%)	790	(48.60)	812	(54.60)	*0.121*	644	(50.90)	657	(51.90)	*0.021*	767	(51.60)	812	(54.60)	*0.061*	1,603.00	(52.00)	1,580.90	(51.60)	*0.008*
Race (%)					*0.041*					*0.034*					*0.037*					*0.003*
African	584	(35.90)	531	(35.70)		478	(37.80)	460	(36.40)		545	(36.70)	531	(35.70)		1,113.20	(36.10)	1,109.80	(36.20)	
Coloured	31	(1.90)	33	(2.20)		25	(2.00)	28	(2.20)		30	(2.00)	33	(2.20)		58.8	(1.90)	58.3	(1.90)	
Indian	110	(6.80)	88	(5.90)		84	(6.60)	82	(6.50)		97	(6.50)	88	(5.90)		204.3	(6.60)	203.6	(6.60)	
White	900	(55.40)	834	(56.10)		678	(53.60)	695	(54.90)		814	(54.80)	834	(56.10)		1,705.10	(55.30)	1,690.00	(55.20)	
Authority (%)					*0.224*					*0.072*					*0.154*					*0.057*
CMB	15	(0.90)	25	(1.70)		14	(1.10)	18	(1.40)		15	(1.00)	25	(1.70)		39.6	(1.30)	41.3	(1.30)	
EC	26	(1.60)	22	(1.50)		17	(1.30)	15	(1.20)		24	(1.60)	22	(1.50)		45	(1.50)	42.1	(1.40)	
FRC	4	(0.20)	2	(0.10)		1	(0.10)	2	(0.20)		3	(0.20)	2	(0.10)		5.7	(0.20)	5.1	(0.20)	
FS	27	(1.70)	32	(2.20)		24	(1.90)	25	(2.00)		27	(1.80)	32	(2.20)		57.2	(1.90)	57.2	(1.90)	
GP	803	(49.40)	672	(45.20)		613	(48.50)	597	(47.20)		723	(48.70)	672	(45.20)		1,458.90	(47.30)	1,448.50	(47.30)	
IEB	286	(17.60)	357	(24.00)		253	(20.00)	275	(21.70)		285	(19.20)	357	(24.00)		634.2	(20.60)	642	(21.00)	
KZN	97	(6.00)	102	(6.90)		82	(6.50)	85	(6.70)		92	(6.20)	102	(6.90)		198.6	(6.40)	196.8	(6.40)	
LI	155	(9.50)	115	(7.70)		110	(8.70)	101	(8.00)		140	(9.40)	115	(7.70)		271.2	(8.80)	271.6	(8.90)	
MP	134	(8.20)	116	(7.80)		103	(8.10)	105	(8.30)		127	(8.50)	116	(7.80)		251.9	(8.20)	250	(8.20)	
NA	4	(0.20)	0	0.00		0	0.00	0	0.00		0	0.00	0	0.00		4	(0.10)	0	0.00	
NC	10	(0.60)	3	(0.20)		5	(0.40)	3	(0.20)		2	(0.10)	3	(0.20)		12.7	(0.40)	8.9	(0.30)	
NW	56	(3.40)	33	(2.20)		36	(2.80)	32	(2.50)		40	(2.70)	33	(2.20)		88.6	(2.90)	84.3	(2.80)	
WC	8	(0.50)	7	(0.50)		36	(2.80)	32	(2.50)		8	(0.50)	7	(0.50)		13.7	(0.40)	14	(0.50)	
Faculty (%)					*0.110*					*0.058*					*0.094*					*0.063*
EMS	1,328	(81.70)	1,207	(81.20)		1,118	(88.40)	1,097	(86.70)		1,208	(81.30)	1207	(81.20)		2,536.30	(82.30)	2,528.20	(82.60)	
Humanities	18	(1.10)	31	(2.10)		13	(1.00)	18	(1.40)		18	(1.20)	31	(2.10)		52.7	(1.70)	49.9	(1.60)	
Law	1	(0.10)	0	0.00		0	0.00	0	0.00		0	0.00	0	0.00		1	0.00	0	0.00	
Other	20	(1.20)	30	(2.00)		8	(0.60)	7	(0.60)		20	(1.30)	30	(2.00)		28.6	(0.90)	46.8	(1.50)	
NAS	258	(15.90)	218	(14.70)		126	(10.00)	143	(11.30)		240	(16.20)	218	(14.70)		462.8	(15.00)	436.8	(14.30)	
Instruction in home language = 1 (%)	842	(51.80)	707	(47.60)	*0.085*	627	(49.60)	615	(48.60)	*0.019*	738	(49.70)	707	(47.60)	*0.042*	1,524	(49.50)	1,521.20	(49.70)	*0.005*
Grade12 math (mean (SD))	72.50	(9.86)	72.70	(9.18)	*0.021*	72.8	(9.71)	*72.87*	*(9.21)*	*0.008*	72.63	(9.88)	72.7	(9.18)	*0.007*	72.64	(9.94)	72.6	(9.18)	*0.004*
Years prior (mean (SD))	1.73	(1.80)	1.66	(1.45)	*0.041*	1.58	(1.37)	1.63	(1.44)	*0.04*	1.67	(1.67)	1.66	(1.45)	*0.009*	1.73	(1.72)	1.74	(1.63)	*0.005*

**SMD ~ Standardized mean difference; 1 ~ treatment group (2017); and 0 ~ control group (2014).*

Table 5 shows only four baseline covariates that achieved balance at the maximum allowable threshold of 0.1, and these are home language instruction, matric mathematics grade, years prior attempting the module, and student racial demographics, respectively (0.085, 0.021, 0.041 and 0.041). The rest of the baseline covariates therefore lack in balance and justify that matching methods can be utilized. Prior to adjusting on propensity scores five of the baseline covariates differ between the groups, with standardized mean differences greater than 5% as seen in Table 5. These were the proportion of males (12.1%), module repeaters (10.6%), home language instruction at (8.5%), and the combined average authority SMD (22.4%), together with faculty percentages (11%). Post-adjustment cases from the greedy with caliper and optimal matching methods seem to balance almost all the covariates at 5% threshold, except for the following: education authority (7.2% and 15.4%), and faculty (5.8% and 9.4%) for both methods respectively. Optimal matching has male (6.1%) and repeater (8.3%) student proportions as unbalanced.

Applying inverse probability of treatment allocation weighting shows all the good signs of balance, that is all covariates have SMDs less than the 5% threshold. In Table 5 the covariate standardized mean difference (SMD) measures are in brackets and mostly balanced, for the relaxed 10% threshold, except for examination authority (15.4%) under optimal matching. In Figure 13, this is shown as a direct consequence of the imbalanced distribution of student proportions from the independent education examination authority (IEB).

4.3.2 Visual balance diagnostics

The standardized mean difference measures in Table 5 may also be visualized. In order to visually inspect covariate balance between treatment groups, the following sample balance diagnostic plots, also known as love plots, were constructed using packages *tableone* (Yoshida, 2019) and *cobalt* (Greifer, 2020) in R (Core, 2019). Figure 13 below depicts baseline covariate balance visualization, prior and post-matching on all three propensity score adjustment methods.

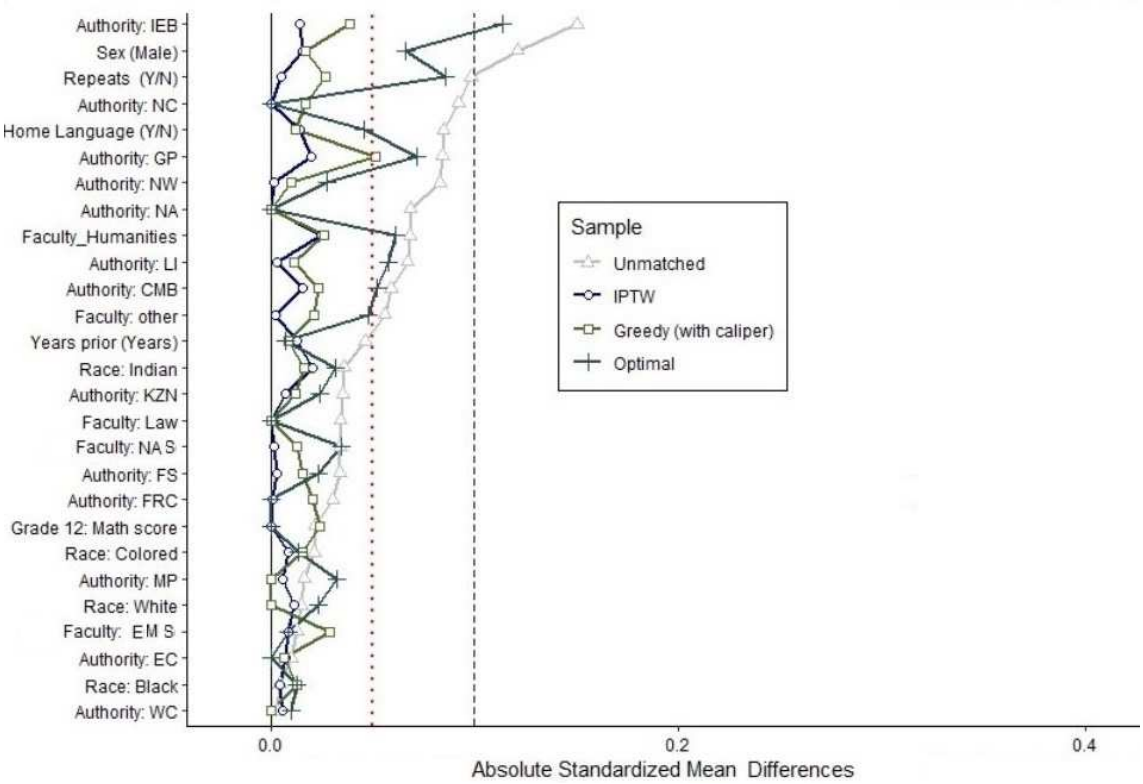


Figure 13 Balance diagnostics all cases (incl. pre-matching)

Figure 13 depicts differences in the ordered standardized mean differences (SMDs) at thresholds 5%, and also at the maximum 10%. All these covariates seem to balance when matching with the greedy with caliper and inverse probability weights matching techniques. Both these techniques' depicts absolute standardized mean differences of less than 5% for the baseline covariates, *i.e.* below the red dotted vertical line.

It is clear that optimal matching lags slightly behind the other two methods, in terms of balancing for some covariates, especially at the 5% threshold but still shows a fairly balanced covariate space for the relaxed 10% threshold. And thus, it is visible from Figure 13 that if optimal matching is an applied technique and that a smaller than the maximum allowable threshold is enforced for balances, a couple of covariates will not meet our high balance standards. Such covariates are male gender representation, repeaters, Humanities enrollees and those matriculated from the independent education sector, Cambridge, Limpopo and Gauteng authorities.

4.3.3 Visualizing individual covariate balance

This subsection continues with covariate balance diagnostics through the plots, but at individual covariate level, to identify improvements in each confounder control per matching technique. We continued using packages from R (Core, 2019), *i.e. tableone* (Yoshida,

2019) and *cobalt* (Greifer, 2020). A good measure of group balance for any given covariate is the degree of overlap for the densities, or proportions in case it were categorical, between the treatment groups (Greifer, 2020).

Distribution balance: Matric mathematics grade

Figure 14 below depicts the distribution of mathematics scores, for students in 2017 (1) and 2014 (0) respectively, prior to their admission to this introductory statistics module.

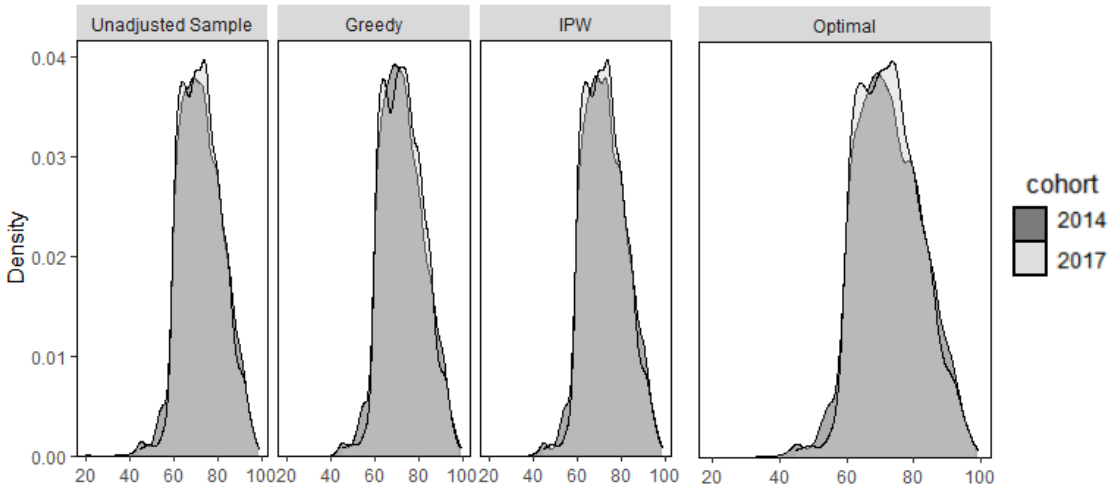


Figure 14 Balance diagnostics: math score or grade

The above plots show a large degree of consistency, in that the university tends to accept students with respective pre-matching mathematics marks of around median and mean of (60, 60.8) and above for the intervention group, across all three matching techniques. Comparatively, these figures were around (59, 60.4) and above for control group subjects. Subsequently this implies that, for the given data set, there is good density overlap for this covariate. Further, one can revisit the numeric balance table, *i.e.* Table 5, in order to validate and relate the mean, standard deviation, and or range, of this covariate, across all three matching methods.

Distribution balance: Race

Figure 15 below is a depiction of student ethnicity distributions across all three matching methods, including the unadjusted or unmatched case.

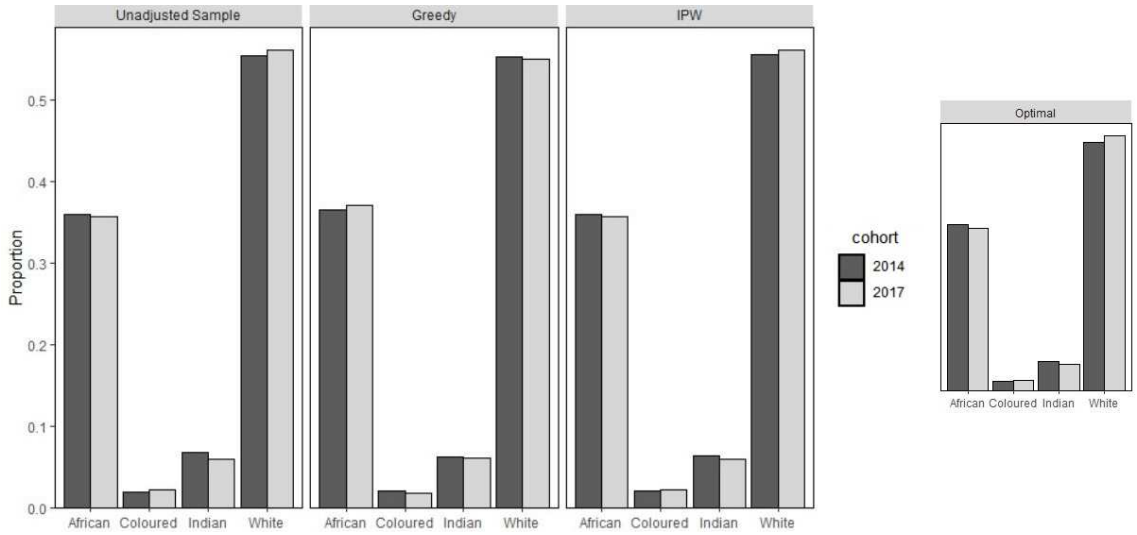


Figure 15 Balance diagnostics: racial demographics

For racial profile balancing, Figure 15 shows that the greedy with caliper and the inverse propensity score weighting methods are improving balance proportions from those prior adjustments. The optimal matching method, on the other hand, does not seem to change much of the visible overlaps prior to adjusting on it. Even though one may argue for fairly balanced, across all four levels, pre-matching racial demographics this balance still gets improved, albeit slightly, by the greedy and inverse probability weighting methods. The love plots in Figure 13 verify the threshold covariate balance for these four category levels post-matching.

Distribution balance: Gender

Figure 16 below visualizes gender distribution between the two enrolment classes across all three matching methods including the unmatched case.

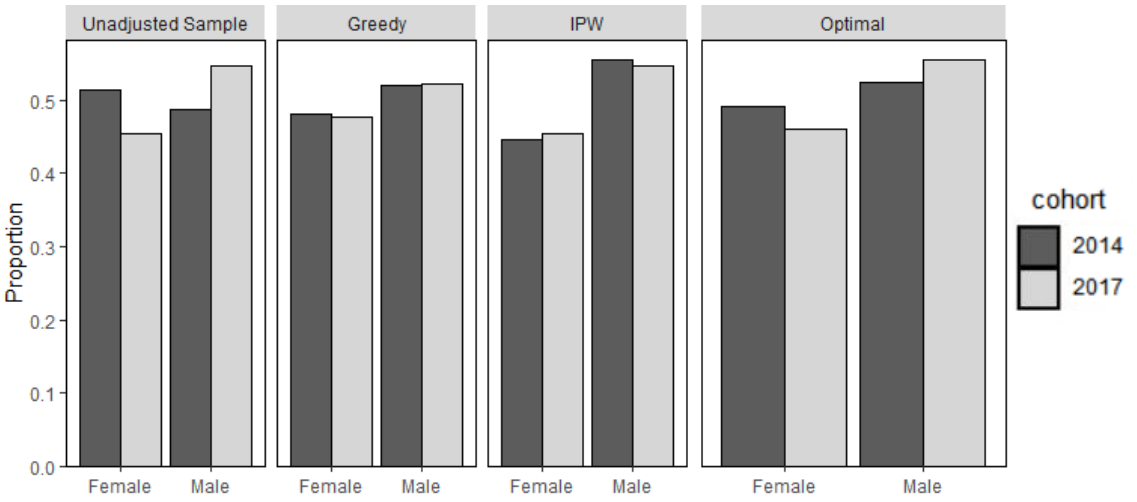


Figure 16 Balance diagnostics: gender representations

It seems that there is not much difference between pre- and post-matching meaning that pre-matching gender was fairly balanced across the two levels, except that greedy matching is seen to be slightly improving upon that balance. Optimal matching has also improved this covariate balance but not as much as the other two techniques, which corresponds to the nature of the love plots in Figure 13. Further note that, the volumes (proportions) for this pre-treatment characteristic, across all three methods and the unadjusted case, can be viewed in Table 5.

Distribution balance: Repeat students

Figure 17 below shows baseline covariate distribution for repeat students across all three matching methods and that of the unadjusted case.

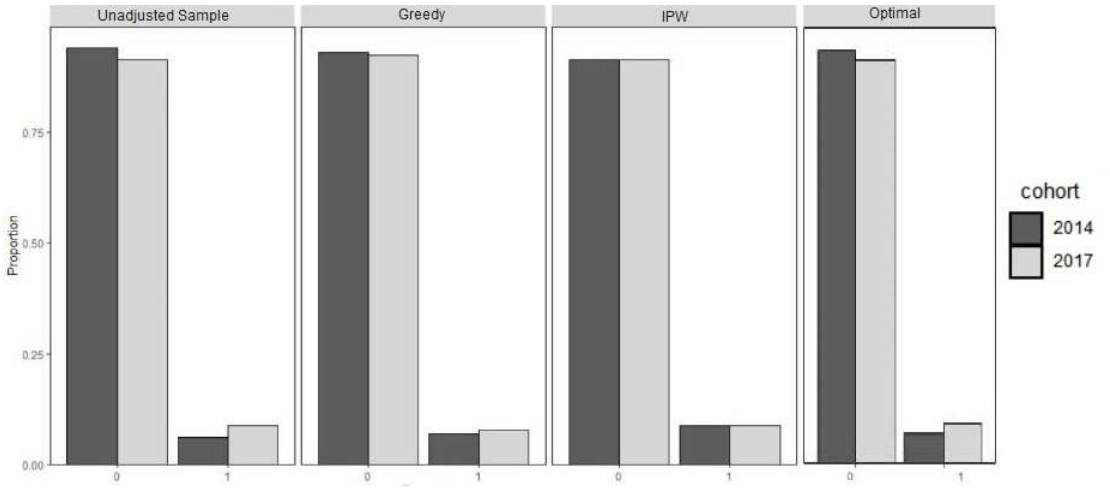


Figure 17 Balance diagnostics: repeating students

Figure 17 show that optimal matching does not change the balance within this covariate space. Even though balance of this covariate looks fairly reasonable, including for the unadjusted scenario on the far left, one can still clearly see some improvements from the other two adjustment techniques in the middle. Whereas, evidently looking on the far right plot such shifts were minor, if there at all, which implies that optimal matching did not improve balance for these category levels. From Table 5, one can view individual category level counts (proportions) for this covariate across all three methods and the unadjusted case.

Distribution balance: Examination authority

Figure 18 below shows baseline covariate distribution for examination authority across all three matching methods and that of the unadjusted case.

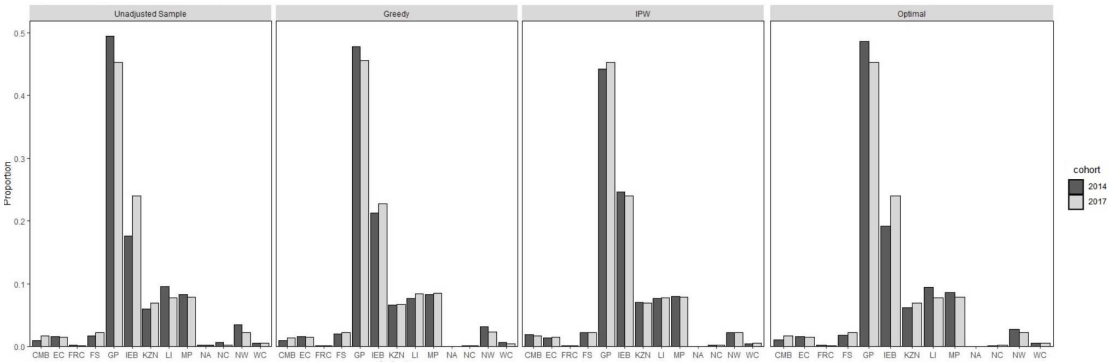


Figure 18 Balance diagnostics: examination boundaries

Distributions of individual category levels for this covariate also depict minor changes, for post-adjustment balance, due to optimal matching. Improvements in balancing due to the greedy with a caliper, and a near-perfect balance with inverse probability weighting are clearly visible though. One could argue that perhaps the 'synthetic' sampling inherent with the inverse probability weighting method allows for better balancing of this category due to matching within a bigger data space.

Distribution balance: Faculty

Figure 19 below shows student faculty covariate distribution across all three matching methods, including the unadjusted case.

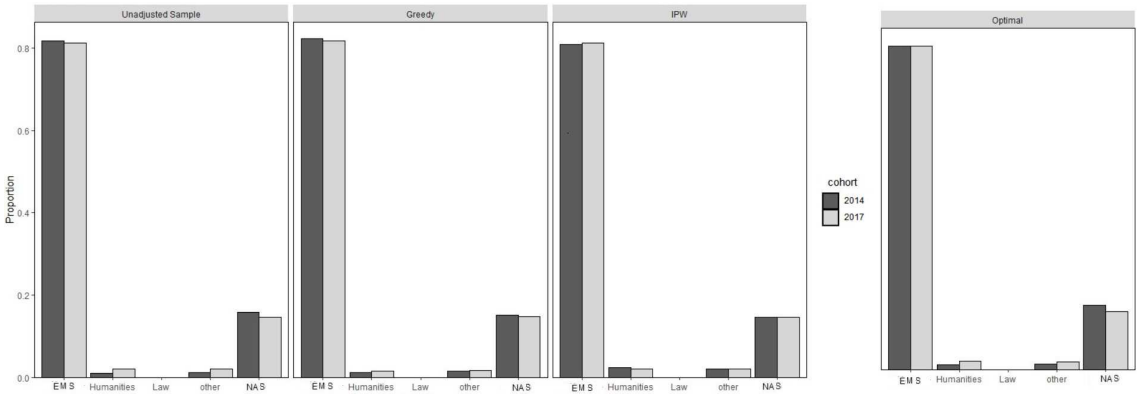


Figure 19 Balance diagnostics: Faculty representations

Few differences are observed in Figure 19 but it can be seen that weighting on the inverse propensity scores evens out the overlap proportions of the four Faculty category levels under this covariate. This is also observed in the love plots of Figure 13 above. Optimal matching, on the far right, shows the least changes or shifts in overlap proportions, not that much visible, in most covariates except for the EMS faculty, from the unadjusted state.

Distribution balance: Years prior attempt

Figure 20 shows the baseline covariate distribution for years prior attempting the module across all three matching methods, including the unadjusted case.

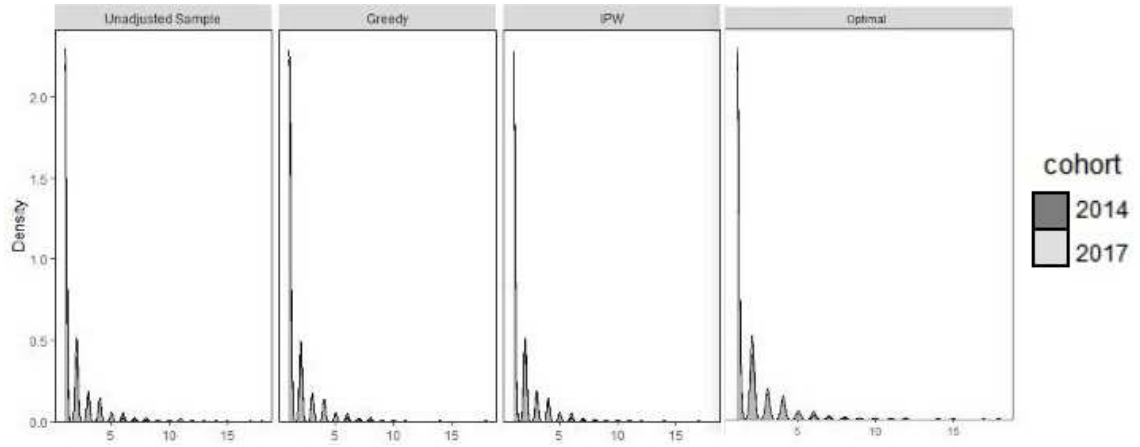


Figure 20 Balance diagnostics: years before attempt

The number of years a student takes from matric prior to attempting this module remains constant across all three techniques for the intervention groups, in terms of distribution. We thus conclude that there were not much balancing changes done through

matching for this covariate. One may argue that this kind distribution makes sense since students will most likely enroll straight after matric. Table 5 gives the mean (standard deviation) cohort breakdown for this covariate across all the matching methods and the unadjusted case.

4.4 Outcome modeling

With the covariate balance confirmed in Section 4.3, the next step is to estimate the treatment effect through an appropriate statistical technique, both the independent samples t-test and regression outcome model on the treatment status are applied here. The main focus of this study is to determine whether introducing clickers, in a blended flipped-classroom setup, impacted the final examination mark of enrolled students. And, in order to accomplish this, we test the hypothesis of no impact from the clickers' intervention against the alternative, as described in Section 3.4. That is,

$$H_0\text{: } \textit{post-clickers average examination mark} = \textit{pre-clickers average}$$

$$H_1\text{: } \textit{post-clickers average examination mark} \neq \textit{pre-clickers average}$$

Here we test for significance in the differences, on average, for examination marks between the cohorts. As mentioned above, an independent samples t-test, together with a causal outcome model, are applied to this end. Tables 6 below depicts all our outcome model results, in order to try keep things cleaner in the paper I listed the t-test results in Table A3 in the Appendix section, from the pre-matching case plus all three applied matching methods. All these are R (Core, 2019) outputs.

Tables 6 give all our outcome model results, across all three methods plus the unadjusted case.

Table 6 Outcome model results (pre- and post-matching)

Matching method	*Effect size*	*(drop)*	*Coefficient*	*Estimate*	*Standard error*	*t value*	*95% CI*	
Unadjusted	$n_1 = 1,486$		(constant)	58.243	0.43	136.97	(57.41, 59.08)	***
	$n_0 = 1,625$		2017 (*Yes*)	1.756	0.62	2.855	(0.55, 2.96)	**
Greedy with a caliper	$n_1 = 1,430$	(251)	(constant)	58.243	0.43	136.97	(57.69, 59.59)	***
	$n_0 = 1,430$		2017 (*Yes*)	2.276	0.69	3.32	(0.93, 3.62)	***
Optimal	$n_1 = 1,486$	(139)	(constant)	58.645	0.49	120.99	(57.41, 59.08)	***
	$n_0 = 1,486$		2017 (*Yes*)	1.756	0.62	2.855	(0.55, 2.96)	**
IP-Weights	$n_1 = 3,062$		(constant)	58.204	0.42	137.23	(57.37, 59.04)	***
	$n_0 = 3,081$		2017 (*Yes*)	1.895	0.63	3.016	(0.66, 3.13)	**

*Signif. codes: 0 ;'***' ~ 0. 001; '**' ~ 0. 01; '*' ~ 0.05; '.' ~ 0.1; and ' ' ~ 1.* n_1 and n_0 are the effective cohort sizes for the exposed (2017) and the control (2014) students, respectively. In particular, dropped student cases are represented by the (*drop*) count variable for each method of adjustment.

4.4.1 The unadjusted case

In Table 6, the causal risk difference is estimated at 1.76 with a subsequent 95% confidence interval of (0.55, 2.96), meaning that we will reject the null hypothesis of no treatment effect. Therefore, we conclude for an impact on the response from this intervention. Furthermore, the high significance of the corresponding probability values at 0.043, of which is less than 0.05, the conventional level of significance for hypothesis testing. The independent samples t-test results for the unadjusted case were that students have mean examination mark at 59.99 in the 2017 or treatment class whereas the baseline control student average was 58.24, a t-statistic of negative 2.85 is within a significant 95% confidence interval of (-2.96,-055), see Table A3. The straightforward outcome model, cf. Table 6, an alternative to the independent-sample t-test, is used to model potential outcomes or the Rubin causal model (Holland, 1986), for this unadjusted case, of the following form:

$$\begin{aligned} \mathbb{E}(Y_z) &= \alpha + \beta z \\ \mathbb{E}(Y_1) - \mathbb{E}(Y_0) &= \alpha + \beta - \alpha \\ &= \beta \end{aligned}$$

$$Causal \quad risk \quad difference \quad = \quad 1.76$$

4.4.2 Greedy matching (with a caliper)

When using the nearest-neighbour with a caliper of 0.2 we got a causal risk difference of 2.28 which is bounded within the 95% confidence interval of (0.93 , 3.62). Now since this interval is also above zero we will conclude a positive impact, or causal risk difference, due to the intervention on the outcome. The probability value below 0.001 implies that the null hypothesis can be rejected with high confidence. Finally, Table 6 shows that the mean examination mark is up to 60.52 for students enrolled after the clickers' intervention from 58.24 at baseline (t–test results in Table A3 shows them as 60.92 and 58.24 respectively). Besides a two-sample t-test one can also estimate, for the greedy method, potential outcomes from an outcome model of the form:

$$\begin{aligned} \mathbb{E}(Y_z) &= \alpha + \beta z \\ \mathbb{E}(Y_1) - \mathbb{E}(Y_0) &= \alpha + \beta - \alpha \\ &= \beta \end{aligned}$$

$$Causal \quad risk \quad difference \qquad = 2.28$$

4.4.3 Optimal matching

For the optimal adjustment, Table 6 gives a 95% confidence interval for the causal difference as (0.55, 2.96), the intervention had a positive impact of 1.76. Note that, this particular risk difference measure is similar or equivalent to the one for the unadjusted case which corresponds to the covariate balancing observed in Section 4.3. A corresponding low probability value of 0.043 (<5%), again equaling that for the no matching case, confirms rejection of the null hypothesis of no treatment effect. Finally, the mean examination mark, after adjusting with optimal matching is estimated at 60.40 for the post-clickers class of 2017 up from the 58.65 of the prior or 2014 class (t–test results in Table A3 shows them as 59.99 and 58.24 respectively). Again, one can also estimate the causal risk via the Rubin causal, or potential, outcome model (Holland, 1986) as:

$$\begin{aligned} \mathbb{E}(Y_z) &= \alpha + \beta z \\ \mathbb{E}(Y_1) - \mathbb{E}(Y_0) &= \alpha + \beta - \alpha \\ &= \beta \end{aligned}$$

i.e.

$$Causal \quad risk \quad difference \quad = \quad 1.76$$

This is noted as an, coincidental, equivalent to that of the unadjusted case above.

4.4.4 Inverse probability of treatment weighting

When adjusting with the inverse probability weighting method, the post-clickers students got a mean exam mark estimated up to 60.10 up from 58.20 prior (t–test results in Table 6 shows them as 59.99 and 58.20 respectively). This method evaluates to a causal risk difference of 1.90 of which is bounded by a 95% confidence interval of (0.66, 3.13). Now, due to this interval being above zero we conclude that the treatment effect is significant. And therefore the implications are that the clickers' intervention had some positive impact. It suffices that the null hypothesis will also be rejected, due to the low probability value of 0.026 (<5%), as it is with all the previous three cases. Utilizing the marginal structural model form for weighting we evaluated the outcome or Rubin causal model, Holland (1986), for this case as follows:

$$\begin{aligned} \mathbb{E}(Y_z) &= \varphi_0 + \varphi_1 z \\ \mathbb{E}(Y_1) - E(Y_0) &= \varphi_0 + \varphi_1 - \varphi_0 \\ &= \varphi_1 \end{aligned}$$

i.e.

$$Causal \quad risk \quad difference \quad = \quad 1.90$$

Short summary:

Causal risk differences are estimated at 1.76, 2.28, 1.76, and 1.90 for the unadjusted case, greedy with a caliper, optimal, and weighting adjustment methods, respectively. Their respective 95% confidence intervals are given as (0.55, 2.96), (0.93, 3.62), (0.55, 2.96), and (0.66, 3.13). We have that all of the probability values are below the 5% level of significance, given by 0.0043, <0.001, 0.0043, and 0.0026, and therefore the null hypothesis is rejected and the conclusion is in favour of clickers, or the classroom response system, having a positive impact on examination grades. Thus, controlling for the baseline confounders gives the same significant evidence to the prior adjustment case that post-clickers intervention students scored higher, on average, for their examination than those in the 2014 cohort.

These minor changes in the causal risk difference imply that pre- and post-matching data do not differ significantly. These results are not surprising since we saw a great deal of overlaps, or common support, in Section 4.1 that may be enough to confirm individual covariate balance. Furthermore, at the individual covariate level, we saw that the matching methods seem to have made little or no big covariate balance difference, post-adjustment, at all, cf. Section 4.3.3. The lack of substantial differences in the causal risk difference, across the matching methods, coupled with the minor shifts in individual pre-treatment

covariate balance is proof that, with so much visible overlap, the two cohorts were not that different to begin with, therefore leading to the consistency of the results. Further, the subsequent minor shifts in the causal risk, for two of the three methods with one not even moving from the unadjusted level, lead us to conclude that maybe matching was not worth the effort for this dataset and that perhaps a simple two-sample comparison would have sufficed.

4.5 Sensitivity analysis

Unlike randomization, adjustments on propensity scores will not necessarily guarantee absolute control for unobserved confounders, and in turn hidden-bias, Glynn, Schneeweiss & Stürmer (2006). Sensitivity analysis is another critical step required in order to satisfy ourselves about the robustness and or generalizability of the applied techniques. We apply *R* (Core, 2019) packages, *i.e. rbounds* (Keele, 2014), for sensitivity analysis in this section to determine if hidden-bias is a factor on our causal models and if so to what extent. This is for our chosen heterogeneity case in the investigation of hidden-bias sensitivity. Different estimates of the Rubin-bounds (Rosenbaum, 2002) from this *rbounds* package are listed in the resultant Table 7 below, for all the three post-adjustment methods. Equation (63), in Section 3.5.1, informs us that low sensitivity to hidden bias post-matching is verified when the Rubin Γ estimate is bounded between $\frac{1}{2}$ and 1.

Table 7 below depicts the Rubin estimates across all the three matching methods, including the unadjusted case, for sensitivity analysis purposes.

Table 7 Gamma values for Rubin bounds

		Rubin estimate Γ
	Unadjusted	5.0401
	Greedy (with a caliper)	1.0003
Matching method	*Optimal matching*	1.0412
	Inverse probability weights	4.8840

In Table 7 the greedy and optimal matching techniques have Rubin estimate values that are within the acceptable interval of Equation (63), Caliendo & Kopeinig (2005) and Rosenbaum (2002). Both methods, in particular, have the implications that from the baseline odds ratio of 1 each pair of enrolled and control students will have a next-to-nothing shift of 0.03% and another of a 4%, respectively, due to any unobserved confounder $\boldsymbol{u}_i$ (Ogutu, Okello, & Otieno, 2014). A technique out of bounds in Table 7, due to its high gamma estimate value of Γ=4.88 and not being closer to 1 and within the required $\frac{1}{2}$ to 2

boundary, is the inverse probability of treatment weighting. Note that, this high gamma estimate is closer to the one for the unadjusted case, of which was Γ=5.04. This implies that the inverse probability weighting method ought to be highly sensitive to hidden-bias in our model almost as if there were no taken control measures, alas the unadjusted case. Subsequently, taken from Ogutu, et al. (2014) again, this implies that from the baseline odds rate a hidden-confounder will move the odds-ratio between a pair of clickers and non-clickers students with an enormous shift effect of 388%, *i.e.* almost a factor four difference in the enrollment odds. Note that, this shift was at 404% for the unadjusted case, meaning that it remains almost similar even after adjustments with the *ipw* method. Therefore, this gives enough confirmation that the inverse probability of treatment weighting method failed abjectly to balance or improve the pre-treatment covariate space.

This outcome, on the higher sensitivity of the inverse probability weights method is plausible in a sense that it negates the fact that it was the one with extremely high covariate balancing compared to the other methods in Section 4.3. All that great balance is, subsequently, meaningless and goes to waste as it did almost nothing to help control for hidden-bias. It is imperative, then, that one takes heed of the fact pertaining to higher balancing not necessarily implying immunity to hidden or unobserved bias. Researchers should keep this at the back of their minds when estimating causal models. Rigorous data mining and understanding of the covariates to include or drop at the design stage is a compulsory requirement, if *weighting* was preferred, since it seems to be highly sensitive to bias from unobserved confounders, at least for these cohorts. It is interesting to note here that even with a somewhat worse covariate balance than the greedy matching and weighting methods, see Figure 13, optimal matching have lower sensitivity (Γ=1.04) to hidden-bias or unobserved confounders than the latter. An implication is thus, for any unobserved confounder, the odds-ratio between any clickers and non-clickers student pair will only shift by 4% to nullify the outcome estimate, subsequently the greedy matching technique will not change (0.03%) much of the odds-ratios at all. Therefore, we conclude in favour of both the greedy and optimal matching methods passing the sensitivity test.

5

Conclusions

Explicit definition of the research question is a critical requirement in order to take advantage of the methods of adjusting observational data using propensity scores (Ali, et al., 2009). The analysis in this dissertation illustrated an effective application, in terms of pre-treatment (or baseline) covariate balance control, using estimated propensity scores. Both the greedy with a caliper and inverse probability of treatment weighting adjustment cases successfully removed covariate differences at 5% threshold significance, whereas adjusting on optimal matching could only balance for racial demographics, home language instruction, matric math score, and years prior attempting the module. Note that, relaxing the threshold at 10% achieved bias reduction, due to controlling for baseline confounders, by all three methods except for optimal matching on the examination authority covariate. This seems to be related to optimal matching not being able to balance proportions for the independent education board level.

With post-matching covariate balance satisfactory, the outcome model was statistically evaluated to determine the causal risk difference or the treatment effect on the treated cohort. In particular, this scenario outcome was concerned with the determination of the impact clickers had on student's examination scores. All post-adjustment causal effect estimates, including for the pre-adjustment case, are bounded within positive 95% confidence intervals therefore implying that clickers may have significantly improved examination performance for the exposed students.

5.1 Results

The two cohorts combine to 3,111 students, of which 1,486 (47.8%) were exposed to clickers. In Table 5 we viewed and then compared the balance of the baseline covariates prior and post adjusting with propensity score methods. Prior to the use of propensity score adjustment techniques we see slightly more proportions of repeat and male students in the clickers group an in addition, proportionally, more students were instructed in

their home language in 2014. There are also some proportional differences in examination authority and faculty numbers between the two cohorts.

Outcome model outputs cf. Table 6; confirm the existence of a treatment effect or causal risk difference on student examination scores. This is also supported by the fact that all our confidence intervals, from the unadjusted to the three matched cases, given in Table 6, are above zero with probability values below the 5% significance level. This led us to reject the validity of the null hypothesis of no treatment effect.

Furthermore, the Rubin bounds in Table 7 shows that only the greedy with caliper and optimal matching methods have Γ estimates bounded in the interval $(\frac{1}{2}, 2)$ and closer to 1, thereby exhibiting relatively lower sensitivity to hidden-bias (Mariani & Pêgo-Fernandes, 2014). We conclude that, at least for our data set-up, the preferred methods of choice for estimating treatment effects are the greedy with a caliper and optimal matching. And, with a Γ value out of bounds, the inverse probability of treatment weights method is dropped since it lacks robustness to handle unobserved confounders due to being highly sensitive. Our conclusion, therefore, is based on both the greedy with a caliper and optimal matching methods, that there is significant evidence that students who enrolled post the clickers' intervention have, on average, better examination scores than the first group.

5.2 Future research

Literature shows a lot of various data scenarios, where one can apply the techniques visited above, refer to Spertus & Normand (2018), Love (2019), Ning et al. (2019) and Imai (2011). I have noted a couple of cases for specific applications of causal inference methods that I would love to pursue further for future research purposes. These include, amongst others, methods relating to different outcome models and their subsequent further techniques in investigating causality (Love, 2019 and Gran, Lie, Øyeflaten, Borgan & Aalen, 2015). In addition, in Chapter 4, I would like to further investigate two outcome scenario, those are continuous using the final mark and binary outcome through the odds of passing cases, for causal modeling. Other interesting outcome scenarios are also discussed in the literature, ranging from repeated measure outcomes which will require using hierarchical or mixed-effect models, to survival outcome scenario that will necessitate Cox proportional hazard modeling, refer to (Love, 2019) where there is an R practical for Right Heart Catheterization treatment data.

In a causal inference scenario where one may need to deal with high dimensional confounders, Spertus & Normand (2018) introduced Bayesian techniques to handle the estimation of propensity scores as a solution. Furthermore, when dealing with spatially-correlated time series data, Ning et al. (2019) suggest using Bayesian methods to balance the counterfactual using the real outcome. They show this through applying the

Expectation-Maximization (EM) algorithm for variable selection in order to pick the significant covariates needed when setting-up a propensity model for an advertisement campaign.

Political scientist Kosuke Imai, on his 2011 summary article *Introduction to the Virtual Issue: Past and Future Research Agenda on Causal Inference* listed a summary of a couple of new techniques that researchers, mostly from the political and social sciences disciplines, are actively pursuing in this interesting causal inference field. Since, as he states further, scholars can no longer 'brush under the table' most of the problems or biases related to causal inference matters. Imai (2011) name drop some of these researchers that are pursuing these techniques, at least, from the political and social science perspective. I have listed two of these as;

- effective design of field experiments through treatment/control ratio adjustment through utilizing placebos in the control (article quoted is David Nickerson's (2005, Vol 13 (3), pp. 233–252), and
- analysis of list experiment which helps scholars elicit thoughtful responses from survey respondents, especially with sensitive issues like racism (and Daniel Corstange's (2009, Vol. 17(1), pp. 45–63) is credited for this work).

Further mentioned were the needs to improve current matching and weighting techniques, ways to identify degrees of benefits or harm on differing subjects by the treatment, design of statistical experiments that will ascertain causal direction, and, finally how time series responses can be handled with causal models. Schweizer, et al. (2016) discusses the idea of interrupted time series designs and analysis as some experimental techniques, and 'will definitely get used a lot', in future research. Note that interested readers are referred to the summary article by Schweizer, et. al (2016).

Authors Austin, Grootendorst, & Anderson, in their (2007) article use Monte Carlo simulation techniques to see if observed confounders may still get balanced, hence reducing biases, especially in cases where researchers are not sure of variables to use for propensity score modeling to begin with. Hirano & Imbens (2001) proposed an interesting technique that aims to analyze causal inference using a regression adjustment combination with inverse probability of treatment allocation weights in order to increase robustness in removing bias. This, they say, works better that just relying on either method for bias reduction on its own. An example, they show, is how an ensemble of both adjustment techniques allows one to perform an exercise in pairwise estimations of treatment effects and hence creating a solution that can be viewed as doubly-robust.

Gran, Lie, Øyeflaten, Borgan, & Aalen, (2015) use causal inference concepts of G-computation together with the inverse probability Cox proportional hazards and Aalen additive hazards outcome models in a multi-state framework for sick-leave causality study.

Post-matching imbalance is looked at using regression adjustment methods of residual balance in Nguyen, et al. (2017), where a simulation study is taken in order to try balancing any of the confounders that remained unbalanced post-matching, in between multiple set-thresholds. Such, they reiterate, have shown to be a success in modeling differences in complex work-benefit data scenarios.

Chih-Lih investigates, in his 2011 PhD dissertation, the use of hierarchical or multilevel models for repeated cross-sectional observational studies, simulating a smoke cessation project pre- and post-ban level for intra-individual and group-wide effects, where the treatment gets repeated onto multiple cohorts and also for the case when the interventions were varied over time. Li, C. (2011), in this sense, got an opportunity expand or flex further the Rubin causal model in order to counter the challenges, due to both the nature of observational design especially when longitudinal data is in the mix, to identify the differing impacts that factoring time tends to cause on participants covariates within such studies. All of the above mentioned methods may be successfully employed on an extension of the applications discussed in Chapter 4.

Bibliography

Agrawal, A. (2017, March 31). Logistic Regression. Simplified. Retrieved June 07, 2019, from *https://medium.com/data-science-group-iitr/logistic-regression-simplified-9b4efe801389*

Ali, M., Prieto-Alhambra, D., Lopes, L., Ramos, D., Silva, N., Ichihara, Y., et al. (2009). Propensity Score Methods in Health Technology Assessment: Principles, Extended Applications, and Recent Advances. *Frontiers in Pharmacology.* **1**(1), 1-19. doi: 10.3389/fphar.2019.00973

Angrist, J. D., & Imbens, G. W. (1995). Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity. *Journal of the American Statistical Association*, **90**(430), 431-442. Applications and Case Studies

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A Comparison of the Ability of Different Propensity Score Models to Balance Measured VariablesBetween Treated and Untreated Subjects: A Monte Carlo Study. *Statistics in Medicine,* **26**(4), 734–753.

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, **46**(3), 399-424. doi: 10.1080/00273171.2011.568786

Austin, P. C., & Stuart, E. A. (2015). Moving Towards Best Practice when Using Inverse Probability of Treatment Weighting (IPTW) Using the Propensity Score to Estimate Causal Treatment Effects in Observational Studies. *Statistics in Medicine,* **34**(28), 3661–3679. doi:10.1002/sim.6607, 3661–3679

Baek, S., Park, S. H., Won, E., Park, Y. R., & Kim, H. J. (2015). Propensity Score Matching: A Conceptual Review for Radiology Researchers. *Korean Journal of Radiology*, **16**(2), 286–296. doi:10.3348/kjr.2015.16.2.286

Barrett, M. (2019). ggdag: Analyze and Create Elegant Directed Acyclic Graphs. *R package version 0.2.1*. https://CRAN.R-project.org/package=ggdag

Barter, R. (2017, July 05). The Intuition Behind Inverse Probability Weighting in Causal Inference. [*Blog post*]. Retrieved from *http://www.rebeccabarter.com/blog/2017-07-05-ip-weighting/*

Baumgarten, M., & Olsen, C. (2004). Confounding in Epidemiology. The Young Epidemiology Scholars Program (YES). *[Lecture notes].* Retrieved from *https://www.unav.edu/documents/6089811/16216616/CONFOUNDING_LECTURE+EXAMPLES_2004.pdf*

Beal, S. J., & Kupzyk, K. A. (2014). An Introduction to Propensity Scores: What, When, and How. *Journal of Early Adolescence*, **34**(1), 66-92.

Becerril, J., & Awudu, A. (2010). The Impact of Improved Maize Varieties on Poverty in Mexico: A Propensity Score-Matching Approach. *World Development*, *Elsevier*, **38** (7), 1024-1035. doi:10.1016/j.worlddev.2009.11.017

Bojinova, E., & Oigara, J. (2011). Teaching and Learning with Clickers: Are Clickers Good for Students. *Interdisciplinary Journal of E-Learning and Learning Objects*. **7**(1), 169-184. doi: 10.28945/1506.

Brame, C. (2013). Flipping the Classroom. Retrieved 05 13, 2020, from *Vanderbilt University Center for Teaching.*: *https://cft.vanderbilt.edu/guides-sub-pages/flipping-the-classroom/*

Cabi, E. (2018). The Impact of the Flipped Classroom Model on Students' Academic Achievement. *International Review of Research in Open and Distributed Learning*, **19**(3), 202-221. *https://files.eric.ed.gov/fulltext/EJ1185114.pdf*

Caliendo, M., & Kopeinig, S. (2005). Some Practical Guidance for the Implementation of Propensity Score Matching. *IZA Institute of Labor Economics Discussion Paper Series.* ssrn: https://ssrn.com/abstract=721907. doi:10.1111/j.1467-6419.2007.00527.x

Caliendo, M., & Kopeinig, S. (2008). Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys*, **22**, 31-72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

Chiba, Y., Azuma, K., & Okumura, J. (2009). Marginal Structural Models for Estimating Effect Modification. *Annals of Epidemiology*, **19**(5), 298-303. doi:10.1016/j.annepidem.2009.01.025

Crouch, C. H., & Mazur, E. (2001). Peer Instruction: Ten years of experience and results. *American Association of Physics Teachers*, **69**(9). *American Journal of Physics*, 970-977. doi: 10.1119/1.1374249

de Luna, X. & Lundin, M. (2009). Sensitivity analysis of the unconfoundedness assumption in observational studies. *Umeå: Department of Statistics, Umeå University, Sweden*

Dehejia, R., & Wahba, S. (2002). Propensity Score Matching Methods for Non-experimental Causal Studies. *The Review of Economics and Statistics*, 151-161.

Du, C. (2011). A Comparison Of Traditional And Blended Learning In Introductory Principles Of Accounting Course. *American Journal of Business Education*, **4**(9), 1-10.
doi: 10.19030/ajbe.v4i9.5614

Fay, M. P. (2016, April 25). cran.rstudio.com. Retrieved August 21, 2019, from
https://cran.rstudio.com/web/packages/exact2x2/vignettes/exactMcNemar.pdf

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. (2010). Impact Evaluation in Practice. Washington DC: *The International Bank for Reconstruction and Development/ The World Bank*. Retrieved June 15, 2019, from The World Bank:
http://siteresources.worldbank.org/EXTHDOFFICE/Resources/5485726-1295455628620 /Impact_Evaluation_in_Practice.pdf

Glynn, R. J., Schneeweiss, S., & Stürmer, T. (2006). Glynn, R. J., SchIndications for Propensity Scores and Review of their Use in Pharmacoepidemiology.
Basic & Clinical Pharmacology & Toxicology, **98**(3), 253–259.
https://doi.org/10.1111/j.1742-7843.2006.pto_293.x

Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky , D. (2018). A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook. *SSRN Electronic Journal*.

Gran, J. M., Lie, S. A., Øyeflaten, I., Borgan, Ø., & Aalen, O. O. (2015). Causal Inference in Multi-state Models–sickness Absence and Work for 1145 Participants After Work Rehabilitation. *BMC Public Health* **15**(1082), 1-16. https://doi.org/10.1186/s

Greifer, N. (2020). WeightIt: Weighting for Covariate Balance in Observational Studies. *R package version 0.9.0*. https://CRAN.R-project.org/package=WeightIt.
https://cran.r-project.org/web/packages/cobalt/vignettes/cobalt_A0_basic_use.html

Greifer, N. (2020). cobalt: Covariate Balance Tables and Plots. *R package version 4.0.0*.
https://CRAN.R-project.org/package=cobalt},

Grilli, L., & Rampichini, R. (2011). Propensity Scores for the Estimation of Average Treatment Effects in Observational Studies. Retrieved May 25, 2019, from [*PowerPoint slides*]:
https://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/prop-scores.pdf

Hair, Jr., J. F., William, C., Black, W. C., & Babin, B. J. (2010). Multivariate Data Analysis, Chapter 6. *Pearson*.

Hansen, B. B., & Klopfer, S. O. (2006). Optimal Full Matching and Related Designs via Network Flows. *Journal of Computational and Graphical Statistics*, **15**(3), 609-627,
doi: 10.1198/106186006X137047

Harris, H., & Horst, S. J. (2016). A Brief Guide to Decisions at each Step of the Propensity Score Matching Process. 21. *A Peer-reviewed Electronic Journal*, **21**(4), ISSN 1531-7714.

Heinrich, C., Maffioli, A., & Vázquez, G. (2010). A Primer for Applying Propensity-Score Matching. *https://www.researchgate.net/publication/235712818*

Hirano, K., & Imbens, G. W. (2001). Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. *Health Services & Outcomes Research Methodology* **2**, 259–278. https://doi.org/10.1023/A:1020371312283

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, **42** (8), 1-28.
url http://www.jstatsoft.org/v42/i08/

Holland, P. W. (1985). Statistics and Causal Inference. *New Jersey*: *Educational Testing Service*, *Princeton.*

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, **81**(396), 945-960. DOI: 10.2307/2289064

Imai, K. (2011). Introduction to the Virtual Issue: Past and Future Research Agenda on Causal Inference. *Political Analysis*, **19**(2), 1-4. doi:10.1017/S104719870001425X

Imbens, G. W. (2019). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. National Bureau of Economic Research.Working Paper **26104,** Working Paper Series (2019, July). doi = 10.3386/w26104

Imbens, G. W. (2019, Jul 16). Cornell University. Retrieved Oct 21, 2019, from arXiv:1907.07271v1 [stat.ME] : https://arxiv.org/abs/1907.07271

Jacovidis, J. N. (2017). Evaluating the Performance of Propensity Score Matching Methods: A simulation Study. *Dissertations*, **149**, https://commons.lib.jmu.edu/diss201019/149.

Joglekar, S. R. (2015, August 16). Logistic Regression (for dummies). Retrieved March 21, 2019, from Logistic Regression (for dummies)[*Blog post*]:
https://codesachin.wordpress.com/2015/08/16/logistic-regression-for-dummies/

Kabunga, N. S. (2014). Improved Dairy Cows in Uganda: Pathways to Poverty Alleviation and Improved Child Nutrition. *IFPRI Discussion Papers,* **1328**. International Food Policy Research Institute (IFPRI).

Keele, L. J. (2014). rbounds: Perform Rosenbaum bounds sensitivity tests for matched and unmatched data. *R package version 2.1*, url: https://CRAN.R-project.org/package=rbounds

Kuang, K., Li, L., Geng, Z., Xu, L., Zhang, K., Liao, B., et al. (2020). Causal Inference. *Engineering*, **6**, 1-11. https://doi.org/10.1016/j.eng.2019.08.016

Kukull, W. A., & Ganguli, M. (2012). Kukull WA, Ganguli M. Generalizability: The Trees, the Forest, and the Low-hanging Fruit. *Neurology*, **78** (23), 1886–1891.
https://doi.org/10.1212/WNL.0b013e318258f812

Kuo, C. L., Duan, Y., & Grady, J. (2018). Unconditional or Conditional Logistic Regression Model for Age-Matched Case-Control Data?. *Frontiers in public health*, **6**, 57. doi:10.3389/fpubh.2018.00057.

Lage, M. J., Platt, G. J., & Treglia, M. (2000). Inverting the Classroom: A Gateway to Creating an Inclusive Learning Environment. *The Journal of Economic Education*, **31**(1), 30-43. doi 10.2307/1183338

Lane, F., To, Y., Kyna, S., & Robin, H. (2012). An Illustrative Example of Propensity Score Matching with Education Research. *Career and Technical Education Research*, **37**, 187-212. 10.5328/cter37.3.187

Lechner, M. (2000). A Note on the Common Support Problem in Applied Evaluation Studies. S*SRN Electronic Journal*, 91–92. doi: 10.2139/ssrn.259239

Li, C. (2011). Propensity Score Matching in Observational Studies with Multiple Time Points. *Dissertations*. Ohio State University, USA

Li, M. (2012). Using the Propensity Score Method to Estimate Causal Effects: A Review and Practical Guide. *SAGE Journals*, 1-39. doi: 10.1177/1094428112447816

Lipton, Z. C., Elkan, C., & Naryanaswamy, B. (2014). Optimal Thresholding of Classifiers to Maximize F1 Measure. ECML PKDD (Conference) (pp. 225–239). *Machine Learning and Knowledge Discovery in Databases: European Conference*, ECML PKDD.

Love, T. E. (March 21, 2019). Some Propensity Ideas and the Right Heart Catheterization Data. Retrieved November 11, 2019, from *https://rstudio-pubs-static.s3.amazonaws.com /476262_07fbb0872de14b9e8335fe230c155ddd.html*.

Lunt, M. (2013). Selecting an Appropriate Caliper can be Essential for Achieving good Balance with Propensity Score Matching. *American Journal of Epidemiology*, **179**(2), 226–235. doi:10.1093/aje/kwt212

Mariani, A. W., & Pêgo-Fernandes, P. M. (2014). Observational Studies: Why are they so Important? Sao Paulo Medical Journal. *Sao Paulo Medical Journal*, **132**(1), São Paulo. https://doi.org/10.1590/1516-3180.2014.1321784

Matsumoto, E., & Del-Moral-Hernandes, E. (2013). Using Neural Networks Committee Machines to Improve Outcome Prediction Assessment in Nonlinear Regression. *Proceedings of the International Joint Conference on Neural Networks.*
10.1109/IJCNN.2013.6707023

McLeod, S. (2019). Extraneous Variable. Simply Psychology. Retrieved from
https://www.simplypsychology.org/extraneous-variable.html

Nguyen, T. L., Collins, G., Spence, J., Daurès, J. P., Devereaux, P. J., Landais, P., et al. (2017). Double-adjustment in Propensity Score Matching Analysis: Choosing a Threshold for Considering Residual Imbalance. *BMC Medical Research Methodology* **17**(78), 1-8.
https://doi.org/10.1186/s12874-017-0338-0

Ning, B., Ghosal, S., & Thomas, J. (2019). Bayesian Method for Causal Inference in Spatially-Correlated Multivariate Time Series. *Bayesian Analysis*, **14**(1), 1-28. doi: 10.1214/18-ba1102

Nouri, J. (2016). The Flipped Classroom: For Active, Effective and Increased Learning – Especially for Low Achievers. *International Journal of Educational Technology in Higher Education*, **13**(33), 1-10. doi 10.1186/s41239-016-0032-z

Ogutu, S. O., Okello, J. J., & Otieno, D. J. (2014). Impact of Information and Communication Technology-Based Market Information Services on Smallholder Farm Input Use and Productivity: The Case of Kenya. *World Development*. **64**, 311-321. ISSN 0305-750X
https://doi.org/10.1016/j.worlddev.2014.06.011

Olmos, A., & Govindasamy, P. (2015). Propensity Scores: A Practical Introduction Using R. *Journal of MultiDisciplinary Evaluation*, **11**(25), 68-88.

Pan, W., & Bai, H. (2015). Propensity Score Analysis Concepts and Issues. Retrieved March 14, 2019, from Propensity Score Analysis Concepts and Issues:
http://people.duke.edu/˜wp40/sample_files/chapter%201.pdf

Peikes, D. N., Moreno, L., & Orzol, S. M. (2008). Propensity Score Matching. *The American Statistician*, **62**(3), 222-231. doi: 10.1198/000313008X332016

Posner, M., & Ash, A. (2012). Comparing Weighting Methods in Propensity Score Analysis. *Unpublished Working Paper*, *Columbia Statistics*, *Columbia University.*

R Core Team (2019). R: A language and environment for statistical computing. *R Foundation* for (2019).

Raymond, S. (2014). The Effectiveness of the Flipped Classroom. *Honors Projects*. **127**.
https://scholarworks.bgsu.edu/honorsprojects/127

Reyneke, F., & Fletcher, L. (2013). Investigating Success Rates of First Level Statistics Students in the new Millennium. *Pretoria: Department of Statistics, University of Pretoria*, *South Africa.*

Reyneke, F., Fletcher, L., & Harding, A. (2018). The Effect of Technology-based Interventions on the Performance of First Year University Statistics Students. *African Journal of Research in Mathematics, Science and Technology Education*, **22**(2), 231-242. doi: 10.1080/18117295.2018.1477557

Ridgeway, G., McCaffrey, D., Morral, A., Burgette , L., & Griffin, B. A. (2017, 07 01). Toolkit for Weighting and Analysis of Nonequivalent Groups: A tutorial for the twang package.

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, **70**(1), 41-55.

Rosenbaum, P. R. (2002). Covariance Adjustment in Randomized Experiments and Observational Studies. *Statistical Science*, **17**(3), 286-327.

Roy, J. (2018). Coursera. Retrieved May 25 , 2019, from Causality - Inferring Causal Effects from Data.: *https://www.coursera.org/learn/crash-course-in-causality*

Saxena, S. (2018, May 11). towardsdatascience.com. Retrieved May 03, 2019, from precision-vs-recall: *https://towardsdatascience.com/precision-vs-recall-386cf9f89488*

Scheines, R. (1997). An Introduction to Causal Inference. doi=10.1.1.118.3002
from http://citeseerx.ist.psu.edu/viewdoc/download, 185-200.

Schroeder, D. A. (2010). Accounting and Causal Effects. *Springer Series in Accounting Scholarship*, **5**. New York, NYC: Springer.

Schroeder, K., Jia, H., & Smaldone, A. (2016). Which Propensity Score Method Best Reduces Confounder Imbalance? An Example From a Retrospective Evaluation of a Childhood Obesity Intervention. *Nursing Research*, **65**(6), 465–474.
https://doi.org/10.1097/NNR.0

Schweizer, M. L., Braun, B. I., & Milstone, A. M. (2016). Research Methods in Healthcare Epidemiology and Antimicrobial Stewardship—Quasi-Experimental Designs. *Infection Control & Hospital Epidemiology*, **37**(10), 1135–1140. https://doi.org/10.1017/ice.2016.117

Siau, K., Sheng, H., & Nah, F. F. (2006). "Use of a Classroom Response System to Enhance Classroom Interactivity," in IEEE Transactions on Education. *IEEE Transactions on Education.* **49**(3), 398-403. doi: 10.1109/TE.2006.879802

Sperandei, S. (2014). Understanding Logistic Regression Analysis. *Biochemia Medica*, **24**(1), 12–18. https://doi.org/10.11613/BM.2014.003

Spertus, J. V., & Normand, S. T. (2018). Bayesian Propensity Scores for High-dimensional Causal Inference: A Comparison of Drug-eluting to Bare-metal Coronary Stents. *Biometrical Journal. Biometrische Zeitschrift*, **60**(4), 721–733. doi:10.1002/bimj.201700305

Spirtes, P., & Scheines, R. (1993). Causation, Prediction, and Search. *Carnegie Mellon University*: doi: 10.1007/978-1-4612-2748-9

Spirtes, P., Glymour, C., & Scheines, R. (1993). Causation_Prediction_and_Search. doi: 10.1007/978-1-4612-2748-9

Starks, H., Diehr, P., & Curtis, J. R. (2009). The Challenge of Selection Bias and Confounding in Palliative Care Research. *Journal of Palliative Medicine*, **12**(2), 181–187. https://doi.org/10.1089/jpm.2009.9672

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, **25**(1), 1-21. doi:10.1214/09-STS313

Stuart, E. A., Bradshaw, C. P., & Leaf, P. J. (2015). Assessing the Generalizability of Randomized Trial Results to Target Populations. *Prevention Science: The Official Journal of the Society for Prevention Research*, **16**(3), 475–485. doi:10.1007/s11121-014-0513-z

Thavaneswaran, A., & Lix, L. (2008). Propensity Score Matching in Observational Studies. *https://www.umanitoba.ca/faculties/health_sciences/medicine/units/chs/departmental_units /mchp/protocol/media/propensity_score_matching.pdf*

van der Wal, W. M., & Geskus, R. B. (2011). ipw: An R Package for Inverse Probability Weighting. *Journal of Statistical Software*, **43**(13), 1-23. url http://www.jstatsoft.org/v43/i13/

Williams, T. D., Tolusso, D. V., Fedewa, M. V., & Esco, M. R. (2017). Comparison of Periodized and Non-Periodized Resistance Training on Maximal Strength: A Meta-Analysis. *Sports Med. 2017*, **47**(10), 2083-2100. doi:10.1007/s40279-017-0734-y

Xie, Y., Brand, J. E., & Jann, B. (2012). Estimating Heterogeneous Treatment Effects with Observational Data. *Sociological Methodology*, **42**(1), 314–347.

Yoshida, K. (2019). tableone: Create 'Table 1' to Describe Baseline Characteristics. *R package version 0.10.0*., url: https://CRAN.R-project.org/package=tableone

Zaga Szenker, D. (2015). The impact of Three Mexican Nutritional Programs: The Case of DIF-Puebla. doi: 10.13140/RG.2.1.1857.9926

Zainuddin, Z., & Halili, S. H. (2015). Flipping The Classroom: What We Know And What We Don't. *The Online Journal of Distance Education and e-Learning*, **3**(1), 15-22. www.tojdel.net

Zainuddin, Z., & Halili, S. H. (2016). Flipped Classroom Research and Trends from Different Fields of Study. *International Review of Research in Open and Distance Learning* **17**(3), 313-340. doi: 10.19173/irrodl.v17i3.2274

Zhang, Z., Kim, H. J., Lonjon, G., & Zhu, Y. (2019). written on behalf of AME Big-Data Clinical Trial Collaborative Group. Balance Diagnostics after Propensity Score Matching. *Annals of translational medicine*, **7**(1), 16. https://doi.org/10.21037/atm.2018.12.10

Appendix 1: Proofs

Theorem 1

From Rosenbaum & Rubin, (1983) The propensity scores or allocation probability given the observed pre-treatment covariate set is conditionally independent to the treatment status, also taken from Equation (1) and (18); given the propensity scores

$$X \perp\!\!\!\perp z \Big| p(\boldsymbol{X}), \tag{A.1}$$

where $p(\boldsymbol{X}_i)$ - is the the treatment allocation probability for subject i. Further note that in

Rosenbaum & Rubin (1983), Theorem 1 is declared to be the special case of Theorem 2 from Cochran & Rubin (1979).

Theorem 2

Let the balancing function be $b(\boldsymbol{X})$, such that $b(\boldsymbol{X})$ it is a function of $\boldsymbol{X}$, that is; $\boldsymbol{X} \perp\!\!\!\perp z \Big| b(\boldsymbol{X})$ if and only if (iff) for some function f, $p(\boldsymbol{X}) = f\{b(\boldsymbol{X})\}$, $b(\boldsymbol{X})$ is finer than $p(\boldsymbol{X})$

Proof:

$$Case \qquad 1^{\circ}$$

Suppose $b(\boldsymbol{X})$ was finer than $p(\boldsymbol{X})$, it then follows that showing that $b(\boldsymbol{X})$ is a balancing score will suffice.

Note that, it is sufficient to show that $P\{z = 1 \Big| b(\boldsymbol{X})\} = p(\boldsymbol{X})$ in order for $b(\boldsymbol{X})$ to be a balancing score

Recalling the definition of the propensity score:

$$p(\boldsymbol{X}) = P\{z = 1 \Big| b(\boldsymbol{X})\} = P\{p(\boldsymbol{X}) \Big| b(\boldsymbol{X})\}, \tag{A.2}$$

And since, $b(\boldsymbol{X})$ is finer than $p(\boldsymbol{X})$ by supposition above, then the right hand side of equation. (A.2) will evaluate to the propensity scores - $p(\boldsymbol{X})$, as required. Thus meaning;

$$\mathbb{E}\{p(\boldsymbol{X}) \Big| b(\boldsymbol{X})\} = p(\boldsymbol{X})$$

Therefore, $b(\boldsymbol{X})$is the balancing score

Conversely,

$$Case \qquad 2^{\circ}$$

Suppose $b(\boldsymbol{X})$ is a balancing score and that it is not finer than the propensity scores $p(\boldsymbol{X})$ so that we have a pair of features X_1 and X_2 such that $p(X_1) \neq p(X_2)$ but $b(X_1) = b(X_2)$ Recalling the definition of the propensity score, again;

$$p(\boldsymbol{X}) = P\{z = 1 \Big| \boldsymbol{X}\}$$

$P\{z = 1 \Big| X_1\} = P\{z = 1 \Big| X_2\}$, so that z and $\boldsymbol{X}$ are not conditionally independent given $b(\boldsymbol{X})$, and that $b(\boldsymbol{X})$ is a balancing score

Therefore; $b(\boldsymbol{X})$ must then be finer than $p(\boldsymbol{X})$ to cause such balancing!

This completes the proof! ■

Theorem 3

Given some $b(\boldsymbol{X})$ a balancing score, Rosenbaun & Robin(1983) notes that if the treatment allocation is strongly ignorable given $\boldsymbol{X}$ then so should it be when the condition is on $b(\boldsymbol{X})$. Mathematically, that is;

$$(Y_1, Y_0) \perp\!\!\!\perp z \Big| \boldsymbol{X} \text{ and } 0 < P(z = 1 \Big| \boldsymbol{X}) < 1, \forall \boldsymbol{X}, \text{ meaning that}$$

$$(Y_1, Y_0) \perp\!\!\!\perp z \Big| b(\boldsymbol{X}) \text{ and } 0 < P\{z = 1 \Big| b(\boldsymbol{X})\} < 1, \forall b(\boldsymbol{X})$$

Proof:

The inequality given $b(\boldsymbol{X})$ is a consequence of that one given $\boldsymbol{X}$. Thus, it suffices to show that;

$$P\{z = 1 \Big| Y_1, Y_0, b(\boldsymbol{X})\} = Pr\{z = 1 \Big| b(\boldsymbol{X})\}$$

Of which is equal to $p(\boldsymbol{X})$ by Theorem 2, *i.e.* $P\{z = 1 \Big| b(\boldsymbol{X})\} = p(\boldsymbol{X})$. Therefore showing;

$$P\{z = 1 \Big| Y_1, Y_0, b(\boldsymbol{X})\} = p(\boldsymbol{X})$$

Now,

$$P\{z = 1 \Big| Y_1, Y_0, b(\boldsymbol{X})\} = \mathbb{E}\{P[z = 1 \Big| Y_1, Y_0, \boldsymbol{X}] \Big| Y_1, Y_0, b(\boldsymbol{X})\}$$

This, by the assumption, equals;

$$\begin{aligned} P\{z=1\Big|Y_1,Y_0,b(\boldsymbol{X})\} &= \mathbb{E}\{P(z=1\Big|\boldsymbol{X})\Big|Y_1,Y_0,b(\boldsymbol{X})\} \\ &= \mathbb{E}\{p(\boldsymbol{X})\Big|Y_1,Y_0,b(\boldsymbol{X})\} \end{aligned}$$

Equalling the propensity score as required, since $b(\boldsymbol{X})$ is a finer balancing score than $p(\boldsymbol{X})$, and therefore:

$$\mathbb{E}\{p(\boldsymbol{X})\Big|Y_1,Y_0,b(\boldsymbol{X})\} = p(\boldsymbol{X})$$

Which completes our proof! ■

Appendix 2: Miscellaneous tables and figures

Propensity scores estimator

Table A1 Logistic regression estimates results

Covariates	*Estimate*	*Standard error*	*z - value*	*Prob value*	
(Intercept)	0.46770	0.4414	1.06	0.28934	
Repeater:1	0.53025	0.1505	3.523	0.00043	***
Gender:Male	0.27632	0.0759	3.64	0.00027	***
Race:Coloured	0.33341	0.2722	1.225	0.22069	
Race:Indian	0.04949	0.1779	0.278	0.78090	
Race:White	0.11746	0.0986	1.192	0.23332	
Education authority: EC	-0.77638	0.4455	-1.743	0.08138	
Education authority: FRC	-1.39389	0.9334	-1.493	0.13535	
Education authority: FS	-0.42325	0.4255	-0.995	0.31988	
Education authority: GP	-0.82981	0.3393	-2.446	0.01446	*
Education authority: IEB	-0.37556	0.3445	-1.09	0.27570	
Education authority: KZN	-0.57829	0.3636	-1.591	0.11172	
Education authority: LI	-1.02858	0.3579	-2.874	0.00405	**
Education authority: MP	-0.85067	0.3588	-2.371	0.01776	*
Education authority: NA	-14.00313	266.6420	-0.053	0.95812	
Education authority: NC	-1.90035	0.7468	-2.545	0.01094	*
Education authority: NW	-1.25950	0.4022	-3.132	0.001734	**
Education authority: WC	-0.90637	0.6236	-1.454	0.14608	
Faculty: Humanities	0.55737	0.3065	1.819	0.06894	
Faculty: Law	-13.83068	535.4112	-0.026	0.97939	
Faculty: other	0.47129	0.2957	1.594	0.11098	
Faculty: NAS	-0.13539	0.1053	-1.286	0.19857	
Insruction in home language	-0.40567	0.0934	-4.343	< 0.001	***
Math score (Grade12)	0.00313	0.0040	0.789	0.43036	
Years prior attempt	-0.05642	0.0249	-2.265	0.02349	

*—Signif. codes: 0 ;'***' ~ 0. 001; '**' ~ 0. 01; '*' ~ 0.05; '.' ~ 0.1; and ' ' ~ 1.*

And the resulting estimation equation, from the significant parameters, is:

$$\begin{aligned} log(p_i) \quad = \quad & +0.53 \cdot Repeater + 0.28 \cdot Gender : Male + (-0.83) \cdot Authority : GP + \\ & (-1.03) \cdot Authority : LI + (-0.85) \cdot Authority : MP + (-1.90) \cdot Authority : NC \\ & + (-1.26) \cdot Authority : NW + (-0.41) \cdot HomeLanguage \end{aligned}$$

Logistic diagnostics

Table A2 Logistic regression diagnostics

Measure Name		*Measure*	
Accuracy	:	0.576	
95% CI	:	(0.5584, 0.5935)	
No Information Rate	:	0.5223	
P-Value [Acc > NIR]	:	< 0.001	***
	:		
Kappa	:	0.1429	
	:		
Mcnemar's Test P-Value		< 0.001	
	:		
Sensitivity	:	0.6868	
Specificity	:	0.4549	
Pos Pred Value	:	0.5794	
Neg Pred Value	:	0.5705	
Prevalence	:	0.5223	
Detection Rate	:	0.3587	
Detection Prevalence	:	0.6191	
Balanced Accuracy	:	0.5708	
	:		
Positive' Class	:	F	

*—Signif. codes: 0 ;'***' ~ 0. 001; '**' ~ 0. 01; '*' ~ 0.05; '.' ~ 0.1; and ' ' ~ 1.*

t-test

Table A3 Independent samples *t*-test model results

Method	*Effect size*	*(drop)*	*2014 (mean)*	*2017 (Mean)*	*t statistic*	*df*	*95% CI*
Unadjusted	$n_1 = 1,486$ $n_0 = 1,625$		58.24	59.99	-2.85	3057.8	(-2.96, -0.55)**
Greedy matching	$n_1 = 1,430$ $n_0 = 1,430$	(251)	58.24	60.92	-3.32	2531.3	(-3.62, -0.93)***
Optimal matching	$n_1 = 1,486$ $n_0 = 1,486$	(139)	58.24	59.99	-2.85	3057.8	(-2.96 -0.55)**
Weighted matching	$n_1 = 3,062$ $n_0 = 3,081$		58.204	59.99	-2.85	3057.86	(-2.96, -0.55)**

*Signif. codes: 0 ;'***' ˜ 0. 001; '**' ˜ 0. 01; '*' ˜ 0.05; '.' ˜ 0.1; and ' ' ˜ 1.*
n_1 and n_0 are the effective cohort sizes for the exposed (2017) and the control (2014) students, respectively. In particular, dropped student cases are represented by the (*drop*) count variable for each method of adjustment.

Appendix 3: R code

1. Data manipulation

```
##------------------------------------------------------ (START!)**

####()**   ...   i. Data manipulations
##-------------------------------------------------**
#*  ***********************************************

#*
#*           * created by: Mxolisi Msibi, Mr.
#*           *organisation: University of Pretoria
#**          ** degree: Magister Scientiae
#*           *       create datae: 2019.11.25
#*           *      update1: 2019.12.05
#*           *      update 2: 2019.12.27
#*           *      update 3: 2020.01.02
#*           *      update 4: 2020.01.08
#*           *      update 5: 2020.03.18
#*
#*           *     update 7: 2020.05.21
#*           *     update 8: 2020.05.22
#**
#*           *     update 9: 2020.06.04

##-------------------------------------------------**
#*  ***********************************************


#++++++++++++++++++++++++++++++++++++++++++++++++++++  ++
##---                                              ---**
#*   **  I. Data Manipulations (   (step)   )**
##---                                              ---**
#++++++++++++++++++++++++++++++++++++++++++++++++++++  ++
```

```
##-------------------------------------------------**
#*   get required libRaries (in)**
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*   0.0. installin' required packages

install.packages( c( #*
#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%**
#1.Data manipulations
"sqldf",  # do sql man (in R)*   ...
"janitor",
"naniar",
"readxl", #get data from excel "sheets"
          #*
"tidyverse",#*( for 'pippin') and datamanipulations
'vctrs',
#*
"broom",
#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%**
#2.Propensity score analysis
####  i.estimation
"lme4",# i.glms ( with logit)
#*  ii.matching
      "Matching",####  ii.matching
      "MatchIt",#for matching
      "ipw",#inverse treatment probability
      'optmatch',### for full matchin'
      "WeightIt",  # ipw weights ( weer weer)
      "CBPS",#require(cbps)###citation("CBPS")
      "cobalt",#install.packages("cobalt")
#*  iii.covariate balance
"tableone",  ####  iii.covariate balance
"sandwich", #for robust variance estimation
"survey",
#*  iv.outcome modeling
      "survival",   #*
      "lme4",
####  v.sensitivity analysis
"rbounds",   ##  #*v.sensitivity analysis
"rgenoud",
```

```
#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%**
#3.Miscellaneous
      "dplyr",#*impute missings'
      "ggpubr",         ###*()**
   #   ggplot()
"ggplot2", "cowplot",#install.packages("ggplot2")
      "ggridges" #plots (and themes)
#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%**
))
#* install package by uncommenting "at top"
#( if commented out)
#*(these are the required packages that you
#might require for analysis)...
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   0.1. loadin' packages
library(janitor)
library(naniar)#*
library(broom)
require(sqldf)#doin'  sql (in R*)  ...
library(tidyverse)#(for 'pippin') and datamanipulations
library(rgenoud)#> library(rgenoud)
#----------------*()**   ...
##2.  Match your data (with these...)
library('optmatch')#* for full matchin'
#load packages#*()**
library(Matching)##; for matching techniques
library(MatchIt)#conventional matching package
library('optmatch')### for full matchin'
library(tableone)
library(ipw)#inverse treatment probability matching
library(sandwich)#robust variance estimation library(survey)
#----------------*()**   ...
##3.  Assess balance
library(tableone)#; balance diagnostic
                 #tools ( incl. in this package)
library(cobalt)  #;*   ... (for) calculatin SMDs and other
#balance measures.
library(survey)  # bala\nce diagnostics ( too   )
#----------------*()**   ...
##4.  Outcome modelin'
library(lme4)  #*   ... (for) linear and hierarchical models
library(survival) #*   .... (for) time-to-event outcomes
```

```
#----------------*()**   ...
##5.  Asses hidden-bias
library(rbounds)#Sensitivity analysis (Rubin bounds)
library(rgenoud)# sensitivity ansalysis (nb)**???...
library("xlsx") #imorting excel files (with "readxl" package)
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   0.2.   create logit and sigmoid functions'
sigmoid<- function(x){1/(1+exp(-x)) }#*.sigmoid fun'
  logit<- function(p){log(p)-log(1-p)}#*.(ze)log-odds
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   0.3. lread the four 'excel-sheets' using
#"readxl" package
library("readxl")  #*get data from excel "sheets"
library(tidyverse)
#*( for) 'pippin' you know/ data manipulations

##   #   ...   *()**   ...
setwd("C:\\Users\\mxmsibi\\Downloads\\Biz.Related
\\ms␣dissertation
\\psm_5th.draft_2019.08.27␣-␣due␣end␣-␣Oct")
getwd()

#*   1.    get all data sets (in)**
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   1.1. read the four sheets using "readxl"
#package import data off both (the) groups***....

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   1.1.1.   read into a list object
lst <- lapply(1:4, function(i) read_excel(
        "1.␣-␣Praticals/data/Mxo␣data␣-␣v3.1..xlsx"
                                , sheet = i))
####(home)**
tail(lst[[1]])
head( lst[[1]] ,11)
####()**...
lapply(excel_sheets(path), read_excel, path = path)
```

```
lst1 <- lapply( excel_sheets(
"1. - Praticals/data/Mxo data - v3.1..xlsx")
                                , read_excel,
path = "1. - Praticals/data/Mxo data - v3.1..xlsx")

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   1.1.2.   separate the data sets from the list
#* stk data ( for both groups)
my_data_2014<- lst[[1]]; my_data_2017<- lst[[3]]
head(my_data_2014, 11)
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   1.1.3.   separate matric data sets
#* matric information( both groups)
matric_2014<- lst[[2]];
matric_2017<- lst[[4]];
head(matric_2014, 11)
#*
#*  ( 2. )   manipulate your data (hier)**
#   ************************************************
#**   *** (redone)
#*   2.1      *2014 data*...

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   2.1.1. create home language instruction
#flag (2014)
colnames(my_data_2014); colnames(matric_2014)#*()**
colnames(my_data_2017); colnames(matric_2017)
#*(don't forget to) rename the "number" column
#to "st number"
colnames(my_data_2014)[
   colnames(my_data_2014)=="Number"] <- "St number"
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   2.1.2. flag 2014 home language

my_data_2014 %>%
mutate(Home_Language_Instr = ifelse(
   `Home Language Desc` ==
         `Language of Preference Desc`, 1, 0))
my_data_2014x<- my_data_2014 %>%
```

```
mutate(Home_Language_Instr = ifelse(
       `Home Language Desc` ==
         `Language of Preference Desc`, 1, 0))
head(my_data_2014x,7)
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   2.2.1. create the home language instruction
#flag (2014)

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   2.1.3. flag 2014 (as "treatment group~'0'")
my_data_2014x<- my_data_2014x %>%
          mutate(treat = "0"); head(my_data_2014x,7)
head(my_data_2014x$treat,7)
#[1] "0" "0" "0" "0" "0" "0" "0"

#+++++++++++++++++++++++++++++++++++++++++++++++++++  ++
##---                                              ---**
#*   **  2.2      *2017 data*...**
##---                                              ---**
#+++++++++++++++++++++++++++++++++++++++++++++++++++  ++

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   2.2.1. create the home language instruction
#flag (2017)

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   2.2.1. flag 2017 home language
my_data_2017 %>%
   mutate(Home_Language_Instr = ifelse(
    `Home Language Desc` ==
         `Language of Preference Desc`, 1, 0))
my_data_2017x<- my_data_2017 %>%
   mutate(Home_Language_Instr = ifelse(
    `Home Language Desc` ==
        `Language of Preference Desc`, 1, 0))
head(my_data_2017x,11)
head(my_data_2017x$Home_Language_Instr,11)
#**()**    [1] 1 1 0 1 1 1 1 1 0 0 1
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
```

```
##-------------------------------------------------**
#*   2.2.2. flag 2014 (as " treatment group ")
my_data_2017x<- my_data_2017x %>%
                          mutate(treat = "1")
head(my_data_2017x,11)
head(my_data_2017x$treat,11)
#[1] [1] "1" "1" "1" "1" "1" "1" "1" "1" "1" "1" "1"

#*   3.    grade 12 marks manipulator (hier)**
##################################################
#*   3.get grade 12 math marks (in)
#           (both group(s)) (*2014/17 data*)*...

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   3.1. map your matric grades (math)  (2014)
matric_2014 %>%
    filter(`School Subject Desc` == "Mathematics")

merge(my_data_2014,matric_2014,by="St number")
#####()**
merge(my_data_2014,
    matric_2014[,
        c('St number','Final Mark')]
                    ,by="St number")
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   3.1.1. get math marks  (2014)
matric_2014  %>%
     filter(`School Subject Desc` == "Mathematics")
matric_2014x<- matric_2014 %>%
     filter(`School Subject Desc` == "Mathematics")
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   3.1.2. map 2014 math grades
colnames(matric_2014x)[
 colnames(matric_2014x)=="Final Mark"]
    <- "Math_Grade12"; matric_2014x$Math_Grade12;
is.na(matric_2014x$Math_Grade12)
colnames(matric_2014x)[
  colnames(matric_2014x)=="Home Language Desc"]
       <- "Home Language" merge(my_data_2014x
                               ,matric_2014x
```

```
                                  ,by="St␣number")
is.na(merge(my_data_2014x,
            matric_2014x,
            by="St␣number")); is.na(
      (merge(my_data_2014x,matric_2014x,
                  by="St␣number"))$Math_Grade12)
my_data_2014y<- merge(  my_data_2014x,
                   matric_2014x ,by="St␣number")
head(my_data_2014y,11)
#*
dim(my_data_2014y)#** ([1] 1625   32)**
dim(my_data_2014x)#***([1] 1676   23)***
is.na(my_data_2014y$Math_Grade12)
#####all "FAALSE" that's good....
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   3.2. map your matric grades (math)  (2017)
matric_2017 %>%
filter(`School Subject Desc` == "Mathematics")
merge(my_data_2017,matric_2017,by="St␣number")
#*( this1 got the right student number)**
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   3.2.1. get math marks  (2017)
matric_2017 %>%
      filter(`School Subject Desc` == "Mathematics")
matric_2017x<- matric_2017 %>%
      filter(`School Subject Desc` == "Mathematics")
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   3.2.2. map 2017 math grades
colnames(matric_2017x)[
       colnames(matric_2017x)==
                  "Final␣Mark"]<-"Math_Grade12"
matric_2017x$Math_Grade12; is.na(
        matric_2017x$Math_Grade12)
#*(^^^)** ... all "FALSE" that's good....
colnames(matric_2017x)[
   colnames(matric_2017x)==
      "Home␣Language␣Desc"] <-"Home␣Language"
merge(my_data_2017x,
            matric_2017x, by="St␣number")
(merge(my_data_2017x,
```

```
        matric_2017x,
               by="St␣number"))$Math_Grade12;
is.na((merge(my_data_2017x,
             matric_2017x,
            ,by="St␣number"))$Math_Grade12)
####()**  all "FAALSE" that's good....
my_data_2017y<- merge(my_data_2017x,
                      matric_2017x,
                           by="St␣number")
head(my_data_2017y,11)#
#*( 4.)    combine these dataframes (into1)**
################################################
#* 4. append your final data(2014/17 data0)**...

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   4.1. make 2014 and 2017 data (one)
dim(my_data_2014y); dim(my_data_2017y)
#*[1] [1] 1625   32
##       [1] [1] 1486   32
rbind( my_data_2014y, my_data_2017y)
#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   4.1.1.  check if matric marks are valid(???)*....
(rbind(my_data_2014y,
       my_data_2017y))$Math_Grade12; is.na(
             (rbind(my_data_2014y
                   ,my_data_2017y))$Math_Grade12)
####()**  all "FAALSE" that's good....
#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   4.1.2.  check rbinding
dim(rbind( my_data_2014y, my_data_2017y))
####()*[1] 3[1] 3111 32
#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   4.2.  final datasets

rbind( my_data_2014y, my_data_2017y)
#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   4.2.1.  create dataset
data<- rbind( my_data_2014y, my_data_2017y)
```

```
typeof(data)#*  [1] "list"
data$Math_Grade12; is.na(data$Math_Grade12)
####()**  all "FAALSE" that's good....
data$Math_Grade12; unique(
                data$‘Matric Authority Desc‘)
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   4.2.2.  try data type changes
data<- as.data.frame(  rbind( my_data_2014y,
                              my_data_2017y)   )
typeof(data)#[1] "list"
####()* seems type remains the same
data$Math_Grade12; is.na(
                 data$Math_Grade12   )
####()** all "FAALSE" that's good....
data[,ncol(data)];is.na(data)
#check the last column & 'nas' #*
#*( 5. )   data cleansin' process (hier)**
#   ***********************************************
#*   5.1.    feature cleanin'...

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   5.1.1.   matric authority
unique(data$‘Matric Authority Desc‘)
  #*()**unique(data_sampled$‘Matric Authority Desc‘)

require(sqldf)#*   install.packages("sqldf")
#*()**data$Matric_Auth<-   data$‘Matric Authority Desc‘
data<-   data %>%
    mutate(Matric_Auth=‘Matric Authority Desc‘)
sqldf("select␣a.*,
case␣when␣a.[Matric_Auth]='Gauteng␣Educ␣Dept'
␣␣␣␣then␣'GP'
when␣a.[Matric_Auth]='Free␣State␣Dept␣of␣Educ'
␣␣␣␣then␣'FS'
when␣a.[Matric_Auth]='IEB␣-␣Independant␣Exam␣Board'
␣␣␣␣then␣'IE'---*()**␣␣␣'IEB'
when␣a.[Matric_Auth]='KZN␣Dept␣of␣Educ'
␣␣␣␣then␣'KZN'
when␣a.[Matric_Auth]='Limpopo␣Prov␣Dept␣of␣Educ'
␣␣␣␣then␣'LI'
when␣a.[Matric_Auth]='North␣West␣Dept␣of␣Educ'
```

```
␣␣␣␣␣then␣'NW'
when␣a.[Matric_Auth]='Mpumalanga␣Dept␣of␣Educ'
␣␣␣␣␣then␣'MP'
when␣a.[Matric_Auth]='Cambridge'␣then␣'CMB'
when␣a.[Matric_Auth]='Eastern␣Cape␣Dept␣of␣Educ'
␣␣␣␣then␣'EC'
when␣a.[Matric_Auth]='Northern␣Cape␣Dept␣of␣Educ'
␣␣␣␣then␣'NC'
when␣a.[Matric_Auth]='Western␣Cape␣Dept␣of␣Educ'
␣␣␣␣then␣'WC'
when␣a.[Matric_Auth]='Foreign␣Country'
␣␣␣␣then␣'FR'---*()**'FRC'
when␣a.[Matric_Auth]
in␣('NE␣-␣Not␣Applicable','0␣-␣Not␣Provided')
␣␣␣␣then␣'NA'
end␣as␣Authority
from␣data␣a")

datax<-  sqldf("select␣a.*,
case␣when␣a.[Matric_Auth]='Gauteng␣Educ␣Dept'
␣␣then␣'GP'
when␣a.[Matric_Auth]='Free␣State␣Dept␣of␣Educ'
␣␣␣then␣'FS'
when␣a.[Matric_Auth]='IEB␣-␣Independant␣Exam␣Board'
␣␣␣then␣'IEB'---*()**␣␣␣'IE'---*()**␣␣␣'IEB'
when␣a.[Matric_Auth]='KZN␣Dept␣of␣Educ'␣then␣'KZN'
when␣a.[Matric_Auth]='Limpopo␣Prov␣Dept␣of␣Educ'
␣␣␣␣then␣'LI'
when␣a.[Matric_Auth]='North␣West␣Dept␣of␣Educ'
␣␣␣then␣'NW'
when␣a.[Matric_Auth]='Mpumalanga␣Dept␣of␣Educ'
␣␣␣then␣'MP'
when␣a.[Matric_Auth]='Cambridge'␣then␣'CMB'
when␣a.[Matric_Auth]='Eastern␣Cape␣Dept␣of␣Educ'
␣␣␣then␣'EC'
when␣a.[Matric_Auth]='Northern␣Cape␣Dept␣of␣Educ'
␣␣␣then␣'NC'
when␣a.[Matric_Auth]='Western␣Cape␣Dept␣of␣Educ'
␣␣␣then␣'WC'
when␣a.[Matric_Auth]='Foreign␣Country'␣then␣'FRC'
---*()**␣␣␣␣␣␣␣␣␣␣'FR'---*()**␣␣␣'FRC'
when␣a.[Matric_Auth]␣in␣('NE␣-␣Not␣Applicable',
'0␣-␣Not␣Provided')␣then␣'NA'
```

```
end␣as␣Authority
from␣data␣a")
#####    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx    **
##-------------------------------------------------**
#*    5.1.2.    'Academic Plan'
sub("\\:.*", "", datax$'Academic Plan')
#*()**
datax<- datax   %>%
             mutate( Plan =
      sub("\\:.*", "", datax$'Academic Plan') )

head(datax,11)
#####    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx    **
##################################################
#*    5.2.     feature cleanin''...

#####    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx    **
##-------------------------------------------------**
#*    5.2.1.    rename (columns) with "_" - scores
datax %>%
      rename(
Ethnic_Group_Desc = 'Ethnic␣Group␣Desc'#,
#
             ,Semester_Mark  = 'Semester␣Mark'
             ,Exam_Mark  = 'Exam␣Mark'
             ,Final_Mark='Final␣Mark'
             ,  Gender_Desc='Gender Desc'
,Offering_Language_Desc='Offering Language Desc'
              , Academic_Plan='Academic Plan'
, Matric_Authority_Desc='Matric Authority Desc')
datax<- datax %>%
             rename(
 Ethnic_Group_Desc = 'Ethnic␣Group␣Desc'#,
#
,Semester_Mark  = 'Semester␣Mark'
               ,Exam_Mark  = 'Exam␣Mark'
               ,Final_Mark='Final␣Mark'
               ,  Gender_Desc='Gender Desc'
,Offering_Language_Desc='Offering Language Desc'
            , Academic_Plan='Academic Plan'
, Matric_Authority_Desc='Matric Authority Desc'
               ,St_number='St␣number'
                   )
```

```
#####    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx    **
##-------------------------------------------------**
#*    5.2.2.     (shuffle the pack) add Facculty
#                                  or schools'
unique(datax$Plan)

dataxx<-  sqldf("select␣a.*,
case␣when␣a.[Plan]␣like'BSc%'␣then'NAS'---'Science'
␣␣when␣a.[Plan]␣like'BCom%'␣or␣a.[Plan]␣like␣'%Econ%'
␣␣␣␣␣or␣␣a.[Plan]␣like'BCon%'␣then'EMS'---'Commerce'
␣␣when␣a.[Plan]␣like'BA%'
␣␣␣␣␣or␣a.[Plan]␣like'%BPolSci%'
␣␣␣␣␣or␣a.[Plan]␣like'BEd%'
␣␣␣␣␣or␣a.[Plan]␣like'%BSocSci%'then␣'Humanities'
␣␣when␣a.[Plan]␣like'LLB%'␣then'Law'
--*()**␣when␣a.[Plan]␣like␣'BIS%'␣then␣'BIS'
---*()**when␣a.[Plan]␣like␣'BTRP%'␣then␣'BTRP'
--BCon
---when␣a.[Plan]␣like␣'BCon%'␣then␣'Consumer'
--BTown␣and␣Regional␣Planning
---*()***when␣a.[Plan]␣like␣'%BTown␣and␣Regional␣Planning%'
--then␣'Engineering'
else␣'other'␣---*()***␣␣else␣a.[Plan]
end␣as␣Faculty
,␣case␣when␣a.[Gender_Desc]␣=␣'Female'
then␣1␣else␣0␣end␣as␣Gender
--␣␣␣**(Nov␣29,␣2019)***␣␣␣add␣race␣to
---␣␣␣␣␣monitor␣black/␣africans␣students
---***,case␣when␣a.Ethnic_Group_Desc␣is␣'African'␣then␣1
--else␣0␣end␣as␣Race
,case␣when␣a.Ethnic_Group_Desc
in␣('African','Coloured')␣then␣1
else␣0␣end␣as␣black
,case␣when␣a.Ethnic_Group_Desc␣is␣'White'
then␣1␣else␣0␣end␣as␣white
,case␣when␣a.Ethnic_Group_Desc␣is␣'Indian'
then␣1␣else␣0␣end␣as␣india
from␣datax␣a")

head(dataxx,11)
dataxx$Ethnic_Group_Desc;unique(
                  dataxx$Ethnic_Group_Desc   )
#*[1] "White" "Coloured" "African" "Indian"
```

```
#####    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx    **
##-------------------------------------------------**
#*    5.2.3. (data) factorize some factor columns'
as.factor(dataxx$treat);str(
      as.factor(dataxx$treat)); glimpse(
                           as.factor(dataxx$treat))
str(datax$Repeat)
unique( (datax$Repeat) )
#####()**data$treat<- as.factor(data$treat)
typeof(dataxx$treat)
#[1] "character"
dataxx$treat<- as.factor(dataxx$treat)
datax$treat ;unique(
                dataxx$treat)
typeof(dataxx$treat)#
typeof(dataxx$Repeat)#

dataxx$Repeat<- as.factor(dataxx$Repeat)
typeof(dataxx$Repeat)#*[1] "integer"
typeof(as.factor(dataxx$Repeat))#[1] "integer"
typeof(dataxx$`Gender`)#*typeof(dataxx$`Gender_Desc`)

dataxx$Gender<- as.factor(dataxx$`Gender`)
typeof(dataxx$`Gender`);unique(dataxx$`Gender`)
#
unique(dataxx$`Ethnic_Group_Desc`)
##[1] "White" "Coloured" "African"  "Indian"
#> gotta code zis
typeof(dataxx$`Ethnic_Group_Desc`)#
typeof(dataxx$`Gender_Desc`)#[1] "character"

dataxx$Ethnic_Group_Desc<-
          as.factor(dataxx$`Ethnic_Group_Desc`)
typeof(dataxx$`Ethnic_Group_Desc`)#[1] "integer"

unique(dataxx$`Ethnic_Group_Desc`)
#[1] White Coloured African  Indian
##Levels: African Coloured Indian White

typeof(dataxx$`Authority`)
dataxx$Authority<- as.factor(dataxx$`Authority`)
typeof(dataxx$`Authority`)#[1] "integer"    "
```

```
unique(dataxx$`Authority`)
#[1] GP  IEB MP  LI  EC  CMB NW  KZN FRC FS  NC  WC  NA
#Levels: CMB EC FRC FS GP IEB KZN LI MP NA NC NW WC
typeof(dataxx$`Plan`)
dataxx$Plan<- as.factor(dataxx$`Plan`)
typeof(dataxx$`Plan`) #[1] "integer"    "
unique(dataxx$`Plan`);typeof(dataxx$`Faculty`)#*

dataxx$Faculty<- as.factor(dataxx$`Faculty`)
typeof(dataxx$`Faculty`);unique(dataxx$`Faculty`)
#Levels: Commerce Humanities Law other Science

#Home_Language_Instr
typeof(dataxx$`Home_Language_Instr`)

dataxx$Home_Language_Instr<- as.factor(
                dataxx$`Home_Language_Instr`
                                      )
typeof(dataxx$`Home_Language_Instr`)
unique(dataxx$`Home_Language_Instr`)
#++++++++++++++++++++++++++++++++++++++++++++++++++++++++++  ++
##---                                                    ---**
#*   **  x. permute data @random...**
##---                                                    ---**
#++++++++++++++++++++++++++++++++++++++++++++++++++++++++++  ++

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   .3.1.    (shuffle the pack)

#* i. set random seed (so that work is reproducible)

set.seed(1234567)
#* ii. use the sample()function to shuffle row indices

rows <- sample(nrow(dataxx))
#* ii. lastly, use random vector (to reorder the df)

data_sampled <- dataxx[rows, ]
head(data_sampled,11)
#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
```

```
#*   5.3.2 add row_id column

data_sampled <- data_sampled %>%
                    mutate(id = row_number())
#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   .3.1.    (shuffle the pack)


#*   1. set random seed
#              (for reproducibility):

set.seed(1234567)

#*   2.use the sample function
#      (shuffle rowindices):

rows <- sample(nrow(dataxx))

#*   3. lastly, use random vector
#        (to reorder the df):

data_sampled <- dataxx[rows, ]
         head(data_sampled,11)
#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   5.3.2 add row_id column

data_sampled <- data_sampled %>%
                   mutate(id = row_number())

#*(6)    the data clensin' process continues (hier)**
##################################################
#*   6.1.    feature organisin''...

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   6.1.1 importing the data (into data_cleaning)
colnames(data_sampled)
column_types_stk <-   cols(
                 Semester_Mark = "d",
                     Exam_Mark = "d",
                    Final_Mark = "d",
```

```
                  Math_Grade12= "d",
              `Number of Years`= 'd',
           Term = col_factor(c("2014", "2017")),
     Ethnic_Group_Desc = col_factor(
   c("African", "Coloured", "Indian", "White")),
  Offering_Language_Desc =
           col_factor(c("English", "Afrikaans")),
   Gender_Desc = col_factor(c( "Female","Male")),
   Home_Language_Instr = col_factor(c("0", "1")),
Matric_Authority_Desc = col_factor(
               c("Gauteng Educ Dept",
                 "Free State Dept of Educ",
            "IEB - Independant Exam Board",
                      "KZN Dept of Educ",
               "Limpopo Prov Dept of Educ",
                 "North West Dept of Educ"
                ,"Mpumalanga Dept of Educ",
                          "Cambridge",
                "Eastern Cape Dept of Educ"
               ,"Northern Cape Dept of Educ",
                 "Western Cape Dept of Educ",
                       "Foreign Country"
                ,"NE - Not Applicable",
                 "0 - Not Provided"  )),
Authority= col_factor(c("CMB","ECC","FRC",
                    "FS","GP","IEB","KZN",
          "LI","MP","NA","NC","NW","WC")),
   Faculty= col_factor(c("EMS",
                         "Humanities",
                                "Law",
                              "other",
                               "NAS")),
                Plan= col_factor(c("BCom ",
               "BSc Information Technology",
                           "BSc",
                    "BSc(Computer Science)",
                           "BIS",
                  "BSocSci", "BA",
            "BSc (Construction Management)",
                        "BConsumer Science",
              "BTown and Regional Planning",
                "BConSci", "BScAgric","BEd",
               "BAdmin", "BTRP", "BComHons",
```

```
                "Econ␣and␣Man␣Sc␣UG", "LLB␣",
                          "BPolSci",
                        "BISHons")),
treat  = col_factor(   c("0", "1")   ),
Gender = col_factor(   c("0", "1"))   )
####()**   Warning message:

data_cleaning <- data_sampled%>%
dplyr::select(   St_number,
                 treat,
                 Final_Mark,
                 Repeat,
                 Gender,
                       Ethnic_Group_Desc,
                 Plan,
                       Home_Language_Instr
                , Math_Grade12,
                          Authority,
                            Faculty,
                   Matric_Authority_Desc ,
                                Plan,
                             Gender_Desc,
                                  Term ,
                 Offering_Language_Desc ,
                          Semester_Mark ,
                         Exam_Mark,
                         `Matric Year`
                   ,`Number of Years`   )   %>%
   mutate(Pass= ifelse(Final_Mark>49,"Yes","No")
       #*add normal pass and or distinctions...
       ,Pass50= ifelse(Final_Mark>49,"Yes","No")
       ,Pass75= ifelse(Final_Mark>74,"Yes","No"))
data_cleaning <-
sqldf("Select␣St_number,
␣␣␣␣␣␣␣␣treat,
Final_Mark␣,Repeat,␣␣Gender,
Ethnic_Group_Desc,␣␣---Plan,
Home_Language_Instr,
Math_Grade12,
Authority,␣Faculty,
Matric_Authority_Desc,␣Plan,
Gender_Desc,Term,
Offering_Language_Desc,
```

```
Semester_Mark,␣Exam_Mark,
`Matric␣Year`,`Number␣of␣Years`
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣from␣data_sampled" )%>%
    mutate(Pass= ifelse(Final_Mark>49,"Yes","No")
          ### add normal pass and or distinctions...
          ,Pass50= ifelse(Final_Mark>49,"Yes","No")
          ,Pass75= ifelse(Final_Mark>74,"Yes","No")
                )
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   6.1.2. get head'

data_cleaning head(data_cleaning,7)
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   6.2. Check for any missingness?

data_cleaning %>%
           miss_var_summary()
#+++++++++++++++++++++++++++++++++++++++++++++++++++  ++
##---                                              ---**
#*II. final set-up before propensity scores (methods)
##---                                              ---**
#+++++++++++++++++++++++++++++++++++++++++++++++++++  ++

#*(1.)identify your treatment / exposure(s) of interest(s)
##################################################   &&&
#*1.1. the treatment of interest here is ('pre
#            and post-clickerS') enrolments...

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   our treatment studied
data_cleaning %>%
               tabyl(Term)
#*(2.) Outcomes/ responses of interest(s) (hier)
##################################################
###*2.1. student exam, final marks plus a
#              binary outcome('passed')...

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   2.1.1. refactoring outcome & exposure variables'
```

```
#*()**   Refactoring

data_clean <- data_cleaning %>%
        mutate(exposure = as.numeric(Term == "2017"),
                    Pass = fct_relevel(Pass, "Yes"),
                  passed = as.numeric(Pass == "Yes"))
str(data_clean$exposure)
glimpse(data_clean[,c('exposure','Pass','passed')])
#*
#   ************************************************
#*2.2. student final course mark,as
#       (a quantitative outcome: 'Final_Mark'...)

#*Finale<-    data_clean$Final_Mark
#   ************************************************
#*2.2. student exam mark
#           (as second quantitative outcome)

Exam<- #exam mark ( will be used as outcome of choice!)**
                            data_clean$Exam_Mark
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   2.2.1.   Our second outcome is ....
data_clean <- data_clean %>%
              mutate(yrsbefore = Term - `Matric Year`)
#years before attemptin' (the module)  *(nb)**   ...
head(data_cleaning,11)

mosaic::favstats(Final_Mark ~ Term,
            data = data_clean) #all of these
#values looks reasonable
#(in that they are all positive)/no missing(s)...
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   22.2.2. impute missing (for years before)
data_clean %>%
       filter(is.na(yrsbefore))%>%
 dplyr::select(St_number,
               Final_Mark,
               Term,
                yrsbefore)

data_clean <- data_clean %>%
```

```
     mutate(yrsbefore = replace_na(yrsbefore,
                 mean(is.numeric(yrsbefore))))
#*
data_clean %>%
        select(yrsbefore) %>%
                      miss_var_summary()
#*
#*(3.)  the covariates features('confounders') listed...
#   **************************************************
#*   3.1.    the covariates of interest...

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   3.1.1. the covariates/ features
## Covariates
vars<-c( "Repeat",  #*()**  ...
                          "Plan",
          "Home_Language_Instr",
                      "Faculty",
                "Semester_Mark",
                       "Math_Grade12"
                      ,"Number of Years"
                        ,"yrsbefore",
                        "Gender_Desc",
                          "Authority",
                        "Math_Grade12",
                    "Ethnic_Group_Desc"
           )

xvars<- c(     'Repeat',
                               'Gender_Desc',
                    'Ethnic_Group_Desc',
###'Plan',
                                 'Authority',
                     'Faculty',
                     'Home_Language_Instr',
                            'Math_Grade12'
                             ,'yrsbefore'
           )
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   3.1.2.  get a tableOne output for these...
tableOne <-   CreateTableOne(vars = xvars,
```

```
                              strata = "Term",
   #*  CreateTableOne(vars = vars, strata = "Term",
                               data = data_clean,
                               test = FALSE)

print(tableOne, smd = TRUE)

## (count) covariates with important imbalance
addmargins(table(ExtractSmd(tableOne) > 0.1))
###> addmargins(table(ExtractSmd(tableOne) > 0.1))
## ###FALSE  TRUE   Sum
###   4     4     8
##
###Authority (%)~           0.224;Gender_Desc
#= Male (%) ~0.121;  Repeat = 1 (%)~ 0.106; and
##Faculty (%)~        0.110

#*(4.)    4.   set random seed ( for rproducibility)**
#   ***************************************************
#*   4.1.    random seed...

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
###here is our seed (for reproducibillity).

set.seed(1234567)
#*
#   *********************************************************
#*   4.2. gettin my data into environ' (for.better.man).

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#**...
attach(data_clean)# ###    ****

##  ######(Data man steps!)**    ...
##------------------------------------------------------ (EnD!)**
```

2. Propensity score modeling

```
##------------------------------------------------------ (START!)**

####()**   ...   ii. PSM methods
##-------------------------------------------------**
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **

#*            *created by: Mxolisi Msibi, Mr.
#*            *organisation: University of Pretoria
#**           ** degree: Magister Scientiae
#*            *       datae: 2019.11.25
#*            *      update: 2020.01.08
#*            *      update 6: 2020.05.18
#*            *      update 7: 2020.05.21
#*            *     update 8: 2020.05.22/23

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**

##-------------------------------------------------**
#*   get required libRaries (step)**
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*  installin' (any) packages

#install.packages(c("any1","any2","..."))
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*     load packages

library(MatchIt)##install.packages('MatchIt')
library('optmatch')#*install.packages("optmatch")
#relaxinfo()
library("RItools")#*install.packages('RItools')
require("vctrs")
library(Hmisc)#*(Mxova's)*.*install.packages("Hmisc")
#install.packages("WeightIt")
library(WeightIt) #citation("WeightIt")
library(ggplot2)
library(cowplot)#install.packages("cowplot")
library(ggridges)##install.packages("ggridges")
##-------------------------------------------------**
```

```
#*        create that match_data for opt
##-----------------------------------------------**

#####*    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx    **
#*       some final data man' steps*    ...
match_data<- data_clean[,c( "Repeat"
                            , "Ethnic_Group_Desc",
              "Plan"                               ,
              "Home_Language_Instr"
                                  , "Math_Grade12"
                                     , "Authority"
                                       , "Faculty"

                                        , "Gender"
                                   , "Gender_Desc"

                                    , "yrsbefore" ,
                                         'passed',
                                       'exposure'
                                   , 'Final_Mark',
                                      'Exam_Mark'
                                    , 'St_number')]
head(match_data,7)#*...Jan 05, 2020*
#####*    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx    **
#*       some final data man' steps*    ...
dataxx[rowSums(is.na(dataxx))==0,]
dataxx[,  c("Repeat" ,
                         "Gender",
               "Ethnic_Group_Desc",
#*()**    "Offering_Language_Desc",
#*()**
                            "Plan",
             "Home_Language_Instr",
                   "Math_Grade12"
##    #,"Matric_Auth"
                      ,'Authority',
'Faculty')][rowSums(is.na(
         dataxx[,
                 c( "Repeat" ,
                                      "Gender" ,
                            "Ethnic_Group_Desc" ,

#*()**    "Offering_Language_Desc",
```

```
#*()**
                                          "Plan" ,
                            "Home_Language_Instr" ,
                                   "Math_Grade12" ,
##   #,"Matric_Auth"
                                       'Authority',
                                  'Faculty')]))==0,]
#> 2986+125
#[1] 3111
2986+125 #*(   > 2986+125   [1] 3111)**

v <- data.frame(colnames(match_data));v
#data.frame(colnames(match_data));v

v1 <- data.frame(old = colnames(match_data), #*
                 new = colnames(match_data),
                 new = c('Repeats',
                         'Race',
                       'Academc Plan',
                     'Home language Instruction',
                       'Grade 12 Math',
                       'Authority',
                       'Faculty',
                       'Sex',
                          'Gender Description',
                        'Years before attempt'
                          ,'Pass', 'Exposure',
                                 'Final Mark',
                                 'Exam Mark'
                                ,'Student No')
                 );v1

xvars1<- c('Repeat','Gender_Desc',
              'Ethnic_Group_Desc'
           ,'Authority','Faculty',
            'Home_Language_Instr'
           ,'Math_Grade12',
            'yrsbefore')
v2<- data.frame(old = xvars1, ###

                 new= c('Repeats',
                                'Sex',
                               'Race',
```

```
                            'Authority',
                              'Faculty',
                          'Home language',
                          'Grade 12 Math',
                           'Years prior' )
);v2;####()**  ...  (^   )
new.names<- c(Authority_IEB = "Authority: IEB",
              Authority_NC = "Authority: NC",

               Authority_GP = "Authority: GP",

               Authority_NW = "Authority: NW",
               Authority_NA = "Authority: NA",
                Authority_LI= "Authority: LI",
              Authority_CMB = "Authority: CMB",
              Authority_KZN = "Authority: KZN",
               Authority_FS = "Authority: FS",
              Authority_FRC = "Authority: FRC",
               Authority_MP = "Authority: MP",
               Authority_EC = "Authority: EC",
               Authority_WC = "Authority: WC"
            ,  yrsbefore= "Years prior (Years)"
         ,  Gender_Desc_Male= "Sex (Male)"
#,  Gender_Desc_Male= "Sex (F/M)"
                  ,  Repeat= "Repeats  (Y/N)"
,Faculty_Science= "Faculty: NAS",#"Faculty: Science",
               Faculty_Law= "Faculty: Law"
,Faculty_Commerce= "Faculty: EMS",#"Faculty: Commerce",
             Faculty_other= "Faculty: other"
        ,Home_Language_Instr= 'Home Language (Y/N)'
,  Ethnic_Group_Desc_African= 'Race: Black'
   ,Ethnic_Group_Desc_Indian= 'Race: Indian' ,
     Ethnic_Group_Desc_White= 'Race: White'  ,

  Ethnic_Group_Desc_Coloured= 'Race: Colored'
             ,  Math_Grade12= "Grade 12: Math score"
)
#######################################################  ++
##---                                                 ---**
#*   **  I. Estimate propensity scores ((step)**
##---                                                 ---**
#######################################################  ++
```

```
##-------------------------------------------------**
#*          logistic regression*...
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*      *set seeds (for reproducibility)
set.seed(1234567)###
##-------------------------------------------------**
#*   fit our propensity score model
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*       ps estimation via logistic regression
ps_model<-glm( (Term==2017)~Repeat+
                          Ethnic_Group_Desc+
                    Plan+#*...
                        Home_Language_Instr+
                         Math_Grade12+
                                 Authority+
                              Faculty+
                                Gender_Desc+
#           'Number of Years'+
                                    yrsbefore
          ,   family = binomial(link = "logit") ,
                data = match_data
              ####data_clean ##
)
summary(ps_model) ###>
##-------------------------------------------------**
#*   data man' (on pscores)*...
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*   -attach the predicted scores to datafile
match_data$pscore <- predict(
                           ps_model,
             type= "response")
#*(May 20, 2020)**match_data<- data_clean
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*   -concatenate pscores and linpscores to data
pscores<- fitted(ps_model);  linpscores =
                     ps_model$linear.predictors
```

```
data_cleaner <- data_sampled %>%
                #data_cleanComplete%>%
                     mutate(   ps = pscores,
                               linps = linpscores)

max(data_cleaner$linps)#*()**[1] 16.25023
#[1] 16.23288##*[1] 16.20938
head(data_cleaner, 7)

data_cleaner %>%
#dplyr::select(St_number, ps, linps) %>%
                  dplyr::select( ps, linps) %>%
                               head(7)

#*OR
sqldf("Select␣St_number,
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣ps,
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣linps
␣␣␣␣␣␣␣␣from␣␣data_cleaner")%>%
#*sqldf("Select St_number,ps,linps from  data_clean")
                           head(7)

attach(data_cleaner)#*attachin the clean dataset*...
##-------------------------------------------------**
#*  causal assumptions*...
##-------------------------------------------------**

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*    plot these scores (to Verify 'common-support')

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*        get min max values
min((data_cleaner%>%subset(treat=1))$ps)
#*[1] 9.950226e-08
min((data_cleaner%>%subset(treat=0))$ps)
#*[1] 9.950226e-08

max((data_cleaner%>%subset(treat=1))$ps)
#*[1] 0.9999999
max((data_cleaner%>%subset(treat=0))$ps)
#*[1] 0.9999999
#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*     get min/ max values  (via sql)
```

```
min(sqldf("select␣*
␣␣␣␣␣␣␣␣␣␣␣from␣data_cleaner
␣␣␣␣␣␣␣␣␣␣␣where␣Term=2017")$ps)#[1]0.09684227
max(sqldf("select␣*
␣␣␣␣␣␣␣␣␣␣␣from␣data_cleaner
␣␣␣␣␣␣␣␣␣␣␣where␣Term=2014")$ps)#[1]0.9043355
###*min ~ [1] 0.06895231 ###*max ~ [1] 0.8983455
##-------------------------------------------------**
#*        the Common-Support region
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*         plot data
ggplot(data_cleaner, aes(x = ps, fill = treat)) +
                      ###*()**
   geom_density(alpha = 0.5) +
   theme_bw()+
      scale_fill_grey()+
theme_classic()+##*()** change my scale grey...
#changin my line type, color and size(common-support)
     geom_vline(xintercept = min(
sqldf("select␣*␣from␣data_cleaner␣where␣Term=2017")$ps),
   linetype="dotted",  color = "red", size=1.0)+
   geom_vline(xintercept = max(
sqldf("select␣*␣from␣data_cleaner␣where␣Term=2014")$ps),
     linetype="dotted", color = "red", size=1.0)
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#   #*       common- support (panoramic view) plot
ggplot(data_cleaner,aes( x= treat, y= ps)) +
                               ###*()**
geom_violin(aes(fill = treat)) +
                    scale_fill_grey()+
     theme_classic()+#*()**  channge scale grey.....
   geom_boxplot(width = 0.2) +
  guides(fill = FALSE) +
   coord_flip() +
    theme_bw()  +
# changin' the line type, color and size(common-support)
     geom_hline(yintercept = min(
sqldf("select␣*␣from␣data_cleaner␣where␣Term=2017")$ps),
    linetype = "dotted",
color = "red", size=1.0)    +
     geom_hline(yintercept = max(
```

```
 sqldf("select␣*␣from␣data_cleaner␣where␣Term=2014")$ps),
   linetype ="dotted",
color = "red", size=1.0)#*this plot stays!!!!!!!   **
##-------------------------------------------------**
##   # *craete prediction data (Apr 07, 2020)**
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#   #*       create pred data (to rank asc)*
predData<- data.frame(
             probTreat = ps_model$fitted.values,
                 treat = data_clean$treat)
##-------------------------------------------------**
#*   *- loglikelihood ratio tests
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*  calc the log-likelihood ("just for control")*
ll_proposed<- ps_model$deviance*((-2)^-1)
## residual dev ~   psmodel$deviance
ll_null<- ps_model$null.deviance

(ll_null-ll_proposed)/ll_null
###[1] 0.0225006###*[1] 0.07615674#*[1] 1.459262
#*^  pseudo R^2 1  -  pchisq(
2*(ll_proposed - ll_null),
 df=(length(ps_model$coefficients)-1)
)
#*[1] 1.461126 #[1] 1.00967e-10
#i.e. ~ 0
#*.. [1] 0 #*[1] 1

#*()** Apr 08..
pred_data<- predData[
               order(predData$probTreat,
                       decreasing = F),  ]
pred_data$rank<- 1:nrow(pred_data)
pred_data$rank#*ggplot()
##-------------------------------------------------**
##   # *-plot the sigmoid curve
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
```

```
#   #*       prep pred data (for plot)*
pred_data<- pred_data%>% #as.factor(treat)
mutate(treated= ifelse(treat== 1,"yes", "no"))
#*
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#   #*       create the plot object*
sigPlot<- ggplot(data= pred_data,
                 aes(x=   probTreat, y=   rank))+
#geom_point()
  #aes(x=   rank, y=   probTreat))+###geom_point()
   geom_point( aes(color= treated) ,
                   alpha=1, shape=4, stroke=2)+
   scale_color_manual(values =
         c("yes" = "darkslategrey", "no" = "grey"))+
   ylab("index")+
   xlab("treatment␣allocation␣probability")+
   theme_classic() +# Classic theme (wanted*)
theme(#Hide panel borders and remove grid lines
     panel.border = element_blank(),
     panel.grid.major = element_blank(),
     panel.grid.minor = element_blank()
) # +
###geom_hline(yintercept = min(
#sqldf("select *
#from data_cleaner where Term=2017")$ps),
#linetype="dotted",
#color = "red", size=1.0)
#scale_fill_grey()+ colors = c("grey", "navy")

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#   #*       plot the plot object*
sigPlot +
#*mark the common-support region
geom_vline(xintercept = min(
   sqldf("select␣*
␣␣␣␣␣␣␣␣␣␣from␣data_cleaner
␣␣␣␣␣␣␣␣␣␣where␣Term=2017")$ps),
     linetype ="dotted",
       color = "red", size=.70)  +
#scale_fill_grey()+ colors = c("grey", "navy") +
    geom_vline(xintercept = max(
    sqldf("select␣*
␣␣␣␣␣␣␣␣␣␣from␣data_cleaner
```

```
␣␣␣␣␣␣␣␣␣␣where␣Term=2014")$ps),
     linetype = "dotted",
     color = "red",
     size = .70)#  +
#scale_fill_grey()+ colors = c("grey", "navy")
#++++++++++++++++++++++++++++++++++++++++++++++++++++++++++  ++
##---                                                    ---**
#*   **  II. Propensity scores match((step)**
##---                                                    ---**
#++++++++++++++++++++++++++++++++++++++++++++++++++++++++++  ++

#(0)*
##-------------------------------------------------------**
#*         quantitative outcome: (final mark)
##-------------------------------------------------------**


#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*      *an un-matched scenario (hier)**...
table(Final_Mark, treat)
t.test(   Final_Mark~   (treat==1)   )#

table(Exam_Mark, treat)
t.test(   Exam_Mark~   (treat==1)   )
#(1)*
##-------------------------------------------------------**
##   # *binary outcome: (passed)
##-------------------------------------------------------**

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*      *an un-matched scenario (hier)**...
table(Pass, treat)
chisq.test(table(   Pass, treat   ))

Performance <-  matrix(table(Pass, treat),
                  nrow = 2,
              dimnames = list("Passed"=
                           c("Yes", "No"),
                           "Treated" =
                         c("2014", "2017")))
Performance

Performance
```

```
mcnemar.test(Performance)

(1298-247)^2/(1298+247)#[1] 714.9521
#(2)*
##-------------------------------------------------**
##   # *...  greedy matching (with caliper)*
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*      ...   with response variable (Fina mark)*
#**ps;linps
match.nnc<-
     Match(Tr = (Term == 2017),#Need to be in 0,1
#*logit of PS,i.e.,log(PS/(1-PS)) as matching scale
                   X = log(   ps / (1 - ps)),
              ## 1:1 matching
                  M = 1,
              ##caliper = 0.2 * SD(logit(PS))
                  caliper  = 0.2,
                  replace  = FALSE,
                  ties     = TRUE,
                  version  = "fast")
## Extract matched data
clicker_nnc.data <- data_clean[unlist(
   match.nnc[
      c("index.treated","index.control")]), ]

summary(match.nnc)###*()**

psens(match.nnc, Gamma = 2, GammaInc = 0.1)
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*      ...   with response variable (Fina mark)*
#*ps;linps
match.nncF <- Match(Tr =
                   (Term == 2017),#Need to be in 0,1
## logit of PS,i.e., log(PS/(1-PS)) as matching scale
                            X = log(   ps / (1 - ps)),
              ## 1:1 matching
                            M = 1,
                            Y = Final_Mark,
              ## caliper = 0.2 * SD(logit(PS))
                         caliper  = 0.2,
                         replace  = FALSE,
```

```
                          ties   = TRUE,
                         version  = "fast")
## Extract matched data
 clicker_nncF.data <- data_cleaner[
              unlist(match.nncF[
 c("index.treated","index.control")]), ]
summary(match.nncF)#*()**

psens(match.nncF, Gamma = 2, GammaInc = 0.1)
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*      ...   with response variable (Exam mark)*
#***ps;linps
#match.nncE <- Match(Tr = treat#*(Term == 2017),
# Need to be in 0,1 #                    ,
#     ## logit of PS,i.e., log(PS/(1-PS)) as matching scale
#     #              X = log(   ps / (1 - ps)),
#                     ## 1:1 matching
#     #              M = 1,
#     #               Y= Exam_Mark,#na.omit(Exam_Mark)
#     #              ## caliper = 0.2 * SD(logit(PS))
#                      #               caliper  = 0.2,
#                     replace  = FALSE,
#                     ties     = TRUE,
#                     version  = "fast")
#*Error in Match(Tr = (Term == 2017),
#*X = log(ps/(1 - ps)), M = 1, Y = Exam_Mark,  :
#*             Match(): input includes NAs
## Extract matched data
#clicker_nncE.data <- data_cleaner[
#               unlist(#*data_clean[unlist(#
#  match.nncE[c("index.treated","index.control")]), ]
#summary(match.nncE) ##
#*(..)**psens(match.nncE, Gamma = 2, GammaInc = 0.1)
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*      ...   with response variable (Pass)*
#*()  ps;linps
match.nncP <- Match(Tr = (Term == 2017),# Need to be in 0,1
      ## logit of PS,i.e., log(PS/(1-PS)) as matching scale
                     X = log(   ps / (1 - ps)),
                ## 1:1 matching
                   M = 1,
                    Y= Pass,###Y= Final_Mark,
                ## caliper = 0.2 * SD(logit(PS))
```

```
             caliper  = 0.2,
             replace  = FALSE,
             ties     = TRUE,
             version  = "fast")
#Extract matched data
clicker_nncP.data <- data_cleaner[unlist(#
             match.nncP[
     c("index.treated","index.control")]), ]
#
summary(match.nncP) ##
#*()**   psens(mDW, Gamma = 2, GammaInc = 0.1)
psens(match.nncP, Gamma = 2, GammaInc = 0.1)
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*       plot data
#sqldf("select a.*,b.ps from clicker_nncP left join ")
ggplot(clicker_nncP.data,
                   aes(x = ps, fill = treat)) +#*()**
  geom_density(alpha = 0.5) +
       theme_bw()+
                 scale_fill_grey()+
  theme_classic()+#*()**  change my scale grey.....
#changin my line type, color and size(common-support)
       geom_vline(xintercept = min(
       sqldf("select␣*
␣␣␣␣␣␣␣␣␣␣␣␣␣␣from␣data_cleaner
␣␣␣␣␣␣␣␣␣␣␣␣␣␣where␣Term=2017")$ps),
          linetype="dotted",
            color = "red", size=1.0) +
 geom_vline(xintercept = max(
  sqldf("select␣*
from␣data_cleaner
␣␣␣␣␣where␣Term=2014")$ps),
    linetype="dotted",
      color = "red",
        size=1.0)  +
#*geom_title("Post greedy match score distribution")
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#   #*   (panoramic view) plot of data
ggplot(clicker_nncP.data,aes(x=treat,y=ps))+####()**
      geom_violin(aes(fill = treat)) +
                        scale_fill_grey()+
theme_classic()+#*()**  change scale grey...
      geom_boxplot(width = 0.2) +
```

```
                  guides(fill = FALSE) +
coord_flip() +
              theme_bw()  +
#changin' the line type, color and size(common-support)
   geom_hline(yintercept = min(
                      sqldf("select␣*
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣from␣data_cleaner
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣where␣Term=2017")$ps),
              linetype = "dotted",
                 color = "red",
                         size=1.0)    +
   geom_hline(yintercept = max(
               sqldf("select␣*
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣from␣data_cleaner
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣where␣Term=2014")$ps),
    linetype ="dotted", color = "red", size=1.0)
#*(^^^)**this plot stays!!!!!!!!!!!!!!!!*
#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*         plot data
#sqldf("select a.*,b.ps from clicker_nncP left join ")
ggplot(clicker_nncF.data, aes(x = ps, fill = treat))+#*
      geom_density(alpha = 0.5) +
      theme_bw()+
            scale_fill_grey()+
  theme_classic()+###()**   change my scale grey...
#changin my line type, color and size (common-support)
       geom_vline(xintercept = min(
            sqldf("select␣*
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣from␣data_cleaner
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣where␣Term=2017")$ps),
    linetype="dotted",
      color = "red",
       size = 1.0)  +
geom_vline(xintercept = max(
       sqldf("select␣*
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣from␣data_cleaner
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣where␣Term=2014")$ps),
   linetype="dotted",
  color = "red",
  size = 1.0)  #+
#*geom_tile("Post greedy match score distribution")
#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#   #*          (panoramic view) plot of data
```

```
ggplot(clicker_nncF.data,aes(x=treat,y=ps))+#*()**
       geom_violin(aes(fill = treat)) +
               scale_fill_grey()+
       theme_classic()+#*()**change scale grey...
           geom_boxplot(width = 0.2) +
       guides(fill = FALSE) +
                    coord_flip() +
                          theme_bw()  +

#changin' the line type, color and size(common-support)
   geom_hline(yintercept = min(
                           sqldf("select␣*
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣from␣data_cleaner
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣where␣Term=2017")$ps),
      linetype = "dotted",
         color = "red",
           size=1.0)    +
      geom_hline(yintercept = max(
                                 sqldf("select␣*
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣from␣data_cleaner
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣where␣Term=2014")$ps),
linetype ="dotted",
   color = "red", size=1.0)
#*()**this plot stays!!!!!!!!!!!!!!!!*
##------------------------------------------------**
#*(OR) 1:2 greedy matching on linear pscore(w/replacement)
##------------------------------------------------**

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#   #*          (panoramic view) plot of data
X <- linps##data_clean$linps
Tr <- as.logical(Term == 2017)
#*as.logical(data_clean$Term == 2017)

match.nnc_linps <- Match(Tr = Tr,
                         X = X,
                         M = 2,
                         estimand = "ATT",
                          replace = TRUE,
                          caliper = 0.2,
                             ties = FALSE)
summary(match.nnc_linps) #*
```

```
##Extract matched data
clicker_matched_linps <- data_cleaner[unlist(
               match.nnc_linps[
         c("index.treated","index.control")]), ]
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*         plot data
#sqldf("select a.*,b.ps from clicker_nncP left join ")
ggplot(clicker_matched_linps,
                  aes(x = ps, fill = treat)) +#*()**
      geom_density(alpha = 0.5) +
      theme_bw()+
                   scale_fill_grey()+
      theme_classic()+##*()**  change my scale grey.....
#changin my line type, color and size(common-support)
        geom_vline(xintercept = min(
sqldf("select␣*
␣␣␣␣␣␣␣from␣data_cleaner
␣␣␣␣␣␣␣where␣Term=2017")$ps),
      linetype="dotted",  color = "red", size=1.0)+
      geom_vline(xintercept = max(
sqldf("select␣*
␣␣␣␣␣␣␣from␣data_cleaner
␣␣␣␣␣␣␣where␣Term=2014")$ps),
     linetype="dotted",
       color = "red",
         size=1.0)  #+
###*geom_title("Post greedy match score distribution")
#(^^^) linearps_post.matching.psdistribution
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#   #*        (panoramic view) plot of data
ggplot(clicker_matched_linps,aes(x=treat,y=ps))+#*()**
       geom_violin(aes(fill = treat)) +
  scale_fill_grey()+
theme_classic()+#*(change scale grey...)**
 geom_boxplot(width = 0.2) +
 guides(fill = FALSE) +
coord_flip() +
theme_bw()  +
#changin' the line type, color and size(common-support)
  geom_hline(yintercept = min(
    sqldf("select␣*
␣␣␣␣␣␣␣from
data_cleaner
```

```
␣␣␣␣␣␣␣␣where␣Term=2017")$ps),
     linetype = "dotted",
   color = "red", size=1.0)    +
     geom_hline(yintercept = max(
   sqldf("select␣*
␣␣␣␣␣from␣data_cleaner
␣␣␣␣␣␣␣␣␣where␣Term=2014")$ps),
   linetype ="dotted",
      color = "red",
        size=1.0)#*(this plot stays!!!!!!!!!!!!)***
##-------------------------------------------------**
#* (OR) a 1:1 greedy matching method using linear
#pscore, (with replacement)
##-------------------------------------------------**

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#   ##*        set up the linps matching
X <- linps##data_clean$linps
Tr <- as.logical(Term == 2017)

match.nnc_linps <- Match(Tr = Tr,
                         X = X,
                         M = 1,
                  estimand = "ATT",
                   replace = TRUE,
                                caliper = 0.2,
                    ties = FALSE)
summary(match.nnc_linps)
## Extract matched data
clicker.nnc_linps.data <- data_cleaner[unlist(
                match.nnc_linps[
         c("index.treated","index.control")]), ]

matchingGreedyCaliper.data <-
                  match.data(match.nnc_linps)
##-------------------------------------------------**
##   # *(OR) 1:1 greedy Matching on the linear ps, & ps
#                         (with replacement)
##-------------------------------------------------**

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*        set up (both)* the linps and ps matching
X<- cbind(linps,ps)##data_clean$linps
```

```
Tr<- as.logical(Term == 2017)

match.nnc_both <- Match(   Tr = Tr,
                            X = X,
                            M = 2,
                        estimand = "ATT",
                         replace = TRUE,
                         caliper = 0.2,
                            ties = FALSE)
summary(match.nnc_both)
#
match.nnc_bothF <- Match(   Y= Final_Mark,
                         Tr = Tr,
                          X = X,
                          M = 2,
                            estimand = "ATT",
                             replace = F,
                             caliper = 0.2,
                                ties = FALSE)

psens(match.nnc_bothF, Gamma= 2,
                       GammaInc = 0.05)
#*()**
data_cleanComplete <- data_clean %>%
#MatchIt does no allow missing values
    dplyr::select( Final_Mark,
                #
 #**Exam_Mark,#* (Jun 06, 2020)
                #
                       (passed),
                          Term ,
                 one_of(xvars)) %>%
                             na.omit()

#*OR
sqldf("Select␣St_number␣from␣data_clean␣")
match.nncModel<- matchit(  (treat) ~
                                  Repeat +
                               Gender_Desc +
                          Ethnic_Group_Desc +
                                  Authority +
                                      Faculty
```

```
                          +Home_Language_Instr
                                +Math_Grade12+
                                     yrsbefore
                           ,method = "nearest",
                                caliper = 0.25
                    ,data = data_cleanComplete)
head(match_data,7)
summary(match.nncModel)
##-------------------------------------------------**
#*      -match using nearest-neighbor  with
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*   greedy with a caliper of 0.2( using "MatchIt" )
m.nnc <- matchit(    treat  ~     Repeat  +
                          `Gender_Desc`  +
                     `Ethnic_Group_Desc` +
                              Authority  +
                               Faculty   +
                     Home_Language_Instr +
                         Math_Grade12    +
                                  yrsbefore,
     data = data_cleanComplete,#*match_data
#**()** , mydata,#
                      method = "nearest",
                     caliper = .2)

summary(m.nnc)
##-------------------------------------------------**
#*      jitter plot for nearest neighbor
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#   #*      (for )greedy with a caliper of 0.2
plot(match.nncModel, type="jitter")
##
###*  > plot(match.nncModel, type="jitter")
#**[1]"To identify the units, use first mouse
#**...button; to stop, use second."
#[1]  627  693 1253 1877 1959
##-------------------------------------------------**
#*           *Optimal matching**   ...
##-------------------------------------------------**
```

```
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#   #*          check columns with missingness'
allmisscols <- sapply(data_clean,
        function(x) all(is.na(x)|x == ''))
allmisscols
allmisscols <- sapply(data_cleaner,
         function(x) all(is.na(x)|x == ''))
allmisscols
##-------------------------------------------------**
##   # *create that match_data for opt and greedy
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*Figurex presents the commands to estimate(Optimal)
#distance using MatchIt.
#-???--???--???-Optimal matching with 1:1 ratio
match.optModel<- matchit(    (treat) ~
                                 Repeat       +
                             `Gender_Desc`     +
                             `Ethnic_Group_Desc`
                                #*   ...+ Plan
                             +    Authority    +
                              Faculty          +
                Home_Language_Instr           +
                              Math_Grade12     +
                                       yrsbefore
, data = data_cleanComplete
#* ...   match_data###   data_clean
                                , method = "optimal"
                                , ratio  =  1)
E#*   ...  Warning message:
#In optmatch::fullmatch(d, min.controls = ratio,
#max.controls = ratio, :
#   Without 'data' argument the order of the match
#is not guaranteed to be the same
#         as your original data.#*()**
summary(match.optModel)

#Extract matched data
clicker_opt.data <- data_cleaner[unlist(
      match.optModel[
            c("index.treated","index.control")]), ]
```

```
##-----------------------------------------------**
##                -(optimal match) jitter plot
##-----------------------------------------------**


###**   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#   #*          optimal matching jitter plot
plot(match.optModel, type="jitter")
#> plot(match.optModel, type="jitter")
#**[1] "To identify the units, use first mouse button;
#**to stop, use second."
#[1]  506  572  693 1959 2010 3016
##-----------------------------------------------**
##   ##* * -create that match_data for opt and greedy
##-----------------------------------------------**

###**   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*create a dataset(df), for the optimal (& nnc)

matchOpt.data <- match.data(match.optModel)
#*matchGreedyC.data<-  match.data(match.nncMod)#
   matchGreedy.data<-  match.data(match.nncModel)

#$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$   (Jun 06, 2020)
###**   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*            plot data
head(match.data(match.optModel), 7)
matchOpt.data

mean(ps)#*[1] 0.4776599
sum(   is.na(sqldf("select␣a.*,
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣b.ps,
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣b.treat
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣from␣matchOpt_data1␣a
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣left␣join␣data_cleaner␣b
␣␣␣␣␣␣␣␣␣␣␣␣on␣b.St_number=a.subclass")$ps)
)#*[1] 266    #**  ...
matchOptPlot.data<- sqldf("select␣a.*,
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣b.ps␣as␣pscore,
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣b.treat
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣from␣matchOpt_data1␣a
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣left␣join␣data_cleaner␣b
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣on␣b.St_number=a.subclass
```

```
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣where␣a.ps␣>␣0")
matchOptPlot_data<-matchOptPlot.data
#$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$   (Jun 06, 2020)

##------------------------------------------------**
##   # *-optimal matching case
##------------------------------------------------**

###**   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*   -???--???--???-Optimal matching with 1:1 ratio
m.om<- matchit(   treat ~    Repeat +
                             Gender_Desc  +
                            ‘Ethnic_Group_Desc‘ +
#*
                                        +
#*
                                   Authority  +#
                          Home_Language_Instr  +
                                 Math_Grade12  +
                           ###*(nb)
                                    Faculty+
                                   yrsbefore,
            data = data_cleanComplete#*match_data
                 ,###data,
                   method = "optimal",
                   ratio  =     1)

summary(m.om)
##------------------------------------------------**
##   # *inverse probability of treatment weights**...
##------------------------------------------------**

###**   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*             new dataset
#   #*      -set possible binary outcome responses
treat<- data_clean$exposure; Finale<-
   data_clean$Final_Mark;Pass<-
                as.factor(data_clean$passed)
Exam<-#exam mark(used as ’outcome’ of interest!)**
         data_clean$Exam_Mark
Pass50<- as.factor(   data_clean$Pass50   )
Pass75<- as.factor(   data_clean$Pass75   )
#*
```

```
mydata<-cbind(data_clean[,xvars],
                                treat,
             Finale                  ,
                                Exam ,
              Pass                   ,
              Pass50                 ,
                               Pass75,
                                linps,
                          ps )###*()**   ...
head(mydata,7)

mydata<-data.frame(mydata)
str(mydata); glimpse(mydata)
##-------------------------------------------------**
##   # *create inverse propensity weights
##-------------------------------------------------**

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#   #*     -get weights (for the ip weighted matching)

         weight<- ifelse(treat==1,1/(ps),1/(1-ps))

max(weight); min(weight)#*> max(weight); min(weight)
#[1] 14.50278
#[1] 1                    #*
mean(weight)#[1] 1.978607
sd(weight)#[1] 0.7216573
mean(weight)+sd(weight)#[1] 2.700264
mean(weight)-sd(weight)#[1] 1.256949
summary(weight)
#   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#1.000   1.610   1.868   1.979   2.202  14.503
##-------------------------------------------------**
#*          apply weights to data
##-------------------------------------------------**

weighteddata<-   svydesign(ids = ~ 1,
                           data = mydata
                      , weights = ~ weight)
summary(weighteddata)
##-------------------------------------------------**
#*       -inverse probability weights case
##-------------------------------------------------**
```

```
#####    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx    **
#    ##       -get weights* for the weighted adjustment
weights.out <- weightit(  (treat) ~     Repeat  +
                                        Gender_Desc  +
                                  Ethnic_Group_Desc  +
                                          Authority  +
                                            Faculty  +
                                  Home_Language_Instr+
                                         Math_Grade12+
                                             yrsbefore
                        ,data = data_cleaner###mydata
                        , estimand = "ATT"
                        , method = "ps" )
#+++++++++++++++++++++++++++++++++++++++++++++++++++++++++  ++
##---                                                  ---**
#*    **  III. Balance diagnostics (step)****
##---                                                  ---**
#+++++++++++++++++++++++++++++++++++++++++++++++++++++++++  ++

#####    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx    **
#*       *numeric balance / SMDs (hier)**...
#*    tableone for 'greedy-caliper' matching...
#*              unadjusted (table.one)

print(tableOne, smd = TRUE)
#####    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx    **
#    ##     greedy with caliper (table.one)
tableOne_Greedy <- CreateTableOne(
                             vars = xvars,
                           strata = "Term",
data = matchGreedy.data#clicker_nnc.data
                                  ,
##clicker_nnc,##clickrMatched,
# data = clicker_nnc.data_linps,
#data = clickrMatched,
                            test = FALSE)
print(tableOne_Greedy, smd = TRUE)##
CreateTableOne(
            vars = xvars,
          strata = "Term",data = matchGreedy.data,
#data= clicker_nnc.data_linps,#data= clickrMatched,
            test = FALSE)
```

```
print(CreateTableOne(vars = xvars,
                     strata = "Term",
                       data = matchGreedy.data,
                       test = FALSE),
                        smd = TRUE)
##-------------------------------------------------**
#*   (&) tableone for optimal matching...
##-------------------------------------------------**

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#   ##*    optimal matching (table.one)
tableOne_Optimal <- CreateTableOne(
                              vars = xvars,
                            strata =   "Term",
                              data = matchOpt.data,
                              test = FALSE)
print(tableOne_Optimal, smd = TRUE)
#*(^^^) iyang'dena nou(???)
##-------------------------------------------------**
#*       weighted tableOne (ipw)
##-------------------------------------------------**

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#   ##*    weightedtable (as a table.one object)
tableOne_IPW <-svyCreateTableOne(
                              vars = xvars,
                            strata = "treat"
,                             data = weighteddata
,                             test = FALSE)
## Show table with SMD
print(tableOne_IPW, smd = TRUE)
#*(   )* ^^^ these are the after iptw adjustment SMDs
#*(^^)*.Error: covs must be a data.frame(of covariates.
#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#   ##*    weightedtable (as is!)
weightedtable <-svyCreateTableOne(
                                vars = xvars,
                              strata = "treat",
                                data = weighteddata,
                                test = FALSE)
## Show table with SMD
print(weightedtable, smd = TRUE)
##-------------------------------------------------**
```

```
##   # *visualize balance / (Love)* plots**...
##-----------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*       -3.2.0. love plot for Match m.nnc
#*  ---visual diag(s): love plot for greedy match(1)
#*(May 21, 2020)**love.plot(match.nncBal,
#*           threshold = .1,
#*                size = 1.5,
#*          var.order  = "unadjusted",
#*title = "Standardized Differences(Greedy
#*              Match with 'caliper')")+
#*    theme_bw()###*()**   ...  xxx
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*      -visual diag(s): love plot for Match m.nnc
#*-visual diag(s): love plot for greedy ('matchit')
#Repeat  +   `Gender_Desc` +
#`Ethnic_Group_Desc` +   Authority + Faculty+
#Home_Language_Instr +  Math_Grade12  +yrsbefore
love.plot(m.nnc,
###Figure 4(7) - Balance diagnostics love plots..
#nnc match case
#*(vs. pre-matching) _new.sh1t_(@treshold~ .05)
         drop.distance = TRUE,
             var.order = "unadjusted",
                   abs = TRUE,
                  line = TRUE,
             threshold = .050,#
             var.names = new.names,
        colors = c("navy", "grey"),
#*c("grey", "navy"),
        shapes = c("circle␣filled",
                   "triangle␣filled"),
sample.names = c("Greedy␣(with␣caliper)",
                 "Unadjusted"),
#*c("Unmatched", "Greedy (with caliper)"),
      limits = c(0, .82),
    position = c(.45, .80)) +###
theme(legend.box.background = element_rect(),
   legend.box.margin = margin(1, 1, 1, 1))   +
ggtitle("Greedy␣match␣(with␣caliper)
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣covariate␣balance")+
# Fixed values
```

```
geom_vline(  xintercept =  .050,
                linetype= "dotted",
             color =  "red", size=1.0)#*   xxxxx**
##-------------------------------------------------**
##   # *   ... for optmatch
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*   -visual diag(s): love plot for optimal match(1)

summary(tableOne_Optimal)
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*-visual diag(s): love plot for optimal ('matchit')
love.plot(m.om,
#*Figure 4(7) - Balance diagnostics love plots..
#*opt match case
#*(vs. pre-matching) _new.sh1t_ (@treshold~ .05)
         drop.distance = TRUE,
             var.order = "unadjusted",
                  abs = TRUE,
                 line = TRUE,
           threshold = .050,#
            var.names = new.names,
               colors = c("navy", "grey"),
         shapes = c("circle␣filled",
                    "triangle␣filled"),
   sample.names = c("Optimal␣matching",
                    "Unmatched"),
         limits = c(0, .82),
       position = c(.45, .80)) +#
theme(legend.box.background = element_rect(),
      legend.box.margin = margin(1, 1, 1, 1))+
  ggtitle("Optimal␣matching␣covariate␣balance")+
# Fixed values
     geom_vline(xintercept =  .050,
                linetype = "dotted",
                   color = "red",
                    size = 1.0)
##-------------------------------------------------**
##   # *   ... for ipw matching
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
```

```
#*       weightedtable
summary(weightedtable)
#*(888)
#####    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx    **
#*        -visual diag(s): love plot for ipw (weightit)
love.plot(  weights.out,
            drop.distance = TRUE,
             var.order = "unadjusted",
                   abs = TRUE,
                  line = TRUE,
         threshold = .05,#*threshold = .1,
            var.names = new.names,
             colors = c("navy", "grey"),#
             shapes = c("circle␣filled",
                        "triangle␣filled"),#
             sample.names =   c("Weighted",
                                "Unweighted"),
#c("IPW matching", "Pre-match"),
#*c("Unweighted", "PS Weighted"),
                  limits = c(0, .82),
                position = c(.75, .65)) +
 theme(legend.box.background = element_rect(),
       legend.box.margin = margin(1, 1, 1, 1))   +
ggtitle("Probability␣weights␣covariate␣balance")+
# Fixed values
geom_vline(xintercept = .030,#*xintercept = .050,
             linetype = "dotted",
          color = "red",
          size = .85)
#*   c("red", "blue"),colors = c("red", "blue"),
##-------------------------------------------------**
#*           combined balanced view
##-------------------------------------------------**

#####    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx    **
#*   combining weightit and matchit cov balance plots
love.plot(   (treat)   ~
                            Repeat+
                         Gender_Desc+
                      Ethnic_Group_Desc+
                           Authority+
                            Faculty+##
##Figure 4(5) - Balance diagnostics all match cases
```

```
#(incl. pre-matching) _new.sh1t_ v2.0.
#(@treshold~0.05)
               Home_Language_Instr+
                      Math_Grade12+
                         yrsbefore,
data = mydata,
            estimand = "ATT",
             weights = list(
                  w1 = get.w(   (weights.out)   ),
                  ###*()**   ...
                  w2 = get.w(   m.nnc  )
                 ,w3 = get.w(   m.om   )
            ),###get.w(m.nnc)),
           var.order = "unadjusted",
                abs = TRUE,
               line = TRUE,
           threshold = .1,
           var.names = new.names,
colors = c("grey", "navy", "darkolivegreen", 'darkslategrey'),#
shapes = c("triangle filled", "circle filled","square filled"
  ,'plus'),##
sample.names = c("Unmatched", "IPTW", "Greedy (with caliper)",
'Optimal'),#
          limits = c(0, .82)) +
theme(legend.position = c(.75, .75),###c(.75, .3),
      legend.box.background = element_rect(),
      legend.box.margin = margin(1, 1, 1, 1))+
ggtitle("Covariate balance: all cases")+
# Fixed values
geom_vline(xintercept =    .10,
              linetype="dotted",
                color = "navy",
                  size=0.000000000025)+
geom_vline(xintercept =.050,
              linetype="dotted",
                color = "red",
                  size=1.0)
#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%***

##------------------------------------------------**
##   # *individual balance plots (hier)**...
##------------------------------------------------**
```

```
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   greedy ( is '   with "ipw"  ')
#*( Matric math)**
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*          Matric math score
bal.plot(   (treat)    ~ Repeat+
               Gender_Desc+
          Ethnic_Group_Desc+
            Authority+ Faculty+
             Home_Language_Instr+
              Math_Grade12+
             yrsbefore
      ,data = mydata, #
weights = data.frame(Greedy = get.w(m.nnc),
                         IPW = get.w(weights.out)),
       method = c("matching", "weighting"),
          var.name = "Math_Grade12",
             which = "both") +
 scale_fill_grey() +
ggtitle('Grade␣12␣Math␣score␣distributional␣balance')
#*(Years prior)**
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*        Years prior attempting module
bal.plot(      (treat)    ~ Repeat+
                Gender_Desc+
                 Ethnic_Group_Desc+
                         Authority+
                           Faculty+
               Home_Language_Instr+
                      Math_Grade12+
                          yrsbefore
        , data = mydata, #
     weights = data.frame(
               Greedy = get.w(m.nnc),
                  IPW = get.w(weights.out)),
            method = c("matching", "weighting"),
          var.name = "yrsbefore", which = "both")+
scale_fill_grey()+  #
ggtitle('Years␣prior␣distributional␣balance')
#*(Gender)**
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*                 Sex distribution
```

```
bal.plot(   (treat)    ~ Repeat+
                 Gender_Desc+
           Ethnic_Group_Desc+
            Authority+ Faculty+
           Home_Language_Instr+
              Math_Grade12+
          yrsbefore
, data = mydata, #bal.plot(treat ~ new.names,
                 #data = data_clean,
         weights = data.frame(
                 Greedy = get.w(m.nnc),
                    IPW = get.w(weights.out)),
            method = c("matching", "weighting"),
          var.name = "Gender_Desc", which = "both") +
 scale_fill_grey()
#*(Race)**
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*              Race distribution
bal.plot(   (treat)    ~ Repeat+
             Gender_Desc+
             Ethnic_Group_Desc+
             Authority+ Faculty+
             Home_Language_Instr+
             Math_Grade12+
             yrsbefore
                 , data = mydata,
#bal.plot(treat ~ new.names, data = data_clean,
            weights = data.frame(
                   Greedy = get.w(m.nnc),
                      IPW = get.w(weights.out)),
         method = c("matching", "weighting"),
         var.name = "Ethnic_Group_Desc",
            which = "both") +
 scale_fill_grey()   +
 ggtitle('Race␣distributional␣balance')
#*(Authority)**
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*          Authority distribution
bal.plot(   (treat)    ~ Repeat+
                    Gender_Desc+
              Ethnic_Group_Desc+
             Authority+ Faculty+
            Home_Language_Instr+
```

```
                    Math_Grade12+
                    yrsbefore
      ,data = mydata,
      weights = data.frame(
                     Greedy = get.w(m.nnc),
                        IPW = get.w(weights.out)),
         method = c("matching", "weighting"),
       var.name = "Authority", which = "both") +
 scale_fill_grey()   +
ggtitle('Exam␣authority␣distributional␣balance')
#*(Faculty)**
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*       Faculty distribution
bal.plot(   (treat)    ~ Repeat+
              Gender_Desc+
            Ethnic_Group_Desc+
             Authority+ Faculty+
              Home_Language_Instr+
              Math_Grade12+
            yrsbefore
      , data = mydata, #
      weights = data.frame(
                    Greedy = get.w(m.nnc), #
                       IPW = get.w(weights.out) ),
        method = c("matching", "weighting"),
      var.name = "Faculty", which = "both") +
                 scale_fill_grey()   +
   ggtitle('Faculty␣distributional␣balance')
#*(Sex)**
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*              Sex distribution
bal.plot(   (treat)    ~ Repeat+
             Gender_Desc+
        Ethnic_Group_Desc+
         Authority+ Faculty+
            Home_Language_Instr+
            Math_Grade12+
            yrsbefore
       , data = mydata,
#bal.plot(treat ~ new.names, data = data_clean,
          weights = data.frame(Greedy = get.w(m.nnc),
                                ###
                       IPW = get.w(weights.out) ),
```

```
         method = c("matching", "weighting"),
             var.name = "Gender_Desc", which = "both") +
scale_fill_grey()   +
 ggtitle('Sex␣distributional␣balance')
#*(Repeat)**
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*          Repeat distribution
bal.plot(   (treat)    ~ Repeat+
              Gender_Desc+
        Ethnic_Group_Desc+
         Authority+ Faculty+
           Home_Language_Instr+
              Math_Grade12+
           yrsbefore
, data = mydata, #
    weights = data.frame(
                 Greedy = get.w(m.nnc),
                    IPW = get.w(weights.out) ),
            method = c("matching", "weighting"),
          var.name = "Repeat", which = "both") +
                 scale_fill_grey()   +
# +    xlabel(c("No","Yes"))+
ggtitle('Repeats␣distributional␣balance')
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   optimal ( is 'lonely")*   ...

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*         Sex
bal.plot(m.om, "Gender_Desc", which = "both",
        which.treat = c("prior", "posterior"),
sample.names = c("Unadjusted",
                 "Optimal␣Matching"))+
   scale_fill_grey()+
ggtitle('Optimal␣␣match:␣Sex␣distributional␣balance')
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*          Race
bal.plot(m.om, "Ethnic_Group_Desc", which = "both",
       which.treat = c("prior", "posterior"),
sample.names = c("Unadjusted",
                 "Optimal␣Matching"))+
    scale_fill_grey()+
ggtitle('Optimal␣␣match:␣Race␣distributional␣balance')
```

```
#####    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx    **
#*       Authority
bal.plot(m.om, "Authority", which = "both",
       which.treat = c("prior", "posterior"),
sample.names = c("Unadjusted",
                 "Optimal␣Matching"))+
   scale_fill_grey()+
ggtitle('Optimal␣␣␣match:␣Exam␣Board␣distributional
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣balance')
#####    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx    **
#*       Reapeats
bal.plot(m.om, "Repeat", which = "both",
       which.treat = c("prior", "posterior"),
sample.names = c("Unadjusted",
                 "Optimal␣Matching"))+
             scale_fill_grey()+
ggtitle('Optimal␣␣␣match:␣Repeats␣distributional
#␣␣␣␣␣␣␣balance')
##
##   #Faculty
#####    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx    **
#*       Faculty
bal.plot(m.om, "Faculty", which = "both",
      which.treat = c("prior", "posterior"),
sample.names = c("Unadjusted",
                  "Optimal␣Matching"))+
   scale_fill_grey()+
ggtitle('Optimal␣␣␣match:␣Faculty
distributional␣balance') ##
####(Math grades)**   ...
#####    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx    **
#*         Matric math grade*
bal.plot(m.om, "Math_Grade12", which = "both",
       which.treat = c("prior", "posterior"),
      sample.names = c("Unmatched", "Matched"))+
                  scale_fill_grey()+
ggtitle('Optimal␣␣␣match:␣Matric␣math␣distributional
␣␣␣␣␣␣␣␣balance')
#####    xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx    **
#*  Years prior attempting the module (from matric)*
bal.plot(m.om, "yrsbefore", which = "both",
        which.treat = c("prior", "posterior")
   , sample.names = c("Unmatched", "Matched"))+
```

```
                  scale_fill_grey()+
ggtitle('Optimal␣␣match:␣Years␣prior␣distributional
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣balance')
#$%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%$  **

#++++++++++++++++++++++++++++++++++++++++++++++++++++++  ++
##---                                                 ---**
#*   **  IV. Outcome model(step)**
##---                                                 ---**
#++++++++++++++++++++++++++++++++++++++++++++++++++++++  ++

#***(4)*
##--------------------------------------------------**
#* continuous outcome (hier)**...
##--------------------------------------------------**

###**   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*    get causal risk difference (greedy matching)
#*   (greedy)  outcome model(s): two sample t-test
with(data_cleanComplete,
       t.test(Final_Mark ~ (Term== 2017)))
with(data_clean,
      t.test(Final_Mark ~ (Term== 2017)))

ttest.Unadj<- with(data_cleanComplete,
    t.test(Final_Mark ~ (Term== 2017)))
               ttest.Unadj
with(data_clean,
     t.test(Final_Mark ~ (Term== 2017)))#
ttestExam.Unadj<- with(data_cleanComplete,
     t.test(Exam_Mark ~ (treat)))#
            ttestExam.Unadj
with(data_clean,t.test(Exam_Mark ~ (treat)  ))#
#*(... )**
#*(OR)**   ...   Or we can use OLS with/out covariates:
linmod.Unadj <- lm(Final_Mark ~ (treat),#(Term== 2017),
                   data = data_cleanComplete)
summary(linmod.Unadj)
linmod.Unadj$coefficients;confint(linmod.Unadj)#
#*(... )**
linmod.UnadjEx<- lm(Exam~ (Term== 2017),#*(treat),
                   data = data_cleanComplete)
summary(linmod.UnadjEx)
```

```
linmod.UnadjEx$coefficients;confint(linmod.UnadjEx)#*
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*      est. treatment effects (outcome model)
#**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*get causal risk difference (greedy matching)_2*...
#*    estimating treatment effects for
match.nncMod ttest.Greedy<- with(matchGreedy.data,
              t.test(Final_Mark ~ (Term== 2017)))
ttest.Greedy##
greedy_ttest_e1<- with(clicker_nnc.data,#*
                     t.test(Exam_Mark~ (treat)))
greedy_ttest_e1##

#*(OR)*Or we can use OLS with/out covariates:
linmod.Greedy <- lm(Final_Mark ~ (Term== 2017),
                    data = matchGreedy.data)
summary(linmod.Greedy)
linmod.Greedy$coefficients;confint(linmod.Greedy)
#*
#** linmod.GreedyEx <- lm(
             Exam_Mark ~ (treat),#(Term== 2014),
                       data =clicker_nnc.data
summary(linmod.GreedyEx)

linmod.GreedyEx$coefficients;confint(linmod.GreedyEx)
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#* get causal risk difference (optimal matching)* ...
#*     (opt) estimating treatment (with a t-test)
ttest.Optimal<- with(matchOpt.data,
        t.test(Final_Mark ~ (Term== 2017)))
#* we'll fail to reject hier!!! ttest.Optimal

ttest.OptimalEx<- with(matchOpt_data,
                       #matchOpt.data,
     t.test(Exam ~ (treat) ))#*(Term== 2017)))
#* we'll reject (the null) hier!!!
#      {& conclude existence of a (+ve) impact}
ttest.OptimalEx#*
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*   get causal risk difference (optimal matching)
linmod.Optimal <- lm(
```

```
            formula = (   (Final_Mark)   )
                          ~ (Term) == 2017,
                    data    = matchOpt.data)
summary(linmod.Optimal) #
linmod.Optimal$coefficients#
confint(linmod.Optimal)#

#**
linmod.OptimalEx <- lm(
        formula = ( (Exam_Mark) ) ~ (treat)
                  ,#(Term) == 2017,
                  data    = matchOpt.data)
summary(linmod.OptimalEx) ###

linmod.OptimalEx$coefficients ###
confint(linmod.OptimalEx) ##
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*  get causal risk difference (IPTW matching)*  ...
#*             (ipw)  outcome model(s)
ttest.IPTW<- with(weighteddata,
        t.test(Final_Mark ~ (Term== 2017)))
#* we'll fail to reject hier!!!
ttest.IPTW; confint(ttest.IPTW)
#* ttest.IPTWEx<- with(weighteddata,
            t.test(Exam_Mark ~ (Term== 2017)))
#* we'll fail to reject hier!!!
ttest.IPTWEx; confint(ttest.IPTWEx)

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
####             get causal risk difference (ipw)
##### glm.obj<-glm(   (Finale)   ~   treat,
#            weights = weight,
#             family = quasibinomial(link="identity"))
#*Error in eval(family$initialize) :
#y values must be 0 <= y <= 1
lm.obj<- lm(   (Finale)   ~ Term==2017,
                  weights = weight,
                          data = mydata )
lm.obj#
#*
lm.objex<- lm(   (Exam)   ~ Term==2017, weights = weight,
               data = mydata )
lm.objex#
```

```
summary(lm.objex);confint(lm.objex)
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
###*                  (ipw) weights model
weightmodel<- ipwpoint(exposure = treat,#*(treat)**,
                         family = "binomial",
                           link ="logit",
                       denominator = ~ Repeat+
                                 Gender_Desc+
                             Ethnic_Group_Desc+
                                     Authority+
                                       Faculty+
                            Home_Language_Instr+
                                   Math_Grade12+
                                      yrsbefore
                            , data= mydata)
summary(weightmodel$ipw.weights)

 #*
sd(weightmodel$ipw.weights)
#*
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
###*..y.       plot of weights
ipwplot(weights = weightmodel$ipw.weights,
       logscale = FALSE,
   main = "inverse␣prob␣weights␣(trunc)",
   xlim = c(0, 52), xlab = 'ip␣weights')
#**
mydata$wt<-weightmodel$ipw.weights
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*.y.          original weights
ipwplot( weights = weight,

         logscale = FALSE,
            main = "inverse␣prob␣weights␣(orig)",
            xlim = c(0, 52), xlab = 'ip␣weights')
####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*y.the split in weights (per treatmnt)* (truncated)
mydata$treat<- as.factor(treat)
ggplot(mydata, aes(x = wt, fill = (treat)   )) +
  geom_density(alpha = 0.5,
               colour = "grey50") +
  geom_rug() +
  scale_x_log10(breaks = c(1, 5, 10, 20, 40)) +
```

```
ggtitle("Distribution␣of␣inverse
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣probability␣weights␣(trunc)")+#
#xlabel = 'ip weights'
  scale_fill_grey()+
      theme_classic()
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*.y.the split in weights (per treatmnt)* (originals)
mydata$weight<- weight
ggplot(mydata, aes(x = weight, fill = (treat))) +
    geom_density(alpha = 0.5,
                 colour = "grey50") +
    geom_rug() +
     scale_x_log10(
             breaks = c(1, 5, 10, 20, 40)) +
ggtitle("Distribution␣of␣inverse
␣␣␣␣␣␣␣␣␣␣probability␣weights␣(pre-trunc)")+
#ggtitle("Distribution of inverse
#probability weights (orig)")+#
        #xlabel = 'ip weights'
                  scale_fill_grey()+
       theme_classic()
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*z.fit a marginal structural model(risk difference)
msm <- (svyglm(Finale ~ treat,
               design = svydesign(~ 1,
                                  weights = ~wt,
                                  data = mydata)))
#*msm;
summary(msm)
#**
coef(msm) ; confint(msm)
#*msm
summary(msm)
#*
confint(msm); smsm$coefficients
#***
msm.ex <- (svyglm(Exam ~ treat,
                  design = svydesign(~ 1,
                           weights = ~wt,
                              data = mydata)))
#**msm;
summary(msm.ex)
#*
```

```
coef(msm.ex); sconfint(msm.ex)
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#* ..xz.  (try) showin all these results together

xwhy<-   (with(matchGreedy.data,
             t.test(Final_Mark ~ (Term== 2017)))  )
xwhy$statistic;xwhy$estimate#*
#*mean in group FALSE  mean in group TRUE
#           60.32168            60.82448
#*( 60.39538  60.8026)** xwhy$conf.int
xwhy$conf.int #*
xwhy$alternative;xwhy$parameter;xwhy$p.value#
xwhy$data.name;xwhy$stderr
#*
xyzee<-   (   with(matchGreedy.data,
              t.test(Exam_Mark ~ (treat))
                  #*(Term== 2017)
                    )
          )
xyzee$statistic;xyzee$estimate;xyzee$conf.int
xyzee$null.value;xyzee$alternative;
xyzee$parameter;xyzee$p.value;xyzee$data.name
#*
xyzee$stderr#
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#* .xxx.  all results together (Final mark outcome)
resultLinearModels<- list(
#Unmatched = (xwhy$statistic  )
            Unadjusted= linmod.Unadj$coefficients,
     GreedyMatching  = linmod.Greedy$coefficients,
     OptimalMatching = linmod.Optimal$coefficients,
         IPWMatching = (msm$coefficients)
#ipwCI= confint(msm),
                          )
print(resultLinearModels, quote = FALSE)#*(... )**
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#* ..xxy.  all results together (Exam mark outcome)
resultLinearModels.Ex<- list(
          #Unmatched = (xwhy$statistic  )
           Unadjusted=linmod.UnadjEx$coefficients#
                           ,#
        GreedyMatch  = linmod.GreedyEx$coefficients,
        OptimalMatch = linmod.OptimalEx$coefficients,
```

```
            ipwMatch = (msm.ex$coefficients    )
 #ipwCI= confint(msm),
                                )
print(resultLinearModels.Ex, quote = FALSE)#*(...)**

#95% CI's   ...
#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*  ..xx.z.  all CIs together (Final mark outcome)
resultCIs<- list(#Unmatched = (xwhy$statistic  )
          UnadjustedCI=confint(linmod.Unadj),#
                 ,   #
            GreedyCI  = confint(linmod.Greedy),
            OptimalCI = confint(linmod.Optimal),
               IPTWCI = confint(msm) )

print(resultCIs, quote = FALSE)
#*
#* #####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
#*   ..xx.xy.  all results together (Exam mark outcome)
resultExCIs<- list(#Unmatched = (xwhy$statistic  )
UnadjustedCI=confint(linmod.UnadjEx),#  ,#   ##  ,
                #   ##
                GreedyCI  = confint(linmod.GreedyEx),
                OptimalCI = confint(linmod.OptimalEx),
                   IPTWCI = confint(msm.ex) )

print(resultExCIs, quote = FALSE)
#++++++++++++++++++++++++++++++++++++++++++++++++++++++++++  ++
##---                                                    ---**
#*    V. Sensitivity analysis(i.e. 'hidden-bias' step)*
##---                                                    ---**
#++++++++++++++++++++++++++++++++++++++++++++++++++++++++++  ++

#***(5)*
##-------------------------------------------------**
#*   *heterogeinity case***...   ...
##-------------------------------------------------**

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
###*            *unadjusted (or unmatched) case*...
#*Rubin Rule: result must ideally be near 0(post-
# matching), certainly in the interval(-50,+50)...
unadjRubin<- with(data_clean,
```

```
            abs(100*(mean(linps[Term=="2017"])
                                -
                        mean(linps[Term=="2014"]))
                                /
                            sd(linps)))
unadjRubin

unadjRubin1<- with(data_clean,
  var(linps[Term=="2017"])/var(linps[Term==2014]))
unadjRubin1#*(  > unadjRubin1   [1] 4.884034)**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
####              *greedy with caliper*  ...
greedy.caliperRubin<-with((matchGreedy.data),
                abs(100*(mean(linps[Term==2017])
                                    -
                            mean(linps[Term== 2014]))
                                    /
                                sd(linps)))
greedy.caliperRubin

greedy.caliperRubin1 <- with((matchGreedy.data),
  var(linps[Term==2017])/ var(linps[Term==2014]))

greedy.caliperRubin1#*([1] 1.000279)**Rubin Rule
#2: the result after matching must ideally be
#near 1 (& certainly bounded in (1/2,2).
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
####                *optimal matching   *...
optimalRubin <- with(   (matchOpt.data)
                                ,
      abs(100*(mean(linps[Term==2017])
                                -
               mean(linps[Term== 2014]))
                                /
                            sd(linps)))
optimalRubin
#*
optimalRubin1 <- with((matchOpt.data),
     var(linps[Term==2017])/var(linps[Term==2014]))
optimalRubin1
#*
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
```

```
###*                 *ipw matching   *...
ipwRubin <- with(   (weighteddata),
          abs(100*(mean(linps[Term== 2017])
                                    -
                   mean(linps[Term== 2014]))
                                    /
                             sd(linps)))
ipwRubin
###*
ipwRubin1 <- with((weighteddata),
      var(linps[Term==2017])/var(linps[Term==2014]))

ipwRubin1
#*
psens(match.nncF, Gamma = 2, GammaInc = 0.1)

psens(match.nncP, Gamma = 2, GammaInc = 0.1)#* #>    >
psens(match.nnc_bothF, Gamma= 2,
                  GammaInc = 0.05) #&&&&&&&*****   -----**

##  ######(PSM steps!)**    ...
##------------------------------------------------------------ (END!)**
```

3. DAGs vizualisations

```
##------------------------------------------------------ (START!)**
####()**   ...   iii. Causal graphing (DAGs)
#  ##########################################################**

#*            *  created by: Mxolisi Msibi, Mr.
#*            *organisation: University of Pretoria
#*            *       datae: 2019.11.25
#*            *      update: 2020.01.08
#*

#  ##########################################################**

##-------------------------------------------------------**
#*   1.  read psm text (in)**
##-------------------------------------------------------**

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------------**
#*   1.1. installin' required packages

install.packages( c(
#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%**
#1.Data manipulations
"sqldf",# do sql man (in R)*   ...
"janitor", "naniar",
"readxl",#get data from excel "sheets"
"tidyverse",#*(for 'pippin') and datamanipulations
'vctrs',  "broom",
"tidytext" #*need this for text minin'
'wordcloud'#*wordclouds (text minin')
'igraph'#*math graph-theory for DAGs
'ggraph'#*math graph-theory

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------------**
#*   1.2.  required libraries( off top)
library(tidyverse)#*should be the tidyverse
library(tidytext)#*need this for text minin'
library(readxl)# Super simple excel reader
library(ggplot2)#* graphics & eda
library(stringr)#string manipulator
```

```
library(wordcloud)#plot text wordclouds
library(tokenizers)#sentencing
library(igraph)#graph theory
library(ggraph)#plot the igraph object#*
require(lubridate)##*install.packages('janitor')
library(janitor)#*

#   ##*(DAGs!)**    ...
library(dagitty)
library(ggdag)#*install.packages(c('dagitty','ggdag'))
#*

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   1.3. sourcing available functions
source(
'C:\\Users\\mxmsibi\\Downloads/Biz.Related/
␣␣␣␣␣ms␣dissertation/R_functions/muliplot.R')

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   1.4. source the textfile

##-------------------------------------------------**
#*   2.  Directed acyclic graphs (graphics)**
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   2.1. unconfoundedness
dagified<- dagify(z ~ x,
                  y ~ z,
            exposure = "z",
             outcome = "y")

tidy_dag<- tidy_dagitty(dagified)
ggdag(tidy_dag, layout = "circle")
#*
#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   1.0. confoudedness
dagified1<- dagify(  z ~ x,
                     y ~ x,
```

```
                      y ~ z,
                    exposure = "z",
                     outcome = "y")

tidy_dag1<- tidy_dagitty(dagified1)
ggdag(tidy_dag1, layout = "circle")

dagify(z ~ x,
       y ~ x + z) %>%
ggplot(aes(  x = x,
             y = y,
          xend = xend,
          yend = yend)) +
       geom_dag_point() +
          geom_dag_edges_arc() +
                       geom_dag_text() +
theme_dag()
#   ##*(DAGs!)**    ...

##------------------------------------------------------------ (END!)**
```

4. Wordcloud vizualisations

```
##------------------------------------------------------ (START!)**
####()**   ...   iv. text mine (WordClouds)
#  ##########################################################**

#*                      *  created by: Mxolisi Msibi, Mr.
#*                      *organisation: University of Pretoria
#*                      *       datae: 2019.11.25
#*                      *      update: 2020.01.08
#*

#  ##########################################################**

##-------------------------------------------------------**
#*   1.  read psm text (in)**
##-------------------------------------------------------**

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------------**
#*   1.1. installin' required packages
install.packages( c(
#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%**
#1.Data manipulations
"sqldf",# do sql man (in R)*   ...
"janitor", "naniar",
"readxl",#get data from excel "sheets"
"tidyverse",#*( for 'pippin') datamanipulations
'vctrs', "broom",
"tidytext" #*need this for text minin'
'wordcloud'#*wordclouds (text minin')
'igraph'#*math graph-theory for DAGs
'ggraph'#*math graph-theory

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------------**
#*   1.2.  required libraries( off top)
library(tidyverse)#*should be the tidyverse
library(tidytext)#*need this for text minin'
library(readxl)# Super simple excel reader
library(ggplot2)#* graphics & eda
library(stringr)#string manipulator
library(wordcloud)#plot text wordclouds
```

```
library(tokenizers)#sentencing
library(igraph)#graph theory
library(ggraph)#plot the igraph object#*
require(lubridate)##*install.packages('janitor')
library(janitor)#*

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   1.3. sourcing available functions
source('C:\\Users\\mxmsibi\\Downloads/Biz.Related/
␣␣␣␣␣␣␣␣ms␣dissertation/R_functions/muliplot.R')

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   1.4. source the textfile
read.csv("psm_textfile__nonblanks.csv", header = TRUE)
psm_text<- read.csv(
       "psm_textfile__nonblanks␣Apr␣27,␣2020.csv",
                    header= TRUE)
#read.csv("psm_textfile__nonblanks.csv", header= TRUE)
head( psm_text)
##-------------------------------------------------**
#*   2.  data manipulations**
##-------------------------------------------------**

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   2.1.   read the (text)   file
######()  already done above   )
require(tidyverse)############
#########
#There were 26 warnings (use warnings() to see them)
warnings()#*

#####   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   2.2.  views first few rows head(psm_text,7)
head(psm_text) tail.matrix(psm_text)
head(   as.data.frame(psm_text)   )
psm_text1<- as.tibble(psm_text)
#  ########################   ################### **
library(lubridate)
# # date formattin' ##* (using lubridate::ymd_hms )
```

```
library(lubridate)
##-------------------------------------------------**
#*   3.  text mining**
##-------------------------------------------------**

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   3.1. we break the text (into individual tokens)

library(tidytext)#*tidytext needed (for stop_words)
#*

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   3.2. callin on the default stop words dataset ...
data(stop_words)  #*

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   3.3. tokenize the text
psm_text1$x<- as.character(psm_text1$x)
                  (psm_text1)%>%
         unnest_tokens( word, x )
         #*YESSSSSSSSSSSSSSSSSSS!!!!*...

(psm_text1)%>%
   unnest_tokens( word, x )%>%
#unnest_tokens works better nou(NB)*
          anti_join(stop_words)
#*exclude stop words(on the text!)*
psm_text2<- (psm_text1)%>%
   unnest_tokens( word, x )%>%
#unnest_tokens works better nou(NB)*
    anti_join(stop_words)
#*excl. stop words(on the text!)**
#*

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   3.4. Counts and plots
psm_text2 %>% count(word, sort = TRUE)#*

##-------------------------------------------------**
```

```
#*   4.  wordclouds**
##-------------------------------------------------**

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   4.1. lookin @ the most common words
#(in the dissertation)
library(wordcloud) #*
library(ggplot2)

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   4.2. set seed for reproducibility set.seed(1234)
# to always get the same WordCloud and
#         (for better reproducibility)

head(psm_text2)
tail(psm_text2)#*

#####*   xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx   **
##-------------------------------------------------**
#*   4.3. plotting (the wordcloud)
psm_text2 %>%
        anti_join(stop_words) %>%
                          count(word) %>%
     filter(n > 17) %>%#only those above 17 times
with(   wordcloud(word, n#min.freq=2,
#                 ,vfont=c("serif", "plain"))   )
#load("mxos_reviews&psm_Des2019.RData")###*()**...
#wordcloud(clean_words, min.freq=2,
#vfont=c("script", "plain"))

#  ######*(TextMinor!)**     ...

##------------------------------------------------------- (END!)**
```