
**Working Paper in Economics**

**# 202225**

# Causal Inference with Observational Data: A Tutorial on Propensity Score Analysis

**Kaori Narita**
**J.D. Tena**
**Claudio Detotto**

https://www.liverpool.ac.uk/management/people/economics/

# Causal Inference with Observational Data: A Tutorial on Propensity Score Analysis

Kaori Narita* J .D. Tena† Claudio Detotto‡

November 12, 2022

**Abstract**

When treatment cannot be manipulated, propensity score analysis provides a practical approach to making causal claims. However, it is still rarely utilised in leadership and applied psychology research. The purpose of this paper is threefold. First, it explains and discusses the application of the method with a particular focus on propensity score weighting. This approach is readily implementable since a weighted regression is available in most statistical software. Moreover, using a double robust estimator can offer protection against the misspecification of the model by including confounding variables both in the treatment and response equations. A second aim is to discuss how propensity score analysis has been conducted in recent management studies and examine future challenges. Finally, we illustrate the method by showing how it can be employed to estimate the causal impact of leadership succession on performance using data from Italian football. The case also exemplifies how to extend the standard single treatment analysis to estimate the separate impact of different managerial characteristic changes between the old and the new manager.

**Keywords:** causality, propensity score, leadership succession, observational data, football

**JEL Codes:** C31, J24, J63, M51, Z22

*Corresponding author. Email: K.Narita@liverpool.ac.uk. Address: University of Liverpool Management School, Liverpool, L69 7ZH, United Kingdom.

†Email: jtena@liverpool.ac.uk. Address: University of Liverpool Management School, Liverpool, L69 7ZH, United Kingdom.

‡Email: detotto_c@univ-corse.fr. Address: Centro Ricerche Economiche Nord Sud (CRENoS), Sassari, 07100, Italy

# 1 Introduction

Causal claims are present in most empirical research reported in the leadership literature. For example, analysts are interested in knowing the consequences of rewards (Fest et al., 2021), traits (Rockey et al., 2021; Kiss et al., 2021), emotions (Sy et al., 2018) or previous experience (Zhang et al., 2021; Hopp and Pruschak, 2020). However, while randomisation provides a failsafe way to provide causal evidence, this is not always possible in social science. In particular, it is challenging to operationalise complex constructs, such as leadership, in laboratory settings (Wofford, 1999) or in some cases to find situations in which key variables such as perceptions, choice, emotions or behaviours are manipulated in natural experiments. Therefore, non-experimental designs are sometimes presented as the only way to conduct research in social science. In this setting, propensity score analysis (PSA) allows for counterfactual comparisons under the strong ignorability assumption, which implies that conditional on observable variables, the potential outcomes are independent of treatment[1] status (Rosenbaum and Rubin, 1983). The application of PSA relies on the estimation of the probability of receiving treatment, or propensity score (PS). The two most common PSA approaches are propensity score matching (PSM) and propensity score weighting (PSW).[2] They differ in the way they transform the sample to be used in causal analysis. While PSM use PSs to form analogous treated and untreated observations, dropping non-matched observations, PSW uses all individuals in the original sample but weights them according to their PSs.

Despite the arguments for its use, PSA has been elusive in management research (Connelly et al., 2013; Schmidt and Pohler, 2018). This apparent absence of interest was highlighted in Li (2013, p. 209): "To my knowledge, no publications in the management field have implemented the PSM in an empirical setting, yet other social science fields have empirically applied the PSM", and Connelly et al. (2013, p. 416): "... most organizational researchers who conduct quasi-experiments are generally not familiar with propensity scoring and have not generally considered using this technique in their research." Li (2013) and Connelly et al. (2013) provide comprehensive and insightful introductions of these methods to management scholars. However, almost one decade later, PSA is still rarely used in either the management or applied psychology literature. In this respect, Schmidt and Pohler (2018) indicate that econometrics, or statistical methods developed/used in economics, have been somewhat separated from other social sciences and underutilisation of PSA is perhaps one example of such. They also attribute the unpopularity of PSA to the late arrival of statistical packages to deal with non-binary treatment variables, which have only recently become available. Even with the help

[1] The word "treatment" originates from medical trials, where a certain treatment is given to the treated group, and no treatment (or a placebo) is given to control or untreated group.

[2] Another approach known as propensity score stratification splits the treatment and control samples into similar groups according to the distribution of PSs (Thoemmes and Kim, 2011). This method keeps the essence of PSM as it uses PSs to match treated and untreated participants.

of statistical packages, understanding of sophisticated automatic algorithms and programming knowledge would still be required.

This paper supplements the previous tutorials in three ways. First, we explain practical issues associated with the application of PSA in management, whilst primarily focusing on the application of PSW. This method was initially proposed by Imbens (2000) and has been used in a variety of contexts, see Wooldridge (2010). PSW uses the inverse of the PS as a weight to apply to each treated unit and the inverse of one minus the PS as the weight to apply to each control unit (Imbens, 2000). Rather than relying on statistical packages with matching algorithms, the implementation of PSW only requires the application of weighted linear regression, which is readily available in most statistical software. However, despite its simplicity, most of the applications related to propensity scores are in matching (Thoemmes and Kim, 2011).

A second aim of the paper is to show and discuss research examples in the recent literature in management and applied psychology where PSA can be used. This represents another additional contribution compared to Li (2013) and Connelly et al. (2013) where these applications were not present yet. We also discuss previous studies that employ PSW with non-binary treatments.

Finally, our third purpose is to provide a practical example of how PSW can be used to study a leadership topic. In particular, we estimate the consequences of involuntary within-season managerial change in top-tier Italian football (*Serie A*) during seasons 2004/2005 - 2017/2018. Two aspects of this tutorial case are of special relevance for management researchers. First, the example is written as a guide to implementing a double robust procedure that uses both PSW and regression adjustment to mitigate bias due to observables (Funk et al., 2011). In the standard PSW procedure, the treatment effect can be estimated within the weighted regression framework, where the weights are based on the estimated PSs, in order to control for the pre-treatment differences between clubs which dismissed managers and those which did not. In the weighted regression model where outcome variable is regressed with treatment variable, additional factors that can affect the outcome can also be included. However, an important limitation of PSW is that it is very sensitive to misspecification of the PS model (Freedman and Berk, 2008; Stone and Tang, 2013). Moreover, PSW does not perform well with small samples (Raad et al., 2020). Thus, to account for these concerns, our tutorial example employs a double robust procedure that increases protection against model misspecification by including the determinants of PSs in the weighted regression (Funk et al., 2011).

A second relevant aspect of the tutorial is that it adapts the approach to deal with multidimensional treatment in PSW. In particular, we extend the analysis by considering leadership succession as simultaneous changes in the different dimensions of managerial characteristics. Fourteen main managerial characteristics are considered. They are related to age, experience, association with the organisation, most recent activity (employment) status, and background. Our analysis shows that a positive outcome is expected following

particular managerial characteristic changes. This highlights the importance of considering the different dimensions in which treatment is operationalised by management researchers.

This paper proceeds as follows. The following section explains the principles of PSA and how to conduct this type of research. Section 3 presents and discusses examples of the use of PSA in recent management and applied psychology research. Section 4 provides the illustrative case on the causes and consequences of head coach turnovers in Italian football. Some discussion on causal analysis in qualitative studies and the role of PSA to complement this approach is presented in Section 5. Finally, we offer ideas for future work and some concluding remarks.

# 2 Principles of propensity score analysis

## 2.1 The strong ignorability assumption

Causal analysis estimation would be straightforward in an ideal situation where we could observe the outcome of a subject $i$ when receiving the treatment, $Y_i(1)$, *and* not receiving the treatment, $Y_i(0)$. This causal effect for unit $i$ is defined as:

$$c_i = Y_i(1) - Y_i(0). \tag{1}$$

The challenge in identifying $c_i$ in social science stems from the fundamental missing data problem as we can only observe one response per unit (Holland, 1986). Thus, the observed outcome for individual $i$ becomes $T_iY_i(1)-(1-T_i)Y_i(0)$, where $T_i$ is a binary variable indicating treatment allocation. In this case, if treatment is randomly allocated, we at least can obtain an unbiased estimate of the treatment effect, averaged over the trial sample as treatment is unrelated to each person's attributes and, therefore, independent of the potential outcomes $(Y(1), Y(0))$ (Fisher, 1935).

Even when random treatment allocation is not possible, quasi-experimental designs allow for causal analysis by manipulating the treatment variable. These designs include, but are not limited to, simultaneous equation, regression discontinuity, difference in difference and selection (Antonakis et al., 2010). The description and discussion of these methods are out of the scope of the present tutorial, and the interested reader is referred to Antonakis et al. (2010) and Cook et al. (2002) among others.

Our goal is to introduce readers to the intuition and the assumptions of PSA in observational studies. Attention is focused on the endogeneity associated with treatment allocation. Other validity threats are discussed in Section 2.4.

To make causal analysis possible, Rosenbaum and Rubin (1983) pointed out the need to assume strong

ignorability, which requires the fulfilment of the following two conditions:

$$(Y(1), Y(0)) \perp T|X, \tag{2}$$

$$0 < \Pr[T = 1|X] < 1. \tag{3}$$

Expression (2) is the unconfoundedness assumption which states that potential outcomes (Y(1), Y(0)) are not affected by (or are independent of) treatment assignment, conditional on a set of observable variables ($X$). This property (also referred to as ignorability, conditional independence or selection on observables) is fundamental to the statistical estimation of causal effects. For this condition to be fulfilled, it is necessary to assume that there are no observable variables other than $X$ simultaneously affecting the treatment assignment and the outcome variable. However, it is not possible to test directly whether treatment assignment is "ignorable" (Guo and Fraser, 2014). Thus, researchers must identify the right covariates based on theoretical and empirical grounds.

Condition (3) is the overlap assumption. It means that every individual has a positive probability of being assigned to the treated and control group conditional on $X$. Under the strong ignorability assumption, even though randomisation is not possible, it is credible to remove pretreatment differences between the treated and the control subjects in a sort of virtual randomisation (Rosenbaum and Rubin, 1983). The following section explains how to apply this approach.

## 2.2 Steps in the analysis

The PS is the *ex-ante* probability of a treatment assignment conditional on a collection of observed baseline variables (Rosenbaum and Rubin, 1983), which is estimated via prediction models for treatment allocation. PSs can be used to identify individuals who are similar in terms of pre-treatment condition, but only differ in treatment assignment (treated or control). Based on this, different types of PSA can be applied to adjust a sample so that the covariates are more similar ("balanced") between the treated and control groups, as though the treatment were randomly allocated.

PSA typically comprises four steps, as illustrated in Figure 1. In the first step, a PS model is specified as a function of observed variables related to pre-treatment conditions. Probit and logit models are the usual approaches to estimate treatment probabilities (Caliendo and Kopeinig, 2008). A common practice is to assess the accuracy of the estimated PSs predictions using a receiver operating characteristics (ROC) curve and the area under the ROC curve (AUC). In a binary classification model, the ROC curve plots the true positive rate (true positives divided by true positives plus false negatives) against the false positive rate

(false positives divided by false positives plus true negatives).[3] The AUC measures how well the probabilistic model discriminates between treatment and control individuals, see DeFond et al. (2017) as an example of the use of this measure in management studies. A greater AUC indicates a better predictive performance. The precision-recall (PR) curve is an alternative to the ROC curve. It plots the precision (the number of true positives divided by the total number of true and false positives) versus the recall (the number of true positives divided by the total number of true positives and false negatives). PR curves are recommended for imbalanced data, where the distribution of classes is severely skewed towards one or the other. In this case, ROC curves may provide an excessively optimistic view of performance (Branco et al., 2017).

The second step differs between PSM and PSW, the two different "virtual randomisation" strategies. The former employs an algorithm to find pairs of individuals in the treatment and control groups with similar PSs. Several alternative algorithms can be used for this purpose. As indicated in Figure 1, PSM links $n$ individuals in the treatment group to their closest $m$ individuals in the control group according to their estimated PS. One of the most popular, nearest neighbour matching, finds one or more units with the closest PS within the control group for each treated individual (i.e. 1:1 or 1:m). The process is repeated until no observations are left in the treatment or control group. An alternative approach, optimal matching, uses the whole sample to determine matched observations with the smallest average within-pair absolute PS differences (Rosenbaum, 1989). Genetic matching considers not only PS but also specific covariates to determine the set of weights (Diamond and Sekhon, 2013). Figure 1 also indicates that these algorithms can differ in other dimensions. For example, a match could be with or without replacement depending on whether or not the controlled matched subject is reintroduced to the control group for next-round matching to be used again. Matching can be restricted to not surpass a maximum PS distance for matched pairs. This maximum distance is denoted as the caliper width. One can also consider a pair of 1:1 matching or any other n:m matching (where n and m are integers).

Matching algorithms become especially cumbersome in the case of multiple or continuous treatments. In the former case, it is still possible to estimate PSs using multinomial logit or probit models and make paired comparisons with a reference treatment group. For example, Hopp and Pruschak (2020) deal with this problem by separately estimating the effect of each treatment using PSM, and they explain that results are robust to a multinomial treatment estimation of PSs. A problem with this approach is that, because some observations are dropped during matching, each paired comparison may be based on different individuals. In the case of continuous treatment, Hirano and Imbens (2004) present a matching approach based on estimating the treatment dose rather than its PS.

Another potential issue with PSM is that it requires many individuals, especially in the control group.

[3] A true positive (negative) occurs when the model correctly predicts treatment (control).

Moreover, certain matching schemes may not use a large number of observations. Contrarily, PSW, in principle, retains all the observations (Guo and Fraser, 2014). A second advantage is its simplicity. In the PSW approach, a weight allocated to each individual is defined by the inverse of the estimated PS for a realised treatment status. Intuitively, a treated unit with a low probability of being treated is given a high weight, and a control unit with a high probability of being treated is also given a high weight. In doing so, the distribution of the *ex-ante* probabilities of being treated become similar across the treated and control groups, as though the treatment were assigned randomly. Therefore, the second step in PSW only involves obtaining these weights to be employed in a weighted regression, similar to the application of sample or survey weights, commonly used in social sciences. Furthermore, this approach can be relatively easily generalised to multi-treatment cases, as in Schmidt and Pohler (2018) and Love et al. (2017). While weighted regressions are readily implementable with most of the statistical software available, applying matching algorithms typically requires becoming familiar with specialist tools such as matchIt in R or psmatch2 in Stata.

The third step is common to PSM and PSW and consists of testing for balance in covariate distributions between the treatment and the control groups. The general idea of these checks is to compare differences between the treated and the control group before and after matching or weighting. Two common approaches are (1) the standardised bias, which assesses the distance in marginal distribution of the X variables, and (2) a two-sample t-test to check whether there are significant differences in covariate means for both groups (Rosenbaum and Rubin, 1985). If these tests are not completely successful, some remedial measures are advised, such as including interactions terms in the PS estimation (Caliendo and Kopeinig, 2008).

The final step in PSA consists of estimating the impact of treatment on the variable of interest. Although the pre-treatment conditions between treated and control groups are balanced through the previous step, it is customary to conduct this estimation in a multiple regression analysis. This regression has two purposes. First, it can be used as a double robust procedure where treatment determinants are included to further protect against the bias due to observables (Funk et al., 2011). Furthermore, it also allows controlling for additional factors that can potentially impact the response variable after treatment. As noted above, in case of PSW, the step follows weighted regression, which directly applies the weight defined in step 2.[4] In general, two main definitions of treatment effects are considered: the average treatment effect (ATE) and the average treatment effect on the treated (ATT). The decision on which of the two causal effects are estimated depends on the researcher's interest and the PSA method employed. For instance, consider a PSM design such that, for all treated individuals, the closest individual in the control group is matched. By averaging the differences in the outcomes of these two groups, we would estimate the ATT. However, by evaluating

[4]Weights obtained from PS estimates should not be confused with weights in survey sampling. While the former tries to address endogeneity in the treatment assignment, the latter is intended to adjust the sample data to reflect population attributes.

the impact of treatment on the whole weighted sample, PSW provides an estimate of ATE.

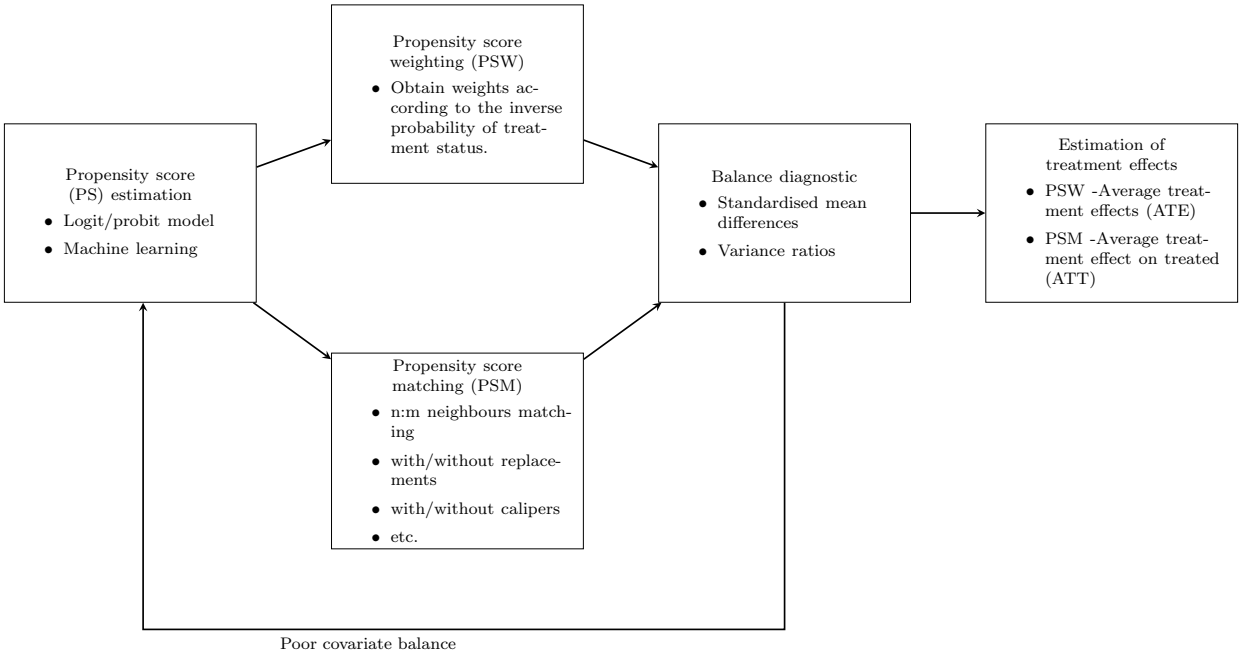


Figure 1: Steps in propensity score analysis

## 2.3 A simulation example

The approach described in the previous section only provides reliable causal estimates under strong restrictions. This section considers a simple simulation to illustrate better the importance of the strong ignorability assumption in causal analysis under PSW. We assume that the outcome variable depends on the impact of treatment $c_i$ and an idiosyncratic component $\gamma_i$ according to the following function:

$$Y_i = c_i T_i + \gamma_i, \tag{4}$$

However, rather than imposing the same treatment effect on all individuals in the sample, we assume that the sample population is divided into two groups. We also assume that treatment has a more favourable impact for the first than for the second group. For example, Connelly et al. (2013) studied the effect of test coaching on SAT scores. They explain that low performing students are more likely to benefit from test coaching (and therefore more likely to seek out it) than high performing ones. Similarly, Schmidt and Pohler (2018) note that observed employee satisfaction affects the level of interest in high-performance work systems investments. The assignment of treatment is not random, hence the characteristics that determine the treatment assignment has to be taken into account for the unconfoundedness assumption to be satisfied.

In our simulation, for a sample of size $N$, there is an equal number of individuals ($N/2$) in each group. We also assign values 10 and -1 to $c_i$ for groups 1 and 2, respectively. The idiosyncratic parameter is assumed to follow a normal distribution with zero mean and unit variance. According to these figures, the true ATE is 4.5. However, we simulate the model under three different scenarios to evaluate the importance of the unconfoundedness assumption. The first one assigns treatment with a probability of 0.5 to each individual in the sample. The second scenario assumes that, because treatment is more appealing to the first group, individuals in the first and second groups choose treatment with probabilities of 0.8 and 0.2, respectively. Finally, the third scenario considers the same probabilities as the previous one but weights treated and untreated outcomes by the inverse of their respective PSs. Regardless of the scenario, ATE is computed as the difference between the average outcome of those who receive treatment and control.

Table 1.A. shows the median and 95% intervals of the 10,000 ATE estimates in each of the three scenarios described above. It stands out that, if no correction is made, a nonrandom treatment allocation generates a biased ATE estimate (scenario 2). This is because, in this case, we are not taking into account that treatment is endogenous, and the first group prefers it. However, the ATE estimate is close to the real one under random allocation. Moreover, the estimation precision increases with the number of observations. When we turn our attention to the third scenario, we observe similar results to those obtained under random allocation. However, we must highlight that we get this result under two strong assumptions. First, we know treatment probabilities in each case, ignoring how confounders have been used to obtain them. The second assumption is that, to apply the correction, it is necessary to impose the overlap assumption, i.e. each individual has a strictly positive probability of being treated and not treated.

Table 1: Estimation of average treatment effects (ATE) using Monte Carlo simulation. (True ATE = 4.5)

**A. ATE estimates with known treatment assignment probabilities**

| | **Randomised assignment** | | | | **Endogenous Assignment** | | | | **Inverse PSW** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Observations | Median ATE | MCE$^{(I)}$ | %Bias$^{(II)}$ | 95% Coverage$^{(III)}$ | Median ATE | MCE$^{(I)}$ | %Bias$^{(II)}$ | 95% Coverage$^{(III)}$ | Median ATE | MCE$^{(I)}$ | %Bias$^{(II)}$ | 95% Coverage$^{(III)}$ |
| $N = 100$ | 4.489 | 0.602 | -0.299 | 0.950 | 7.824 | 0.551 | 73.964 | 0 | 4.545 | 0.874 | 2.600 | 0.946 |
| $N = 1,000$ | 4.498 | 0.185 | -0.048 | 0.950 | 7.802 | 0.173 | 73.415 | 0 | 4.504 | 0.265 | 0.293 | 0.951 |
| $N = 10,000$ | 4.501 | 0.059 | 0.005 | 0.951 | 7.800 | 0.056 | 73.324 | 0 | 4.500 | 0.085 | 0.005 | 0.951 |

*Notes:* The number of Monte Carlo simulations is 10,000 in all cases. Randomised assignment treats each individual with a probability of 0.5. Endogenous assignment treats individuals with probabilities of 0.8 and 0.2 in groups 1 and 2. Inverse propensity score weights outcomes of treated and untreated individuals with the inverse of their respective probabilities. (I) $MCE = \sqrt{Var(\widehat{ATE}_R)}$; (II) $\%Bias = \frac{1}{R}\sum_{r=1}^{R} \frac{\widehat{ATE}_R - ATE}{ATE} * 100$; (III) $coverage = \frac{1}{R}\sum_{r=1}^{R} I[\widehat{ATE}_R - 1.96\widehat{se}(\widehat{ATE}_R) \leq ATE \leq \widehat{ATE}_R + 1.96\widehat{se}(\widehat{ATE}_R)]$.

**B. ATE estimates with estimated treatment assignment probabilities**

| | **Relevant covariates** | | | | **Non informative covariates** | | | | **All covariates** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Observations | Median ATE | MCE$^{(I)}$ | %Bias$^{(II)}$ | 95% Coverage$^{(III)}$ | Median ATE | MCE$^{(I)}$ | %Bias$^{(II)}$ | 95% Coverage$^{(III)}$ | Median ATE | MCE$^{(I)}$ | %Bias$^{(II)}$ | 95% Coverage$^{(III)}$ |
| $N = 100$ | 4.668 | 0.830 | 3.793 | 0.939 | 7.822 | 0.567 | 73.949 | 0 | 4.703 | 0.963 | 4.272 | 0.938 |
| $N = 1,000$ | 4.514 | 0.259 | 0.145 | 0.948 | 7.804 | 0.174 | 73.433 | 0 | 4.526 | 0.269 | 0.248 | 0.951 |
| $N = 10,000$ | 4.503 | 0.080 | 0.021 | 0.951 | 7.800 | 0.055 | 73.336 | 0 | 4.503 | 0.081 | 0.021 | 0.945 |
| | **Stepwise selection** | | | | **Lasso selection** | | | | | | | |
| Observations | Median ATE | MCE$^{(I)}$ | %Bias$^{(II)}$ | 95% Coverage$^{(III)}$ | Median ATE | MCE$^{(I)}$ | %Bias$^{(II)}$ | 95% Coverage$^{(III)}$ | | | | |
| $N = 100$ | 4.688 | 0.910 | 4.000 | 0.939 | 5.415 | 0.639 | 20.862 | 0.719 | | | | |
| $N = 1,000$ | 4.525 | 0.267 | 0.249 | 0.952 | 4.734 | 0.252 | 4.992 | 0.858 | | | | |
| $N = 10,000$ | 4.503 | 0.081 | 0.021 | 0.946 | 4.570 | 0.082 | 1.542 | 0.864 | | | | |

*Notes:* The number of Monte Carlo simulations is 10,000 in all cases. PSs are estimated using logit models. Relevant covariates: estimate PS including group information and treatment+noise. Stepwise and lasso selection estimate PSs using group information, treatment + noise and three additional noise variables. All covariates estimate PSs, including the five covariates. Noise is always generated from standardised normal processes. (I) $MCE = \sqrt{Var(\widehat{ATE}_R)}$; (II) $\%Bias = \frac{1}{R}\sum_{r=1}^{R} \frac{\widehat{ATE}_R - ATE}{ATE} * 100$; (III) $coverage = \frac{1}{R}\sum_{r=1}^{R} I[\widehat{ATE}_R - 1.96\widehat{se}(\widehat{ATE}_R) \leq ATE \leq \widehat{ATE}_R + 1.96\widehat{se}(\widehat{ATE}_R)]$.

In the second set of simulations, we address the issue of how confounders can be used to estimate treatment probabilities under different strategies. In particular, we assume that the analyst can observe five variables. The first two variables are: (1) the group and (2) the treatment decision contaminated with noise (a standardised normal variable). Additionally, the analyst observes three standardised normal variables that do not report any information about the probability of treatment. Treatment probabilities are estimated with logit models under four different strategies. The first one is a model that only includes the two informative variables. The second and third models consider the five confounders and select them according to a stepwise regression based on AIC and Lasso approach. The final strategy includes the five variables in the model. The estimated values under each method are used to weight treated and untreated outcomes by the inverse of their respective probabilities.

Table 1.B. shows the results of this simulation exercise. It can be noted that estimating PSs reduce the precision of the ATEs estimates in all the approaches. However, if the model includes the relevant variables, final estimates converge to the real ones when $N$ increases. When we compare the two model selection strategies, the stepwise based on AIC outperforms the lasso procedure. Of course, we cannot draw firm conclusions from this simple exercise. However, simulations in the previous literature also highlight the importance of over-specifying the propensity score (Millimet and Tchernis, 2009; Millimet et al., 2010). Overall, these results indicate that PS estimation negatively affects the estimation precision. However, when $N$ increases, if the model includes the true determinants of treatment, the estimated ATE will converge to the real one.

## 2.4 Other validity concerns

In the previous sections, we have argued that PSA can remove endogeneity associated with treatment allocation if we assume strong ignorability. However, other validity concerns may remain in the analysis even if this assumption is satisfied. A definition of the most prominent causal threats can be found, for example, in Cook and Campbell (1976) and Crano et al. (2014). Podsakoff and Podsakoff (2019) summarise the main validity threats and provide illustrations from the leadership literature. These concerns can be split into internal and external validity threats. Internal validity requires correctly attributing differences in the dependent variable to treatment variations. That is, Podsakoff and Podsakoff (2019) identifies potential validity threats due to selection, history, maturation, testing, instrumentation, regression, mortality and selection by maturation interactions.

Depending on the characteristics of the observational sample, some internal validity threats can be particularly relevant in PSA. In particular, the history threat is a consequence of the external events affecting

individuals over time between the impact of the treatment and the instant when the dependent variable is observed. Unlike history, maturation is related to external events but to the way individuals evolve over time. For example, they may become older, more tired or less motivated than at the time of treatment. History and maturation threats increase the larger the length of time between the treatment and the measurement of the response variable(s) (Podsakoff and Podsakoff, 2019). For example, testing the long term consequences of educational decisions (Hopp and Pruschak, 2020) or previous military experiences (Zhang et al., 2021) requires dealing with history and maturation threats. However, looking at short-term reactions of the dependent variable(s) could be also problematic if the analysis involves situations where the treatment requires some time before having its effect. For instance, some time is needed to assess the final impact of training on workers' productivity. Therefore, coping with this problem requires estimating the sensitivity of estimates to using different periods and controlling for all the possible factors affecting the dependent variable. These issues are particularly worrying when there is an interaction of selection with history and maturation. Another example of an internal validity concern often present in observational samples is attrition. The Heckman two-step model can be used to mitigate this threat (Heckman et al., 1999).

However, even where these internal validity conditions are fulfilled, a fundamental research question is the generalisability of causal effects to other settings. In particular, two main external validity concerns are: (1) generalisability of operationalisations and (2) generalisability of results to other places and participant populations. Validity of operationalisations concerns the correct identification of the treatment and response variables and the underlying relationship between them. A "treatment" could have many different meanings. In PSA, this concern requires estimating the different impacts of different treatment intensities or subgroups in observational samples. Section 3 shows examples of multi-treatment situations. For example, Boivie et al. (2016) and Hopp and Pruschak (2020) show how to conduct such analysis using PSM while Schmidt and Pohler (2018), Love et al. (2017) and the tutorial case in Section 4 employ a PSW design.

Generalisability can also be an issue under these designs as results could depend on the specific sample used in the analysis. As in experiments, a way to overcome this problem is to repeat the estimation analysis with different samples (Li et al., 2021) or combine experimental and non-experimental designs (Carton et al., 2014).

# 3 PSA in management and psychology research

Propensity scoring is still rarely used in the management and psychology literature. To illustrate this issue, we explored the same top-tier journals surveyed by Antonakis et al. (2010) in their review of causal analysis: Academy of Management Journal, Journal of Applied Psychology, Journal of Management, Journal

of Organizational Behavior, The Leadership Quarterly, Organizational Behavior & Human Decision Processes and Personnel Psychology. We add The Strategic Management Journal to this search as it is an FT50 journal that contains some examples of PSA in management.

Initially, for the purpose of comparison, we considered 4,330 abstracts in these journals from 2015 to 2022[5]. In this search, the term "propensity score" appeared in 8 abstracts. Additionally, we accounted for the possibility that papers may have employed PS in causal analysis without necessarily using the term "propensity score" in the abstracts. Consistent with this possibility, we found 40 instances where the word "matching" appeared without "propensity score". However, in these cases, only three papers had conducted PSA. This amounts to a total of 11 papers (0.25 percent) that refer to PSA in the abstract compared, for example, to 47 and 70 for "laboratory experiments" and "field experiments" respectively. In this group of 11 papers we could only find three examples of PSW studies. However, only two of them use PSW in their core analysis because Rocha and Van Praag (2020) employ PSW as one of three alternative methods to deal with endogeneity in a robustness exercise.

To identify and discuss more specific examples of PSW in the extant management literature, in addition to the previous search, we account for the possibility that papers could still employ PSA without referring to it in the abstract. Thus, first, we searched the term "propensity score" in the text of the 4,330 articles.[6] Then, in a second step, we visually inspected the selected cases to identify 25 additional studies that conduct PSA as part of the main econometric analysis. However, an important issue in this search is that PSM (rather than PSW) is becoming the more popular approach, with 21 (out of 25) papers. Table 2 summarises the main characteristics of the 9 PSA studies from the abstract search while we have selected the 6 (2+4) examples of the use of PSW in causal analysis for a more detailed discussion in the following.

[5]The search took place on 22/03/2022. We explored all publications from Scopus between 2015 and 2022 after removing editorials (93), errata (67) and one retraction. The terms "propensity score" and "matching" were employed to identify potential PS studies.

[6]Our search took place on Google Scholar on 09/04/2022. We explored the presence of the term "propensity score" within the document (abstract, main text and references) for all publications between 2015 and 2022 in the selected journals. By entering these search criteria in the Google Scholar database, a total of 79 papers were recorded and downloaded: 26 from the Academy of Management Journal, 6 from the Journal of Applied Psychology, 21 from the Journal of Management, 1 from the Journal of Organizational Behavior, 23 from the Leadership Quarterly, and 2 from the Organizational Behavior and Human Decision Processes.

Table 2: Examples of PSA and matching for causal studies in the management and leadership literature

| Article | Background | Methodology | Strengths | Limitations |
|---|---|---|---|---|
| Chen (2015) | Chen (2015) estimates the implications of the initial compensation of CEOs hired in turnaround situations on their subsequent initiatives. | They employ PSM to match each CEO hired in turnaround situations with a CEO hired in non-turnaround situations in firms with similar characteristics. | They test whether results are robust to a different event window and compensation type. They also tested that their results are not general but specific to CEOs hired in turnaround situations. | Controlling for the potential selection bias in the selection of CEOs. - Impossibility of controlling for the successor's prior pay and industry growth. |
| Boivie et al. (2016) | They estimate the effect of serving on boards on different aspects of executives' professional careers. | PSM to get 1,052 directorships and 1,052 counterfactual executives without directorships. | They explore different types of response variables (promotions) in different models. | Board appointments and subsequent promotions could reflect two stages of the promotion process. |
| Bechtoldt et al. (2019) | They tested whether women are more likely to access leadership roles in precarious circumstances. They also estimate differences in shareholders' reactions to appointed women compared to men. | Using PSM, they match each firm that appoints women to their management boards to a sample of companies that promote men. They obtain a final sample of 42 men and 42 women. | They consider an alternative approach based on instrumental variables. They test their hypotheses in two studies. | The possibility of attrition as low performing firms can drop during the analysis period. Other characteristics such as age, religion or cultural aspects could explain the different reactions. |
| Gupta et al. (2020) | They study the influence of CFO gender on financial misreporting and how governance mechanisms moderate this effect. | PSM to create 1,545 comparable samples of firms with male and female CFOs. | Results are robust to different econometric specifications. | The response variable does not directly measure CFO ethical attitudes. The estimated gender effects could be associated with different attributes. |
| Rocha and Van Praag (2020) | The authors estimate how the gender of founders can determine future entrepreneurial career choices of their male and female joiners. | In a robustness exercise they use PSW to account for selective matching based on gender. | They show the estimation results are robust to different subsamples and estimation methods. They compare the influence of same-gender on other social interactions. | The authors acknowledge the impossibility of inferring the motivations of both joiners and founders driving their match. |
| Li et al. (2021) | They explore how the transition from employee to leader fosters growth in contentiousness and emotional stability. | PSM is used to match each individual who became a leader with two non-leader individuals. | The consideration of two different databases confers external validity to the causal analysis. | The subjective number of personality dimensions. Self-report measures of personality. Maturation. |
| Vitanova (2021) | She studies the impact of overconfidence on performance. | She uses PS to match overconfident leaders with a control group. The final sample contains 793 treated and 793 control observations. | The use of a longitudinal sample allows for addressing potential reverse causality problems. | Binary treatment. The possible presence of unobserved differences in the treatment and control groups. |
| Hopp and Pruschak (2020) | They estimate the effect of having had specific leadership roles at high school on earnings and several individual characteristics. | They consider three probit models to assess the probability of being president and captain, captain only and president only. Using these probabilities, they match each of the three cases with the control group. | They use a regression with instrumental variables to tackle the problem of unobserved omitted variables. | Maturation. Generalisability of findings to other samples. |

| | | | | |
|---|---|---|---|---|
| Ong (2021) | The author tests whether women experience loneliness and less authenticity than men when occupying a leadership role. | The author matches individuals according to the estimated probability of becoming a leader. He gets a final sample of 204 observations. | Ong (2021) tests his hypothesis in three different studies. One of them is a randomised control trial. | The sample used in the PSM only includes participants who have completed responses (potential selection bias). - Bias due to unobserved characteristics associated with gender. |

*Note:* The papers in the table belong to the following journals: Academy of Management Journal, Journal of Applied Psychology, Journal of Management, Journal of Organizational Behavior, The Leadership Quarterly, Organizational Behavior & Human Decision Processes, Personnel Psychology and Strategic Management Journal. It includes papers that conduct PSA and contain the terms "Propensity score" or "Matching" in the abstract and employ a PSA design in the main analysis.

## Example 1: The importance of CEO-CFO social interaction to explain outcomes for the CFO and the organisations

**Background:**

Shi et al. (2019) examine the role of CEO-CFO interactions to explain outcomes for the CFO and organisations. More specifically, they measure the level of CEO-CFO verbal mimicry from common function words (e.g., articles, pronouns, auxiliary verbs, and conjunctions) observed in conference calls in the context of firm mergers and acquisitions. They denote this measure as CEO-CFO language style matching (CEO-CFO LSM). Using different regression analyses, they find that CEO-CFO LSM explains CFO compensation, the likelihood of the CFO becoming a board member, and the number and value of mergers and acquisitions.

**Methodological design:**

To deal with the fact that the level of CEO-CFO LSM is not randomly assigned but selected by the CEOs, the authors implement a PSW analysis. First, they estimate a probit model predicting the probability of a firm having a high or low level of CEO-CFO LSM. They code it as a binary variable using the median value of CEO-CFO LSM. In the probit model, they include firm-level variables and previous information about the CFO. Then, they use the inverse of the PS calculated from the probit regression as a weight in regressions for different outcomes. The focus variable in these regressions is the level of CEO-CFO LSM, but it also controls for other firm and CFO characteristics as well as other CEO-CFO similarities.

**Strengths and limitations:**

The paper addresses a relevant question in the leadership literature, the role of social interaction to explain firm outcomes. It presents the PSW regression to complement previous regressions that do not explicitly deal with the endogeneity of CEO-CFO LSM. A first limitation is that the paper does not provide information about whether the application of PSW makes the sample more balanced in terms of observable variables. This is an essential consideration when interpreting PSW results. A second limitation of the PSW approach as used here is that transforming their continuous treatment variable (CEO-CFO LSM) into a binary one is arbitrary. Results could be different if another transformation rule were used.

## Example 2: The role of leader behaviour in understanding the effect of HPWS on employee and consumer satisfaction.

**Background:**

Schmidt and Pohler (2018) use PSW to estimate the causal impact of high-performance work systems (HPWS) on employee and customer satisfaction using longitudinal data from a financial service organisation in Canada. They also study whether this relationship could be explained as a result of reverse causality or a consequence of a commonly omitted variable such as leader behaviour.

**Methodological design:**

The paper employs a non-conventional PSW method. In particular, as Schmidt and Pohler (2018) indicate, transforming a continuous variable into a dichotomous variable consisting of treatment and control conditions is problematic. It requires arbitrary judgment that results in a loss of information and could generate model specification problems. Therefore, they use the covariate balancing propensity score for a continuous treatment proposed by Fong et al. (2018). This procedure assigns a weight to each observation, minimising the association between treatment and covariates. Using these weights, they specify regression models to estimate the two-way causality between HPWS and employee and customer satisfaction and the role of leader behaviour as an omitted variable in the causal relationship.

**Strengths and limitations:**

Estimation results by Schmidt and Pohler (2018) are suggestive as they only find a significant effect of HPWS on consumer and employee satisfaction in meta-analytic and cross-sectional studies. However, the significant impact disappears when using the covariance balancing propensity score method. Moreover, they found that it is leader behaviour -rather than HPWS- that is the primary driver of consumer satisfaction. Overall, the paper provides an interesting example of the relevance of adequately accounting for selection bias and simultaneity problems in causal analysis. More importantly, it also provides a way to deal with a continuous treatment variable in causal analysis.

Focusing on the methodology employed, a potential caveat of this analysis is that treatment allocation may depend on unobservable variables. Although the authors indicate that an instrumental variable analysis would be a way to tackle this concern, they do not pursue this approach as "it is very difficult to find a justifiable instrumental variable in survey-based research" (Schmidt and Pohler, 2018, p. 1013).

## Example 3: how internet activism affects the speed of donations in firms.

**Background:**

Using information from 613 large publicly listed Chinese firms, Luo et al. (2016) study how internet activism, and its interaction with other firm indicators, affected the speed of donations after the 2008 earthquake in the Sichuan Province of China.

**Methodological design:**

The study uses continuous-time event history design to estimate how quickly companies reacted to the 2008 earthquake with donations. The dependent variable is hazard rate of donation. Luo et al. (2016) employ a wide set of independent variables that include measures of internet activism, media coverage, reputation, political status of top executives and indicators for state-controlled or belonging to a culpable industry. Given that firms that donate and do not donate are not comparable, Luo et al. (2016) estimate the PS for donation prior to the earthquake using a probit model. In a second step, they adjust the event history regression through PSW. The paper does not provide detailed information about the PS specification or balance tests. However, a relevant aspect of the research is that they use the weighted regression to estimate the impact of different independent variables on the hazard rate of donation.

**Strengths and Limitations:**

Luo et al. (2016) show the contribution of different sets of variables to the regression of the speed of donation by adding these variables in sequential steps. They also show that results are robust to employing an OLS regression and a Heckman regression model that corrects for potential selection bias in non-donating firms.

Two potential limitations acknowledged by the authors is that relevant features of online media (such as the number of times an article was forwarded) or a wide range of online tactics are ignored. Also, the paper does not report nor mention balance tests. Nevertheless, Luo et al. (2016) provide a novel and interesting example of the use of PSW to study the determinants of firm donation decisions.

## Example 4: The role of partner's administrative controls to explain knowledge transfer.

**Background:**

Devarakonda and Reuer (2018) analyse how partners' administrative controls in nonequity collaborations affect knowledge transfer across partners. They postulate that technology overlap and the value of the

partners' knowledge drive the degree to which partners build upon each other's knowledge. They also hypothesise that this effect is moderated by steering committees.

**Methodological design:**

Devarakonda and Reuer (2018) estimate the impact of a set of independent variables including technology overlap and a steering committee indicator on cross-citations in publications by the client and R&D firms. For this analysis, they use pooled cross-sectional data from alliances in the biopharmaceutical industry. They address the problem that the choice of using a steering committee is not random by implementing a PSW analysis. They first estimate the PS for steering committees. Then, they weight observations with the inverse of the PS and estimate the determinants of cross-citations by client and R&D firms in a negative binomial framework. It is worth noticing that the output regression includes information about alliance experience and alliance citations that is not included in the PS specification.

**Strengths and limitations:**

Devarakonda and Reuer (2018) employ a number of robustness exercises to study the effect of steering committees on knowledge flows in nontechnological areas finding similar results for the R&D firm but not for the client firm. The paper is also an interesting example of how to apply PSW to a case where the output regression is non-linear. Given that the study is focused on alliances in the biotechnology sector, the authors acknowledge that a potential limitation of the paper is its lack of generalisability to other industries. They also explain that their model does not control for the dynamic effects of the steering committee that could be just responding to an incipient problem of misappropriation of knowledge.

## Example 5: The influence of CEOs on corporate reputation.

**Background:**

Love et al. (2017) study how CEOs influence corporate reputation. In particular, they hypothesise that companies whose CEOs receive more media attention will have a stronger reputation and this effect will be stronger the more positive the amount of media attention is. They state that a stronger reputation can be also explained by CEOs having outsider standing or having received industry awards.

**Methodological design:**

The authors test the hypotheses using separate models for each of the independent variables. They had two main issues to address. The first one is the potential endogeneity of the independent variable that motivates

the use of the PSA. The second problem stems from the fact that some of the independent variables are not dichotomous. The authors dealt with these two issues by using a weighting scheme based on the generalised propensity score technique (Imbens, 2000). Thus, in the first step, they run multinomial logit regressions to estimate PS for each category of the independent variable conditional to firm and CEO characteristics. They use specific control variables in each PS regression. Then, PSs are used to weight each category and estimate the impact of the different treatments on the measure of firm reputation (the dependent variable).

**Strengths and limitations:**

The methodological part of the paper shows how to use PSW to conduct causal analysis in settings with non-dichotomous treatment variables. The article also shows that their results are robust to the use of time-series output regressions with fixed effects.

As is common in PSA, a general concern is the endogeneity of the treatment variable because it might be explained by omitted variables. Love et al. (2017) address this issue using a cumulative count of awards as an instrumental variable. However, the authors acknowledge not being able to find valid instruments for media coverage and outsider status. Other concerns, also mentioned by the authors, are that the study uses a short time period (from 1991 to 1997) and that some relevant CEO characteristics could be omitted.

## Example 6: how do political and executive ties affect the sell-off strategy of firms?

**Background:**

Zheng et al. (2017) appraise the relevance of political and executive ties to affecting the sell-off strategy of firms in emerging markets. They also study how this effect is moderated by capital market and how developed the legal system is.

**Methodological design:**

They use a categorical indicator with three outcomes (sell-off, dissolution, or survival) as the dependent variable and indicators of political ties and institutional development as explanatory variables. Because firms with and without political ties are not comparable, the authors estimate the propensity to establish political ties by means of a probit regression. This PS regression includes five additional control variables not employed in the output regression that "may influence the formation of political ties but are not directly associated with sell-offs." (Zheng et al., 2017, p. 2021). Then, the second step uses the estimated PS to reweight the sample and estimate the likelihood of sell-offs employing a multinomial logit regression.

**Strengths and limitations:**

The paper provides an interesting example of the use of the PSW to conduct causal analysis in a multinomial logit output regression. Another strength of the study is that it explores alternative explanations for the results and alternative methods. Regarding the former, they tested whether political ties lead to poorer firm performance, finding non-significant results. They estimated the interaction between political ties and state ownership, finding that state ownership decreases the effectiveness of political ties but not legislative ties. Concerning alternative methods, they mention that they repeated the estimation but including a variable indicating the number of political ties (instead of a dichotomous variable), finding similar results. Although the paper refers to the use of the number of political ties in a robustness exercise, it does not provide detail on how such analysis is conducted using PSW. The authors also acknowledge as two potential limitations that the study does not account for unofficial ties and does not explore the mechanism of transmission.

## General discussion

The discussion above and the examples in Table 2 show that PSA could be a valid alternative to laboratory, field or natural experiments in causal analysis. Still, two main concerns can be mentioned. First, papers must provide enough detail about how the research is conducted. In the particular case of PSM, knowing the characteristics of the matching algorithm employed is critical to ensure replicability. Moreover, showing that the matching algorithm significantly improves covariate balance is essential to know whether PSA makes the treatment and control groups comparable in terms of observable variables. A second concern is that, even if PSA is rigorously conducted, it might not be enough to infer causality. More specifically, some internal validity threats are also present in many of the examples due to pre-treatment differences, maturation or history and attrition, among others. The papers show different ways to deal with these concerns. One possibility is to check how changes in the methodological design within a given study affects results. For example, while PSA relies on observed variables, estimating an instrumental variable regression (Gupta et al., 2017; Hopp and Pruschak, 2020) is a way to control for the impact of omitted variables, though suitable instruments are often lacking in the data set. Another relevant approach is to estimate causal effects in a regression model that permits double control for treatment predictors and/or other determinants of the response variable (Vitanova, 2021; Boivie et al., 2016; Schmidt and Pohler, 2018; Love et al., 2017). Furthermore, PSA can also suffer from external validity concerns as its results depend on the specific sample used in the analysis. A way to overcome this problem is to consider alternative settings (Bechtoldt et al., 2019; Li et al., 2021) to check how results depend on the particular conditions of a given study.

A challenge in future research would be to adapt the PSA to explore better how treatment is oper-

ationalised in a multi-treatment setting. One possibility is to estimate the different impacts of different treatment levels rather than dichotomising the treatment variable. Another option is to decompose the treatment variable into different sub-treatments to study the specific effect of each of them. In this regard, the example in Hopp and Pruschak (2020) shows an illustration of how one can implement such analysis using PSM. However, due to its simplicity, PSW provides an alternative way to deal with the multi-treatment extension as it only requires weighting observations according to the inverse of the PSs for each treatment level. Love et al. (2017), Schmidt and Pohler (2018) and the tutorial in the following section are examples of the use of such an approach for multivariate treatments.

# 4 A tutorial on PSW: Leadership succession effects

Leadership replacements are crucial decisions that can shape the performance of many organisations. Given its relevance, the matter has attracted the attention of researchers from different fields and with diverse backgrounds and interests. For instance, Berns and Klarner (2017) provide a complete review of the factors affecting the impact of CEO succession in publicly traded firms, Farah et al. (2020) extend their discussion to leadership changes in privately owned businesses and political organisations. Among these studies, the field of professional sports is particularly well suited to study leadership succession by offering stronger internal validity (Giambatista et al., 2005; Rowe et al., 2005). Regarding this aim, event studies, the most common methodological approach in this context, requires a precise definition of event dates, confounding factors, and event windows (de Jong and Naumovska, 2016). Great interest among the public in professional sports means that the dates of and reasons for head coach[7] replacements are widely covered by the media. Second, the firm objective, sporting success, is clearly defined, and such performance is frequently and regularly documented. Finally, we can clearly identify confounding variables such as the characteristics of a club and the difficulty of a match.

The following empirical example shows how to estimate the consequences of involuntary within-season managerial change in top-tier Italian football (*Serie A*) during seasons 2004/2005-2017/2018. An essential identification issue in such analysis is that managerial dismissal is not a random event. For example, it tends to occur particularly when a club is performing poorly. Estimation results can be biased if this issue is not properly accounted for in the model. Numerous studies analyse this issue using regression models that include previous performance information among the regressors (Audas et al., 2002; Tena and Forrest, 2007). However, a problem with regression analysis is that it is not informative on whether there is overlapping between treated and control observations. More recent research has employed matching methods to find

[7]The head coach/manager occupies a role akin to that of a CEO in other organisations (Hughes et al., 2010).

comparable counterfactuals in terms of observable variables. For example, Muehlheusser et al. (2016) use previous performance as the matching variable. van Ours and van Tuijl (2016) consider control groups formed with counterfactual observations that followed a similar path of cumulative surprise[8] but where the clubs did not replace their manager.

Our tutorial example shows how to address the question of the effect of head coach replacement by employing PSW. The PS is estimated as a function of multiple variables related to indicators of recent match outcomes, relative performance compared to expectations, position in the league, and recent performance in other competitions. The method used in the exercise is a double robust estimator as we control for determinants of managerial dismissals in two regressions, one for treatment assignment and another for the outcome variable (Funk et al., 2011). Such an approach offers protection for mismodelling as only one of the two specifications needs to be correct.

A second aim of the example is to show how to address a critical challenge faced by empirical researchers, the simultaneous estimation of the causal impact of multiple treatments. As discussed in the previous section, a limited number of papers consider a non-binary treatment (Hopp and Pruschak, 2020; Schmidt and Pohler, 2018). In this example, rather than just focusing on the aggregate impact of a head coach dismissal on future performance, we explain how to estimate the effect of a set of changes in managerial characteristics. Again, the proposed setting is particularly appropriate for this type of analysis as the natural time for changing leadership in sports clubs is at the end of the season (Tena and Forrest, 2007). This implies that the possibility of endogenously selecting a new head coach, let alone each of their different characteristics, in a within-season football turnover is limited in terms of available candidates and time to reach an agreement. Nonetheless, we also explain in Section 4.3 how to adapt the PSW analysis to deal with the possible endogeneity of the similarity in characteristics between dismissed and new coaches. Our proposed estimation is simple and built upon the standard regression analysis. This means that no statistical package specialised in PSA is not required for this estimation.

## 4.1 Data

We collected club-match level data from the top tier of the Italian professional football leagues (*Serie A*) for the seasons 2004/2005 - 2017/2018, which gives a total of 10,640 observations (5,320 matches). Throughout a season, each club competes against all others once at their home stadium and once away. In each match, a club is awarded 3, 1, or 0 points for a win, draw, or loss, respectively. At the end of the season, the club with the highest accumulated points wins the championship title, whilst the three lowest placed clubs are relegated

[8] In particular, the authors employ the cumulative sum of the difference between the actual and expected points measured by betting odds.

to the lower-tier league (*Serie B*). The league publishes official match reports. They contain, for example, the names of each club in the match, the respective managers and the outcome of the match. Additional sources used are provided below, together with the descriptions of (1) treatment variable, (2) variables that explain treatment assignment, and (3) outcome variables and additional control variables associated with the outcome.

#### 4.1.1 Treatment variable

Our treatment variable *New coach*$_t$ takes the value 1 if *Head coach*$_t \neq$ *Head coach*$_{t-1}$, where *Head coach*$_t$ is the name of the head coach who was in charge of the club in the match that took place in round $t$.[9] Note that our analysis focuses on dismissals and does not consider cases of termination by mutual consent or voluntary quit by the old coach. Moreover, any match managed by a temporary caretaker manager is discarded from the analysis. Given this, we identified 157 relevant cases during the seasons 2004/2005 - 2017/2018 by a careful inspection of the archives from the official websites of the league and individual clubs, as well as the two most-read national sports newspapers in Italy, *Corriere dello Sport-Stadio* and *La Gazzetta dello Sport*.[10]

Given that each case of changing the leadership in an organisation introduces simultaneous changes in managerial characteristics, investigating whether such changes can account for the effectiveness of replacement is a relevant issue in leadership succession. Therefore, we collected additional information related to the individual manager's characteristics from the League Managers Association (LMA)[11] and Transfermarkt.[12] These include important indicators of leadership characteristics previously identified in the literature.[13] A first set of managerial characteristics are related to the individual manager's previous experience as a head coach in professional football leagues; experience in years (*Experience in years*), dummy variables that indicate whether: a manager had previously held a relevant role within *Serie A* (*Experience Serie A*), this is his first employment in the relevant role (*No previous experience*), a manager has a previous experience in top tier professional league abroad (*Experience abroad*). The second set of dummy variables are related to a manager's background as a professional player, which indicate whether: a manager is a former professional football player (*Former player*), a manager is a former player in *Serie A* (*Former player Serie A*), and a

[9] The league currently features 20 clubs, yielding the total number of matches played by an individual club in a given season of 38. Round $t$, therefore, corresponds to the $t$-th match in a particular season.

[10] During the relevant seasons, 15 cases of voluntary departures of head coaches were identified. For the same period of time, there were eight caretaker managers who were in charge during the transition between outgoing and incoming head coaches.

[11] Available at https://leaguemanagers.com/.

[12] Available at https://www.transfermarkt.com/.

[13] See, for instance, Bolton et al. (2013), Bridgewater et al. (2011), Dawson and Dobson (2002) and Detotto et al. (2018) for the managerial characteristic indicators related to professional sports. In more general setting, the effect of CEO characteristics on corporate performance have also been studied (Kaplan et al., 2012).

manager is a former defender or goalkeeper (*Former defender/goalkeeper*). The third set of indicators relate to a manager's association with the club. They indicate whether: the manager is a former vice coach of the club (*Former vice coach*), he is a former player of the club (*Former player club*), and the club is the last club with which he has been a player (*Last club as a player*). Another couple of dummy variables associated with recent employment status in the relevant role are considered. In particular, one takes a value equal to 1 if a manager was not employed in a relevant role in any club in the immediately preceding season, and 0 otherwise (*Absent last season*). The other indicator associated with recent activity indicate whether a manager was active or employed at any club participating in *Serie A* in the immediately preceding season (*Active Serie A last season*). The final set of variables are related to a manager's personal features: a manager's age in years (*Age in years*) and an indicator that takes a value equal to 1 if a manager is Italian, and 0 otherwise (*Italian nationality*).

Again, since each case of managerial succession results in changes in these managerial characteristics which could also affect post-succession performance, we take into account differences between the new and old managers. That is, for each characteristic variable $h_t$, we take the difference in the value of the variable between the manager in place at time $t$ and the manager who had been in place at time $t-1$, i.e. $\Delta h = h_t - h_{t-1}$. Where a characteristic variable $h_t$ is binary, as is the case for many of them, $\Delta h$ is tertiary and takes values $\{-1, 0, 1\}$. Effectively, $\Delta h = 0$, where there was no managerial succession, or no difference between the new and old managers in the respective characteristic. Given this, Table 3 provides the summary statistics of the characteristic change variables for the 157 cases of managerial change considered in the analysis.

Table 3: Summary statistics of differences in managerial characteristics

| **Variable** | **Difference between new and dismissed coaches** ($\Delta h$) | | |
|---|---|---|---|
| **Binary indicators** | **-1** | **0** | **1** |
| Former player | 18 (11%) | 120 (76.43%) | 19 (12.1%) |
| Absent last season | 18 (11%) | 100 (63.69%) | 39 (24.84%) |
| Former defender/goalkeeper | 34 (22%) | 99 (63.06%) | 24 (15.29%) |
| Former vice coach | 7 (4%) | 136 (86.62%) | 14 (8.92%) |
| Italian nationality | 12 (8%) | 132 (84.08%) | 13 (8.28%) |
| Experience Serie A | 25 (16%) | 106 (67.52%) | 26 (16.56%) |
| No previous experience | 8 (5%) | 133 (84.71%) | 16 (10.19%) |
| Former player Serie A | 36 (23%) | 88 (56.05%) | 33 (21.02%) |
| Former player club | 18 (11%) | 113 (71.97%) | 26 (16.56%) |
| Last club as a player | 7 (4%) | 141 (89.81%) | 9 (5.73%) |
| Experience abroad | 25 (16%) | 100 (63.69%) | 32 (20.38%) |
| Active Serie A last season | 41 (26%) | 87 (55.41%) | 29 (18.47%) |
| **Continuous variables** | **Min.** | **Mean** | **Max.** |
| Age in years | −26 | 1.080 | 29 |
| Experience in years | −31 | 0.760 | 33 |

*Notes:* Table shows the summary statistics of managerial characteristics change variables for the 157 replacements included in the analysis. By construction, the difference variables for a binary characteristics indicator takes the value of -1, 0, and 1, where frequencies of each value together with the percentage of the all the cases are reported. Whilst those for continuous variables are also continuous for which the maximum, mean, and minimum values are presented.

The Table provide some insight into the sort of changes made in managerial succession. In most cases, the value of $\Delta h$ is equal or close to 0, implying that the new and old managers share a similar respective characteristic, hence suggesting a tendency. This may be because many clubs have a vision of what the ideal profile of a manager would be. However, there are also many cases of changes in the values of the characteristic variables. Therefore, in the subsequent analysis we estimate the individual effect of changes in specific characteristics, other things being equal, on post-succession performance.

### 4.1.2 Variables related to treatment assignment

In order to estimate the propensity scores, a number of covariates which may affect the likelihood of treatment are considered for inclusion in the treatment assignment model. These are identified in another strand of literature, for instance, Bryson et al. (2021) and references therein. The main cause of within-season managerial dismissal is related to the club's recent on-field performance. We measure this by the average number of points earned over the last four matches (*Points last four matches*) and a dummy variable to indicate a loss in the most recent match (*Loss last match*). In addition, we include a dummy variable to indicate whether a defeat was at the club's home stadium (*Loss last match at home*), to account for the possibility that this event brings more pressure on a club than an away defeat. It has also been shown that performance relative to expectations matter. To take this into account, we include a measure of "surprise" accumulated over the relevant season (*Cumulative surprise*). Following van Ours and van Tuijl

(2016), surprise is measured by the deviation of actual points from expected points for each match, where expected points are obtained using the *ex ante* probabilities of win, draw, and loss for each match based on the closing odds available from various bookmakers.[14] We also consider the current league position relative to the final position in the previous season (*Relative standing*), which captures performance against subjective expectation by the fans. Furthermore, the current situation of a club is captured by two variables indicating whether a club is in the relegation zone (*Relegation zone*) and current position in the league (*Standing*), respectively. Whilst this study focuses on performance in the domestic league (*Serie A* in our case), it is possible that performance in other competitions could affect the prospect of a manager being dismissed. In particular, unfavourable outcomes, particularly critical ones, in other important competitions can impose extra pressure on the job security of a manager. To take this into account, we consider three binary variables indicating whether a club had been eliminated from UEFA Champions League (*Eliminated Champions League*), UEFA Europa League (*Eliminated Europa League*), or Coppa Italia (*Eliminated Coppa Italia*), between two *Serie A* matches $t$ and $t-1$.

Additional variables considered in the treatment assignment model are an indicator of whether the club had already replaced a manager in the particular season (*Having dismissed this season*) and the number of days between two matches (*Days between matches*), which could potentially affect the decision of within-season managerial dismissals. Finally, as previous studies have shown, see Muehlheusser et al. (2016) for instance, within-season dismissals occur more frequently in mid-season. To capture this effect, we include round fixed effects in the treatment assignment model.

#### 4.1.3 Outcome variables and additional control variables associated with the outcome

To measure club performance following treatment assignment, we construct outcome variables based on average points obtained in subsequent matches. For robustness, we obtain these values using up to five matches (*Points five matches*), 10 matches (*Points ten matches*), and all of the remaining matches in the season (*Points rest of season*) or until the next managerial change, whichever occurs earlier.

In our outcome model, we include additional control variables that can affect post-treatment performance. First, a variable *Home advantage* controls for home advantage measured by the proportion of the matches that took place at the home stadium, out of the matches with which we measure the outcome variable. In addition, the ability level of the club (*Club ability*), and that of opponents (*Opponent ability*) are controlled by the ability indicator constructed in the following manner. First, we take a club's final position in the league table in the preceding season, reversing the order so that, for example, the top club was assigned the value 20 (and the bottom club would be assigned the value 1). The order is reversed to ensure that the

[14] Available at https://www.football-data.co.uk/.

variable increases with club ability as captured by its performance in the preceding season. In cases where a club had not played in the top division in the preceding season, it was assigned the value 1 (i.e. treated as having been equivalent to the bottom club in the top-tier). We obtain these values for the final positions over the past four seasons, then take the weighted average with higher weights given to the more recent seasons for each club.[15] The variable *Opponent ability* is the average value of the ability indicator for the opponents in the subsequent matches with which the outcome is measured.

## 4.2 Methodology

We estimate the following outcome model to analyse the consequence of involuntary head coach replacements on $y_{its}$, our measure of the performance of club $i$, at round $t$ in season $s$. Specifically, it is defined as:

$$y_{its} = \delta \; New \; coach_{\,its} + \gamma' X_{its} + \varepsilon_{its}, \tag{5}$$

where $New\ coach_{\,its} = 1$ if club $i$ has replaced its manager prior to round $t$, and $New\ coach_{\,its} = 0$ otherwise; $X_{its}$ is a vector of control variables. In particular, it includes variables related to managerial dismissal as well as variables associated with match outcome.[16] A coefficient $\delta$ and a vector $\gamma$ are parameters to be estimated. Finally, $\varepsilon_{its}$ is a stochastic error component.

Our focus is to obtain the estimate of $\delta$, which, if the treatment ($New\ coach_{\,its}$) were randomly allocated, should capture the Average Treatment Effect (ATE) of managerial change, if any. Model (5) is subsequently augmented to account for the various changes that may have been made with respect to the managerial characteristics of the head coach. Such possible changes are captured by a set of indicators defined as differences in managerial characteristic variables between replaced and appointed coaches, as explained in the previous section. Therefore, our extended model is specified as follows:

$$y_{its} = \delta \; New \; coach_{\,its} + \beta' \Delta H_{its} + \gamma' X_{its} + \varepsilon_{its}, \tag{6}$$

where $\Delta H_{its}$ is a vector of the managerial characteristic change variables, and $\beta$ is its associated vector of parameters. Variables in $\Delta H_{its}$ take value zero when there was no managerial change prior to match $t$, or when no change was made with regards to the particular feature of the manager. Therefore, if parameters in vector $\beta$ are significantly different from zero, this implies that differences in the characteristics between

[15] More precisely, the weights given to the seasons $s-1$, $s-2$, $s-3$, $s-4$ are 0.5, 0.3, 0.15, and 0.05, respectively, where $s$ represents the current season. As reported in Dixon and Coles (1997), a club's ability is better measured by recent performance with increasing weights on the more recent information.

[16] Including determinants of match outcome observed after the treatment is relevant in this setting as match score is also affected by home advantage and ability measures of both teams. All these variables can be considered exogenous as they occur in a quasi-random fashion.

outgoing and incoming managers do matter for the successful implementation of managerial change.

However, an important concern in the estimation of models (5) and (6) is that head coach changes are not random events since they tend to occur more frequently with exceptionally low performing clubs. Note that the inclusion of determinants of managerial dismissals allows us to control for different characteristics of treated and untreated teams. However, a simple OLS regression is not informative on whether these two groups are comparable in terms of their observable characteristics, threatening the causal interpretation of the estimation results. Under the assumption that we can observe the main determinants of managerial dismissal, PSA can be used to obtain counterfactuals that allow for a causal estimation. A characteristic of our setting is that a head coach dismissal is a sporadic event, in the sense that 157 club-match observations out of 10,344 were followed by managerial replacement.[17] Thus, it is essential to find comparable counterfactuals in terms of observable variables for each treated observation. For this example, we choose PSW based on its simplicity and because it can include the whole sample in the estimation. Thus, since our treatment is binary, this implies that treated observations are weighted with the inverse of the probability of being treated, while control observations are given weights defined by the inverse of (1 − the probability of being treated). As a result, the distribution of propensity scores, i.e. the *ex-ante* probabilities of being treated, becomes similar between the treatment and control groups, as though the treatment were allocated randomly.

In our case, an observation is considered treated when a newly assigned manager is in charge at time $t$ following the dismissal of a previous manager. Therefore, the likelihood of treatment assignment depends on the information related to performance that has been realised prior to time $t$. We estimate propensity scores by means of logistic regression. The model selection follows stepwise regression with a sequential replacement algorithm. The sequential replacement combines forward and backward selections, where the predictors proposed in Section 4.1 are iteratively added and removed until the lowest predictive error is achieved.[18] Given the set of selected covariates, $Z_{its}$, we obtain the predicted values $\hat{p}_{its} = \Pr[New\ coach_{its} = 1|Z_{its}]$, then the inverse propensity score weights are defined as follows:

$$w_{its} = \begin{cases} \frac{1}{\hat{p}_{its}}, & \text{if } New\ coach_{its} = 1, \\ \frac{1}{1-\hat{p}_{its}}, & \text{if } New\ coach_{its} = 0. \end{cases} \tag{7}$$

These weights may now be used in the weighted regression analysis to obtain the parameter estimates of models (5) and (6); see Guo and Fraser (2014) and Morgan and Todd (2008) for the use of inverse propensity

[17] This low number of treated observations is also common in previous PSA papers (Bechtoldt et al., 2019; Li et al., 2021; Ong, 2021)

[18] We use the most common measure of the predictive error, Akaike Information Criterion (AIC). See Bruce and Bruce (2017) for the details of stepwise regressions.

score weights in the estimation of linear models. Moreover, a set of covariates selected in the treatment assignment model will be included in the outcome models, according to the doubly robust estimation procedure. In the following sections, we follow the steps described in Section 2 and Figure 1 to estimate first the treatment assignment model, then the outcome models (5) and (6).

## Step 1: Propensity score estimation

As described in Section 2, the initial step in PSW is to estimate the treatment assignment models. Following Section 4.2, we estimate the logistic regression with step wise selection, where we consider the set of covariates presented in Section 4.1 in the initial step.[19] The set of covariates selected in the final model and estimation results are reported in Table 4.

Table 4: Stepwise regression[a] results for treatment assignment

| | *Dependent variable:* |
|---|---|
| | New coach |
| (Intercept) | $-5.356^{***}$ (0.363) |
| Cumulative surprise | $-0.253^{***}$ (0.025) |
| Days between matches | $0.062^{***}$ (0.020) |
| Points last four matches | $-0.800^{***}$ (0.190) |
| Loss last match | $1.424^{***}$ (0.252) |
| Relegation zone | $0.393^{**}$ (0.186) |
| Eliminated Europa League | $1.309^{*}$ (0.769) |
| Observations | 10,344 |
| Log Likelihood | -615.719 |
| Akaike Inf. Crit. | 1,245.438 |

*Notes:* [a]The stepwise regression with the lowest AIC as a stopping criterion. $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01. Robust standard errors in parentheses.

The estimated coefficients of the selected covariates present expected signs,[20] being in line with previous findings. The probability of managerial change increases when a club has: lost the last match, performed poorly in the previous four games, and suffered negative surprising results during the season. In addition, the likelihood of turnover is higher when there are more days available between the last and current match. A recent elimination from the Europa League as well as a threat of relegation also contribute to a higher probability of managerial change.

To assess the separability of the model, Figure 2 plots the Precision-Recall curve (PRC) and reports the

[19]Thoemmes and Ong (2016) indicate that PSW in the longitudinal case should repeat the process of weighting at every single point. The idea is to make treatment dependent only on information occurring before this decision, including also previous treatment decisions. In this example, we keep that spirit as model covariates only contain information that precedes treatment decisions. Moreover, a variable *Having dismissed this season* captures previous decision to dismiss a manager in the same season.

[20]Note that these estimates could be dependent on the selection method employed. For example, using Lasso results in a slightly different set of selected covariates and associated coefficients. However, as shown in Section 4.4, the final estimation of ATE remains similar.

Area Under the PRC (AUC-PR). The value of AUC-PR (.12786) is larger than that of baseline (.01518). This indicates the separability of the model, i.e. the model is better able than the empirical probability to classify observations as treated or untreated.

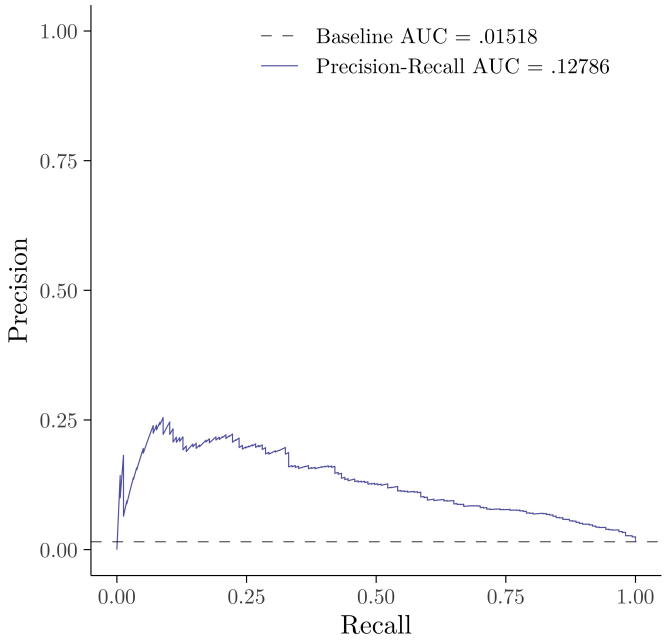


Figure 2: Precision-recall curve and area under the curve

*Notes:* The Figure plots the combinations of precision (y-axis) and recall (x-axis) for different thresholds for predicted classification (PR curve). The baseline PR curve is defined by the proportion of true positive cases out of the whole sample. Also indicated in the Figure are the area under the PR curve (Precision-Recall AUC) and the area under the baseline PR curve (Baseline AUC).

## Step 2: Obtaining PS weights

Having estimated the treatment assignment model and checked the model performance in terms of separability, we will now have a closer look at the distribution of predicted values. First, the average predicted probabilities of treatment ($\hat{p}_{its} = \Pr[New\ coach_{its} = 1|Z_{its}]$) are 0.0141 for those who did not change the manager (control group) and 0.0834 for those who actually did change their manager (treatment group). Estimates ranged from almost nil ($4.395 \times 10^{-6}$) to 0.8191 for the former group, and from 0.0016 to 0.5460 for the latter. Note that all the treated cases are contained within the common support, i.e. where the ranges of propensity scores for treated and control groups overlap. We now compute the weights according to the weighting function defined in (7). To show the distribution of predicted values for each group before and after weighting, Figure 3 plots the kernel density of log propensity scores for each group.

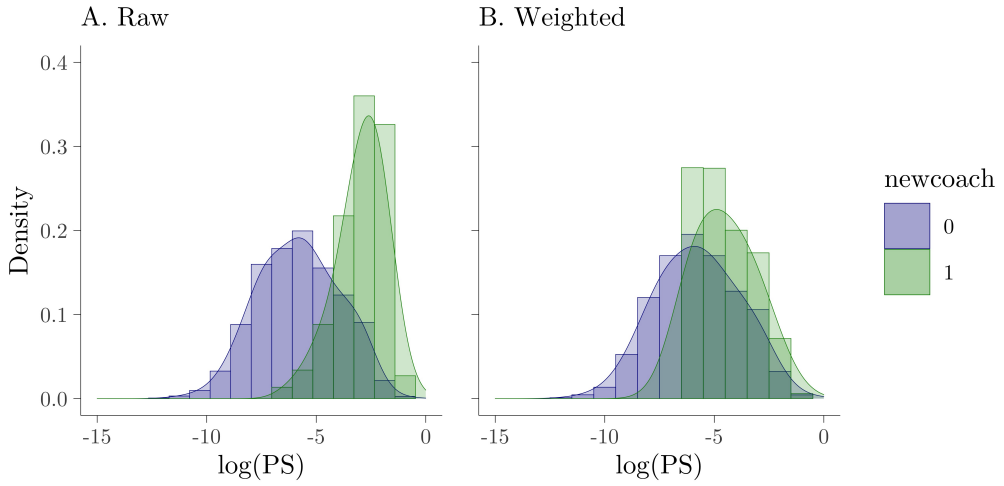


Figure 3: Kernel density of log propensity scores before/after weighting

*Notes:* Figure presents the Kernel densities for the *ex-ante* probability distributions of treatment before weighting (panel A) and after weighting (panel B) for the two treatment groups ($New\ coach = 0$ and $New\ coach = 0$).

## Step 3: Balance diagnostic

We can now check whether the PSW can reduce the imbalancedness of the covariates included in the treatment assignment model. To do so, following Austin and Stuart (2015) and Morgan and Todd (2008), we compare the average value of absolute standardised mean differences (SMD) between treated and control group for each covariate. The standardised difference of the mean for a covariate $z$ is calculated as:

$$\frac{|\bar{z}_{i,d_i=1} - \bar{z}_{i,d_i=0}|}{\sqrt{\frac{1}{2}\text{Var}[z_{i,d_i=1}] + \frac{1}{2}\text{Var}[z_{i,d_i=0}]}}, \tag{8}$$

where $\bar{z}_{i,d_i=1}$ and $\bar{z}_{i,d_i=0}$ are the means for those in treatment group ($d_i = 1$) and control group ($d_i = 0$), respectively, and $\text{Var}[z_{i,d_i=1}]$ and $\text{Var}[z_{i,d_i=0}]$ are the respective variances. The measure reflects the distance between the two groups in terms of the covariate that affects the treatment assignment. Table 5 present these values for (1) raw sample, (2) weighted sample, and (3) weighted sample within common support. The average values of absolute SMDs are 0.796, 0.312, and 0.127 in raw sample, weighted sample, and weighted sample within common support, respectively. This suggests significant reductions in imbalancedness due to weighting, and further improvement in balance within common support. The latter point is consistent with Heckman et al. (1999) who note that a further improvement in covariate balance is obtained by restricting the estimation sample to observations with a positive probability of being both participants and non-participants. Therefore, we use the sample within the common support to estimate the consequences of head coach turnover

in the following subsection.

Table 5: Covariate balance table (before/after weighting, all/common support)

| | Raw | | Weighted | | Weighted (CS) | |
|---|---|---|---|---|---|---|
| Covariate | SMD | P-value | SMD | P-value | SMD | P-value |
| Cumulative surprise | 1.265 | 0.000 | 0.658 | 0.000 | 0.199 | 0.199 |
| Days between matches | 0.284 | 0.004 | 0.107 | 0.271 | 0.041 | 0.654 |
| Eliminated Europa League | 0.074 | 0.500 | 0.064 | 0.000 | 0.077 | 0.000 |
| Points last four matches | 1.107 | 0.000 | 0.622 | 0.000 | 0.120 | 0.350 |
| Loss last match | 1.044 | 0.000 | 0.133 | 0.402 | 0.268 | 0.082 |
| Relegation zone | 1 | 0.000 | 0.290 | 0.070 | 0.058 | 0.671 |
| Mean SMD | 0.796 | | 0.312 | | 0.127 | |
| N (Treated) | 157 | | 157 | | 157 | |
| N (Control) | 10187 | | 10187 | | 6218 | |

*Notes:* Table reports the absolute values of standardised mean differences (SMD) between the treatment and control groups before and after weighting.

## Step 4: Estimation of treatment effects

The final step is to estimate the treatment effects by applying the defined weights through weighted regression analysis. Before we proceed, however, we briefly discuss the possible consequences of not addressing the imbalancedness detected in the previous steps. In particular, the differences in preceding performances between the treated and control groups are large, implying that involuntary managerial changes are not random events. However, theory does not provide a clear indication of how ignoring such differences can affect conclusions on the impact of replacing a manager. On the one hand, one can argue that poorly-performing teams may revert to their mean performance levels regardless of whether they replace their head coaches. On the other hand, it is also plausible to assume that some poorly-performing teams are more likely to carry on with this negative inertia due to persistent issues, such as long-term injuries or conflicts among players, even if they replace their head coach. Figure 4 plots the average values of performance in the post-treatment periods (between treatment assignment and the end of the respective season) for treated and control groups, at a given level of the club's ability in the raw sample. The initial look of the Figure suggests that performance is increasing in a club's ability; however, the *prima facie* difference between treatment and control groups is not evident in the raw sample.

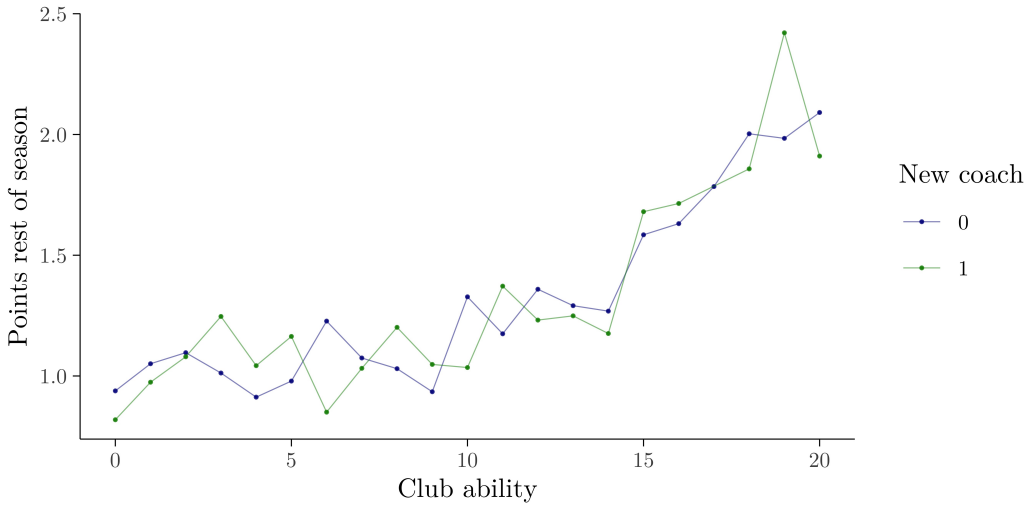


Figure 4: Mean value of outcome variable (*Points rest of season*) for each ability and treatment group

*Notes:* The x-axis represents the ability of the club (*Club ability*), computed based on the weighted average of the final league positions in the preceding four seasons, with the value 1 being the lowest ability and 20 being the highest. the y-axis measures the mean values of an outcome variable (*Points rest of season*), the average points obtained following assignment or non-assignment of the treatment for each treatment group ($New\ coach = 0$ and $New\ coach = 1$).

The OLS estimates of model (5) suggest some weak evidence of detrimental effects from managerial change.[21] Of course, this approach is not robust to the potential selection bias discussed above since it does not focus on comparable treated and control groups in terms of observable characteristics.

Now we revert to the estimation of our outcome models (5) and (6), using PSW. The estimation results of these models are shown in Table 6, where outcome variables are the average points obtained in the post-treatment matches, where we include up to 5 matches, 10 matches, and the rest of the season.

[21] The details of OLS estimation are available in Table 9 in Appendix.

Table 6: Double robust estimates of outcome models

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Points five matches | | Points ten matches | | Points rest of season | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| New coach | 0.139 | 0.102 | 0.184* | 0.137** | 0.180** | 0.117** |
| | (0.104) | (0.063) | (0.097) | (0.061) | (0.090) | (0.056) |
| Former player | | 0.120 | | 0.048 | | 0.049 |
| | | (0.105) | | (0.110) | | (0.103) |
| Absent last season | | 0.163 | | 0.256** | | 0.288*** |
| | | (0.107) | | (0.102) | | (0.097) |
| Age in years | | −0.009 | | −0.019* | | −0.020** |
| | | (0.011) | | (0.011) | | (0.010) |
| Experience in years | | −0.007 | | 0.009 | | 0.013 |
| | | (0.010) | | (0.010) | | (0.009) |
| Former defender/goalkeeper | | 0.164* | | 0.242*** | | 0.186** |
| | | (0.094) | | (0.091) | | (0.082) |
| Former vice coach | | −0.454*** | | −0.170 | | −0.210 |
| | | (0.174) | | (0.188) | | (0.182) |
| Italian nationality | | 0.198 | | 0.171 | | 0.068 |
| | | (0.124) | | (0.154) | | (0.143) |
| Experience Serie A | | 0.006 | | −0.035 | | −0.126 |
| | | (0.117) | | (0.117) | | (0.114) |
| No previous experience | | 0.055 | | −0.085 | | −0.109 |
| | | (0.198) | | (0.165) | | (0.152) |
| Former player Serie A | | −0.365*** | | −0.204* | | −0.193** |
| | | (0.099) | | (0.105) | | (0.090) |
| Former player club | | 0.033 | | −0.065 | | 0.026 |
| | | (0.150) | | (0.155) | | (0.136) |
| Last club as a player | | 0.523** | | 0.665*** | | 0.529*** |
| | | (0.232) | | (0.216) | | (0.191) |
| Experience abroad | | −0.110 | | −0.069 | | −0.061 |
| | | (0.105) | | (0.104) | | (0.096) |
| Active Serie A last season | | 0.094 | | 0.075 | | −0.012 |
| | | (0.086) | | (0.098) | | (0.091) |
| Home advantage | 1.101*** | 0.849*** | 1.057*** | 0.773*** | 1.029*** | 0.802*** |
| | (0.230) | (0.139) | (0.261) | (0.149) | (0.298) | (0.151) |
| Club ability | 0.057*** | 0.038*** | 0.059*** | 0.043*** | 0.061*** | 0.043*** |
| | (0.015) | (0.004) | (0.013) | (0.004) | (0.013) | (0.004) |
| Opponent club ability | −0.032** | −0.040*** | −0.039* | −0.046*** | −0.031 | −0.035** |
| | (0.015) | (0.010) | (0.022) | (0.014) | (0.023) | (0.015) |
| Constant | 0.638*** | 0.807*** | 0.797*** | 0.949*** | 0.724*** | 0.809*** |
| | (0.155) | (0.117) | (0.195) | (0.131) | (0.229) | (0.152) |
| Observations | 6,375 | 6,375 | 6,375 | 6,375 | 6,375 | 6,375 |
| Log Likelihood | -7,998.567 | -7,261.832 | -7,361.029 | -6,685.688 | -7,036.235 | -6,288.864 |
| Akaike Inf. Crit. | 16,019.130 | 14,573.670 | 14,744.060 | 13,421.380 | 14,094.470 | 12,627.730 |

*Notes:* *p<0.1; **p<0.05; ***p<0.01. Robust standard errors in parentheses. The coefficients for the covariates from treatment assignment model are not reported.

The estimates for model (5)[22] are reported in columns (1), (3), and (5), for the respective outcome variables. No significant treatment effects at the 10% significance level are detected in the short run (first five matches). Still, a positive and significant impact at the 5% significance level is evident once a longer run of post-treatment matches is considered.

[22]Following Ridgeway et al. (2021), the estimation of outcome models are obtained using "svyglm" function within "survey" package in R, which is commonly used for survey sample analysis and automatically produces robust standard errors.

Columns (2), (4), and (6) in Table 6 present the estimated parameters for our extended model (6), where the additional variables that capture the characteristic difference between new and dismissed managers are included. Including these variables does not affect the sign of the binary treatment effect (*New coach*) in the corresponding baseline models (1), (3), and (5), respectively, whilst the size of such is smaller. The results suggest that the changes in particular characteristics of managers affect the post-treatment performance. For instance, when a new manager was absent (not employed as a head coach elsewhere) in the previous season, this tends to have a positive impact on post-succession outcome. On the other hand, older replacement managers tend to achieve a negative treatment effect. The variables that capture the changes associated with experience, such as experience in years, experience abroad, experience in *Serie A*, and no previous experience, do not show significant effects to explain the post-succession performance at the 10% significance level. Similarly, having been employed at a *Serie A* club in the immediate previous season is not a significant variable at the 10% significance level.

A new manager's background as a professional player relative to that of a dismissed manager in general does not have a significant impact at the conventional significance levels, whilst a positive outcome is expected if a manager played more defensive role as a player. However, when a new manager is a previous *Serie A* player, where a dismissed one is not, the succession tend to have a negative effect holding other things constant. A speculative explanation for this is that becoming a manager in a new market (where they did not participate as a player) indicates desirable managerial skills. The positive and significant coefficient of *Last club as a player* implies that a manager with stronger association with the club (when one finished their playing career at the club) can positively influence the post-succession performance, whilst merely being a former player of the club (*Former player club*) has no significant effect at the 10% significance level. However, replacing a manager with a former vice coach at the club tend to have a detrimental effect, particularly in the short term. Finally, changes in nationality, i.e. being Italian, do not show any significant impact at the conventional levels.

In all the models, the coefficient estimates on all the control variables have expected signs; the percentage of home matches and club ability both have a significant positive effect on match outcomes, a club's performance is negatively correlated with the average ability of their opponent clubs.

### 4.3 Extension: endogeneity of similarity in coach characteristics

As we have previously discussed in Section 3, a common problem in empirical research is dealing with a multi-treatment situation. However, in the last step of our previous analysis, we did not account for the potential endogeneity of changes in managerial characteristics in the causal estimation. This decision could

be justified in this context because the scope for selecting each dimension of managerial characteristics is limited given the limited time and candidates available in the within-season setting. Nevertheless, we can consider the possibility of a club endogenously choosing a similar or dissimilar replacement in terms of overall characteristics. Therefore, in this example, our purpose is to illustrate how to modify the analysis to consider a multi-level treatment where a club can decide further whether the replacement should be similar or dissimilar to the dismissed manager.

To define the similarity between the new and dismissed manager, we cluster managers using the characteristic variables used in the previous analysis. In particular, we employ the Partitioning Around Medoid (PAM) algorithm[23] to group the managers into clusters based on the similarities in terms of their characteristics. We then define the treatment as "similar" if dismissed and appointed managers are in the same cluster, and "dissimilar," if they are in distinct clusters. More formally, we create an additional dummy variable *Dissimilar coach*, where *Dissimilar coach* $= 0$ if the new coach is "similar" to the dismissed according to the above definition, and *Dissimilar coach* $= 1$ if the new coach is "dissimilar." Based on this definition, we identify 112 cases out of 157 cases of the replacements as "dissimilar" changes and 45 cases as "similar" change. To incorporate this additional layer into the decision problem, we consider a nested logistic regression to obtain the probabilities of no change, similar change, and dissimilar change. The first nest models the decision regarding whether to replace a manager (*New coach* $= 1$) or not (*New coach* $= 0$), as considered in the previous section. The model of the second nest estimates the probability of a dissimilar replacement (*Dissimilar coach* $= 1$), within the treated observations (*New coach* $= 1$). A graphical representation of the nested logit model is given in Figure 5, which also illustrates the three treatment types and associated probabilities.

[23]This method for clustering is suitable for our context, where a mix of continuous and categorical variables are to be considered. The algorithm identifies the optimal number of clusters based on "silhouette widths", a measure of relative similarity to the members in the same group compared to those in the other group. See, der Laan et al. (2003) for the details.

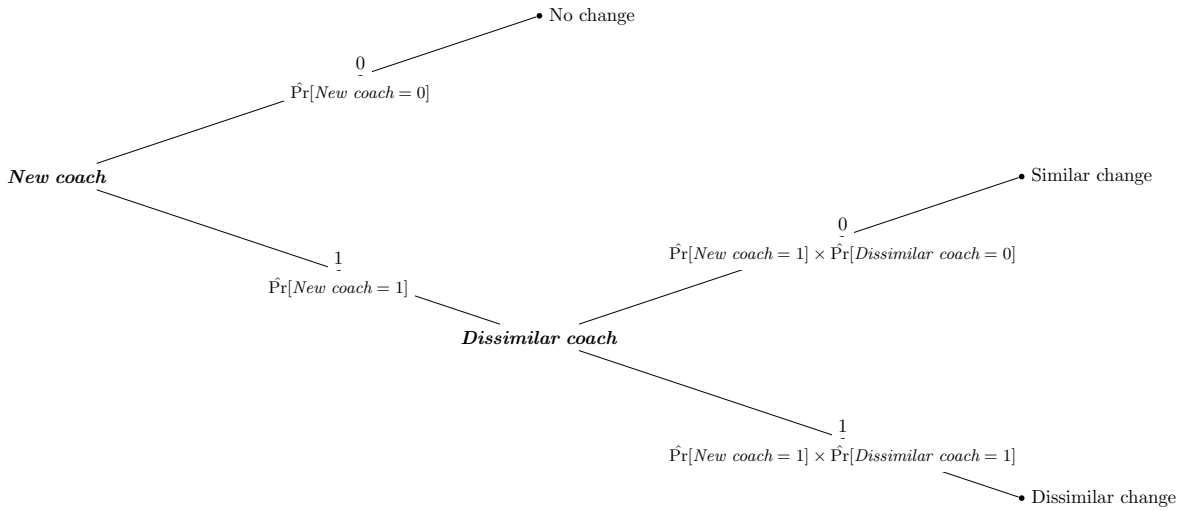


Figure 5: Nested logit model

*Notes:* The Figure illustrates the nested logit model, where the first nest classifies the cases into $New\ coach = 0$ or $New\ coach = 1$, and the second nest further classifies cases with $New\ coach = 1$ into $Dissimilar\ coach = 0$ or $Dissimilar\ coach = 1$, resulting in the three possible outcomes (No change, similar change, and dissimilar change). Corresponding probabilities of each outcome are obtained using the predicted values resulting from the estimation of logistic regression of each nest.

The procedure, akin to Step 1 in the previous section, can be applied to estimate the probability of dissimilar change, i.e. $\hat{\Pr}[Dissimilar\ coach = 1|Z]$. The estimation results are quite different from those in Table 4, where only a few covariates were selected: *Relegation zone*, *Days between matches*, and *Standing*. The associated AUC-PR is .82301, which is larger than the baseline of .71338. However, the AUC-PR relative to baseline in this case (1.154) is considerably smaller than that of the first nest model (8.422). This indicates that the separation is more challenging in the former case than the latter.

The flip side of this is, however, that the underlying imbalancedness is less severe. In fact, the SMDs of the selected covariates between the dissimilar and similar changes are not significant at 5% significance level in the raw sample, and the average value of the absolute SMDs is .217. Nevertheless, a significant reduction in the SMDs is achieved between the similar and dissimilar changes by applying the weights defined by the inverse of respective propensity scores; the average value of the absolute SMDs is .026 in the weighted sample.

Based on this, we extend our previous model to assess the effectiveness of the three possible treatments, (1) no change, (2) similar change, and (3) dissimilar change. As explained in Wooldridge (2010), regression adjustment in the multiple treatment case is an obvious extension of the case where treatment is binary. Therefore, we weight the sample with the inverse of the *ex-ante* probability of an actual treatment status, as depicted in Figure 5. Then, we estimate the outcome model (5) with an additional treatment variable *Dissimilar coach*, together with the control variables associated with the outcome and the covariates selected

in the treatment assignment model. The results are reported in Table 7. The estimated coefficients of *New coach* and *Dissimilar coach* indicate that replacement with a similar manager has no statistically significant effect at the 10% significance level, whilst the appointment of a new manager who has a different profile than the old is associated with an improvement in the five and ten following matches at 10% and 5% significance levels, respectively.

Table 7: Double robust estimates of outcome model with dissimilar treatment

| | *Dependent variable:* | | |
|---|---|---|---|
| | Points 5 matches | Points 10 matches | Points rest of season |
| | (1) | (2) | (3) |
| New coach | $-0.084$ | 0.004 | 0.060 |
| | (0.121) | (0.075) | (0.066) |
| Dissimilar new coach | $0.247^{*}$ | $0.195^{**}$ | 0.117 |
| | (0.147) | (0.090) | (0.084) |
| Home advantage | $1.080^{***}$ | $0.985^{***}$ | $0.894^{***}$ |
| | (0.253) | (0.249) | (0.259) |
| Club ability | $0.036^{***}$ | $0.028^{***}$ | $0.030^{***}$ |
| | (0.010) | (0.007) | (0.007) |
| Opponent club ability | $-0.026^{*}$ | $-0.028$ | $-0.014$ |
| | (0.014) | (0.020) | (0.021) |
| Constant | $1.286^{***}$ | $1.982^{***}$ | $1.873^{***}$ |
| | (0.415) | (0.211) | (0.242) |
| Observations | 6,375 | 6,375 | 6,375 |
| Log Likelihood | -8,557.107 | -7,595.530 | -7,228.201 |
| Akaike Inf. Crit. | 17,140.210 | 15,217.060 | 14,482.400 |

*Notes:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01. Robust standard errors in parentheses. The coefficients for the covariates from treatment assignment model are not reported.

## 4.4 Robustness exercise

We conducted two groups of robustness exercises, attempting to address the relevance of the PS specification and the uncertainty that stems from using a two-step procedure in the final ATE estimation. Regarding the first aim, the previous literature (Millimet and Tchernis, 2009; Millimet et al., 2010) and our simulation exercise in Section 2 suggest the benefits of using an approach that does not impose high penalties for including variables. However, we check the sensitivity of our ATE estimation to four alternative model specification techniques. The first two utilise two machine learning approaches, Lasso and Gradient Boosting Machine (GBM), respectively, based on their parsimony and flexibility. The other two methods explore the relevance of relying upon the lowest AIC as a stopping criterion in the stepwise regression. Although we prefer the AIC over the BIC alternative because the former uses lower penalty terms for additional parameters, we appraise the importance of this decision by estimating the PS by (1) using the three models with the lowest AIC in the stepwise procedure and weighting the respective predicted values by the Akaike weight (Burnham and Anderson, 2004) and (2) using the lowest BIC as a stopping criterion in the stepwise approach.

The second exercise appraises the uncertainty generated due to PSW involving two estimation steps. This means that any misspecification error in the PS estimation will affect the estimation of the ATE. To take this

into consideration, we obtain a bootstrap estimation using 1,000 replicate datasets of the original sample size (10,344 club-match observations) generated by random sampling with replacements. The estimation of PS and ATE is repeated for each replicate dataset and we obtain the mean value of the ATE and the associated standard deviation to estimate standard errors.

Table 8 shows the ATE estimation under the original and the five alternative strategies described in this section. It can be observed that the strategy to select the PS model does not alter the qualitative results about the influence of a new head coach on performance. The causal effect estimates of the fourteen head coach characteristics were also qualitatively similar across the different approaches but are not reported for the sake of conciseness.

Table 8: Robustness check. Estimate of the ATE of *New coach* using different methods

| | *Dependent variable:* Points | | | | | |
|---|---|---|---|---|---|---|
| | 5 matches | | 10 matches | | rest of season | |
| | ATE | SE | ATE | SE | ATE | SE |
| Stepwise (AIC)$^{(I)}$ | 0.139 | (0.104) | 0.184* | (0.097) | 0.180** | (0.090) |
| Lasso$^{(II)}$ | 0.173 | (0.108) | 0.212** | (0.096) | 0.212** | (0.093) |
| GBM$^{(III)}$ | 0.105* | (0.063) | 0.114** | (0.054) | 0.107** | (0.052) |
| Stepwise (BIC)$^{(IV)}$ | 0.118 | (0.097) | 0.157* | (0.091) | 0.157* | (0.086) |
| Stepwise (Akaike weights)$^{(V)}$ | 0.155* | (0.094) | 0.199** | (0.085) | 0.191** | (0.080) |
| Bootstrap$^{(VI)}$ | 0.115 | (0.072) | 0.142** | (0.067) | 0.140** | (0.064) |

*Notes:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01. PSs are estimated using the stepwise regression with the lowest AIC as a stopping criterion (I), Lasso (II), GBM (III), the stepwise regression with the lowest BIC as a stopping criterion (IV). For (V), PS is obtained using three specifications with the lowest AICs in the stepwise procedure and computing weighted sum of the predicted values by Akaike weights explained in Burnham and Anderson (2004). For (VI), the ATE and SE are based on the mean and standard deviation of ATEs obtained using 1,000 bootstrap samples for which the estimations of PS (the stepwise regression with the lowest AIC as a stopping criterion) and ATE are repeated.

## 4.5 Discussion

Estimation results reported in Tables 6 indicate that replacement of a head coach has, on average, a positive impact on subsequent performance once we correct for the probability of treatment using a double robust approach. Moreover, we show how to extend the standard binary analysis by decomposing a head coach replacement into changes in different managerial attributes between the old and the new manager in a way that we can assess their separate impact.

The example shows that taking into account the differences between the new and dismissed coaches does provide further insights into the effectiveness of leadership change. For example, when a new manager has a stronger association with the club, indicated by the manager having finished his playing career at the club, this can positively influence post-succession performance. The negative (and significant in the short term at conventional levels) coefficients on *Former vice coach* imply that internal succession is expected to worsen a club's performance. This can be partly explained by the view that the internal succession may involve more minor strategic change due to the cognitive and psychological attachment to the existing strategy (Farah

et al., 2020). The analysis also shows that appointing a new head coach who had not been in employment as a coach in the preceding season could be effective. Recent absence could be a desirable managerial characteristic since engaging in activities outside coaching and reflecting on their working methods may help them adopt a broader perspective.

To appraise the magnitude of the estimation results, they can be compared with the marginal effects of the club strength variables. According to the last column of Table 6, the estimated effect of a new manager who had not been on managerial duty in the immediately preceding season is $\hat{\beta}_{Absent\ last\ season} = 0.288$ per match in the post-succession period. This is, relative to the effect of a one unit increase in our ability measure, $\hat{\beta}_{Club\ ability} = 0.043$, $0.288/0.043 = 6.698$ times larger. Recall that our measure of a club's ability takes the minimum and maximum values of 1 and 20, respectively, and is increasing in their strength. Suppose a club's ability is at the 1st quartile of the distribution (= 5.75). Then, the change is equivalent to an increase in the club's ability of $5.75 + 6.698 = 12.448$, which is in between the 2nd and 3rd quartiles. That is, the magnitude is approximately equal to a move from the bottom 25% in terms of the ranking in the league, to between the top 25% and 50%. Similarly, the expected effect of a new manager with a strong association with the club (having finished their career as a player at the club) relative to the marginal effect of *Club ability* is $0.529/0.043 = 12.302$. Therefore, the impact of the change is similar to that associated with an improvement in the measure of a club's ability from 5.75 to (the 1st quartile) to 18.052 (above the 3rd quartile = 15.25).

# 5 Causal analysis in qualitative research designs

Proponents of qualitative methods have argued the advantages of studying causality using case studies (Maxwell, 2013, 2012). In this respect, in an enlightened discussion, Maxwell (2012) distinguishes between two original philosophical approaches to causation. On the one hand, the qualitative view builds on the "regularity" or "successionist" theory of causality. Thus, it accepts that we cannot directly perceive causal relationships but only observe a conjunction of events. This theory contrasts with a quantitative method based on a theoretical estimation of the relationship between two variables. This approach studies how a change in the first (independent variable) is followed by a change in the second (dependent variable). While the previous dichotomy diminishes the role of qualitative research in causal analysis, Maxwell (2012) advocates an alternative view in which causal analysis focuses on identifying the (observed) consequences of causal variables that resulted in a specific outcome in a particular context (see, for example, Little (2010)). Therefore, although qualitative analysis may lack a particular elaborated design, its advantage is that it is conducted as a reflexive process operating through every stage of the project (Hammersley and Atkinson, 2019; Maxwell, 2013).

Despite the discussion above, Antonakis et al. (2010) criticise qualitative methods in causal analysis because, unlike laboratory experiments, case studies do not have complete control of other variables affecting the outcome of the case being studied. In this regard, synthetic control methods (SCM), developed in Abadie and Gardeazabal (2003) and Abadie et al. (2010, 2015), provide a way to bridge quantitative and qualitative designs. The general idea of SCM is that, to assess the causal effects of, for example, policy decisions (California's Tobacco Control Program in Abadie et al. (2010)) or events (terrorist acts in Abadie and Gardeazabal (2003)), it is necessary to compare the studied case with a relevant counterfactual. This counterfactual is intended to show the hypothetical state of the institution being analysed if it had not been subject to treatment. However, the observations in the control group do not provide a suitable counterfactual because they can be different from the treated organisation in ways that are relevant to outcomes, which can bias the analysis. Thus, SCM uses a data-driven procedure for constructing a synthetic counterfactual case by combining observations in the control group based on their similarity with the case of interest. Then, estimating the treatment effect is simply a matter of comparing the treated institution with the synthetic counterfactual. Based on one comparison, SCM does not allow for classical inference. However, it is still possible to perform falsification tests estimating the impact of treatment on populations that are not affected by treatment.

The philosophy of SCM is similar to PSM in the particular case where there is only one treated observation to match. Thus, SCM does not estimate propensity scores as the treated individual is determined in the natural experiment, however, similarly to PSM, it matches treated and control observations based on observable variables. The use of SCM can overcome the criticism of qualitative research in Antonakis et al. (2010) by comparing the case of interest to a synthetic counterfactual case. However, it should be noted that SCM is only applied where there is a natural experiment, while qualitative research can generally study the consequences of endogenous behaviour and decisions. PSA could still complement rigorous case studies by providing general estimates of causal effects in the latter case. For example, our tutorial case studies the separate impact of different types of characteristic changes associated with managerial turnover. In doing that, we control for the different variables that explain the decisions to replace a head coach. Moreover, the contribution of each managerial characteristic is studied in a ceteris paribus analysis. This type of research could complement case studies that reflect on the consequences of a particular type of managerial change. Overall, while qualitative studies reflect on the effects of actions in a specific case and context, PSA can show whether these results can be generalised to other settings.

# 6 Lessons, limitations and implications for future research

Randomised control trials could be unfeasible or even non-natural when empirical research involves the analysis of behaviour, emotions or decisions. In this paper we explain how to conduct a PSA and discuss the implementation of this approach in recent papers in the leadership literature. PSA is illustrated with a tutorial case that estimates the causes and consequences of head coach changes in Italian football. The example presented in this paper also illustrates how to extend the analysis to estimate how different types of managerial dismissal affect post-succession performance. The tutorial approach is simple as it only requires the use of propensity scoring in weighted regression. It can also be easily adapted to study how managerial turnover operationalises, such as different leader-characteristic changes.

Although the example presented is specific to the sports industry, the particular nature of professional sports facilitates tackling internal validity concerns typically present in causal analysis. Giambatista et al. (2005) noted that while it is unclear whether results for specific sectors could be generalised elsewhere, non-sports contexts in the literature are also concentrated in very specialised settings such as manufacturing enterprises. Therefore, given the advantages of transparency in organisational objectives and measures of performance, they recommend researchers to continue exploiting sports data to investigate issues around managerial succession. We hope this tutorial contributes to incentivising the use of sports data in future management and leadership research.

Three future lines of research can be proposed based on this study. The first possibility is to employ more advanced methodologies such as machine learning (Doornenbal et al., 2021) for causal analysis. PS estimated with a machine learning approach can be easily integrated into the estimation process described in the tutorial without the need to wait for statistical packages that include the new methods in the matching algorithm. A second possibility is to explore further how managerial change is operationalised. Thus, future research could study, for example, how changes in head coach characteristics interact with organisational and environmental attributes or extend the set of managerial characteristics to include charisma indicators (Tur et al., 2021). A third possible future line of research is to use PSA to explore critical questions in the leadership literature, such as the effect of awards on performance or the impact of different types of leader decisions. The joint consideration of PSA and sports data seem, in principle, a promising avenue for future research.

# Appendix

Table 9: OLS estimates of outcome models

| | *Dependent variable:* | | |
|---|---|---|---|
| | Points 5 matches | Points 10 matches | Points rest of season |
| | (1) | (2) | (3) |
| New coach | −0.051 | −0.067 | −0.063 |
| | (0.053) | (0.047) | (0.044) |
| Home advantage | 0.594*** | 0.609*** | 0.648*** |
| | (0.044) | (0.045) | (0.045) |
| Club ability | 0.051*** | 0.051*** | 0.050*** |
| | (0.001) | (0.001) | (0.001) |
| Opponent club ability | −0.042*** | −0.042*** | −0.041*** |
| | (0.002) | (0.003) | (0.003) |
| Constant | 0.955*** | 0.936*** | 0.885*** |
| | (0.034) | (0.036) | (0.036) |
| Observations | 10,344 | 10,344 | 10,344 |
| $R^2$ | 0.215 | 0.249 | 0.261 |
| Adjusted $R^2$ | 0.215 | 0.248 | 0.261 |
| Residual Std. Error (df = 10339) | 0.654 | 0.582 | 0.551 |
| F Statistic (df = 4; 10339) | 709.474*** | 855.274*** | 912.136*** |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

# References


Abadie, A., Diamond, A., , and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's Tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative Politics and the Synthetic Control Method. *American Journal of Political Science*, 59(2):495–510.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque country. *American Economic Review*, 93(1):113–132.

Antonakis, J., Bendahan, S., Jacquart, P., and Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21(6):1086–1120.

Audas, R., Dobson, S., and Goddard, J. (2002). The impact of managerial change on team performance in professional sports. *Journal of Economics and Business*, 54(6):633–650.

Austin, P. C. and 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.

Bechtoldt, M. N., Bannier, C. E., and Rock, B. (2019). The glass cliff myth? – Evidence from Germany and the U.K. *The Leadership Quarterly*, 30(3):273–297.

Berns, K. V. and Klarner, P. (2017). A Review of the CEO succession Literature and a Future Research Program. *Academy of Management Perspectives*, 31(2):83–108.

Boivie, S., Graffin, S. D., Oliver, A. G., and Withers, M. C. (2016). Come Aboard! Exploring the Effects of Directorships in the Executive Labor Market. *Academy of Management Journal*, 59(5):1681–1706.

Bolton, P., Brunnermeier, M. K., and Veldkamp, L. (2013). Leadership, coordination, and corporate culture. *Review of Economic Studies*, 80(2):512–537.

Branco, P., Torgo, L., and Ribeiro, R. P. (2017). A survey of predictive modelling under imbalanced distributions. *ACM Computing Surveys*, 49(2):1–50.

Bridgewater, S., Kahn, L. M., and Goodall, A. H. (2011). Substitution and complementarity between managers and subordinates: Evidence from British football. *Labour Economics*, 18(3):275–286.

Bruce, P. and Bruce, A. (2017). *Practical statistics for data scientists*. O'Reilly, California, USA.

Bryson, A., Buraimo, B., Farnell, A., and Simmons, R. (2021). Time To Go? Head Coach Quits and Dismissals in Professional Football. *De Economist*, 169(1):81–105.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33(2):261–304.

Caliendo, M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72.

Carton, A. M., Murphy, C., and Clark, J. R. (2014). A (blurry) vision of the future: How leader rhetoric about ultimate goals influences performance. *Academy of Management Journal*, 57(6):1544–1570.

Chen, G. (2015). Initial compensation of new CEOs hired in turnaround situations. *Strategic Management Journal*, 36(12):1895–1917.

Connelly, B. S., Sackett, P. R., and Waters, S. D. (2013). Balancing Treatment and Control Groups in Quasi-Experiments: An Introduction to Propensity Scoring. *Personnel Psychology*, 66(2):407–442.

Cook, T. and Campbell, D. (1976). The design and conduct of true experiments and quasi-experiments in field settings. In Dunnette, M., editor, *Handbook of Industrial and Organizational Psychology*, pages 223–326. Rand McNally, Chicago.

Cook, T. D., Campbell, D. T., and Shadish, W. (2002). Quasi-experimental designs that either lack a control group or lack pretest observations on the outcome. In *Experimental and quasi-experimental designs for generalized causal inference*, chapter 4, pages 103–134. Houghton Mifflin, Boston, MA, 2nd edition.

Crano, W. D., Brewer, M. B., and Lac, A. (2014). *Principles and methods of social research*. Routledge, New York, 3rd edition.

Dawson, P. and Dobson, S. (2002). Managerial efficiency and human capital: An application to English association football. *Managerial and Decision Economics*, 23(8):471–486.

de Jong, A. and Naumovska, I. (2016). A note on event studies in finance and management research. *Review of Finance*, 20(4):1659–1672.

DeFond, M., Erkens, D. H., and Zhang, J. (2017). Do client characteristics really drive the big N audit quality effect? New evidence from propensity score matching. *Management Science*, 63(11):3531–3997.

der Laan, M. J. V., Pollard, K. S., and Bryan, J. (2003). A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, 73(8):575–584.

Detotto, C., Paolini, D., and Tena, J. D. (2018). Do managerial skills matter? An analysis of the impact of managerial features on performance for Italian football. *Journal of the Operational Research Society*, 69(2):270–282.

Devarakonda, S. V. and Reuer, J. J. (2018). Knowledge sharing and safeguarding in R&D collaborations: The role of steering committees in biotechnology alliances. *Strategic Management Journal*, 39(7):1912–1934.

Diamond, A. and Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945.

Dixon, M. J. and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 46(2):265–280.

Doornenbal, B. M., Spisak, B. R., and van der Laken, P. A. (2021). Opening the black box: Uncovering the leader trait paradigm through machine learning. *The Leadership Quarterly*, 101515, In press.

Farah, B., Elias, R., De Clercy, C., and Rowe, G. (2020). Leadership succession in different types of organizations: What business and political successions may learn from each other. *The Leadership Quarterly*, 31(1):101289.

Fest, S., Kvaløy, O., Nieken, P., and Schöttner, A. (2021). How (not) to motivate online workers: Two controlled field experiments on leadership in the gig economy. *The Leadership Quarterly*, 32(6):101514.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Fong, C., Hazlettand, C., and Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *Annals of Applied Statistics*, 12(1):156–177.

Freedman, D. A. and Berk, R. A. (2008). Weighting regressions by propensity scores. *Evaluation Review*, 32(4):392–409.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767.

Giambatista, R. C., Rowe, W. G., and Riaz, S. (2005). Nothing succeeds like succession: A critical review of leader succession literature since 1994. *The Leadership Quarterly*, 16(6):963–991.

Guo, S. G. and Fraser, M. W. (2014). *Propensity score analysis: Statistical Methods and Applications*. SAGE Publications, Inc, Thousand Oaks, CA, 2nd edition.

Gupta, V. K., Han, S., Mortal, S. C., Silveri, S., and Turban, D. B. (2017). Do women CEOs face greater threat of shareholder activism compared to male CEOs? A role congruity perspective. *Journal of Applied Psychology*, 103(2):228–236.

Gupta, V. K., Mortal, S., Chakrabarty, B., Guo, X., and Turban, D. B. (2020). CFO gender and financial statement irregularities. *Academy of Management Journal*, 63(3):802–831.

Hammersley, M. and Atkinson, P. (2019). *Ethnography: Principles in practice*. Routledge, New York, 4th edition.

Heckman, J. J., Lalonde, R. J., and Smith, J. A. (1999). The Economics and Econometrics of Active Labor Market Programs. In *Handbook of Labor Economics, Volume 3, Part A*, pages 1865–2097. Elsevier.

Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. In Gelman, A. and Meng, X.-L., editors, *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, pages 73–84. John Wiley & Sons, Ltd.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Hopp, C. and Pruschak, G. (2020). Is there such a thing as leadership skill? – A replication and extension of the relationship between high school leadership positions and later-life earnings. *The Leadership Quarterly*, 101475, In press.

Hughes, M., Hughes, P., Mellahi, K., and Guermat, C. (2010). Short-term versus Long-term Impact of Managers: Evidence from the Football Industry. *British Journal of Management*, 21(2):571–589.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

Kaplan, S. N., Klebanov, M. M., and Sorensen, M. (2012). Which CEO Characteristics and Abilities Matter? *Journal of Finance*, 67(3):973–1007.

Kiss, A. N., Cortes, A. F., and Herrmann, P. (2021). CEO proactiveness, innovation, and firm performance. *The Leadership Quarterly*, 101545, In press.

Li, M. (2013). Using the Propensity Score Method to Estimate Causal Effects: A Review and Practical Guide. *Organizational Research Methods*, 16(2):188–226.

Li, W. D., Li, S., Feng, J. J., Wang, M., Zhang, H., Frese, M., and Wu, C. H. (2021). Can becoming a leader change your personality? An investigation with two longitudinal studies from a role-based perspective. *Journal of Applied Psychology*, 106(6):882–901.

Little, D. (2010). *New Contributions to the Philosophy of History*. Springer, New York.

Love, E. G., Lim, J., and Bednar, M. K. (2017). The face of the firm: The influence of CEOs on corporate reputation. *Academy of Management Journal*, 60(4):1462–1481.

Luo, X. R., Zhang, J., and Marquis, C. (2016). Mobilization in the internet age: Internet activism and corporate response. *Academy of Management Journal*, 59(6):2045–2068.

Maxwell, J. A. (2012). The importance of qualitative research for causal explanation in education. *Qualitative Inquiry*, 18(8):655–661.

Maxwell, J. A. (2013). *Qualitative research design: an interactive approach*. SAGE Publications, Thousand Oaks, Calif., 3rd edition.

Millimet, D. L. and Tchernis, R. (2009). On the specification of propensity scores, with applications to the analysis of trade policies. *Journal of Business and Economic Statistics*, 27(3):397–415.

Millimet, D. L., Tchernis, R., and Husain, M. (2010). School nutrition programs and the incidence of childhood obesity. *Journal of Human Resources*, 45(3):640–654.

Morgan, S. L. and Todd, J. J. (2008). 6. A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects. *Sociological Methodology*, 38(1):231–282.

Muehlheusser, G., Schneemann, S., and Sliwka, D. (2016). The impact of managerial change on performance: The role of team heterogeneity. *Economic Inquiry*, 54(2):1128–1149.

Ong, W. J. (2021). Gender-contingent effects of leadership on loneliness. *Journal of Applied Psychology*, Advance online publication.

Podsakoff, P. M. and Podsakoff, N. P. (2019). Experimental designs in management and leadership research: Strengths, limitations, and recommendations for improving publishability. *The Leadership Quarterly*, 30(1):11–33.

Raad, H., Cornelius, V., Chan, S., Williamson, E., and Cro, S. (2020). An evaluation of inverse probability weighting using the propensity score for baseline covariate adjustment in smaller population randomised controlled trials with a continuous outcome. *BMC Medical Research Methodology*, 20(1):1–12.

Ridgeway, G., McCaffrey, D., Morral, A., Cefalu, M., Burgette, L., Pane, J., and Griffin, B. A. (2021). Toolkit for weighting and analysis of nonequivalent groups: A guide to the twang package. *vignette, July*, 26.

Rocha, V. and Van Praag, M. (2020). Mind the gap: The role of gender in entrepreneurial career choice and social influence by founders. *Strategic Management Journal*, 41(5):841–866.

Rockey, J. C., Smith, H. M., and Flowe, H. D. (2021). Dirty looks: Politicians' appearance and unethical behaviour. *The Leadership Quarterly*, 33(2):101561.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):42–52.

Rosenbaum, P. R. and 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. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, 39(1):33–38.

Rowe, W. G., Cannella, A. A., Rankin, D., and Gorman, D. (2005). Leader succession and organizational performance: Integrating the common-sense, ritual scapegoating, and vicious-circle succession theories. *The Leadership Quarterly*, 16(2):197–219.

Schmidt, J. A. and Pohler, D. M. (2018). Making stronger causal inferences: Accounting for selection bias in associations between high performance work systems, leadership, and employee and customer satisfaction. *Journal of Applied Psychology*, 103(9):1001–1018.

Shi, W., Zhang, Y., and Hoskisson, R. E. (2019). Examination of CEO–CFO social interaction through language style matching: Outcomes for the CFO and the organization. *Academy of Management Journal*, 62(2):383–414.

Stone, C. A. and Tang, Y. (2013). Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations. *Practical Assessment, Research and Evaluation*, 18(13):1–12.

Sy, T., Horton, C., and Riggio, R. (2018). Charismatic leadership: Eliciting and channeling follower emotions. *The Leadership Quarterly*, 29(1):58–69.

Tena, J. D. and Forrest, D. (2007). Within-season dismissal of football coaches: Statistical analysis of causes and consequences. *European Journal of Operational Research*, 181(1):362–373.

Thoemmes, F. and Ong, A. D. (2016). A primer on inverse probability of treatment weighting and marginal structural models. *Emerging Adulthood*, 4(1):40–59.

Thoemmes, F. J. and Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate behavioral research*, 46(1):90–118.

Tur, B., Harstad, J., and Antonakis, J. (2021). Effect of charismatic signaling in social media settings: Evidence from TED and Twitter. *The Leadership Quarterly*, 101476.

van Ours, J. C. and van Tuijl, M. A. (2016). In-season head-coach dismissals and the performance of professional football teams. *Economic Inquiry*, 54(1):591–604.

Vitanova, I. (2021). Nurturing overconfidence: The relationship between leader power, overconfidence and firm performance. *The Leadership Quarterly*, 32(4):101342.

Wofford, J. C. (1999). Laboratory research on charismatic leadership: Fruitful or futile? *The Leadership Quarterly*, 10(4):523–529.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press, London.

Zhang, Z., Zhang, B., and Jia, M. (2021). The military imprint: The effect of executives' military experience on firm pollution and environmental innovation. *The Leadership Quarterly*, 33(2):101562.

Zheng, W., Singh, K., and Chung, C. N. (2017). Ties to unbind: Political ties and firm sell-offs during institutional transition. *Journal of Management*, 43(7):2005–2036.