
Essays in Economics of Managers

Insights from Professional Football Leagues

By

KAORI NARITA

Department of Economics
UNIVERSITY OF LIVERPOOL MANAGEMENT SCHOOL

Thesis submitted in accordance with the requirements
of the University of Liverpool for the degree of DOCTOR
IN PHILOSOPHY.

SEPTEMBER 2022



THE UNIVERSITY
of LIVERPOOL

Essays in Economics of Managers

Insights from Professional Football Leagues

Kaori Narita

Abstract

This thesis presents a collection of papers that address issues related to leadership succession and the role of managers in firm production by employing data from professional football leagues. Professional sports produce valuable data that can be seen as an instrument to test economic and management theories. Leadership research can particularly benefit from such data since the role played by a manager in professional sports clubs are similar to a leader's role in a more general context, such as that of CEOs in corporations.

However, an important methodological concern related to observational studies is the fact that events such as leadership succession do not occur randomly. One of the remedies for this problem is a propensity score analysis. The use of the method is yet sparse in leadership research and related fields despite its applicability. Chapter 2, therefore, discusses the method and demonstrates how to implement the method using the real-world example of head coach changes in Italian professional football.

The previous studies in leadership succession in professional sports primarily focus on establishing whether managerial replacements improve a club's performance. However, conditions under which such decisions can bring about favourable outcomes are not well understood. Therefore, the empirical example included in Chapter 2 adds to the previous analysis by viewing a leadership change as simultaneous changes in leader characteristics, such as their age, experience, and association with the organisation. The study finds that differences between new and dismissed managers in certain characteristics do affect the effectiveness of managerial succession.

Furthermore, Chapter 3 analyses the decision to replace a manager more than once within a season in order to understand whether the frequency of replacements can determine the effectiveness of such decisions. In particular, we separately estimate the causes and consequences of the first and second dismissals of a manager in a given season by employing machine learning and the inverse propensity score weighting method. The results suggest that clubs may benefit from the first replacement, whilst the second replacement has no significant effect on performance, despite the latter decision being made more cautiously.

Chapter 4 diverts from the specific issues related to leadership succession and examines

the role of leaders themselves in firm production. To do so, we estimate a football club's production function, where labour and capital inputs are explicitly measured using individual players' historical performance and a club's estimated transfer budgets. This allows us to quantify how much an individual manager adds to a club's performance, given the resources at his disposal. In addition, we take into account the randomness of the outcome by employing event data within individual matches. Our analysis shows that having different managers can have both economically and statistically significant impacts on a club's outcomes.

Declaration

I declare that the thesis has been composed by myself and that the work has not be submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored papers has been included. My contribution and those of the other authors to this work have been explicitly indicated within the authorship declaration forms attached at the end of the thesis. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to Dr Juan De Dios Tena Horrillo and Dr Babatunde Buraimo for their supervision, encouragement, and patience throughout my PhD studies, without which production of this thesis would not have been possible. This thesis has also hugely benefited from valuable advice from Professor David Forrest, Professor Oliver De Groot, Professor Ian McHale, and Dr Ian Burn, to whom I am very grateful. Financial assistance from the University of Liverpool Management School and the Economic and Social Research Council (ESRC) is gratefully acknowledged.

I am also thankful to all my friends and PhD colleagues – Anneke, Francis, Sheana, Yigit, Hyacinthe, Ruoxi, Tien, and Carol, for their encouragement and support all through my studies. Finally, I must thank my family, Ben, and his family. Without their unwavering support and belief in me, it would have been impossible to complete my study.

Contents

1	Introduction	2
2	Causal Inference with Observational Data: A Tutorial on Propensity Score Analysis	7
2.1	Introduction	7
2.2	Principles of propensity score analysis	10
2.2.1	The strong ignorability assumption	10
2.2.2	Steps in the analysis	11
2.2.3	A simulation example	14
2.2.4	Other validity concerns	17
2.3	PSA in management and psychology research	19
2.4	A tutorial on PSW: Leadership succession effects	28
2.4.1	Data	30
2.4.2	Methodology	34
2.4.3	Extension: endogeneity of similarity in coach characteristics	44
2.4.4	Robustness exercise	46
2.4.5	Discussion	47
2.5	Causal analysis in qualitative research designs	48
2.6	Lessons, limitations and implications for future research	50
3	Causes and Consequences of Recurrent Managerial Changes	52
3.1	Introduction	52
3.2	Related literature and hypotheses	54
3.2.1	Causes of dismissal	54
3.2.2	Consequences of dismissal	56
3.3	Data	58
3.3.1	The treatment variable and treatment groups	58
3.3.2	Determinants of dismissals	61

3.3.3	Outcome and control variables	66
3.4	Methodology	67
3.4.1	Estimation of outcome model	68
3.4.2	Estimation of treatment assignment models	69
3.4.3	Imbalancedness of classes	70
3.5	Results	72
3.5.1	Predictors of single and multiple dismissals	73
3.5.2	Estimation of propensity scores and covariate balance	76
3.5.3	Average treatment effects (ATE) on field performance	79
3.6	Conclusion	83
4	Rating Football Managers with Match-Day Contribution to Performance	88
4.1	Introduction	88
4.2	Related literature	91
4.3	Data	94
4.3.1	Match and manager data	94
4.3.2	Event data	96
4.3.3	WhoScored rating	97
4.4	Methodology	98
4.4.1	xG model	99
4.4.2	Adjustments of WS ratings	100
4.4.3	Production function	102
4.5	Results	103
4.5.1	xG and naive rankings	103
4.5.2	League coefficients and league adjusted player ranking	108
4.5.3	Estimated production function	111
4.5.4	Estimated manager coefficients	113
4.5.5	Case study	118
4.6	Conclusion	120
5	Conclusion	122
	Bibliography	126

List of Figures

2.1	Steps in propensity score analysis	14
2.2	Precision-recall curve and area under the curve	38
2.3	Kernel density of log propensity scores before/after weighting	39
2.4	Mean value of outcome variable (<i>Points rest of season</i>) for each ability and treatment group	41
2.5	Nested logit model	45
3.1	Kernel density of <i>Cumulative surprise</i> in treated and control groups	62
3.2	Standardised mean differences (SMD) in predictors in full and SMOTE samples	72
3.3	Relative importance of predictors with non-zero influences	76
3.4	Standardised mean differences of selected covariates before and after weighting	79
3.5	ATE under different estimation strategies	83
3.6	Receiver operating characteristics (ROC) curves and area under the curve (AUC)	85
3.7	Covariate balance with logistic regression	86
4.1	Transfer budget for selected clubs (2014/2015-2020/2021)	95
4.2	Estimated xG values for home and away clubs	105
4.3	Manager coefficients with 95% confidence intervals	115
4.4	Naive rankings and manager fixed effects	117
4.5	Cumulative expected points (xP) for Everton F.C. (2020/2021)	119

List of Tables

2.1	Estimation of average treatment effects (ATE) using Monte Carlo simulation. (True ATE = 4.5)	16
2.2	Examples of PSA and matching for causal studies in the management and leadership literature	20
2.3	Summary statistics of differences in managerial characteristics	32
2.4	Stepwise regression ^a results for treatment assignment	37
2.5	Covariate balance table (before/after weighting, all/common support)	40
2.6	Double robust estimates of outcome models	42
2.7	Double robust estimates of outcome model with dissimilar treatment	46
2.8	Robustness check. Estimate of the ATE of <i>New coach</i> using different methods	47
2.9	OLS estimates of outcome models	51
3.1	Number of treated and control units by seasons	61
3.2	Descriptive statistics of treatment predictors	65
3.4	The number of units in full and SMOTE samples	71
3.5	Logit estimates of treatment assignment models	75
3.6	Summary of estimated propensity scores	78
3.7	PSW estimates: ATE of single dismissal on points and goal differences . . .	81
3.8	PSW estimates: ATE of multiple dismissal on points and goal differences . .	81
3.9	Balanced accuracy of classification models	85
3.10	OLS estimates: ATE of single dismissal on points and goal differences	87
3.11	OLS estimates: ATE of multiple dismissal on points and goal differences . .	87
4.1	Number of matches used in the analysis by leagues	94
4.2	Descriptive statistics for shot data	96
4.3	Number of WS ratings by leagues	98
4.4	Estimated xG model parameters	104
4.5	Naive rankings (winning percentage and xG differential)	107

4.6	Estimated league coefficients	108
4.7	Weighted average of WS and LAWS ratings	110
4.8	Estimated parameters for production functions	112
4.9	F-test for significance of manager effects	112
4.10	Estimated manager coefficients	114
4.11	Estimated parameters for ordered logit model	118

Chapter 1

Introduction

The neoclassical theory of firms undermines the role of managers by assuming them as homogeneous inputs. This implies that with given levels of other inputs and technology, a firm's output will remain constant at the profit maximising level, whoever manages the firm. However, this view has been challenged by empirical findings, such as a study by Bertrand and Schoar (2003). Their findings are one of the earliest to provide empirical evidence that the presence of different CEOs can explain the heterogeneity in firm productivity to a great extent. Further studies in management and economics research, including Siebert and Zubanov (2010); Bloom et al. (2014, 2013), agree that managerial inputs are important to determine firm performance.

The relationship between a firm owner and managers is often described in the principal-agent model, where the former delegates day-to-day management tasks to the latter, who are typically more specialised in such tasks and have superior knowledge in the respective industry. However, the nature of managerial tasks, such as monitoring and motivating workers, makes the evaluation of managers rather challenging since these are not easily quantified. A hidden information problem is also present in the sense that it is not easy to disentangle the contributions of different inputs. Some managers may seem to be more competent than others, however, this may as well be the case that they are fortunate to have high-quality workers.

Consequently, scholars have turned to professional sports as a field to investigate economics and management issues since data produced in this industry can overcome some limitations of more conventional data. Firms, or sports clubs, in the industry pursue very similar goals, and their performance is highly publicised, easily measured, and regularly updated. Similarly, managers or head coaches, who are hired by club owners, play an identical role in maximising field performance with a given set of resources at their disposal. In this sense, their role is also comparable to that of CEOs (Frick and Simmons, 2008; Pieper et al.,

2014). For this reason, a head coach in a professional sports club is seen as a valuable instrument for studying leadership in general. For instance, a seminal work in leadership succession by Grusky (1963) is in the context of field manager changes in Major League Baseball (MLB). Myriad of studies have followed his work in the context of other sports (Rowe et al., 2005) and beyond sports (Berns and Klarner, 2017).

This thesis presents three papers that separately address various issues in leadership research and the economics of managers, using data from professional football leagues. As mentioned above, the causes and consequences of leadership succession have been investigated in the context of professional sports, including professional football. In particular, Audas et al. (1999, 2002) present the first empirical studies to identify the factors of head coach dismissals and their implications on club performance in the context of professional football. Whilst their study is based on English football leagues, subsequent studies have investigated these issues in professional football leagues in other countries, such as Spain (Tena and Forrest, 2007), Argentina (Flores et al., 2012), the Netherlands (van Ours and van Tuijl, 2016), and Germany (Muehlheusser et al., 2016).

Those are observational studies in nature, implying that some events, such as leadership succession, do not occur randomly. The authors indeed identify the significant differences between clubs that dismiss managers and those that do not in terms of field performance leading up to the dismissal (or non-dismissal) and other characteristics such as being in the relegation zone and even managerial characteristics. This is in contrast to the randomised control trial, where a “treatment” is allocated to the subjects randomly, hence any difference between the treated and control groups after the treatment assignment can be attributed to the treatment effect.

The previous studies employ different empirical strategies to cope with the non-randomness of managerial dismissal. Earlier studies, such as Audas et al. (2002) and Tena and Forrest (2007), employ multiple regression, where a post-succession performance is regressed with a set of indicators for managerial change in the preceding periods, together with the variables that capture the pre-succession performance. More recently, matching techniques have been employed to identify “counterfactual” observations, which are similar to the actual treated observations in terms of a certain pre-treatment characteristic, yet are not allocated the treatment. For instance, van Ours and van Tuijl (2016) use a variable that indicates a club’s performance relative to expectation to identify counterfactual cases to managerial dismissals. That is, they find clubs which do not change a manager despite the fact that they follow a similar path of performance (normally poor performance) to that of the clubs which dismiss a manager. Then, they compare the post-treatment performance in the actual and counterfactual cases with that of the control observations. In this case, they find that both

actual and counterfactual cases present a more favourable outcome compared to the control group. Their findings, therefore, imply that when a club is experiencing poor performance, it eventually reverts back to its mean level, with or without changing its manager.

In Chapter 2, we discuss an alternative method that can be useful in identifying a treatment effect of managerial succession and estimating a causal effect of other events/phenomena in observational studies in general. In particular, the Chapter provides a review and practical guide on propensity score analysis (PSA), with a particular focus on propensity score weighting (PSW). As the Chapter reveals, the use of this particular method is elusive despite its applicability in social science. Therefore, we build on the previous methodological reviews related to PSA by Li (2013) and Connelly et al. (2013) by presenting examples of the use of PSW in the recent leadership and management literature and demonstrating the implementation of the method using our original study in leadership succession in Italian football.

Whilst Chapter 2 serves as a methodological manual paper that could benefit researchers in social science in general, the original study incorporated in the paper aims to fill some gaps in the leadership succession literature. As mentioned earlier, to the best of our knowledge, this method is yet to be implemented to study the effectiveness of managerial dismissal in professional sports. Furthermore, we expand the analysis by looking at a change in managers as simultaneous changes in managerial characteristics, such as previous experience, connection with a managing club, and playing background. This type of analysis is scarce in the literature since existing studies primarily focus on establishing whether changing a manager can be effective or not, rather than identifying potential conditions under which managerial change can improve the situation, make no difference, or even deteriorate the situation. Our study indeed shows that changes in specific managerial characteristics can positively affect post-succession performance, suggesting that one may benefit from choosing a new manager who has particular attributes rather than simply replacing a manager with another.

Chapter 3 further explores the topic of managerial succession using the data and method introduced in Chapter 2. As mentioned above, one aspect of managerial succession research that deserves more attention is understanding the conditions under which leadership replacement can bring about a favourable outcome as opposed to estimating its average treatment effect. Such information can better support managerial replacement decisions. Whilst our study in Chapter 2 addresses this issue by identifying a new manager's characteristics relative to the dismissed that could have a positive impact on the post-succession performance, Chapter 3 analyses the role of frequency in the effectiveness of managerial changes by estimating the causes and significance of replacements that occur multiple times in a given

season. In particular, we separately identify the determinants of dismissals that happen for the first time in a given season and ones that occur following the first one within the same season and estimate the impact of each case on the subsequent performance.

Such an analysis is economically relevant in at least two ways. First, it helps us to understand how a principal (club) may adjust their expectation with respect to the agent (manager)’s contribution to the productivity of working teams, given an (unfavourable) outcome of the first replacement. Second, it allows us to understand how first and second dismissals are operationalised in a causal analysis. Given that these two decisions could be motivated by different factors, they could also affect team performance differently. Indeed, our analysis shows that the determinants of the two decisions are not identical in that the second replacement is likely to be taken with somewhat greater caution. Although limited, we find some positive effects of first dismissals on subsequent performance, whilst the second dismissals do not appear to make any difference. Given the potentially high costs of replacing a manager, our analysis suggests that another managerial replacement is not rational, even if a first replacement turns out to be ineffective.

Finally, Chapter 4 presents a paper that examines the contribution of managerial inputs to firm production. As discussed above, one of the challenges related to the evaluation of a leader is to disentangle their contributions from those of his/her subordinates as well as other inputs such as financial capital. To take this into account, previous studies have separately estimated managers’ and firms’ fixed effects. For instance, the aforementioned study by Bertrand and Schoar (2003), investigates individual CEO’s effects on different corporate performance measures by disentangling CEO’s fixed effects from those of individual firms’. Again, some researchers exploited data from professional sports to study this issue (Buzzacchi et al., 2021; Muehlheusser et al., 2018), where the contribution of managers to field success is estimated using a similar method.

This paper also utilises the data from professional football, however, we make even better use of such data by exploiting the fact that information on individual workers, i.e. players, is also available in this industry. This allows us to quantify the quality of individual players in an up-to-date manner. Using this, we employ a different approach to disentangle the contribution of managerial input from that of other inputs. In particular, we evaluate individual managers by quantifying how much they contribute to field performance given the resources (players’ quality and financial strengths) at their disposal on a particular match day. To do so, we adopted the historical performance of individual players and an indicator of the club’s financial capital, which varies over time and hence captures more accurate levels of resources at hand. Additionally, we employ a more advanced metric of field performance than those previously employed in order to better reflect a manager’s ability rather than luck. Given

that the football manager's labour market is mobile, we also expand the pool of managers by including multiple leagues rather than one. Our results suggest that the presence of different managers can explain the heterogeneity in club performance after controlling for the level of resources. They also suggest that not taking into account the player's quality and financial strength, as well as the randomness of the outcome, can result in misperceptions of an individual manager's ability. Therefore, the study contributes to the development of the evaluation of football managers while adding to the empirical evidence of managerial inputs in general.

Chapter 2

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

2.1 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 treatment¹ 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

¹The 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.

²Another 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.

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 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 2.3 presents and discusses examples of the use of PSA in recent management and applied psychology research. Section 2.4 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 2.5. Finally, we offer ideas for future work and some concluding remarks.

2.2 Principles of propensity score analysis

2.2.1 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:

$$c_i = Y_i(1) - Y_i(0). \quad (2.1)$$

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 $T_i Y_i(1) - (1 - T_i) Y_i(0)$, 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 2.2.4.

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:

$$(Y(1), Y(0)) \perp T | X, \quad (2.2)$$

$$0 < \Pr[T = 1 | X] < 1. \quad (2.3)$$

Expression (2.2) 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 (2.3) 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.

2.2.2 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 2.1. 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

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

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 2.1, 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 2.1 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. 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 the impact of treatment on the whole weighted sample, PSW provides an

⁴Weights 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.

estimate of ATE.

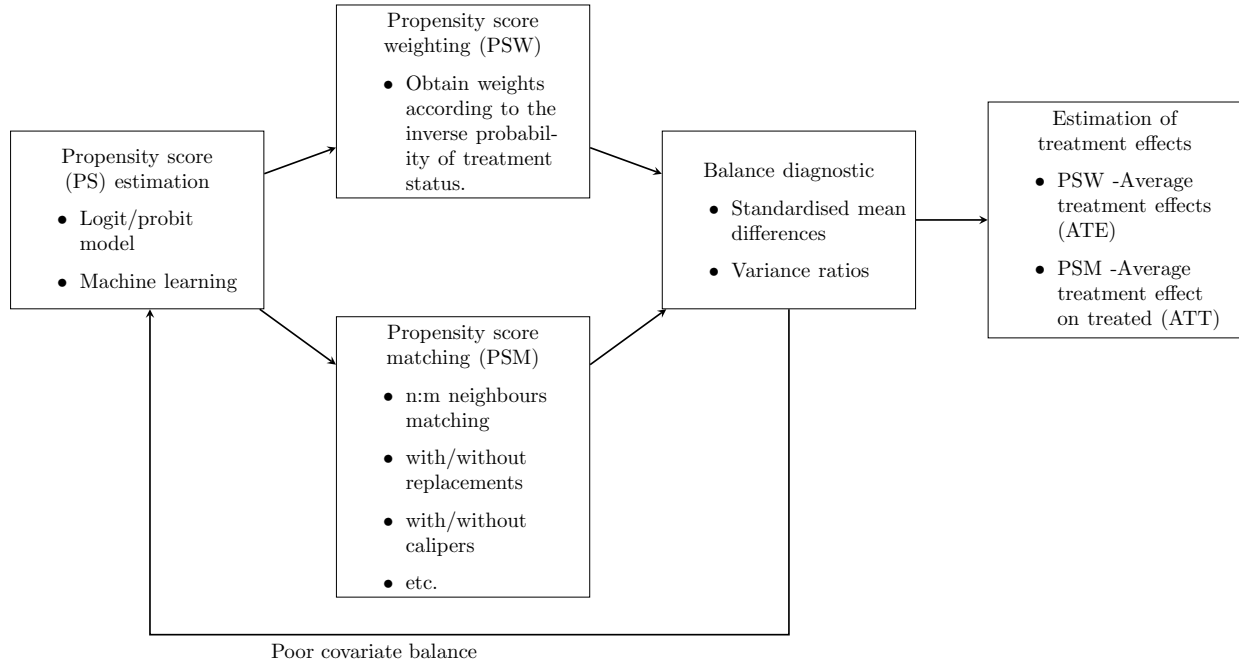


Figure 2.1: Steps in propensity score analysis

2.2.3 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 γ_i according to the following function:

$$Y_i = c_i T_i + \gamma_i, \quad (2.4)$$

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 2.1.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 2.1: Estimation of average treatment effects (ATE) using Monte Carlo simulation. (True ATE = 4.5)

A. ATE estimates with known treatment assignment probabilities

Observations	Randomised assignment				Endogenous Assignment				Inverse PSW			
	Median ATE	MCE ^(I)	%Bias ^(II)	95% Cov- erage ^(III)	Median ATE	MCE ^(I)	%Bias ^(II)	95% Cov- erage ^(III)	Median ATE	MCE ^(I)	%Bias ^(II)	95% Cov- erage ^(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) $\%Bias = \frac{1}{R} \sum_{r=1}^R \frac{\widehat{ATE}_{Rr} - ATE}{ATE} * 100$; (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

Observations	Relevant covariates				Non informative covariates				All covariates			
	Median ATE	MCE ^(I)	%Bias ^(II)	95% Cov- erage ^(III)	Median ATE	MCE ^(I)	%Bias ^(II)	95% Cov- erage ^(III)	Median ATE	MCE ^(I)	%Bias ^(II)	95% Cov- erage ^(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

Observations	Stepwise selection				Lasso selection			
	Median ATE	MCE ^(I)	%Bias ^(II)	95% Cov- erage ^(III)	Median ATE	MCE ^(I)	%Bias ^(II)	95% Cov- erage ^(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) $\%Bias = \frac{1}{R} \sum_{r=1}^R \frac{\widehat{ATE}_{Rr} - ATE}{ATE} * 100$; (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 2.1.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.

2.2.4 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 2.3 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 2.4 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).

2.3 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 2022⁵. 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 2.2 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.

⁵The 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.

⁶Our 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 2.2: 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 2.2 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 operationalised 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.

2.4 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 coach⁷ 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 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 surprise⁸ 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 misspecification 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

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

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

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 2.4.3 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.

2.4.1 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 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.

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*

⁹The 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.

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 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

¹⁰During 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.

¹¹Available at <https://leaguemanagers.com/>.

¹²Available at <https://www.transfermarkt.com/>.

¹³See, 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).

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, Δh 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 2.3 provides the summary statistics of the characteristic change variables for the 157 cases of managerial change considered in the analysis.

Table 2.3: Summary statistics of differences in managerial characteristics

Variable	Difference between new and dismissed coaches (Δh)		
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 Δh 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.

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. (2021b) 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

¹⁴Available at <https://www.football-data.co.uk/>.

occur more frequently in mid-season. To capture this effect, we include round fixed effects in the treatment assignment model.

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 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.

2.4.2 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:

$$y_{its} = \delta \text{New coach}_{its} + \gamma' X_{its} + \varepsilon_{its}, \quad (2.5)$$

¹⁵More 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.

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 δ and a vector γ are parameters to be estimated. Finally, ε_{its} is a stochastic error component.

Our focus is to obtain the estimate of δ , which, if the treatment ($New\ coach_{its}$) were randomly allocated, should capture the Average Treatment Effect (ATE) of managerial change, if any. Model (2.5) 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:

$$y_{its} = \delta\ New\ coach_{its} + \beta' \Delta H_{its} + \gamma' X_{its} + \varepsilon_{its}, \quad (2.6)$$

where ΔH_{its} is a vector of the managerial characteristic change variables, and β is its associated vector of parameters. Variables in ΔH_{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 β are significantly different from zero, this implies that differences in the characteristics between outgoing and incoming managers do matter for the successful implementation of managerial change.

However, an important concern in the estimation of models (2.5) and (2.6) 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

¹⁶Including 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.

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

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 $(1 - \text{the probability of being treated})$. 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 2.4.1 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 $\hat{p}_{its} = \Pr[\text{New coach}_{its} = 1 | Z_{its}]$, then the inverse propensity score weights are defined as follows:

$$w_{its} = \begin{cases} \frac{1}{\hat{p}_{its}}, & \text{if } \text{New coach}_{its} = 1, \\ \frac{1}{1-\hat{p}_{its}}, & \text{if } \text{New coach}_{its} = 0. \end{cases} \quad (2.7)$$

These weights may now be used in the weighted regression analysis to obtain the parameter estimates of models (2.5) and (2.6); see Guo and Fraser (2014) and Morgan and Todd (2008) for the use of inverse propensity 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 2.2 and Figure 2.1 to estimate first the treatment assignment model, then the outcome models (2.5) and (2.6).

Step 1: Propensity score estimation

As described in Section 2.2, the initial step in PSW is to estimate the treatment assignment models. Following Section 2.4.2, we estimate the logistic regression with step wise selection, where we consider the set of covariates presented in Section 2.4.1 in the initial step.¹⁹ The

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

¹⁹Thoemmes 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.

set of covariates selected in the final model and estimation results are reported in Table 2.4.

Table 2.4: Stepwise regression^a results for treatment assignment

	<i>Dependent variable:</i>
	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: ^aThe 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 2.2 plots the Precision-Recall curve (PRC) and reports the 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.

²⁰Note 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 2.4.4, the final estimation of ATE remains similar.

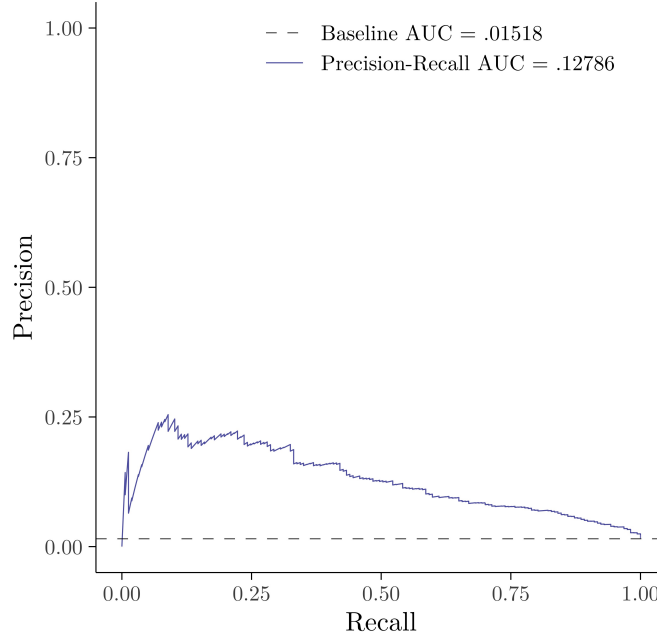


Figure 2.2: Precision-recall curve and area under the curve

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 ($\hat{p}_{its} = \Pr[\text{New coach}_{its} = 1 | Z_{its}]$) 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 (4.395×10^{-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 (2.7). To show the distribution of predicted values for each group before and after weighting, Figure 2.3 plots the kernel density of log propensity scores for each group.

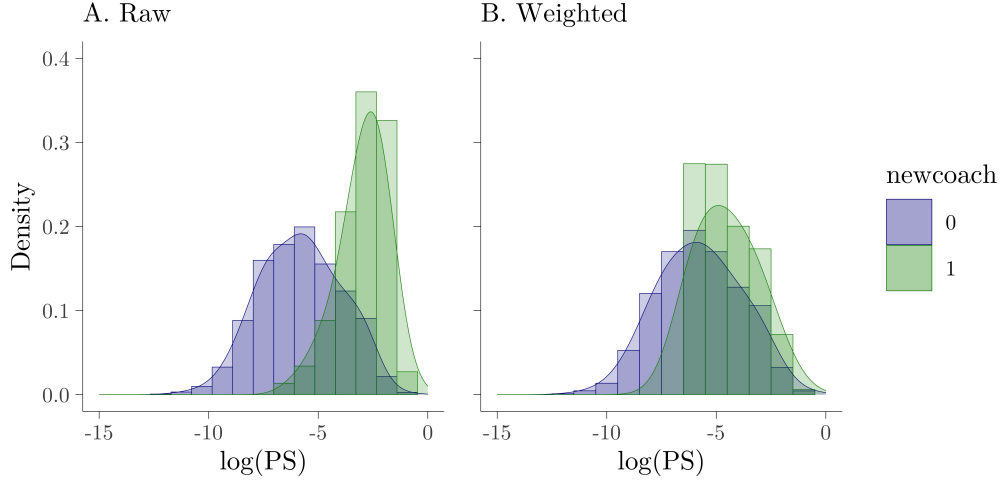


Figure 2.3: Kernel density of log propensity scores before/after weighting

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* = 1).

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:

$$\frac{|\bar{z}_{i,d_i=1} - \bar{z}_{i,d_i=0}|}{\sqrt{\frac{1}{2}\text{Var}[z_{i,d_i=1}] + \frac{1}{2}\text{Var}[z_{i,d_i=0}]}} \quad (2.8)$$

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 2.5 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 2.5: Covariate balance table (before/after weighting, all/common support)

Covariate	Raw		Weighted		Weighted (CS)	
	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 2.4 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.

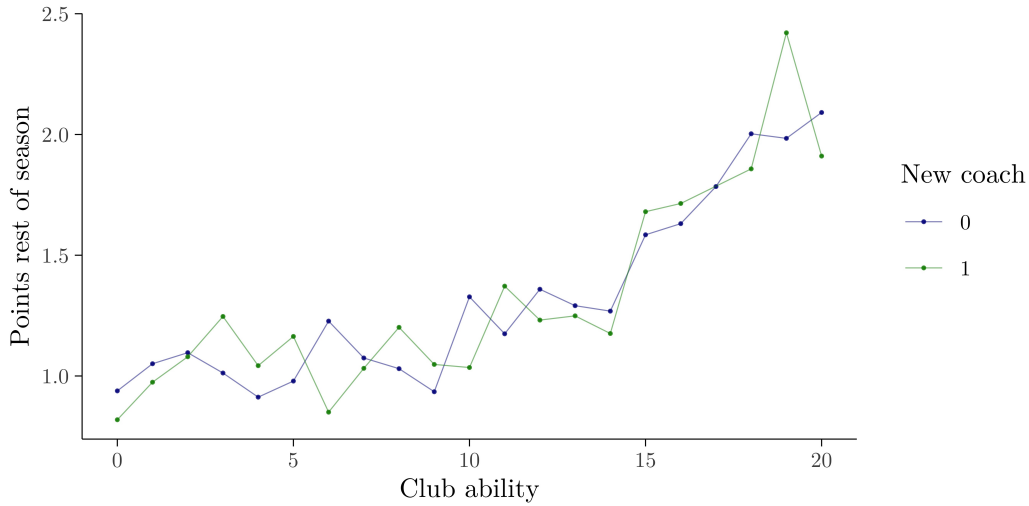


Figure 2.4: Mean value of outcome variable (*Points rest of season*) for each ability and treatment group

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 (2.5) 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 (2.5) and (2.6), using PSW. The estimation results of these models are shown in Table 2.6, 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.

²¹The details of OLS estimation are available in Table 2.9 in Appendix.

Table 2.6: Double robust estimates of outcome models

	<i>Dependent variable:</i>					
	Points five matches		Points ten matches		Points rest of season	
	(1)	(2)	(3)	(4)	(5)	(6)
New coach	0.139 (0.104)	0.102 (0.063)	0.184* (0.097)	0.137** (0.061)	0.180** (0.090)	0.117** (0.056)
Former player		0.120 (0.105)		0.048 (0.110)		0.049 (0.103)
Absent last season		0.163 (0.107)		0.256** (0.102)		0.288*** (0.097)
Age in years		-0.009 (0.011)		-0.019* (0.011)		-0.020** (0.010)
Experience in years		-0.007 (0.010)		0.009 (0.010)		0.013 (0.009)
Former defender/goalkeeper		0.164* (0.094)		0.242*** (0.091)		0.186** (0.082)
Former vice coach		-0.454*** (0.174)		-0.170 (0.188)		-0.210 (0.182)
Italian nationality		0.198 (0.124)		0.171 (0.154)		0.068 (0.143)
Experience Serie A		0.006 (0.117)		-0.035 (0.117)		-0.126 (0.114)
No previous experience		0.055 (0.198)		-0.085 (0.165)		-0.109 (0.152)
Former player Serie A		-0.365*** (0.099)		-0.204* (0.105)		-0.193** (0.090)
Former player club		0.033 (0.150)		-0.065 (0.155)		0.026 (0.136)
Last club as a player		0.523** (0.232)		0.665*** (0.216)		0.529*** (0.191)
Experience abroad		-0.110 (0.105)		-0.069 (0.104)		-0.061 (0.096)
Active Serie A last season		0.094 (0.086)		0.075 (0.098)		-0.012 (0.091)
Home advantage	1.101*** (0.230)	0.849*** (0.139)	1.057*** (0.261)	0.773*** (0.149)	1.029*** (0.298)	0.802*** (0.151)
Club ability	0.057*** (0.015)	0.038*** (0.004)	0.059*** (0.013)	0.043*** (0.004)	0.061*** (0.013)	0.043*** (0.004)
Opponent club ability	-0.032** (0.015)	-0.040*** (0.010)	-0.039* (0.022)	-0.046*** (0.014)	-0.031 (0.023)	-0.035** (0.015)
Constant	0.638*** (0.155)	0.807*** (0.117)	0.797*** (0.195)	0.949*** (0.131)	0.724*** (0.229)	0.809*** (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 (2.5)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.

Columns (2), (4), and (6) in Table 2.6 present the estimated parameters for our extended model (2.6), 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

²²Following 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.

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.

2.4.3 Extension: endogeneity of similarity in coach characteristics

As we have previously discussed in Section 2.3, 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) algorithm²³ 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 2.5, which also illustrates the three treatment types and associated probabilities.

²³This 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.

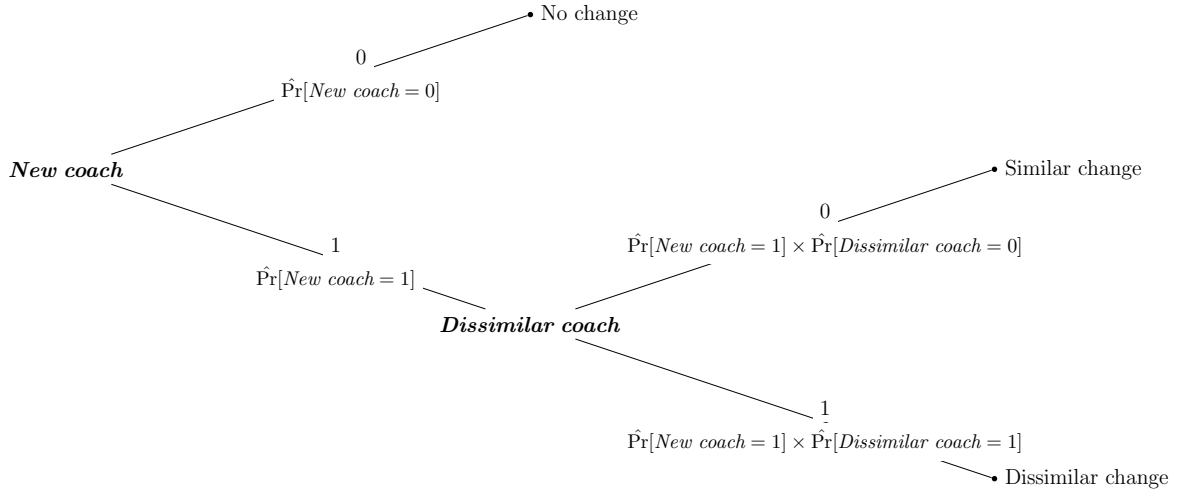


Figure 2.5: Nested logit model

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. $\hat{\Pr}[\text{Dissimilar coach} = 1|Z]$. The estimation results are quite different from those in Table 2.4, 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 2.5. Then, we estimate the outcome model (2.5) 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 2.7. 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 2.7: Double robust estimates of outcome model with dissimilar treatment

	<i>Dependent variable:</i>		
	Points 5 matches (1)	Points 10 matches (2)	Points rest of season (3)
New coach	−0.084 (0.121)	0.004 (0.075)	0.060 (0.066)
Dissimilar new coach	0.247* (0.147)	0.195** (0.090)	0.117 (0.084)
Home advantage	1.080*** (0.253)	0.985*** (0.249)	0.894*** (0.259)
Club ability	0.036*** (0.010)	0.028*** (0.007)	0.030*** (0.007)
Opponent club ability	−0.026* (0.014)	−0.028 (0.020)	−0.014 (0.021)
Constant	1.286*** (0.415)	1.982*** (0.211)	1.873*** (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.

2.4.4 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 2.2 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 2.8 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 2.8: Robustness check. Estimate of the ATE of *New coach* using different methods

	<i>Dependent variable: Points</i>					
	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.

2.4.5 Discussion

Estimation results reported in Tables 2.6 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 2.6, 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 $= 15.25$).

2.5 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.

2.6 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. Giambattista 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 2.9: OLS estimates of outcome models

	<i>Dependent variable:</i>		
	Points 5 matches	Points 10 matches	Points rest of season
	(1)	(2)	(3)
New coach	−0.051 (0.053)	−0.067 (0.047)	−0.063 (0.044)
Home advantage	0.594*** (0.044)	0.609*** (0.045)	0.648*** (0.045)
Club ability	0.051*** (0.001)	0.051*** (0.001)	0.050*** (0.001)
Opponent club ability	−0.042*** (0.002)	−0.042*** (0.003)	−0.041*** (0.003)
Constant	0.955*** (0.034)	0.936*** (0.036)	0.885*** (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

Chapter 3

Causes and Consequences of Recurrent Managerial Changes

3.1 Introduction

Leadership succession is a critical event in any organisation, be it a political party, a corporation, or a sports club. As such, this issue has been intensively studied in many contexts. For instance, as is evident from the reviews by Giambatista et al. (2005) and Berns and Klarner (2017), CEO successions in publicly traded companies have been an active research area in management. Farah et al. (2020) also provides a review of studies related to leadership succession in different contexts, such as privately-owned businesses and political organisations. In addition, professional sports clubs have been a popular field to explore the causes and consequences of leadership succession, where the researchers have examined the impact of replacing a manager or a head coach on a club's performance (Rowe et al., 2005).

This study falls into the last category in that we will provide further evidence of the causes and consequences of head coach dismissals using data from the Italian football league, although our findings can be generalised to broader contexts of leadership succession. The previous studies in the domain of professional football focus on establishing whether replacing a manager improves team performance, however, there is little guidance on how to make such a decision. A recent study by Flepp and Franck (2021) fills this gap by providing evidence that the consequence of dismissal depends on whether the decision was subject to misperceptions of performance. In particular, the study indicates that to implement managerial changes successfully, one needs to identify whether the poor performance is due to low manager ability or bad luck.

To better understand the conditions under which a managerial dismissal can bring about

a favourable outcome, we analyse the decision of replacing a manager more than once within a season. In particular, we separately identify the determinants and estimate the impacts of the two types of managerial dismissals: (1) first dismissals that occur in a given season (*single dismissal*), and (2) second dismissals that happen following the first one within the same season (*multiple dismissal*). Such an analysis is economically relevant in at least two ways. First, it helps us to understand how a principal (club) may adjust their expectation with respect to the agent (manager)’s contribution to the productivity of working teams, given an (unfavourable) outcome of the first replacement. Second, it allows us to understand how first and second dismissals are operationalised in a causal analysis. Given that these two decisions could be motivated by different factors, they could also affect team performance differently.

Our empirical study is based on observational data from the top-tier Italian football league (*Serie A*) for the seasons from 2004/2005 to 2017/2018, where we observe 114 and 36 cases of single and multiple dismissals, respectively. Firstly, we identify the predictors of each type of dismissal by forming appropriate control and treatment groups for single and multiple dismissals and employing classification models. In addition to logistic regression, which is normally employed in this context, we apply a machine learning algorithm, gradient boosting machine (GBM), to identify the most influential variables.

Whilst one can assume that a manager is likely to be dismissed when a club is performing poorly, whether it is the first or second time in a given season, the set of specific predictors for each event could be different. For instance, if a club has already replaced a manager yet the situation has not improved, they may revise their belief on a manager’s role in the adverse outcomes; hence the decision to replace yet another manager may be made with greater caution.

Allowing for the possible differences in the set of predictors between the single/multiple dismissals is also important in order to take into account the pre-treatment differences between the treated and control groups for the respective type of dismissal. A decision to dismiss a manager, like many other managerial decisions, does not occur randomly. For instance, such decisions are more likely to be made when a club is performing poorly. This means that one cannot simply attribute post-treatment differences to the treatment effects of dismissal decisions, without taking into account the differences in the pre-treatment status that could also affect outcomes. Therefore, to estimate the average treatment effects (ATE) of single and multiple dismissals, we employ the inverse propensity score weighting (PSW) method, where such differences in pre-treatment characteristics are taken into account through the weighting based on the propensity score, i.e. the predicted probability of being treated.

Our main findings are the following. First, we find that the predictors of dismissals can depend on the situation, specifically, whether a club has dismissed a manager in the specific season or not. For instance, a rather crude measure of performance such as the average points obtained in recent matches can influence a decision to dismiss a manager for the first time in a given season, as much as relative performance against expectations can. On the other hand, our results suggest that the second within-season dismissal is mostly influenced by relative performance. In addition, whilst the threat of relegation can increase the probability of the first dismissals, it does not appear to affect the second dismissal decision. It is also shown that different sets of managerial characteristics can influence the two types of dismissal. Our study also finds that the consequences of single and multiple managerial changes can be different. In particular, whilst we find a limited boost in club performance after the first managerial change, the second change seems to have no impact on club performance. Therefore, if a club has not dismissed a manager in a given season, making changes could be beneficial, although such positive effects could be short-lived. On the other hand, should a club have changed a manager in a given season, going through yet another change does not make any difference.

The remainder of the paper is structured as follows. Section 3.2 reviews related studies and offers our hypotheses regarding the causes and consequences of single and multiple dismissals. In Section 3.3, we describe our data set and variables used in the analysis. Section 3.4 then describes relevant empirical challenges and our methodology. We present our results in Section 3.5 and conclude our study in Section 3.6.

3.2 Related literature and hypotheses

3.2.1 Causes of dismissal

The relationship between a football club and a head coach can be characterised as a principal-agent (PA) model, just as that of a firm owner and a CEO (Desai et al., 2018). Club owners invest in capital and pay workers, including a head coach and players, to maximise performance. As a sports club, output that will generate monetary revenue is (favourable) performance on the field. A club, therefore, hires a head coach and delegates responsibilities for field management to him/her, who is specialised in training a squad, creating a tactic, and game management. As is commonly the case in any PA relationship, the delegation of tasks entails asymmetric information. In our example, from the point of view of a club owner, who has inferior information in terms of sporting aspects, the effort level of a head coach is unknown since such input is neither directly measurable nor observable. Furthermore,

hidden information arises since the contribution of a head coach towards production (match outcome) cannot be easily disentangled from that of players.¹

Incentive theory suggests that the threat of dismissal improves effort by workers under an environment characterised with asymmetric information, as documented in Kwon (2005) and Sparks (1986). In particular, the use of dismissal threat in compensation of top management such as CEOs is investigated by Wang et al. (2017), Jensen and Murphy (1990a) and Jensen and Murphy (1990b). Indeed, involuntary dismissal of a head coach is a common practice in the world of professional sports. As a club owner, his/her interest is to maximise a manager's effort given the cost that entails such incentive design (firing and hiring a manager). The decision is based on a signal or proxy of managerial contribution, such as realised match outcomes. A head coach, in turn, exerts effort to maximise the probability of winning (hence minimising the probability of being dismissed) with a given set of sub-ordinates.

As shown in the previous studies on the causes of managerial turnover, a football manager's tenure is, in fact, heavily dependent on field performance. The most common factors of dismissals are: falling short of expected performance (van Ours and van Tuijl, 2016; Bryson et al., 2021b), a streak of unfavourable match outcomes (D'Addona and Kind, 2014; van Ours and van Tuijl, 2016), and a threat of relegation (Tena and Forrest, 2007). Although it is suggested to be of less importance, some studies also find managerial characteristics can affect the likelihood of dismissal. For instance, Bryson et al. (2021b) find that a less experienced manager is more likely to be dismissed, after controlling for the recent performance. As is set out in the introduction, one of our objectives in this paper is to identify the causes and consequences of a first within-season replacement (single dismissal) and dismissal of a new manager following the first replacement (multiple dismissal). To the best of our knowledge, this issue has not previously been investigated. We conjecture, however, that the causes of these two types of dismissals may well be different, as we elaborate below.

Traditional theories around managerial dismissal (Grusky, 1963; Gamson and Scotch, 1964) suggest that performance should improve following dismissal *if* an incumbent is to blame for poor performance. When a club dismisses a manager, therefore, the consequence of such an event reveals some information regarding the factors contributing to the underachievement. If a post-dismissal performance bounces back, this may indicate that a

¹There is a strand of literature which looks into the contribution of a manager to organisational success. A study of the largest U.S. firms by Bertrand and Schoar (2003) suggests that a wide range of managerial practices is affected by individual manager effects. In the sports domain, Muehlheusser et al. (2018) and Peeters et al. (2020) provide further evidence that a firm's productivity depends on the quality of an individual manager after controlling for individual firm characteristics. In addition, Frick and Simmons (2008) show that relative coach salaries have a significant impact on team efficiency. Therefore, it is evident that a manager can affect a firm's success. However, imperfect information remains in terms of the extent to which a manager is responsible for a firm's performance.

dismissed manager was indeed responsible for poor performance. On the other hand, if the firm continues to stagnate, the problem may well lie out of a head coach's control, hence they may seek an alternative solution.

Yet, theoretically speaking, a new manager resulting from recent turnover is now under the watchful eyes of a board, and he/she can be fired anytime based on realised match outcomes. Particularly, in the context of professional sport, the majority of variables (e.g. playing talent and club's finance) are fixed or at most adjustable with limitations within a season. Therefore, a head coach may often be replaced with the hope of turning a situation around, and this sometimes happens multiple times within a season. Given the discussion above, however, a club may revise their belief about a manager's role in adverse outcomes based on an outcome of the recent turnover. If a recent dismissal turned out to be unsuccessful, the decision to replace yet another manager might be made with greater caution. Accordingly, we test the following hypothesis regarding the determinants of the single and multiple within-season dismissals:

Hypothesis 1 (H1): *Causes of single and multiple dismissals are not identical, in that multiple dismissal is considered with greater caution.*

Studying this issue is relevant since factors that lead to a certain decision could affect the consequences of such a decision, and therefore they need to be taken into account when evaluating their effectiveness. The relevance is not limited to the methodological point of view; this would add to the literature by providing some empirical evidence on how the outcome of the recent dismissal can influence the motivating factors to dismiss yet another manager within a short period of time.

3.2.2 Consequences of dismissal

Needless to say, one hopes leadership change (as a result of dismissal) will improve club performance. However, most of the previous studies suggest that there are no causal effects of managerial replacement on performance. For instance, Scelles et al. (2020) and van Ours and van Tuijl (2016) estimate the consequences of a managerial change in the French *Ligue 1* and the Dutch *Eredivisie*, respectively. In these studies, the control group is formed with counterfactual observations that followed a similar path of performance against expectation, yet did not replace their manager. Considering all the remaining matches following an actual and counterfactual turnover as treated, they identified a statistically significant improvement in both actual and counterfactual cases. Therefore, these studies suggest that the seemingly positive effect of replacement is, in fact, due to regression to the mean. Similarly, ter Weel (2011) and Bruinshoofd and Ter Weel (2003) compare the subsequent performance of clubs

that replaced their head coach with a control group formed by untreated clubs with similar previous performance in the four post-turnover matches. Results in these papers do not support the hypothesis that a managerial change improves performance.

However, the studies cited above, among others, primarily focus on establishing whether managerial replacements are effective in improving club performance. Knowing specific circumstances or moderating factors that may affect the consequences of dismissals would be useful for decision-makers. Muehlheusser et al. (2016) and Flepp and Franck (2021) fill this gap by identifying some conditions under which a dismissal can bring about favourable outcomes. Muehlheusser et al. (2016) find that a dismissal can positively influence club performance, given that a club is homogeneous, i.e. the ability of the top and bottom players within the squad is rather similar.² Flepp and Franck (2021) also offer further insights into the consequences of managerial dismissals by distinguishing “wise” and “unwise” cases, where the former is led by actual poor performance on the pitch and the latter follows seemingly poor performance due to bad luck. They find that “wise” dismissals can improve performance, whilst “unwise” dismissals have no impact.

This study aims to contribute to understanding factors that may affect the consequence of managerial dismissal by distinguishing single and multiple dismissals as described above. Whilst no previous study compares the consequences of the first and second within-season managerial replacements, there are some studies in other professional sports and organisations that indicate that the frequency of leadership succession can affect the effectiveness of such events. Boyne and Dahya (2002) and Gordon and Rosen (1981)’s hypothesis predicts a negative relationship between the frequency of executive succession and its consequence. The latter states that “too many managerial replacements in too brief a period can be disruptive to a unit” (Gordon and Rosen, 1981, p.238).

Consistent with their theories, Kim et al. (2021) find that high frequencies of CEO turnover have a negative impact on future performance in Korean firms. Khaliq et al. (2006) study CEO turnover in hospitals and conclude that frequent managerial turnover can significantly reduce the morale of the workforce. In sports literature, Hill (2009) finds a non-linear relationship between the number of occurrences of managerial turnover and its impact on field performance in professional baseball teams. The latter study suggests that a moderate incidence of managerial change could be successful, whilst a high frequency of such events entails detrimental effects. It can be argued that constant managerial change can send an adverse signal to workers, undermining their confidence in the firm and their leader, which in turn hinders a new manager from implementing their strategies effectively.

Given the discussion above, one can argue that the first and second within-season dis-

²The authors suggest this is due to more intense competition among players to impress a new manager.

missals can have quite different impacts on subsequent club performance. In particular, our hypothesis is as follows:

Hypothesis 2 (H2): *Consequences of single and multiple dismissals are not identical, in that single dismissal can have a positive impact, whilst multiple dismissal is detrimental to firm performance, or at most ineffective.*

3.3 Data

In order to test our hypotheses, we collected club-match level data from the top-tier Italian football league (*Serie A*) that took place over the seasons from 2004/2005 to 2017/2018. During this period, the 20 most competitive clubs in the country participated in the tournament each season. Akin to most of the domestic professional football leagues in the world, the competition operates in a format of a round-robin tournament. Hence, each competitor meets each other club twice during a campaign, once at its home stadium and once at the opponent's (away) stadium. Therefore, for any given season s , club i appears in each game week t , where $s \in \{1, \dots, 14\}$, $i \in \{1, \dots, 20\}$, and $t \in \{1, \dots, 38\}$. This gives a total number of observations $N = 10,640$.

For each match, three points are awarded to a winner, none to a loser, and one to each competitor in case of a draw. At the end of a campaign, a club with the highest accumulated points receives the championship title, while the bottom three clubs are demoted to the second tier league (*Serie B*), and replaced by the three most competitive clubs promoted from *Serie B*.

Our data set is suitable for the analysis of managerial turnover since the role of a head coach, whose tenure can be terminated at any point during a season by a club owner, is akin to that of a top manager in a corporation (Pieper et al., 2014). In particular, the fact that some clubs replaced their manager multiple times in a given season facilitates testing for the possible differences in causes and consequences of single and multiple dismissals. Moreover, such events are publicly observable with precise timing. This allows us to fairly specifically associate a match outcome with a particular manager's contribution.

3.3.1 The treatment variable and treatment groups

The data set indicates a club i 's manager who was in charge at any specific time, i.e. a game week t in a particular season s , and recorded as the variable $Head\ coach_{its}$. Among other variables that are related to match statistics, this information was obtained from the official website of *Serie A TIM*, which provides a match report for the individual matches. Based

on this, we define our treatment variable $New\ coach_{its}$ as follows.

$$New\ coach_{its} = \begin{cases} 0 & \text{if } Head\ coach_{i,t,s} = Head\ coach_{i,(t-1),s}, \\ 1 & \text{otherwise.} \end{cases} \quad (3.1)$$

Therefore, a club-match observation at time t is deemed to be treated ($New\ coach_{its} = 1$), when club i dismisses its manager after a match $t - 1$ in a season s , followed by a new coach in charge from a match t onwards. In order to make sure any managerial replacement considered in the analysis is due to an involuntary departure of a manager dismissed by a club rather than a voluntary decision or mutual consent between the two parties, we consulted the archives from the official website 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*.³ Furthermore, in case a caretaker manager is introduced between the departure of the outgoing manager and the appointment of a replacement, matches overseen by the interim head coach are excluded from our analysis.⁴

Before presenting descriptive statistics, we form different treatment groups in the following fashion. First, we split the sample into two sub-samples (Sample I and Sample II), for the two treatments, Treatment I and Treatment II, defined as follows. Sample I consists of club-match observations where a match $t - 1$ does not lead to managerial dismissal ($New\ coach_{its} = 0$) or lead to dismissal ($New\ coach_{its} = 1$), given that a club i has not replaced their manager in Season s . Thus Treatment I refers to a dismissal that happens for the first time during Season s . Sample II, on the other hand, is represented by clubs which have already replaced their manager in Season s , and either retain a new manager who replaced a first manager ($New\ coach_{its} = 0$), or replace them with yet another new manager after match $t - 1$ ($New\ coach_{its} = 1$). Therefore, Treatment II is a second dismissal that happens during season s .⁵ Thus, the average treatment effects (ATE) of each treatment are obtained by comparing the post-treatment performance between the treated and control groups within the respective sub-samples. Finally, we refer to treated and control groups in Sample I as Treated I and Control I, respectively, and those in Sample II as Treated II and

³We identified 15 cases of voluntary departures in the 14 seasons analysed in this study. As Bryson et al. (2021b) and Bryson et al. (2021a) show, the motivations and outcomes of quits and dismissals are distinct from one another.

⁴During the relevant seasons, eight individuals served as caretakers amid a transition from an outgoing to an incoming manager.

⁵There are very few cases where a club dismisses a manager more than twice in a given season. We do not consider these cases since it is neither common enough to be of interest nor easy to form a counterfactual control group to evaluate the treatment effect.

Control II.⁶

Table 3.1 summarises the number of club-match observations that led to single and multiple dismissals, and the number of corresponding control observations that did not lead to the respective type of dismissal. During the 14 seasons, we observe a total of 114 single turnovers and 36 multiple turnovers. That is, on average, more than 8 out of 20 clubs dismissed their managers at least once during a season, and almost three clubs go through managerial dismissals twice within a season on average. Both single and multiple treatments are observed in every season in our sample. For instance, AFC Fiorentina underwent managerial change twice during their 2004/2005 campaign. After seven matches in the season, Emiliano Mondonico was dismissed and replaced by Sergio Buso. He was then dismissed after 13 matches, and a new manager Dino Zoff was appointed and was in charge for the rest of the season. Another example is Inter Milan in Season 2011/2012, where we observe the first turnover when Gian Piero Gasperini was dismissed and replaced by Claudio Ranieri after just three matches of the season, and the second time in the dismissal of Ranieri, whose role was taken over by Andrea Stramaccioni, the then vice coach.

However, there are many cases, 16 out of 36 cases to be precise, where a dismissed manager in the first dismissal is re-hired following the second within-season dismissal. For instance, Massimo Ficcadenti was in charge of Cagliari Calcio before the club fired him after the first ten matches of Season 2011/2012. His replacement, Davide Ballardini, was eventually dismissed after 17 matches, which resulted in the re-hiring of Massimo Ficcadenti. The reasons behind the re-hiring of a dismissed manager are likely to be related to unique rules regarding head coach dismissals in *Serie A*. A head coach who is dismissed by a club is not allowed to be hired by another club in the division for the remaining of the season. Meanwhile, a club is obliged to pay the dismissed manager until the end of the season. This implies that the dismissed coach is likely to be available, and the club may have the incentive to bring him back since he continues to be paid regardless. The existence of the unusual rule in Italy could potentially limit the generalisability of our results in terms of the likelihood of re-hiring the dismissed coach following the second dismissal.⁷

⁶In this context, treatment decisions are sequential in the sense that the second managerial change can only happen given the first managerial change has occurred. Therefore, these events cannot be simultaneously captured in a multi-level treatment variable.

⁷However, it is impractical to distinguish the two categories of second dismissal because of low numbers in each category. Therefore, we recommend that future research examines the issue in other countries.

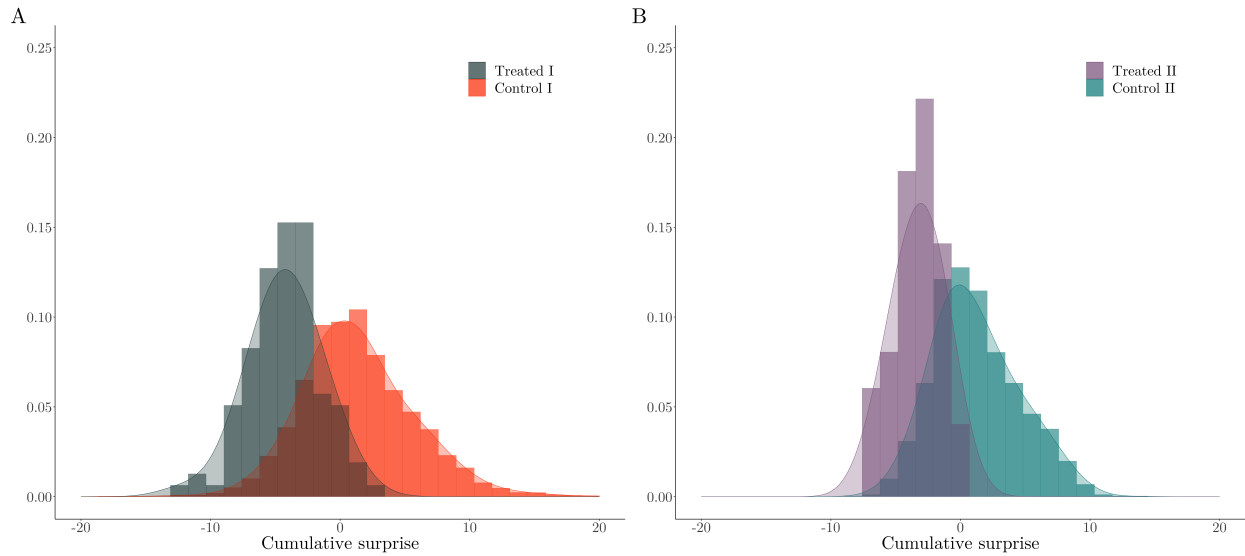
Table 3.1: Number of treated and control units by seasons

	Treated I, N = 114	Control I, N = 7,890	Treated II, N = 36	Control II, N = 1,903
Season				
2004/2005	7	604	1	111
2005/2006	7	588	4	81
2006/2007	9	578	3	128
2007/2008	8	562	5	118
2008/2009	8	578	3	120
2009/2010	11	480	3	203
2010/2011	9	579	2	128
2011/2012	10	478	4	210
2012/2013	8	543	2	157
2013/2014	10	524	3	170
2014/2015	5	623	1	100
2015/2016	9	563	2	122
2016/2017	5	631	2	86
2017/2018	8	559	1	169

3.3.2 Determinants of dismissals

An identification issue that typically arises in analyses of observational data is the endogeneity of treatment assignment. That is, the treatment is not randomly assigned to individuals, hence there are fundamental differences between those sorted to a treated group and those sorted to a control group. Following the previous subsection, we have two types of treatment: the first dismissal that happens within a season (single treatment) and the new manager dismissal after the first dismissal that occurs within the same season (multiple dismissal). Therefore, identifying the determinants of the two treatments is important from the methodological point of view, as further described in Section 3.4. At the same time, our aim is to test the possible differences in the factors leading to these two types of dismissals.

Whilst there is no previous study which has differentiated the causes of the two types of treatments discussed above, there is ample literature that examines the causes of dismissals in general. For instance, using the data from the top-tier football league in the Netherlands (*Eredivisie*) covering the seasons 2000/2001-2013/2014, van Ours and van Tuijl (2016) show

Figure 3.1: Kernel density of *Cumulative surprise* in treated and control groups

Notes: Each figure plots cumulative surprise on the x-axis and density on the y-axis in Sample I (Panel A) and Sample II (Panel B). Each contains the densities for treated and control groups in the respective sample.

that within-season dismissals are most likely to occur at clubs which accumulated negative “match surprise.” Match surprise refers to unexpected outcomes and is captured by the difference between actual and *ex-ante* expected outcomes, where the latter can be measured with, for example, betting odds. To illustrate the pre-treatment differences with respect to this variable, *Cumulative surprise*, the Kernel densities of this variable for treated and control groups for respective treatment types are presented in Figure 3.1. Panel A compares the distributions of *Cumulative surprise* for the treated and control groups for Treatment I (single dismissal), and Panel B compares those for the treated and control groups for Treatment II (multiple dismissal). For both types of treatment, the underlying differences between treated and control groups are noticeable. For both cases, cumulative surprise for the treated units is mostly distributed below zero, whilst that for the control counterpart is distributed more evenly about zero.

There are many other factors that are identified as a predictor of dismissal in the literature. The two main categories are: (1) recent performance variables and (2) characteristics of a manager.

Predictors capturing recent team performance are most commonly found in prior research. The short-term performance history measured with the most recent four matches (van Ours and van Tuijl, 2016) and within a couple of weeks prior to the current match (D’Addona and Kind, 2014) are shown to be significant predictors of involuntary departure of a manager. Unlike the cumulative surprise variable described above, however, these performance

indicators are considered a rather crude measure of performance since they do not take into account expectations. Among others, Bryson et al. (2021b) provide evidence that league position can also determine the likelihood of dismissal. Furthermore, Tena and Forrest (2007) find that the threat of demotion is the most common factor in dismissal decisions. Whilst performance indicators employed tend to vary across the studies, all the evidence points to a negative relationship between recent performance and the probability of dismissal.

Accordingly, in addition to the variable *Cumulative surprise* illustrated above, we examine the following performance indicators: average points obtained in the most recent four matches (*Point last four*) and a club's current position in the league (*Standing*). Also included in this category are three dummy variables to indicate whether the most recent match outcome was loss (*Lost last*), the most recent match outcome was loss and took place at the club's home stadium (*Lost last home*), and whether a club is under threat of relegation (*Relegation*).

The second set of variables that could affect the likelihood of dismissals is the characteristics of a manager. Whilst existing evidence is not as consistent as the one for recent performance, some studies suggest specific managerial characteristics can also affect the chances of dismissal. For instance, Audas et al. (1999) show that a manager's age can influence the survival rate of football managers, whilst others (Frick et al., 2010; van Ours and van Tuijl, 2016) find significant effects of their previous experience. Other managerial characteristics such as their nationality and background have also been tested. Gilfix et al. (2020), for instance, provides some evidence that domestic managers are less likely to be dismissed relative to those who are not nationals of the country.

In our analysis, therefore, we examine a set of managerial characteristics that could potentially affect the probability of dismissal. The information was obtained from various sources, with Transfermarkt (*transfermarkt.com*) being our primary source. Firstly, a manager's general and industry-specific human capital are measured with the number of years of experience (*Experience years*) and a dummy variable to indicate whether a manager has previously worked in *Serie A* (*Experience Serie A*), respectively. We also add the two dummy variables associated with managerial experience, one to indicate a new entrant to the labour market (*No experience*),⁸ and the other to indicate a manager's experience in professional

⁸Peeters et al. (2017) provide interesting findings regarding market failure in the football managers' labour market, where they show that an experienced manager has an excess advantage over a novice manager. By the time the firing decision is made, however, the novice manager's ability would have become public. Therefore, the mere fact that they have no previous experience as a head coach should not affect the likelihood of dismissal, after controlling for the realised performance. Nevertheless, there may be another type of hidden information, in the sense that a new manager's particular ability to cope with the adverse situation is yet to be realised. If a club were unwilling to take the risk of not being able to turn a situation around, or worse, aggravating the situation, it may be more prone to fire a novice coach than the experienced.

leagues abroad (*Experience abroad*). To take into account a manager's recent participation in the professional football manager labour market, we examine indicators of activity status in the most recent season (*Active*), as well as that in *Serie A* (*Active Serie A*) in our models.

Secondly, a set of dichotomous variables regarding a manager's background as a player are considered. Specifically, they indicate whether a head coach had a playing career in professional football (*Player*) and whether that included playing in *Serie A* (*Player Serie A*). Additionally, information regarding the position played is indicated with the variable *Player GK/DF*, where *Player GK/DF* = 1 if a manager is a former defender or goalkeeper and *Player GK/DF* = 0 otherwise.

Thirdly, association with the club is captured by dummy variables to indicate whether a manager has previously served the club as a vice coach (*Former vice*), whether he is a former player of the club (*Player club*), and whether the club is the last club with which he was a player (*Player club last*).

Finally, we examine two personal traits, nationality and age, where *Italian* takes 1 if a manager is Italian and 0 otherwise, and *Age* measures their age in years.

In addition to the two groups of predictors described above, we include a variable *Week*⁹, which indicates how far it is into the season, and a variable *Days*, the number of days between match $t - 1$ and match t . Table 3.2 provides descriptive statistics of the pre-treatment status measured by the variables described above, for the respective treated and control groups for the two types of dismissal.

⁹We also include a squared term of *Week* (*Week squared*) in logistic regression specification, following Tena and Forrest (2007).

Table 3.2: Descriptive statistics of treatment predictors

Variable	Treated I, N = 114	Control I, N = 7,890	Treated II, N = 36	Control II, N = 1,903
Cumulative surprise	-4.27 (2.75)	1.32 (4.36)	-3.27 (1.81)	1.20 (3.29)
Point last four	0.59 (0.41)	1.45 (0.76)	0.47 (0.41)	1.24 (0.73)
Lost last				
0	16 (14%)	5,196 (66%)	5 (14%)	1,121 (59%)
1	98 (86%)	2,694 (34%)	31 (86%)	782 (41%)
Lost last home				
0	71 (62%)	6,906 (88%)	20 (56%)	1,599 (84%)
1	43 (38%)	984 (12%)	16 (44%)	304 (16%)
Standing	15.51 (4.20)	9.07 (5.41)	17.33 (3.51)	14.32 (4.52)
Relegation				
0	63 (55%)	7,226 (92%)	13 (36%)	1,324 (70%)
1	51 (45%)	664 (8.4%)	23 (64%)	579 (30%)
Experience years	11.30 (7.65)	12.23 (7.59)	14.75 (8.56)	13.42 (8.90)
Experience Serie A				
0	25 (22%)	1,228 (16%)	5 (14%)	218 (11%)
1	89 (78%)	6,662 (84%)	31 (86%)	1,685 (89%)
No experience				
0	108 (95%)	7,635 (97%)	34 (94%)	1,789 (94%)
1	6 (5.3%)	255 (3.2%)	2 (5.6%)	114 (6.0%)
Experience abroad				
0	87 (76%)	5,625 (71%)	23 (64%)	1,191 (63%)
1	27 (24%)	2,265 (29%)	13 (36%)	712 (37%)
Active				
0	16 (14%)	800 (10%)	10 (28%)	412 (22%)
1	98 (86%)	7,090 (90%)	26 (72%)	1,491 (78%)
Active Serie A				
0	51 (45%)	2,385 (30%)	17 (47%)	792 (42%)
1	63 (55%)	5,505 (70%)	19 (53%)	1,111 (58%)
Player				
0	12 (11%)	510 (6.5%)	5 (14%)	188 (9.9%)
1	102 (89%)	7,380 (94%)	31 (86%)	1,715 (90%)
Player Serie A				
0	37 (32%)	2,707 (34%)	16 (44%)	661 (35%)
1	77 (68%)	5,183 (66%)	20 (56%)	1,242 (65%)
Player club				
0	97 (85%)	6,630 (84%)	30 (83%)	1,533 (81%)
1	17 (15%)	1,260 (16%)	6 (17%)	370 (19%)
Player club last				

0	108 (95%)	7,237 (92%)	35 (97%)	1,846 (97%)
1	6 (5.3%)	653 (8.3%)	1 (2.8%)	57 (3.0%)
Player GK/DF				
0	74 (65%)	6,067 (77%)	26 (72%)	1,447 (76%)
1	40 (35%)	1,823 (23%)	10 (28%)	456 (24%)
Former vice				
0	109 (96%)	7,659 (97%)	33 (92%)	1,794 (94%)
1	5 (4.4%)	231 (2.9%)	3 (8.3%)	109 (5.7%)
Italian				
0	12 (11%)	778 (9.9%)	2 (5.6%)	186 (9.8%)
1	102 (89%)	7,112 (90%)	34 (94%)	1,717 (90%)
Age	49.33 (6.82)	50.33 (6.75)	52.42 (6.77)	51.10 (7.43)
Days	8.46 (4.56)	7.20 (3.53)	7.39 (3.17)	7.08 (3.14)
Week	16.34 (8.49)	17.12 (10.66)	25.69 (6.54)	24.44 (8.05)

Notes: Mean (SD); Frequency (%)

3.3.3 Outcome and control variables

As discussed in the earlier section, a club owner delegates the tasks that particularly involve sporting aspects of the firm to a head coach, who normally possesses superior knowledge and experience in football. Sporting success is, of course, one of the primary aims of a football club since the number of wins is directly and indirectly translated into a club's revenue. Hence, evaluating the treatment effect of managerial turnover focuses on post-succession field performance. In particular, we measure the performance with average points (*Point*) and goal differences (*Goal dif*) in the subsequent matches.

Whilst the most common practice is to use all the remaining matches in the relevant season to measure the post-treatment performance, it is the authors' discretion to decide how long a manager is to be considered as a "new" manager after their appointment. Since in principle, a club may be taking a decision over whether to retain or replace the manager at any point in the season, a "new" manager is also subject to dismissal from the day they arrive. Indeed, the focus of this study is on cases where a club replaces a manager more than once in a given season. Therefore, we allow flexibility on the time window within which matches are deemed to be treated.

More precisely, we examine the impact of managerial change on performance in the very first match after the change, as well as the average performance in the subsequent 5 matches, the subsequent 10, and for the rest of the season, or until the next managerial turnover, if any, whichever occurs first. Should there be a more limited number of matches than those exact time windows, we consider the average performance of the available matches. Hence,

we denote our outcome variables with a suffix which indicates the maximum number of post-treatment matches considered to measure the performance. For instance, a variable *Point_10* is the average value of points within the subsequent 10 matches in the post-treatment period. However, should a manager be dismissed after 7 post-treatment matches, the variable takes the average points of these matches.

To control for factors that could affect match outcomes other than the treatment assignment, we construct the following control variables. First, a variable *Home* 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. With the example of *Point_10* demonstrated above, therefore, we control for the proportion of matches held at a club's home stadium out of the 10 post-treatment matches (*Home_10*).

In addition, the ability level of the 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 is assigned the value 20 (and the bottom club would be assigned the value 1). The order is reversed to ensure that the 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 is 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.¹⁰ 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. Again, when considering the preceding 10 matches, we take the average of strengths over the 10 corresponding opponents (*Opponent ability_10*).

3.4 Methodology

An important identification issue that arises in observational studies is that assignment of treatment is not random, hence a direct comparison of outcomes in treated and control groups is not valid to establish a treatment effect. This applies to this study where we aim to estimate the effects of single and multiple dismissals since, as we will see, these treatments are more likely to be assigned to, for instance, clubs that are performing poorly. As discussed in Section 3.2, the main approach to this issue has been to apply various matching techniques

¹⁰More 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.

to form comparable counterfactual control groups. Our approach is different from those used in these prior studies as we employ the inverse propensity score weighting (PSW) method (Imbens, 2000). Intuitively, this method virtually “randomises” the treatment assignment by giving higher weights to “unusual” cases in treated and control groups. That is, if a treated unit has a lower predicted probability of being treated, it will be given a higher weight since it resembles a control unit. Similarly, if a control unit has a higher predicted probability of being treated, it will be given a higher weight since, although it is a control unit, it resembles a treated unit. By doing so, the distribution of the predicted probability of being treated, i.e. how likely a unit is being treated *ex-ante*, becomes similar between treated and control groups, as if the treatment was assigned randomly. The *ex-ante* probability of being treated is the so-called “propensity score,” and we obtain these for the two types of treatment using the respective subsamples. In the following subsections, we provide further descriptions of the method and how to estimate outcome and treatment assignment models.

Another common characteristic of observational data is imbalance class. That is, the number of treated and control cases is often not approximately equal. This could be problematic particularly when implementing machine learning techniques to obtain propensity scores. Therefore, in the last part of this section, we explain how to deal with this issue using an over-sampling technique when estimating the treatment assignment models.

3.4.1 Estimation of outcome model

The outcome model includes post-treatment performance as a response variable (y_{its}), which is regressed with a treatment variable ($New\ coach_{its}$) as well as other control variables that can independently influence the outcome variables ($X_{its} = (Home, Ability, Opponent\ ability)$) as follows:

$$y_{its} = \delta\ New\ coach_{its} + \beta' X_{its} + \varepsilon_{its}, \quad (3.2)$$

Since our aim is to estimate the treatment effect that could possibly be different for single and multiple dismissals, we estimate the model (3.2) for the two types of dismissal separately, using the respective treatment and control groups defined in Section 3.3. Then, δ is the ATE of a respective managerial dismissal, β is a vector of parameters also to be estimated, and ε_{its} is a stochastic error component.

Following the discussion above, non-randomness of the treatment assignment means the *ex ante* probability of being treated ($\Pr[New\ coach_{its} = 1]$) varies across individuals.¹¹ Therefore, δ would be biased if the factors that affect such probabilities are not taken into account.

¹¹In contrast, under a randomised control trial, where a treatment is randomly allocated to participants with equal probabilities.

Accordingly, we employ inverse propensity score weighting (PSW) to deal with this endogeneity issue. The method has been used traditionally in medical research (Austin and Stuart, 2015) and more recently in social sciences (Morgan and Todd, 2008). The PSW adjusts the pre-treatment differences between treated and control units by weighting the treated units with the inverse of the predicted probability of receiving the treatment and the control units with the inverse of the predicted probability of *not* receiving the treatment. The predicted probability of receiving the treatment is the so-called propensity score, and the weights assigned to individual units are formally defined as follows:

$$w_{its} = \begin{cases} (P[\text{New coach}_{its} = 1])^{-1}, & \text{if } \text{New coach}_{its} = 1, \\ (1 - P[\text{New coach}_{its} = 1])^{-1}, & \text{if } \text{New coach}_{its} = 0. \end{cases} \quad (3.3)$$

Again, since we estimate two types of treatment, we obtain the weights defined by the equation (3.3) separately using the two subsamples. Once we have obtained each weight (w_{its}) associated with the respective treatment type, we can separately estimate the model (3.2) within the weighted regression framework for each treatment. In addition, we estimate equation (3.2) with several post-treatment time windows as well as different performance measures discussed in the previous sections.

3.4.2 Estimation of treatment assignment models

Following the discussion above, the estimation of the outcome model using PSW requires obtaining the probabilities of receiving a respective treatment. Therefore, we first estimate the treatment assignment models defined as follows:

$$P[\text{New coach} = 1] = f(Z_{its}, \gamma), \quad (3.4)$$

where Z_{its} is a set of covariates that can affect the probability of receiving treatment, and γ is a vector that reveals the importance of each covariate in Z_{its} .

Whilst the estimation of the model (3.4) is a necessary step in the PSW procedure, we use this estimation to test our hypotheses regarding the causes of single and multiple dismissals. To fit the model (3.4), we employ two classes of classification algorithms: (1) logistic regression (LGT) and (2) Gradient Boosting Machine (GBM) with decision trees. While the former is the most commonly used estimation method for propensity scores in non-experimental causal analysis (Dehejia and Wahba, 2002), the use of Machine Learning (ML), including the latter model, has recently been shown to be a competitive option.

In their reviews, Westreich et al. (2011) discuss the advantages and disadvantages of ML

models against a logistic regression in the context of propensity score estimations. Perhaps the most notable advantage of tree-based models is the fact that functional forms are not required to be specified by a researcher. This is particularly attractive since underlying covariates are most likely to interact with each other or have a non-linear effect on the probability of treatment, and misspecification of functional forms may lead to a failure in mitigating the bias. On the other hand, a possible disadvantage of decision tree based models is the fact that the specific interaction of variables or polynomials cannot be identified, often referred to as the “black box” nature of the model class. As noted in Olmos and Govindasamy (2015), however, this drawback is not most concerning in this context, given that the aim of propensity score weighting is to mitigate the underlying covariate imbalancedness, not to identify such functional forms.

GBM employed in this analysis is a machine learning algorithm to assemble a series of weak learners, i.e. a classification model that alone does not have strong predictive power, in order to improve prediction. Here a weak learner used is typically a regression tree, as is the case in this study. Intuitively, decision trees are sequentially built one after another to reduce prediction errors made by a previous learner until no improvement can be made or the specified maximum number of tree M is reached. More specifically, a new learner tries to model errors made by a previous learner and update this to a set of predicted values. By sequentially assembling simple weak learners in this way, the bias can be reduced without compensating for increased variance. As such, GBM models often outperform a single strong learner such as regression, which is subject to trade-offs between estimation errors and biases.

In each iteration, a set of such predicted values, $F_m(z)$ conditioning on covariates z , are obtained such as to minimise a loss function given the previous prediction $F_{m-1}(z)$, where $m = 1, \dots, M$. Since our target value is binary, the loss function can be characterised as a negative value of log-likelihood.¹²

The model complexity, however, increases in the number of decision trees. Therefore, to avoid over-fitting, iterations are stopped where predictive performance is optimised on an independent data set rather than the sample used to fit the models. We implement this early stopping using cross-validation (CV) errors.¹³

3.4.3 Imbalancedness of classes

Another common characteristic of observational data is that the proportion of each class, for instance, treated and control, is often not even. In this study, the number of treated

¹²This corresponds to the parameter estimation by maximum likelihood, where log-likelihood is to be maximised. For the formal presentation of the algorithm, see Friedman (2001).

¹³See, James et al. (2013) for more on early stopping.

units in Sample I for single dismissal is 114, as opposed to that of control units being 7,890. Similarly, the number of treated units in Sample II for multiple dismissal is 36, as opposed to that of control units being 1,903. This is problematic, particularly for ML models, since they can bias toward the majority class.

To cope with the class imbalancedness, therefore, we use a synthesised sample to fit the treatment assignment models. Particularly, we employ the synthetic minority oversampling technique (SMOTE) proposed by Chawla et al. (2002). This approach to imbalanced data involves under-sampling the majority class (control groups) and over-sampling the minority class (treated groups) by synthesising data. More precisely, the latter is done in the following manner. First, k -nearest neighbours with respect to a vector of covariates (possible determinants of dismissals) are selected for a randomly selected member within the minority class. Second, we randomly select one of the k -nearest neighbours and obtain the distance between this and the member of the minority class under consideration. Finally, synthetic data is obtained where the difference is multiplied by a random value within the range of $(0, 1)$. In other words, synthetic data is a convex combination of the two minority observations. This is repeated until the specified number of synthesised minority cases is achieved. Then, randomly selected majority cases are also removed to achieve better class balance.¹⁴

The advantage of synthesising data points rather than purely duplicating members in the under-represented class is that the former can not only enlarge the minority sample size but also introduce variations. As explained in Chawla et al. (2002), the subsampling method can improve predictions in classification models. Table 3.4 summarises the compositions of each group and type of the treatments for all samples, and the resulting synthesised samples. Within both Sample I and Sample II, the class imbalance is significantly reduced in SMOTE samples relative to the respective original samples.

Table 3.4: The number of units in full and SMOTE samples

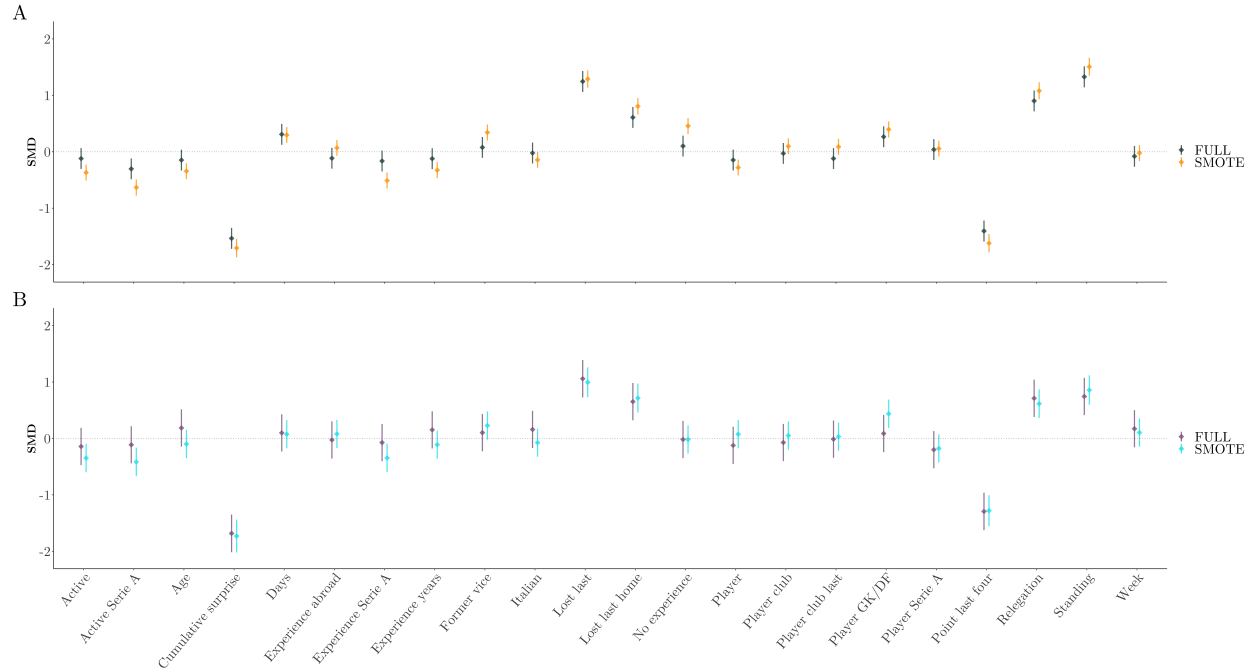
	Sample I		Sample II	
	Treated	Control	Treated	Control
All	114 (1.424%)	7890 (98.576%)	36 (1.857%)	1903 (98.143%)
SMOTE	342 (42.857%)	456 (57.143%)	108 (42.857%)	144 (57.143%)

Notes: Table shows the number of units in the respective treated and control groups in Sample I and Sample II. The numbers are presented for full and synthesised (SMOTE) samples.

¹⁴More specifically, the SMOTE is implemented using the R function SMOTE, which uses $k = 5$ nearest neighbours and synthesises minority cases until the number of extra minority cases generated is equal to $2 \times$ (the number of minority cases in the original sample). In addition, the majority cases are randomly removed so that the number of majority cases is equal to $2 \times$ (the number of synthesised minority cases).

Figure 3.2 visualises the standardised mean differences (SMD) between treated and untreated observations for each treatment (Treatment I and Treatment II) and sampling type (full sample and SMOTE sample). As the figures show, the synthesised sample preserves the underlying differences between treated and control groups for each treatment type. Therefore, our treatment assignment models are fitted using the SMOTE sample.

Figure 3.2: Standardised mean differences (SMD) in predictors in full and SMOTE samples



Notes: Figure shows the SMDs in all the considered predictors for Treatment I (Panel A) and Treatment II (Panel B). Each panel shows the SMDs between treated and control groups with full and synthesised (SMOTE) samples.

3.5 Results

In this section, we present the results from each step of our analysis. That is, we first present and discuss the results related to the causes of the two types of dismissals, i.e. single and multiple dismissals, where the logistic regression and GBM are fitted for treatment assignment models of the two. Then, we use the GBM models¹⁵ to obtain the probabilities of each dismissal, i.e. propensity scores, and shows how well the weighting based on these propensity scores can reduce the pre-treatment unbalancedness between the respective control and treated groups. Finally, we present our results for the PSW estimators of the average treatment effects of the single and multiple dismissals.

¹⁵We select GBM for obtaining the propensity scores since it outperforms logistic regression in terms of balancing covariates, as explained further below.

3.5.1 Predictors of single and multiple dismissals

Following the procedure explained in the previous section, we proceed with the analysis by fitting classification models (3.4). As discussed there, the models for single dismissal are trained with SMOTE sample synthesised from Sample I, and those for multiple dismissal with SMOTE sample synthesised from Sample II. The objective here is two-fold: to test the hypothesis regarding the motivations behind the two types of dismissals (H1) and to obtain the predicted values of being sorted into a realised treatment group, i.e. a propensity score. The classification models are estimated using (1) logistic regression and (2) Gradient Boosting Machines (GBM), in which possible determinants of treatment assignments discussed in Section 3.3.1 are considered.

For logistic regression models, the final set of covariates to be included is selected by means of stepwise regression with a sequential replacement algorithm, which combines forward and backward selection. Hence, the predictors under consideration are iteratively added and removed until the lowest predictive error is achieved. Following Bruce and Bruce (2017), errors are measured with the Akaike information criterion (AIC).

Table 3.5 shows the logit estimates with selected predictors. Included in Column (1) are the parameter estimates for the selected covariates in the treatment assignment model for Treatment I. The results suggest that most of the indicators associated with recent field performance show a strong correlation with the likelihood of Treatment I, the first dismissal that happens during a season. In particular, performance exceeding the expectation (*Cumulative surprise*) and a streak of favourable outcomes in recent matches (*Point last four*) are negatively related to the probability of dismissal. On the other hand, a loss in the most recent match (*Lost last*) is associated with a higher probability of dismissal, with an additional effect when the defeat was at the home stadium (*Last lost home*). Whilst a club's interim league position (*Standing*) itself does not appear in the selected model, there is a positive and significant effect of relegation threat (*Relegation*). These results are fairly consistent with the previous findings.¹⁶

Furthermore, our results suggest that some of the variables related to managerial characteristics can also influence dismissal decisions. For instance, having previous experience as a head coach within *Serie A* (*Experience Serie A*) and being a former player (*Player*) are negatively associated with a manager being dismissed. On the other hand, the probability of dismissal is higher when a manager has a previous experience as a head coach in a foreign professional league (*Experience abroad*) and is a former defender/goalkeeper (*Player GK/DF*). Other managerial characteristics, such as experience in years, are not selected in

¹⁶See Section 3.2 above for a review of related literature.

the model, suggesting that they are not an important predictor of the first dismissal.

The results further indicate that when there are more days between two matches, a club is more likely to dismiss a manager (*Days*), and the coefficients on the variable *Week* and its quadratic term *Week squared* suggest that dismissals occur more often as a season progresses, but such decisions are then made less often towards the end of the season.

The counterpart estimates for the treatment assignment of Treatment II, the second dismissal that happens during the same season, are shown in Column (2), where we observe some inconsistencies. Particularly, the variables *Cumulative surprise* and *Point last four* are selected for this treatment assignment model, however, unlike that for the first dismissals, the predictors *Lost last*, *Lost last home*, and *Relegation* are not selected. This may indicate that the second dismissal decisions are not as sensitive to external pressure as the first dismissals.

Comparing the selected covariates in Columns (1) and (2), some of the managerial characteristics included in the prediction of first dismissals are also present in the treatment assignment model for second dismissals, whilst the effects of each characteristic are not consistent. For example, being a former player (*Player*) and having experience as a head coach abroad (*Experience abroad*) have opposite effects on the likelihood of second dismissals compared to those on the first dismissal decision. The second dismissals are further affected by managerial characteristic variables such as being a former vice coach of the club (*Former vice*), which is suggested to increase the probability of such events. It is not straightforward to decide which decisions are more sensible based on which managerial characteristics affect each decision. However, the fact that different managerial characteristics can influence the first and second dismissal decisions differently may suggest that the club's expectations or preference on particular managerial characteristics may be updated based on the outcome of the recent dismissal.

Whilst the general trend with respect to weeks is again similar to the first dismissal case, the number of days between matches does not seem to have an influence on the second dismissal decision.

Table 3.5: Logit estimates of treatment assignment models

	<i>Dependent variable:</i>	
	New coach	
	(1) First dismissal	(2) Second dismissal
Cumulative surprise	−0.387*** (0.052)	−0.958*** (0.167)
Point last four	−1.262*** (0.318)	−1.235** (0.564)
Lost last	0.962*** (0.285)	
Lost last home	0.680** (0.290)	
Relegation	1.067*** (0.299)	
Experience Serie A	−0.627** (0.296)	
Experience abroad	0.715** (0.285)	−0.810* (0.482)
Active Serie A		−1.020** (0.456)
Player	−0.924** (0.398)	1.567** (0.650)
Player Serie A	0.712*** (0.266)	1.573*** (0.544)
Player GK/DF		−1.390** (0.541)
Former vice		1.327* (0.763)
Days	0.068** (0.031)	
Week	0.299*** (0.058)	0.676*** (0.244)
Week squared	−0.008*** (0.002)	−0.013*** (0.005)
Constant	−2.934*** (0.741)	−9.311*** (2.922)
Observations	798	252
Log Likelihood	−225.789	−71.782
Akaike Inf. Crit.	477.578	165.565

Note:

*p<0.1; **p<0.05; ***p<0.01

Overall, the estimations of logistic regressions to predict the first and second dismissals indicate that the predictors of these two events could be quite different. The findings from the alternative classification model, GBM, are similar in that it also suggest that the factors affecting the single and multiple dismissals can differ. With GBM, the predictive power of each covariate is measured in terms of relative influence. As explained by Breiman (2001), this is computed by a percentage increase in misclassification rate when randomly introducing noise to one variable at a time. These values obtained for each variable are then normalised so that they add up to 100%.

Figure 3.3 illustrates the relative importance of the predictors that are assigned non-zero influence for the predictions of the first dismissal (A) and the second dismissal (B). The most influential predictor of the first dismissal is *Point last four* (36.9%), although *Cumulative surprise* is a close competitor (34.6%). On the other hand, the second dismissal has one dominating predictor, *Cumulative surprise* (53.9%), with the second most influential

predictor being *Point last four* (12.4%). This indicates that the first dismissal is driven by both absolute measure of performance (*Four last point*) and the relative performance against expectation (*Cumulative surprise*), whilst the second dismissal is mostly influenced by the relative performance. It can be argued that the second dismissal decision is less sensitive to the mere fact that a club has poorly performed on average in the recent matches, but such a decision may be made by taking into account realistic expectations.

Consistent with the logistic regression presented above, the first dismissals are affected by the threat of demotion (*Relegation*), as well as the defeat in the most recent match (*Lost last*), which is not the case for the second dismissal. Furthermore, the sets of managerial characteristics that can affect the two events are again slightly different, although those predictors are shown to be of less importance.

The two classification models presented in this section confirm that the factors affecting the likelihood of the first and second within-season dismissals could be quite different. In addition, our findings suggest that the second dismissal decision may be less sensitive to the crude measure of performance, such as average performance in recent matches and defeat in the most recent match, but instead motivated by a fairer measure of performance i.e. performance against expectation. This is consistent with our hypothesis (H1).

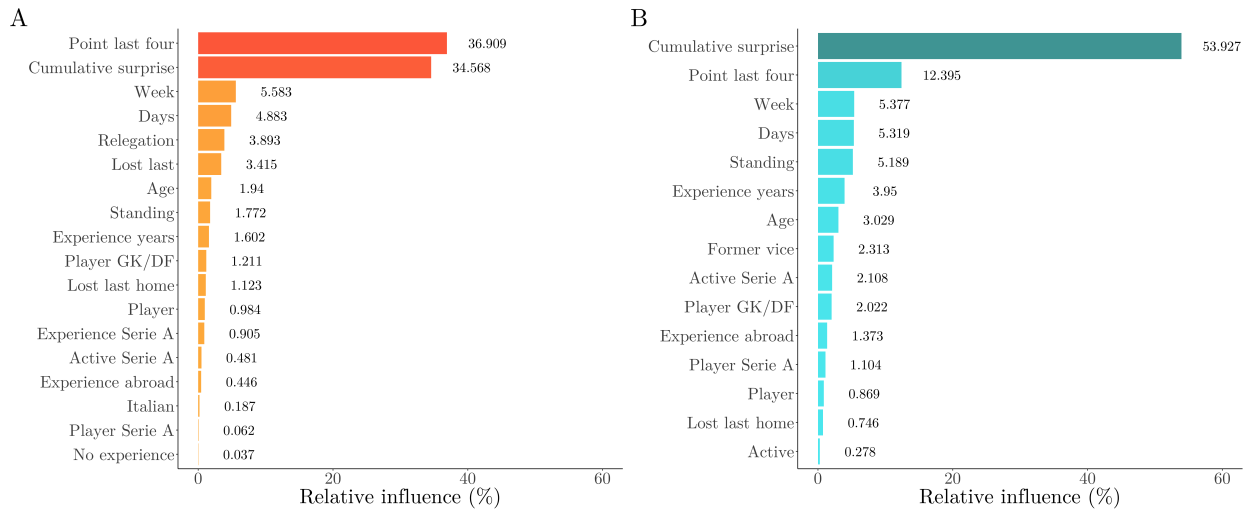


Figure 3.3: Relative importance of predictors with non-zero influences

Notes: Figure shows the relative influence of covariate assigned non-zero influence. GBM models are fitted for Treatment I (Panel A) and Treatment II (Panel B).

3.5.2 Estimation of propensity scores and covariate balance

In the previous subsection, we presented the fitted treatment assignment models to predict the first and second managerial dismissal events using the two classification models, logistic

regression and GBM. The selected predictors vary across the two different dismissal types. On the other hand, the main predictors for each treatment are fairly consistent regardless of the choice of a classification model.

Note that we used synthesised samples to fit the classification models due to imbalanced classification. However, our aim is now to obtain the predicted probability of treatment for each individual unit. Therefore, we now obtain the fitted values using Sample I and Sample II, i.e. the entire samples rather than the SMOTE sample that are used to fit the models.

It turns out that the performance of the two classification models in terms of predictability is not significantly different from each other.¹⁷ Meanwhile, it is found to be the case that GBM outperforms logistic regressions in terms of balancing underlying covariates differences between the treated and control groups for both treatment types. Given that our aim is “virtual randomisation” of each treatment, we, therefore, proceed with our analysis by obtaining the propensity scores with GBM.¹⁸

It is evident from the results presented in the previous subsection that managerial dismissals are not a random event, i.e. there are significant differences between the treated and control observations in terms of the pre-treatment conditions. This implies that any direct comparison between the treated and control observations for any treatment types examined here is subject to selection bias. Therefore, there is a need to adjust for the pre-treatment differences, and we do so using PSW described above.

The mean values and range of predicted values as well as the number of units in each treatment group are summarised in Table 3.6. The first two rows include the statistics for the whole sample (Sample I and Sample II for Treatment I and Treatment II, respectively). In addition, the bottom two rows include the statistics for observations within the common support, i.e. the range where the propensity scores within treated and control groups overlap. As expected, the predicted value of dismissal is higher in the treated groups compared to those in the control groups for the respective treatment types. Furthermore, only one treated observation is outside the common support for Treatment I, and all the treated cases are contained within the common support for Treatment II.

¹⁷This is based on the evaluation metrics, AUC and balanced accuracy. See, Figure 3.6 and Table 3.9 in Appendix for details.

¹⁸We present the standardised mean differences of each covariate before and after applying PSW based on GBM later in this section and those based on logistic regression in Figure 3.7 in Appendix as a comparison.

Table 3.6: Summary of estimated propensity scores

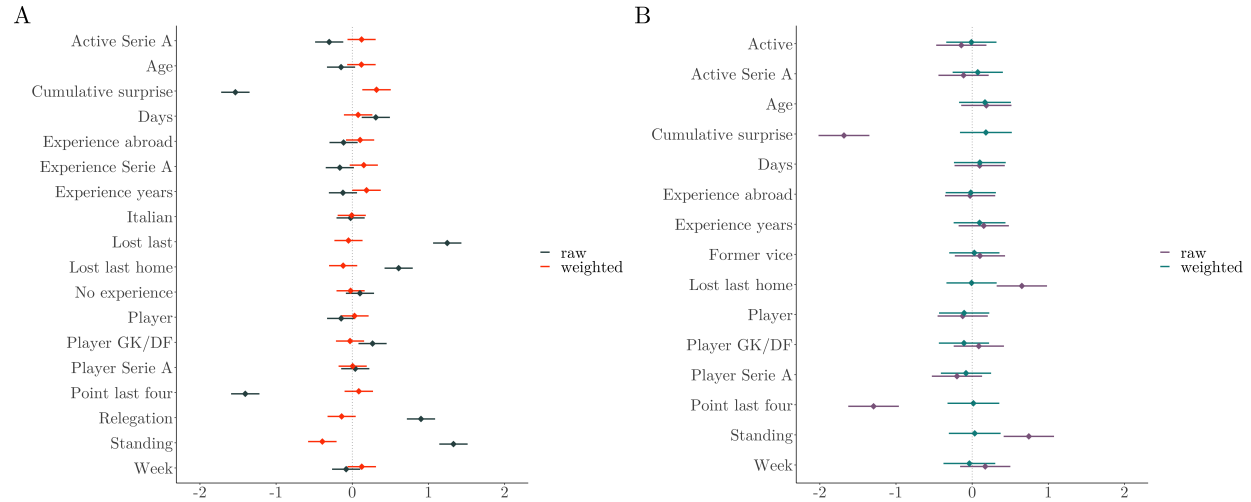
	Treatment I				Treatment II			
	Mean	Min.	Max.	N	Mean	Min.	Max.	N
Treated	.6904	.0270	.9908	114	.7176	.2376	.9769	36
Control	.1577	.0002	.9903	7890	.1648	.0006	.9884	1903
Treated (CS)	.6878	.027	.9900	113	.7176	.2376	.9769	36
Control (CS)	.2779	.027	.9903	4346	.5771	.2395	.9718	433

Notes: Table includes the summary statistics of estimated propensity scores for Treatment I and Treatment II. The first two rows are the statistics related to the full sample, and the last two include those for the observations within the respective common support (CS).

Now that the propensity scores for each treatment have been obtained, we examine how well PSW can reduce the pre-treatment differences between the treated and control groups. Following Austin and Stuart (2015) and Morgan and Todd (2008), we first obtain the standardised mean differences (SMD) in selected covariates between the treated and control groups using a raw sample. Then, we obtain the weighted SMDs using the weights defined by Equation (3.3) and the estimated propensity scores and observations within the common support.¹⁹ Figure 3.4 presents the raw/weighted SMDs in the relevant covariates for Treatment I (Panel A) and Treatment II (Panel B). The Figure suggests that weighting can significantly reduce the pre-treatment differences in relevant characteristics in both cases.

¹⁹Whilst it comes with a cost (efficiency loss), using the observations within common support can further reduce the selection bias (Rosenbaum and Rubin, 1983).

Figure 3.4: Standardised mean differences of selected covariates before and after weighting



Notes: Figure shows the standardised mean differences (SMD) in the selected covariates between treated and control groups for Treatment I (Panel A) and Treatment II (Panel B). Each panel includes SMDs for the respective raw and weighted sample. The horizontal line represents the 95% confidence intervals. The weights used here are based on the propensity scores estimated with GBM for each treatment.

3.5.3 Average treatment effects (ATE) on field performance

It is evident from the previous subsection that PSW can substantially mitigate the unbalancedness in the important pre-treatment characteristics between treated and controlled units for each type of treatment. As explained in Section 3.4.1, we therefore estimate our outcome models (Equation 3.2) by means of weighted regression with PSW in order to obtain the ATE of each treatment. As response variables, average performance in post-treatment matches within different time windows (subsequent one, five, and 10 matches, and all the remaining matches in the relevant season) are obtained for points (*Point*) and goal differences (*Goal dif*). In addition to the treatment and response variables, we control for the strength of a club and their opponent and home field advantage.

The estimation results of ATEs (the coefficients of our treatment variable *New coach*) for single and multiple dismissals are presented in Table 3.7 and 3.8, respectively.²⁰ Table 3.7 suggests that managerial change resulting from the first dismissal in the season has no significant effects on the performance in the first match immediately after the change. This is the case for performance measured with both points obtained (*Point_1*) and goal difference (*Goal dif_1*). However, the first managerial change can improve performance when

²⁰The OLS estimates of ATEs are provided in Table 3.10 and Table 3.11 in Appendix as a comparison. This “naive” approach, where the endogeneity of treatment assignment is not taken into account, generates somewhat different results from the ones presented in Table 3.7 and Table 3.8.

we consider the average performance up to 5 post-treatment matches. In particular, the ATE on the performance measured with average points (*Point_5*) are statistically significant at the 5% significance level; whilst the ATE on the performance measures with average goal differences (*Goal dif_5*) is only significant at the 10% significance level. These results are consistent when we include up to 10 post-treatment matches to obtain the average performance (*Point_5* and *Goal dif_5*). However, the positive ATE of the first dismissal is only significant at the 10% significance level with the response variable *Point_37*, and it is not statistically significant at any conventional significance level with the response variable *Goal dif_37*. Therefore, when we include all the post-turnover matches in the remaining season, the effects of the first dismissal are no longer robust. Overall, whilst the first dismissal does not immediately influence club performance, we observe a limited boost in performance shortly after.

Contrarily, it is evident from Table 3.8 that the ATEs of the second managerial change are not significant at any conventional significance level. Unlike the first turnover, therefore, the second turnover does not influence the club's performance at all. Therefore, the consequences of the first and second managerial turnover are somewhat different, although neither type of dismissal has positive and significant effects in the long run. This partially supports our hypothesis (H2), although the difference is not as clear-cut as we initially expected.

Therefore, these results may justify a club's decision to replace a manager, provided that they have not done so already in the season, although the positive effects are not expected to be persistent. Although our results do not suggest that a consistent managerial change can be detrimental, it still indicates that undergoing yet another managerial change does not make any difference. This is in one sense ironic, given that our results suggest that the second managerial dismissal appears to be made with more caution.

Table 3.7: PSW estimates: ATE of single dismissal on points and goal differences

	<i>Dependent variable:</i>							
	Point_1 (1)	Goal dif_1 (2)	Point_5 (3)	Goal dif_5 (4)	Point_10 (5)	Goal dif_10 (6)	Point_37 (7)	Goal dif_37 (8)
New coach	0.235 (0.240)	0.409 (0.387)	0.361** (0.158)	0.483* (0.250)	0.348** (0.161)	0.470* (0.253)	0.291* (0.174)	0.429 (0.263)
Observations	4,459	4,459	4,459	4,459	4,459	4,459	4,459	4,459
Log Likelihood	−7,990.227	−9,090.177	−5,456.790	−6,802.615	−4,979.050	−6,427.112	−4,759.587	−6,279.965
Akaike Inf. Crit.	15,990.450	18,190.350	10,923.580	13,615.230	9,968.101	12,864.220	9,529.173	12,569.930

Notes: *p<0.1; **p<0.05; ***p<0.01. Robust standard errors in parentheses. All models include control variables (club ability, opponent ability, and home advantage) associated with the response variables.

Table 3.8: PSW estimates: ATE of multiple dismissal on points and goal differences

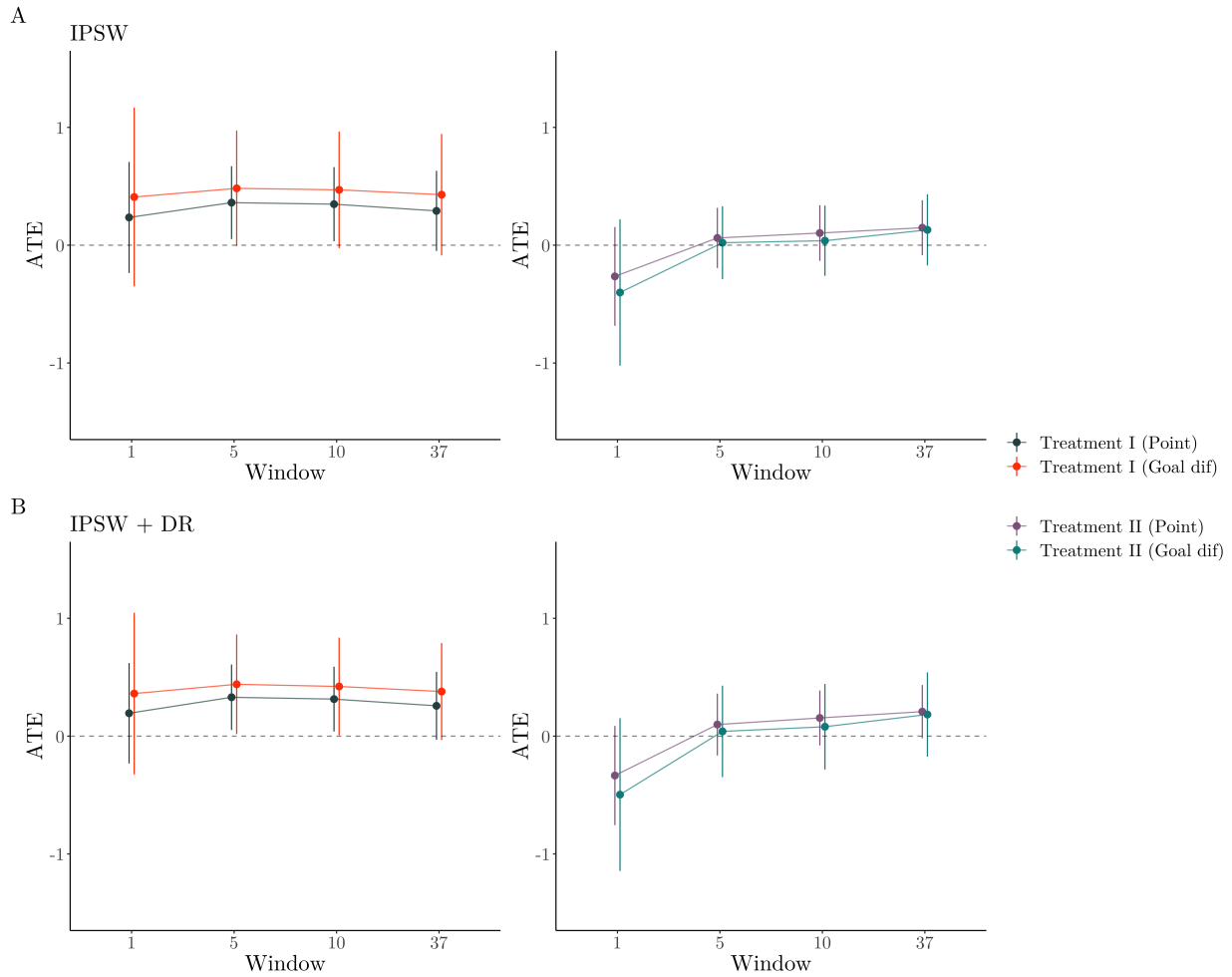
	<i>Dependent variable:</i>							
	Point_1 (1)	Goal dif_1 (2)	Point_5 (3)	Goal dif_5 (4)	Point_10 (5)	Goal dif_10 (6)	Point_37 (7)	Goal dif_37 (8)
New coach	−0.265 (0.214)	−0.402 (0.317)	0.061 (0.130)	0.021 (0.158)	0.103 (0.120)	0.038 (0.152)	0.148 (0.119)	0.130 (0.154)
Observations	469	469	469	469	469	469	469	469
Log Likelihood	−833.997	−964.951	−602.062	−775.277	−561.534	−735.339	−544.708	−720.081
Akaike Inf. Crit.	1,677.994	1,939.901	1,214.125	1,560.553	1,133.068	1,480.678	1,099.416	1,450.162

Notes: *p<0.1; **p<0.05; ***p<0.01. Robust standard errors in parentheses. All models include control variables (club ability, opponent ability, and home advantage) associated with the response variables.

The results discussed above were obtained using inverse propensity scores weighting (PSW). To check the robustness of our results, we compare these results with the double-robust estimator of ATEs. The double-robust estimation follows the PSW procedure, where the outcome model is estimated by means of weighted regression with the weights based on the estimated propensity scores. In addition, this estimation includes the covariates used to obtain the propensity scores in the outcome models, which can increase protection against model misspecifications (Funk et al., 2011).

Figure 3.5 compares the estimated ATEs of single dismissal (Treatment I, left-hand side of each panel) and multiple dismissal (Treatment II, right-hand side of each panel) for PSW (Panel A) and PSW with doubly-robust estimation (Panel B). Outcome variables are again measured with average points (*Point*) and goal differences (*Goal dif*), and the maximum number of post-treatment matches used to obtain the average performance are represented on the horizontal axis. The double-robust estimators of ATEs are very similar to those estimated with PSW. However, the significance of ATEs on some performance measures changes to some degree in that some become slightly more significant. For instance, the ATEs of first dismissals are now positive and significant at the 5% significance level for both measures of outcome (*Point* and *Goal dif*) when we consider up to 5 and 10 post-treatment matches. Overall, however, our conclusion remains: there are limited positive effects of first dismissals, whilst second dismissals barely influence club performance.

Figure 3.5: ATE under different estimation strategies



Notes: Figure presents estimated ATEs for Treatment I (left column) and Treatment II (right columns) with the 95% confidence intervals. The horizontal axis (window) represents the maximum number of post-treatment matches included to obtain response variables. The estimations are obtained with PSW (panel A) and PSW with a double robust estimator (panel B).

3.6 Conclusion

This study examines the causes and consequences of managerial dismissals by differentiating the first and second dismissals that occur within a season. Firstly, in order to identify the causes of the two types of dismissal, we employed logistic regression and GBM. Estimation results of the two classification models indicate that the motivating factors of the two can be different. A first dismissal is likely to have followed a sequence of poor results, measured by points-per-game in recent matches, and a series of matches where outcome has been disappointing relative to expectation. Consistent with previous findings, such as in Tena

and Forrest (2007), the threat of relegation can also influence such a decision. On the other hand, the dominating factor of the second turnover is suggested to be performance against expectation, and the crude measure of performance such as the average points in recent matches has much less influence on a decision to dismiss another manager. In this sense, the second managerial dismissal appears to be considered with greater caution, instead of acting upon the mere streak of unfavourable results. Overall, whilst the predictors of the first dismissal are comparable with the findings from previous studies, for instance, van Ours and van Tuijl (2016), those of the second dismissal could be quite different.

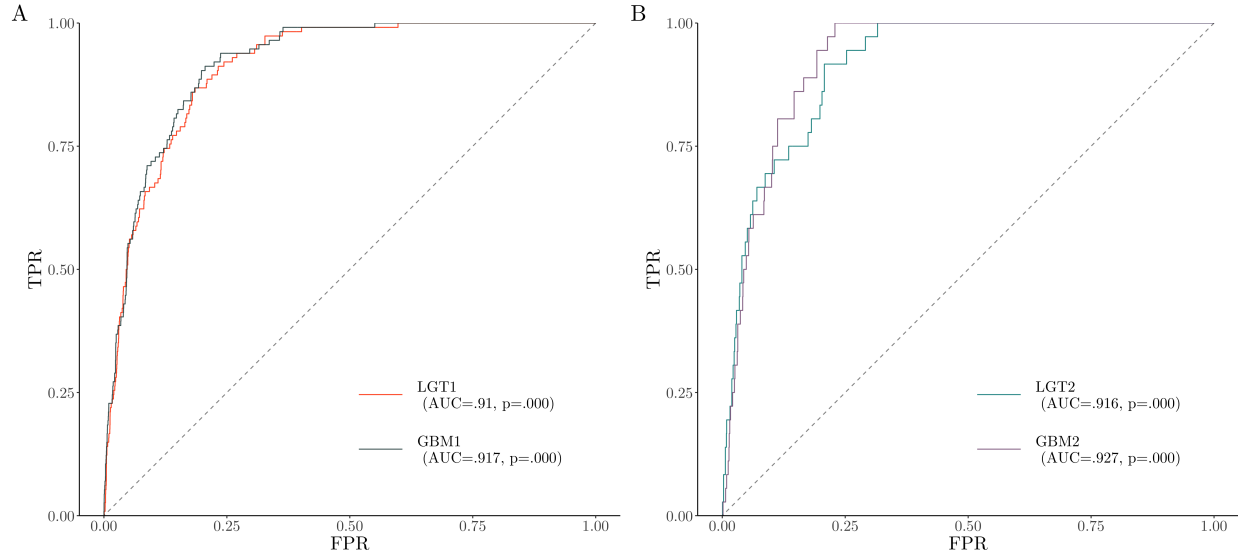
From the methodological point of view, the difference in the factors affecting the likelihood of the two types of dismissal is a relevant concern. Specifically, in order to estimate the effectiveness of managerial dismissals, one needs to take into account the endogeneity of such decisions. To do so, we employ PSW, where the weights based on the propensity score, i.e. the probability of a club dismissing a manager, are applied in order to make the treated and control groups comparable. Estimating the ATEs of the first and second dismissals with PSW, therefore, requires obtaining the probability of the two types of dismissal separately, using the relevant predictors for the respective treatment type.

We find that the consequences of the first and second managerial dismissals are also not identical. The immediate effect of the first turnover is not statistically significant at any conventional significance level, whilst some boost in performance has been observed when we consider up to 5 and 10 matches following the turnover. The effect on the average performance for the rest of the season is, however, non-significant. Despite our findings that indicate the second managerial change is likely to be made with greater caution, the impact of the second dismissal is not statistically significant at any conventional significance level on both immediate and extended post-treatment performance.

These findings add to the previous research on the causes and consequences of managerial dismissals in professional football. The existing studies focus on establishing whether within-season replacements are in general effective or not, and therefore they do not always provide guidance on how to make such decisions. Our findings are more informative since they suggest that managerial dismissals should not be enforced so soon after a recent dismissal since this is unlikely to make any significant difference. This can also add to the understanding of the leadership succession effect in broader management literature. Our findings provide some support for Boyne and Dahya (2002) and Gordon and Rosen (1981)'s hypothesis that frequent leadership change within a short period of time is unproductive. Indeed, whilst it could be tempting to replace leaders of an organisation whenever its performance is not favourable, constant changes in management are unlikely to be beneficial.

Appendix

Figure 3.6: Receiver operating characteristics (ROC) curves and area under the curve (AUC)



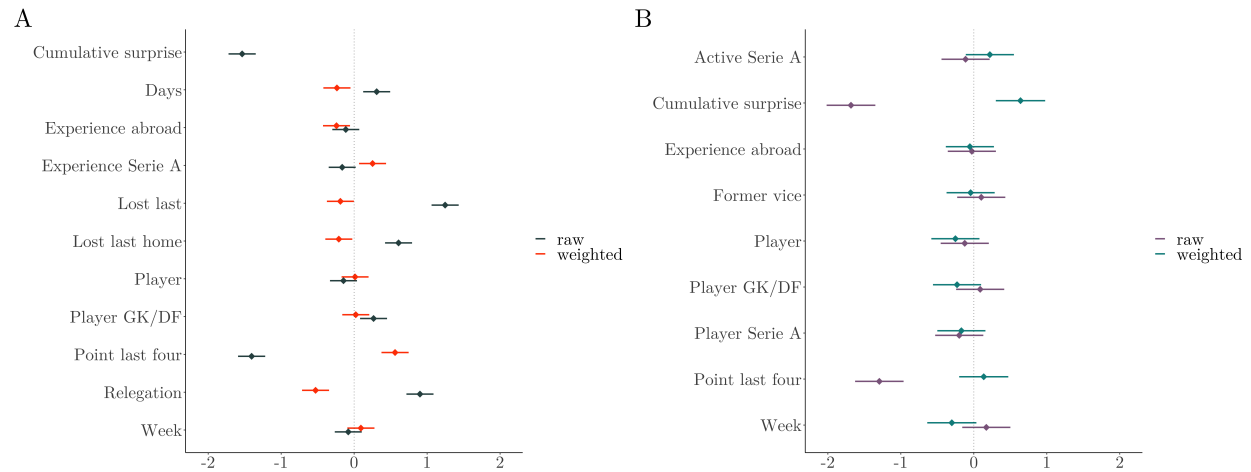
Notes: Figure plots ROC curves for logistic regression for treatment I (LGT1, Panel A) and GBM for Treatment I (GBM1, Panel A), logistic regression for Treatment II (LGT2, Panel B) and GBM for Treatment II (GBM2, Panel B). ROCs represent the true positive rate and the false positive rate at a given discrimination threshold for binary classification problems. A 45-degree line represents the ROC when the binary classification was done by a random selection. Therefore, ROCs with the area under the curve (AUC) significantly larger than that of the ROC of random selection, i.e. 0.5, exhibit a significant separability of the respective models. Also presented in Figure are the values for AUCs as well as p-values associated with the Mann-Whitney-Wilcoxon (MWW) test with $H_0 : AUC = 0.5$ against $H_1 : AUC > 0.5$ for each curve. The AUCs and MWW test suggest that all the models show significant separability. The statistical significance of the differences between the two AUCs between the two competing models for each treatment is tested using DeDong's test, similar to MWW. The associated p-values are $p = 0.162$ for the comparison between LGT1 and GBM1, and $p = 0.2833$ for the comparison between LGT2 and GBM2.

Table 3.9: Balanced accuracy of classification models

	Treatment I	Treatment II
Logit	0.841	0.854
GBM	0.853	0.885

Notes: Balanced accuracy is defined by the arithmetic mean of true positive rate and true negative rate.

Figure 3.7: Covariate balance with logistic regression



Notes: Figure shows the standardised mean differences (SMD) in the selected covariates between treated and control groups for Treatment I (Panel A) and Treatment II (Panel B). Each panel includes SMDs for the respective raw and weighted sample. The horizontal line represents the 95% confidence intervals. The weights used here are based on the propensity scores estimated with logistic regressions for each treatment.

Table 3.10: OLS estimates: ATE of single dismissal on points and goal differences

	<i>Dependent variable:</i>							
	Point_1	Goal dif_1	Point_5	Goal dif_5	Point_10	Goal dif_10	Point_37	Goal dif_37
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
New coach	−0.136 (0.113)	−0.151 (0.145)	−0.034 (0.061)	−0.099 (0.080)	−0.059 (0.054)	−0.123* (0.071)	−0.075 (0.050)	−0.153** (0.068)
Observations	8,004	8,004	8,004	8,004	8,004	8,004	8,004	8,004
R ²	0.156	0.176	0.229	0.252	0.269	0.290	0.286	0.300
Adjusted R ²	0.156	0.176	0.228	0.252	0.269	0.290	0.285	0.300

Notes: *p<0.1; **p<0.05; ***p<0.01. All models include control variables (club ability, opponent ability, and home advantage) associated with the response variables.

Table 3.11: OLS estimates: ATE of multiple dismissal on points and goal differences

	<i>Dependent variable:</i>							
	Point_1	Goal dif_1	Point_5	Goal dif_5	Point_10	Goal dif_10	Point_37	Goal dif_37
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
New coach	−0.333* (0.202)	−0.366 (0.253)	−0.046 (0.111)	−0.131 (0.149)	−0.025 (0.101)	−0.109 (0.135)	0.037 (0.097)	−0.013 (0.129)
Observations	1,939	1,939	1,939	1,939	1,939	1,939	1,939	1,939
R ²	0.116	0.147	0.148	0.155	0.170	0.181	0.181	0.193
Adjusted R ²	0.115	0.145	0.146	0.153	0.168	0.179	0.179	0.191

Notes: *p<0.1; **p<0.05; ***p<0.01. All models include control variables (club ability, opponent ability, and home advantage) associated with the response variables.

Chapter 4

Rating Football Managers with Match-Day Contribution to Performance

4.1 Introduction

The relationship between a firm owner and managers is often described in the principal-agent model, where a firm owner delegates day-to-day management tasks to managers, who are typically more specialised in such tasks and have superior knowledge in the respective industry. Based on the evidence in previous management and economics research (Hendricks et al., 2015; Siebert and Zubanov, 2010; Bloom et al., 2014, 2013), there is probably not much room for debate on the importance of managerial inputs in firm productivity. However, the nature of managerial tasks, such as monitoring and motivating workers, makes the evaluation of managers rather challenging since these are not easily quantified. A hidden information problem is also present in the sense that it is not easy to disentangle the contributions of different inputs. Some managers may seem to be more competent than others, however, this may as well be the case that they are fortunate to have high-quality workers.

This study aims to quantify the managerial contribution to a firm's success in the context of professional football. A sports club is not any different from firms operating in other industries in that it employs managers, labour, and capital to produce output. Frick and Simmons (2008) draw an analogy between professional football managers and firm executives:

Soccer head coaches have roles that resemble that of CEOs. They propose hiring and firing decisions to the board of directors (most often through a Director of Football), and they impose team playing strategies and make tactical adjustments

within games. Head coaches have important motivational roles in trying to raise individual players and team performance. (Frick and Simmons, 2008, p.594)

In addition, top managers in a corporate setting and football managers share similar characteristics such as age, accountability, experience, and resilience to pressure (Pieper et al., 2014).

However, a great deal of publicity at an individual employee level, be it player or manager, is perhaps unique to the professional sports industry. The huge public interest in the industry means the characteristics of individual managers and related events, for instance, managerial turnover, are highly publicised through media. A club's output or performance is clearly and regularly measured. This makes the industry an attractive place to study issues in management and economics. As noted by Kahn (2000), it offers a (natural) laboratory whereby researchers can explore the implications of certain policies or decisions on the outcome of interest. Furthermore, unlike laboratory settings, sports events occur in professional environments with high stakes. Therefore, the sports industry produces valuable data with a sophisticated level of quality and quantity, contributing to the field of management (Day et al., 2012; Lefgren et al., 2015).

At the same time, the conclusions drawn based on football managers are valuable in their own right. In professional football, not only is it an employer who is interested in knowing the abilities of potential employees, but also the stakeholders, including the huge fan base worldwide. In fact, much effort has been devoted to evaluating players by the communities and researchers. For instance, the EA Sports FIFA website, known as *SoFIFA.com*¹ and *WhoScored.com*² offer ratings for the comprehensive sets of players created by scout communities. In academia, Decroos et al. (2020) and Kharrat et al. (2020), among others, contribute to the field by providing objective evaluation systems of players within team sports.

Contrarily, we have not seen much development in the evaluation of individual managers to date. Managers arguably attract as much attention as players; the media covers every event related to managerial change, particularly in the top-tier domestic football leagues. Therefore, advancement in the understanding of managerial contributions and the evaluation system of individual managers would be beneficial to making better-informed decisions regarding the employment of a manager.

This study builds on the previous research on managerial contributions to firm outcomes, including Bertrand and Schoar (2003), Peeters et al. (2020), Muehlheusser et al. (2018), and Buzzacchi et al. (2021), whilst overcoming some of the limitations in these studies. Firstly, the previous studies rely on the separate estimations of firm and leader fixed effects in order

¹Available at <https://sofifa.com/>.

²Available at <https://www.whoscored.com/>.

to disentangle the contributions of managers from firm-specific effects. However, this limits the scope of estimation of individual leaders' effects since this is only feasible for those observed in multiple firms. Therefore, instead of estimating the two sets of fixed effects, we explicitly measure other inputs (labour and capital) to control for resources available for a manager to produce output, whilst capturing individual managers' contributions with their fixed effects.

To do so, we collected individual players' historical performance and clubs' transfer budgets as proxies for labour and capital inputs available for each club. The former is based on player ratings provided by *WhoScored.com* mentioned above, which consists of more than a million data points. We propose some adjustments to the observed ratings of individual players in order to take into account the different levels of difficulty across the leagues. This makes ratings observed in different leagues comparable to each other.

These variables are also allowed to change over time; therefore, we can quantify an individual manager's contribution to the match outcome, given the quality of a squad and financial resources in a particular period of time. This allows us to control for the most up-to-date fitness of the squad and financial strengths of a club, which may not be well-captured by individual club-fixed effects as in the previous studies. We also argue that this is a valid way to evaluate a football manager since the main role of a manager is to optimise a match day outcome given these resources, especially due to the emergence of the role of Director of Football at many clubs, which has taken some control away from a manager in terms of recruitment of players.

Secondly, we extend the studies by Muehlheusser et al. (2018) and Buzzacchi et al. (2021), who provide the rankings of managers within the German and Italian football leagues, respectively, by considering multiple leagues in Europe, the "Big Five" in particular. This is a relevant extension due to the fact that the labour market for football managers is mobile, and many managers are observed across different leagues. The proposed measures of the inputs are also comparable across the leagues, making this extension feasible.

Finally, in order to take into account the randomness of match outcome, we measure a club's performance with expected goals, which capture the quality of chances created, using information related to individual shot data. As shown by recent studies such as Flepp and Franck (2021), this measure of performance is more informative than the conventional crude measure of outcome based on average points as in Muehlheusser et al. (2018) and Buzzacchi et al. (2021).

Employing the advanced measures of output and inputs, our study finds the overall significance of managerial input and significant heterogeneity among managers. Furthermore, our findings confirm that taking into account the strength of players and finance as well as

uncertainties associated with the randomness of match outcome matters in the evaluation of managerial contribution.

The remainder of this paper is organised as follows. The next section provides an overview of related studies that give us some direction to achieve our objectives set out above. Section 4.3 describes data used in this analysis. Section 4.4 explains constructions of the important variables and models to be estimated. Our empirical results are presented in Section 4.5, followed by concluding remarks in Section 4.6.

4.2 Related literature

There are a few strands of literature that can give us some directions to achieve our objectives set out in the introduction; quantifying the managerial contribution to the firm success and the heterogeneity among managers. First, previous studies in economics and management have estimated the contributions of managerial inputs by empirically disentangling the individual “firm” and “manager” effects on production. One of the most cited studies by Bertrand and Schoar (2003) employs a CEO-firm matched data to estimate managers’ fixed effects on the performance and behaviour of firms, separating them from the firms’ fixed effects on the variables of interest. This approach is followed by many others, including Graham et al. (2012) and Lazear et al. (2015), to establish a relationship between the heterogeneity in managerial ability and various measures of corporate outcomes.

In the sports domain, Peeters et al. (2020), Muehlheusser et al. (2018), and Buzzacchi et al. (2021), among others, have applied a similar method to investigate the significance of managers in professional sports clubs. In particular, Muehlheusser et al. (2018) and Buzzacchi et al. (2021) employ data from the top-tier professional football leagues, German Bundesliga and Italian Serie A, respectively. Both studies show that individual managers’ effects are economically and statistically meaningful for explaining the variations in field performance measured by the average points obtained per match per half-season. Buzzacchi et al. (2021) in addition find that financial performance measured by the growth in players’ market values is partly explained by individual manager’s effects.

The studies mentioned above conclude that managerial contributions are statistically significant, that is, managers do matter to determine the firm’s success both in corporate and sports settings. This view is supported by another strand of literature on technical efficiency, which suggests that the extent to which a firm can achieve technical efficiency, i.e. the optimal output given production factors available, is dependent on the quality of managers. For instance, Frick and Simmons (2008) and Dawson et al. (2000) provide empirical evidence that technical inefficiencies are reduced by employing more competent

managers in German Bundesliga and English Premier League, respectively. Such relationship is also established in other sports (Hofler and Payne, 2006; Brian, 2013) and beyond the sports industry (Papadopoulos, 2021).

Therefore, it is not surprising that leaders are rewarded or punished according to the field performance. The most common causes of dismissals of football managers are indeed poor performance in recent matches (van Ours and van Tuijl, 2016; Tena and Forrest, 2007). Nevertheless, the prospect of improvement in post-succession performance is not well-supported by empirical evidence. Employing the data from the Dutch football league, van Ours and van Tuijl (2016) show that seemingly positive effects of managerial change are not due to the succession itself, but merely a regression to the mean performance level. Tena and Forrest (2007) find positive effects of a new manager in Spanish football, however, this is limited to performance in home matches, concluding that such improvement in performance at home may simply reflect a boost in support by the fans. Evidence from National Hockey League (Rowe et al., 2005) suggests that within-season managerial change can deteriorate performance, followed by positive long-term effects.

This unfortunate phenomenon, particularly in the short term, prevails beyond the sports industry. A meta-analysis conducted on the relationship between CEO successions and firm outcomes by Schepker et al. (2017) concludes that leadership succession is on average costly to organisations since it causes disruptions in the short term, and any strategic change could take a while to manifest its effect. In this respect, a firm seeking a new leader to replace the incumbent would need to identify one that can at least partly offset these possible negative effects of succession. According to the studies cited above, however, this proves challenging, since we would otherwise see more successful cases of leadership changes.

One of the aspects that could be hindering an effective evaluation of managers is the randomness of outcomes. In professional football, in particular, any measure of performance that relies on match outcome is particularly susceptible to this concern due to the low-scoring nature of the sport. For instance, findings in a recent study of managerial dismissals in European football leagues by Flepp and Franck (2021) imply that other metrics, such as expected goals that capture the qualities of each shot within a match, can better reflect a manager's performance than conventional measure purely based on match outcome. In particular, they find that a club's performance improves after a streak of under-performance measured by match outcome, regardless of replacing a manager, whereas improvement is only expected with clubs who replace a manager when a club is experiencing under-performance based on expected goals. In the former case, it can be argued that performance simply reverts back to a mean level, since adverse performance leading up to dismissal was merely due to bad luck. On the other hand, dismissing a manager with poor performance based on

expected goals improves the situation since this measure better captures a manager’s actual ability, rather than luck. This suggests that it is important to account for the randomness of match outcome when evaluating a manager’s performance.

The current study extends the previous studies estimating the managerial contribution to firm performance, such as Muehlheusser et al. (2018) and Buzzacchi et al. (2021) in the following ways. Firstly, the past studies disentangle the manager’s effects from other contributors to firm performance by measuring the latter by individual firm effects. In the context of football clubs, these are club-specific fixed effects, which capture heterogeneity among individual clubs, or “how big a club is.” However, these effects are effectively held constant over time. This can be a rather strong assumption since the resources available in a specific club is likely to be time-variant. For instance, the quality of players and a club’s financial strengths can vary over time. In addition, separately estimating the two sets of fixed effects (managers and clubs) could significantly limit the scope of evaluation of individual managers. This is because such estimation requires managers to be observed in more than one club.

Therefore, we will explicitly control for the resources that are available for a manager to work with in order to achieve on-field success. As mentioned in the introduction, there are sophisticated rating systems for individual players in professional football, which can be used as a proxy for the quality of labour input. This measure of labour quality can provide more accurate information on the player’s fitness at a particular point in time since this is updated every time a player appears on a pitch. Furthermore, a club’s financial resources will be controlled with transfer budgets available, which also vary over time.

Second, Muehlheusser et al. (2018) and Buzzacchi et al. (2021) measure field performance with the average points obtained in a given half season. Following the discussion above, however, this measure is less capable of capturing a club’s actual performance. Therefore, we measure a club’s performance using expected goals. Finally, whilst the previous studies focus on a single league, we estimate the individual manager’s contribution using multiple leagues. This is a relevant matter since the labour market for football managers is characterised by great mobility. Our measures of labour and capital inputs are comparable across the leagues, making the analysis with multiple leagues feasible.

These extensions allow us to compare managers who are operating in the different leagues in a way robust to random variation, without limiting the pool of managers to those hired by multiple clubs during the sample period. Therefore, this study contributes to understanding of managerial inputs and the development of the evaluation system of a manager by estimating heterogeneity amongst managers based on their contributions to match-day performance, given the resources at hand in the particular period of time.

4.3 Data

4.3.1 Match and manager data

We collected data from the top-tier football leagues in Spain, France, England, Italy, and Germany, known as the “Big Five”, for seasons from 2014/2015 to 2020/2021. A very similar league format is followed by the five leagues included in our sample. Each league features the 18 or 20 most competitive clubs in the respective countries. A club competes with all others twice during a season, once at its home stadium and once away. With 20 participating clubs, this amounts to 38 matches per club per season or 380 matches per league per season. A club is rewarded with points after each match based on a match outcome; 3, 1, and 0 points for a win, draw, and loss, respectively. At the end of a season, the championship title is awarded to the club with the highest number of cumulative points within the respective league, whilst the weakest clubs are relegated to the second-tier league.³ Table 4.1 provides the number of matches used in the main analysis by leagues.

Table 4.1: Number of matches used in the analysis by leagues

League	Matches
England Premier League	2,658
France Ligue 1	2,554
Germany Bundesliga	2,140
Italy Serie A	2,658
Spain LaLiga	2,658
Total	12,668

Notes: Seasons 2014/2015-2020/2021.

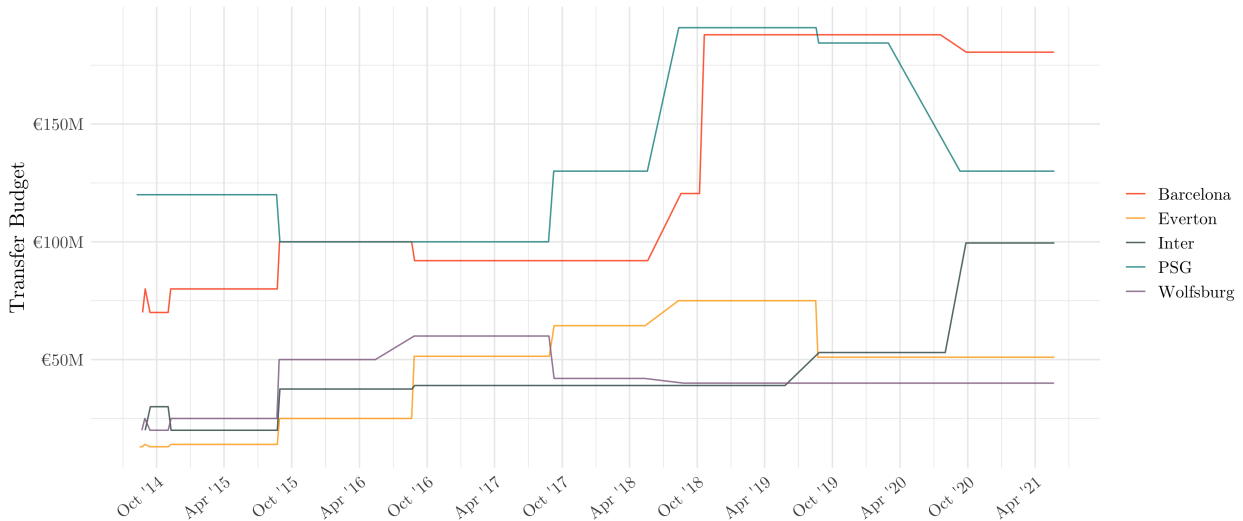
During the sample period, we observe 405 managers across 146 clubs, who together managed 12,668 matches. As mentioned in Section 4.2, the previous studies relied on the “mover condition,” where a manager has to be observed or matched with more than one club, to estimate the individual effects of the managers. We argue that this condition could be quite restrictive since the number of managers whose effects can be estimated would be reduced significantly, which hinders the comparison among different managers. One way to overcome this issue is to expand the number of individual clubs or seasons, across which a manager may be observed.

³In general, the bottom three clubs are relegated, whilst the number could vary slightly since some leagues feature playoffs.

We indeed build on the previous studies by covering more than a single league. This is certainly relevant in the context of European football since the labour market for football managers is mobile, particularly at the top tier. Nevertheless, there is a limit on expanding the time horizon. This is because some data are only available for more recent seasons. Whilst match-level statistics go back to as far as the establishment of the leagues, more advanced data, such as event-level data, have only recently been introduced. As explained later in this section, part of our analysis relies on this type of data. Regardless, the value added by including older seasons can be limited, due to the attrition of managers and the fact that the more recent data carry more useful information to predict future performance.

Imposing the mover condition on our sample would shrink the pool of managers to 142 individuals. To avoid this, we employ alternative measures to control for club heterogeneity, in terms of resources available for a manager to work with. In particular, we collected information on a club's transfer budgets, which change over time. This is obtained from the EA Sports FIFA website, known as *SoFIFA.com*⁴. The website provides individual club profiles that are updated regularly. Another advantage of this measure is that it is directly comparable among clubs that are operating within different leagues. Figure 4.1 plots the transfer budgets over the sample period for the five selected clubs: Barcelona (Spain), Everton (England), Inter (Italy), PSG (France), and Wolfsburg (Germany). The Figure suggests that there is significant heterogeneity in budgets available among clubs and that it varies across time within clubs.

Figure 4.1: Transfer budget for selected clubs (2014/2015-2020/2021)



We do, however, impose a restriction on managers' appearances. In particular, we es-

⁴Available at <https://sofifa.com/>

timate the individual fixed effects of managers who managed more than 10 matches in the sample period. The threshold is used since with our sample this threshold fairly separates permanent managers from caretaker managers. This leaves us with 322 managers. The number of matches managed by the 322 individuals ranges from 11 to 266, with the mean and median values of 77.15 and 53, respectively.

4.3.2 Event data

In addition to the match-level data, shot information is extracted from the Opta F24 feed, so-called “event data.” Within the sample period, 321,074 shots are identified. For each shot, we gathered the outcome (whether it led to a goal), and the following characteristics: location ((x, y) coordinate), body part involved, and situation. The (x, y) coordinate is used to obtain the distance from the centre of the goal (*Distance*) and the angle between the coordinates of the shot origin and the two goal posts (*Angle degrees*). The categorical variable *Body part* indicates whether a shot involved a header or other body parts (right/left foot or other body parts). *Situation* variable is categorised into free-kick, open play, penalty, or corner kick. The descriptive statistics of shot data are presented in Table 4.2.

Table 4.2: Descriptive statistics for shot data

Variable	N = 320,050 ¹
Goal	
0	285,941 (89%)
1	34,109 (11%)
Distance	17 (12, 25)
Angle degrees	20 (15, 31)
Body part	
Foot	265,069 (83%)
Header	54,981 (17%)
Situation	
Corner	46,894 (15%)
Free kick	35,866 (11%)
Open play	233,255 (73%)
Penalty	4,035 (1.3%)

Notes: ¹n (%); Median (IQR).

4.3.3 WhoScored rating

The data for individual players were collected from *WhoScored.com*.⁵ They provide players' ratings (Who Scored (WS) ratings), which reflect the contribution of individual players to the outcome in every match. This means that the rating for a particular individual is updated every time they appear on the pitch. The initial value of the rating is set at 6.0 and the values can increase (up to the maximum value of 10.0) or decrease based on in-match and post-match statistics. In-match statistics are related to their action (dribble, shots, etc.). This is then weighted with its impact measured by the area on the pitch and outcome of the event, and reflected in their rating. At the end of the match, the overall outcome of the match is added to the rating, where such an outcome is weighted based on the individual's influence, determined by playing position, minutes played, etc. The advantages of using WS ratings are that the player's performance is evaluated regularly and that the ratings are obtained for virtually all professional leagues around the world, using the same method. Previous studies, Frick and Simmons (2008), for instance, used players' wage bills to capture labour input. However, this measure is generally only observed when a new contract is agreed,⁶ and is available for limited countries.

The website covers 18 major professional football leagues, and there are 1,048,877 WS ratings available that resulted from 37,442 matches within these leagues over the seasons 2013/2014-2020/2021. The number of WS ratings collected by leagues is summarised in Table 4.3.

⁵Available at <https://www.whoscored.com/>.

⁶Hence it could even be less often than annually.

Table 4.3: Number of WS ratings by leagues

League	N(Ratings)
Argentina Superliga	46,463
Belgium Jupiler Pro League	8,703
Brazil Brasileirão	75,063
England Championship	122,343
England League One	15,610
England League Two	27,556
England Premier League	83,761
France Ligue 1	82,229
Germany Bundesliga	68,963
Germany Bundesliga II	47,442
Italy Serie A	86,121
Netherlands Eredivisie	65,720
Portugal Liga NOS	43,782
Russia Premier League	53,107
Scotland Premiership	5,659
Spain LaLiga	85,878
Turkey Super Lig	63,873
USA Major League Soccer	66,604
Total	1,048,877

Notes: Seasons 2013/2014-2020/2021.

For each WS rating collected, our data set indicates the date on which a rating is observed, i.e. the date of the match on which a rating is based, the name of the player associated with the rating, their date of birth, and the league in which they played on the date. In light of the adjustments we are going to propose in the following section, we identify the players who were observed in more than one league during the sample period and played at least 10 matches within a given league. We identify 3,478 players who satisfy these conditions, and the number of ratings observed for these players amounts to 406,239.

4.4 Methodology

Our ultimate goal is to estimate a club's production function relative to its opponent at a match level and disentangle an individual manager's contributions from those of other inputs, labour and capital, measured at the time when a match takes place. Therefore, using the data presented in Section 4.3, we construct our output and input variables at a match level.

Our output measure is the so-called expected goals (xG). This is based on the probability of a shot leading to a goal, given observed characteristics of the shot. To obtain the output measure at a match level, therefore, these probabilities are aggregated over all the observed shots in a specific match.

As for input measures, we use transfer budgets and individual players' historical performance (*WhoScored* ratings) as proxies for financial strengths and labour quality. The latter is observed at an individual player level, hence we construct a variable to capture the strength of the squad on a particular match day, using the historical performance of the players within the squad. The player ratings used in this study are based on an impact that he makes in a particular match and created at a match-level across most of the major leagues across the world.⁷ However, these ratings are not directly comparable across the leagues in that it is easier for a player to make an impact in lower-level leagues than in higher-level ones. Therefore, the first adjustment we make is to take into account the different levels of difficulty across the leagues. Once we obtain league-adjusted ratings for individual players, we compute the weighted average of ratings observed within two years prior to the match for which we estimate the production function. The weights applied to each rating available in the time window are based on how far it is in the time horizon so that a more recent rating is given a higher weight than an old rating. In addition, the initial value of rating (the rating of a particular player exactly two years prior to the match day) is set at the average value of league-adjusted ratings in the league where the match under consideration is observed. This is so that the weighted average of ratings for players with very few observations is shrunk towards this initial value, rather than skewed towards these few observations. We then obtain the mean value of the weighted average of league-adjusted historical ratings over players in the squad available for the match under consideration to capture the overall strength of the squad on the day.

Finally, we estimate the match-level production function using the output and input metrics computed as above together with the individual manager dummies, where all output and input are measured for a home club relative to an away club. The remainder of this section describes each step in detail.

4.4.1 xG model

The conventional way to measure output is based on the tertiary outcome (win, loss, and draw) or points earned in each match. However, this measure is susceptible to random variation in match outcome. This is particularly relevant due to the low-scoring nature of

⁷See Table 4.3 for the list of the leagues in which the individual players' historical ratings are observed.

association football, where relying on match outcomes that are heavily affected by random forces can lead to misjudgment (Brechot and Flepp, 2020). Therefore, this study measures performance with scoring chance and its quality, which is observed much more frequently than goals. In particular, we adopt the concept of expected goals (xG), which assigns the probability of scoring to each shot that occurred in a match. Whilst the number of studies that employ xG as a measure of performance in professional football is yet sparse (Kharrat et al., 2020; Flepp and Franck, 2021; Brechot and Flepp, 2020), the measure is shown to add to the predictive power for future performance (Brechot and Flepp, 2020).

Using the shot data described in Section 4.3.2, we first estimate the log odds of a shot leading to a goal, given the characteristics of the shot as follows:

$$\text{Ln} \left(\frac{\text{P}[\text{Goal}_{ikts} = 1]}{\text{P}[\text{Goal}_{ikts} = 0]} \right) = f(\text{Angle degrees}_{ikts}, \text{Distance}_{ikts}, \text{Body part}_{ikts}, \text{Situation}_{ikts}), \quad (4.1)$$

where i , k , t , and s represent a club, shot, match, and season, respectively. We estimate the model (4.1) by means of logistic regression, using observations from the first season (2013/2014). Once the model is estimated, we obtain the fitted $\hat{\text{P}}[\text{Goal}_{ikts} = 1]$ for each shot and aggregate these values for home and away clubs in a particular match in the rest of the seasons (2014/2015 - 2020/2021). This gives us the expected goals for home and away clubs in the match.

4.4.2 Adjustments of WS ratings

We propose some adjustments to WS ratings presented in Section 4.3.3 in order to (1) make the ratings comparable across the leagues, (2) mitigate the possible bias towards very few observations for some players, and (3) take into account the fact that newer observations carry more useful information than older ones. Given that WS ratings essentially reflect an individual player's contributions to within and overall match performance, the rating of an individual player is influenced by how difficult it is to make an impact in the particular league in which he plays, because different leagues have different standards of play. For instance, a player in the top-tier English league may have a low rating, however, his rating may well be higher if measured within the second-tier league. To measure the quality of a player on a particular match day, historical ratings of the player prior to the match day will be taken into account. However, a player typically moves across leagues, as his career progresses. Therefore, it is important to adjust the ratings in order to make them comparable across the leagues in which they are observed.

To achieve this, we first identify the degree of heterogeneity among leagues by running the following regression:

$$\text{WS} = \beta_0 + X'\gamma + L'\rho + \beta_1\text{Age} + \beta_2\text{Age}^2 + \varepsilon, \quad (4.2)$$

where X and L are vectors of dummy variables for individual players and leagues, respectively. A variable Age and its quadratic term are included to control for the general progression of a player's quality over their career. Therefore, ρ is the vector of league coefficients of interest. To estimate ρ , we use the observations for raw WS ratings that belong to players who have been observed in at least two leagues where they appeared in a minimum of 10 matches. As described in Section 4.3.3, we identify 3,478 players who meet the criteria, for whom we have 406,239 observed ratings.

Once we obtain the league coefficients, we add the negative value of the respective league coefficient to a realised (raw) WS rating to obtain league adjusted WhoScored (LAWS) rating defined as follows:

$$\text{LAWS} = \text{WS} + (-\hat{\rho}_l), \quad (4.3)$$

where ρ_l is the estimated league coefficient for a league l . The term $(-\hat{\rho}_l)$, therefore, can be interpreted as league strength, or the level of difficulty for a player to make a positive effect on the game.

Using LAWS ratings, we measure a player's ability at time t using all the available LAWS ratings in the preceding periods of two years, $(\text{LAWR}_1, \text{LAWR}_2, \dots, \text{LAWR}_n)$. Previous literature on the prediction of outcome in football (Boshnakov et al., 2017; Kharrat et al., 2020; Dixon and Coles, 1997) suggests that recent performance carries more predictive power for future performance. Accordingly, we apply the following weighting function for each LAWS rating i :

$$w_i = \exp(\xi(\text{Date}_i - \text{Date}_t)/3.5), \quad (4.4)$$

where ξ is the time-weighting parameter, Date_t is the date when the player's ability is measured, and Date_i is the date when LAWS_i is observed. Following the studies mentioned above, time distances are scaled in half-week units ($= 3.5$ days), and set $\xi = 0.002$. Therefore, LAWS_i that are recently observed (closer to time t) are given exponentially higher importance to measure a player's ability at time t .

The number of LAWS ratings available for a certain player at time t , i.e. the number of LAWS_i between time t and $t - (2 \text{ years})$, may vary significantly among players. If a player has very few observations, this could result in a bias towards such observations. To circumvent this, we set the initial values (a rating that would have been observed exactly

two years prior to t) to be the average value of LAWS for a respective league. By doing so, in an extreme case where a player has no previous observation at time t , his ability at time t will shrink towards the mean value in the respective league, weighted according to the time difference of two years. On the other hand, if a player is regularly observed, this shrinkage hardly affects their rating, particularly because the time discount applied to the initial value is large. Therefore, we measure the ability of a player at time t with the following weighted value of LAWS:

$$\overline{\text{LAWS}}_t = \left(\tilde{w}_0 + \sum_{i=1}^n w_i \right)^{-1} \left(\tilde{w}_0 \times \text{LAWS}_l + \sum_{i=1}^n w_i \times \text{LAWS}_i \right), \quad (4.5)$$

where \tilde{w}_0 is the weight with the time distance of two years; n is the number of the player's LAWS ratings available between time t and $t - (2 \text{ years})$; LAWS_l is the average LAWS ratings of the league l to which a player belongs at time t .

We therefore calculate $\overline{\text{LAWS}}_t$ for each available player on match day t using WS ratings observed between $t - (2 \text{ years})$ and the last appearance of a respective player prior to t . We then obtain average values separately for the starting eleven and substitutes, which gives overall strength of the squad available for a manager to work with on a particular match day t . The advantage of measuring the quality of players this way is that it can vary throughout the season, hence allowing it to capture possible drops in the strengths, for instance, due to injuries. The fact that individual players' ability is updated every time they are observed on the pitch means the fluctuation in their performance is also reflected. These are the elements that are not captured in other measures, for instance, their wage bills or even market values.

4.4.3 Production function

We model an individual club i 's output within a match that takes place at time t in season s , (Y_{its}) , as a function of inputs, i.e. labour (L_{its}) and capital (K_{its}) . In addition, we include home advantage (γ_{it}) and managers' fixed effects (μ_i) . Specifically, a club i 's production function is defined as the following Cobb-Douglas production function:

$$Y_{its} = L_{its}^{\beta_l} K_{its}^{\beta_k} \exp(\gamma_{it} \mu_i), \quad (4.6)$$

where the return parameters β_l and β_k measure the impact of labour and capital. Ultimately, a match outcome is dependent on a club i 's output relative to that of its opponent club j ,

therefore, we rewrite the model (4.6) in logs and take the differences between clubs i and j ⁸:

$$y_{its} - y_{jts} = \gamma_{it} - \gamma_{jt} + \beta_l(l_{its} - l_{jts}) + \beta_k(k_{its} - k_{jts}) + \mu_i - \mu_j + \varepsilon_{ijts}, \quad (4.7)$$

where ε_{ijts} is a match-specific error term.

We now look at model (4.7) from a home club's perspective, and consider a match outcome as a home club's log xG net of an away club's log xG ($d_log_xG_{ts}$). Similarly, labour and capital inputs are included as a log-difference form of the measure for the strengths of the squad and transfer budgets, respectively. Furthermore, to obtain an individual manager's coefficients, we include a 322×1 vector M_{gt} of manager dummies m_{ts} , where:

$$m_t = \begin{cases} -1 & \text{if a manager } m \text{ manages the away club in match } t, \\ 0 & \text{if a manager } m \text{ manages neither the home nor away club in match } t, \\ 1 & \text{if a manager } m \text{ manages the home club in match } t. \end{cases} \quad (4.8)$$

The resulting production function is, therefore:

$$d_log_xG_{gt} = (\gamma_h - \gamma_a) + \beta_l(l_{hgt} - l_{agt}) + \beta_k(k_{hgt} - k_{agt}) + M_{gt}'\mu + \varepsilon_{gt}, \quad (4.9)$$

where μ is a 322×1 vector of the coefficients for individual managers, and the first term $(\gamma_h - \gamma_a)$ is interpreted as a home advantage and estimated as an intercept of the model.

4.5 Results

4.5.1 xG and naive rankings

The estimation results of the expected goal model (4.1) using the first season (2013/2014) is summarised in Table 4.4. All the variables are statistically significant, and the estimated parameters have expected signs.

⁸The definition and transformation of the production function for a sports club are similar to those in (Peeters et al., 2014).

Table 4.4: Estimated xG model parameters

	<i>Dependent variable:</i>
	Goal
Angle degrees	0.032*** (0.001)
Distance	−0.084*** (0.004)
Body part	
Header	−1.070*** (0.047)
Situation	
Free kick	0.642*** (0.092)
Open play	0.669*** (0.077)
Penalty	3.275*** (0.129)
Constant	−2.186*** (0.125)
Observations	47,766
Log Likelihood	−12,991.710
Akaike Inf. Crit.	25,997.420
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

We then obtain the fitted values ($\hat{P}[\text{Goal}_{ikts} = 1]$, where i = club, k = shot, t = match, and s = season) for the observations within seasons 2014/2015-2020/2021. Finally, to obtain expected goals per club per match (xG_{its}), we aggregate these fitted values for home and away clubs in a particular match. Figure 4.2 plots the values obtained for xG for home and away clubs in the main sample.

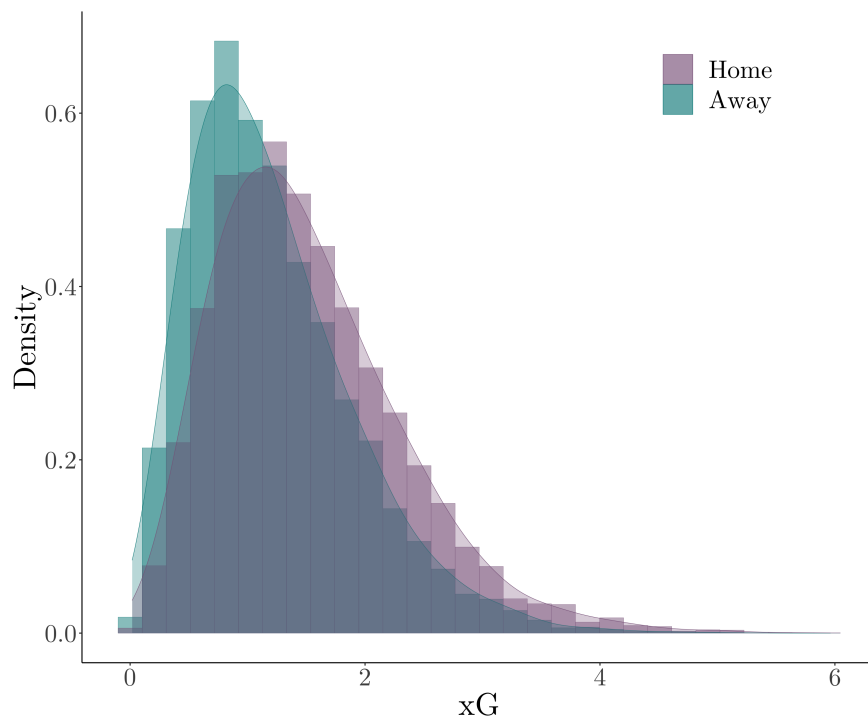


Figure 4.2: Estimated xG values for home and away clubs

One of the conventional ways to measure a football manager’s success is to look at a winning percentage (WP) over their career.⁹ As a comparison, Table 4.5 provides the rankings of managers in terms of WP and the average net xG (d_xG), i.e. xG differentials between the managing club and their opponents, for the top 50 and selected managers. We call these “naive” rankings since they do not take into account the strengths of a squad and the financial resources that a manager has at his disposal. The only difference between the two rankings is the performance (output) metric.

Nevertheless, the Table provides quite different pictures, implying that not taking into account the randomness of the outcome can provide misperception with respect to an individual manager’s performance. For instance, Diego Simeone and Santiago Solari are ranked 10th and 11th, respectively, in terms of the WP. However, the two Argentinean managers are lower-ranked (35th and 39th, respectively) in terms of xG differentials. This may indicate that the WPs of the two reflect good luck to some extent. On the other hand, Stefano Pioli and, more notably, Graham Potter are examples of managers whose abilities may not be reflected in their WPs. Stefano Pioli, a manager of AC Milan (Italy), is ranked 47th and 27th in terms of xG differentials and WP, respectively. Graham Potter, a manager of Brighton & Hove Albion F.C. (England), is ranked much higher in terms of xG differentials

⁹This is obtained by $WP = ((N \text{ of wins}) + 0.5 (N \text{ of draws})) / (N \text{ of matches})$.

(62nd) than WP (173th). Note, however, that both rankings are again naive in the sense that other inputs that would have contributed to either measure of performance are not taken into account.

Table 4.5: Naive rankings (winning percentage and xG differential)

Ranking	Manager (Club)	WP	Ranking	Manager (Club)	d_xG
1	Josef Heynckes (Bayern)	0.870	1	Luis Enrique (Barcelona)	1.483
2	Hans-Dieter Flick (Bayern)	0.836	2	Josep Guardiola (Man City)	1.396
3	Luis Enrique (Barcelona)	0.825	3	Josef Heynckes (Bayern)	1.263
4	Laurent Blanc (PSG)	0.822	4	Hans-Dieter Flick (Bayern)	1.227
5	Massimiliano Allegri (Juventus)	0.821	5	Thomas Tuchel (PSG)	1.146
6	Josep Guardiola (Man City)	0.807	6	Laurent Blanc (PSG)	1.072
7	Zinedine Zidane (Real Madrid)	0.781	7	Zinedine Zidane (Real Madrid)	0.901
8	Thomas Tuchel (PSG)	0.766	8	Frank Lampard (Chelsea)	0.849
9	Antonio Conte (Inter)	0.765	9	Jurgen Klopp (Liverpool)	0.812
10	Diego Simeone (Atletico)	0.736	10	Massimiliano Allegri (Juventus)	0.811
11	Santiago Solari (Real Madrid)	0.735	11	Andrea Pirlo (Juventus)	0.802
12	Andrea Pirlo (Juventus)	0.724	12	Maurizio Sarri (Juventus)	0.731
13	Luciano Spalletti (Inter)	0.720	13	Antonio Conte (Inter)	0.730
14	Jurgen Klopp (Liverpool)	0.707	14	Ralf Rangnick (RBL)	0.728
15	Maurizio Sarri (Juventus)	0.706	15	Luciano Spalletti (Inter)	0.694
16	Carlo Ancelotti (Everton)	0.696	16	Arsene Wenger (Arsenal)	0.680
17	Ralf Rangnick (RBL)	0.691	17	Carlo Ancelotti (Everton)	0.638
18	Mauricio Pochettino (PSG)	0.675	18	Bruno Genesio (Rennes)	0.619
19	Ernesto Valverde (Barcelona)	0.675	19	Edin Terzic (Borussia Dortmund)	0.607
20	Edin Terzic (Borussia Dortmund)	0.674	20	Gian Piero Gasperini (Atalanta)	0.599
21	Jorge Sampaoli (Marseille)	0.673	21	Niko Kovac (Monaco)	0.579
22	Adi Hutter (Eintracht Frankfurt)	0.671	22	Paulo Fonseca (Roma)	0.577
23	Leonardo Jardim (Monaco)	0.669	23	Julen Lopetegui (Sevilla)	0.570
24	Bruno Genesio (Rennes)	0.668	24	Julian Nagelsmann (RBL)	0.563
25	Unai Emery (Villarreal)	0.665	25	Peter Bosz (Leverkusen)	0.545
26	Ole Gunnar Solskjaer (Man Utd)	0.665	26	Unai Emery (Villarreal)	0.535
27	Lucien Favre (Borussia Dortmund)	0.663	27	Stefano Pioli (AC Milan)	0.523
28	Julen Lopetegui (Sevilla)	0.663	28	Rudi Garcia (Lyon)	0.504
29	Gennaro Gattuso (Napoli)	0.659	29	Ole Gunnar Solskjaer (Man Utd)	0.469
30	Arsene Wenger (Arsenal)	0.658	30	Gennaro Gattuso (Napoli)	0.467
31	Andre Villas-Boas (Marseille)	0.656	31	Guus Hiddink (Chelsea)	0.453
32	Rudi Garcia (Lyon)	0.654	32	Ronald Koeman (Barcelona)	0.420
33	Jose Mourinho (Tottenham)	0.649	33	Brendan Rodgers (Leicester)	0.416
34	Louis Van Gaal (Man Utd)	0.638	34	Ernesto Valverde (Barcelona)	0.399
35	Simone Inzaghi (Lazio)	0.637	35	Diego Simeone (Atletico)	0.398
36	Julian Nagelsmann (RBL)	0.636	36	Mauricio Pochettino (PSG)	0.395
37	Hubert Fournier (Lyon)	0.632	37	Simone Inzaghi (Lazio)	0.393
38	Gian Piero Gasperini (Atalanta)	0.620	38	Jorge Sampaoli (Marseille)	0.387
39	Sergio Conceicao (Nantes)	0.614	39	Santiago Solari (Real Madrid)	0.380
40	Niko Kovac (Monaco)	0.613	40	Leonardo Jardim (Monaco)	0.370
41	Paulo Fonseca (Roma)	0.612	41	Hubert Fournier (Lyon)	0.370
42	Christophe Galtier (Lille)	0.608	42	Frank de Boer (Crystal Palace)	0.365
43	Peter Bosz (Leverkusen)	0.604	43	Adi Hutter (Eintracht Frankfurt)	0.353
44	Bo Svensson (Mainz)	0.600	44	Marco Rose (Borussia M.Gladbach)	0.351
45	Marcelino Garcia Toral (Athletic Bilbao)	0.597	45	Roger Schmidt (Leverkusen)	0.311
46	Ronald Koeman (Barcelona)	0.596	46	Franck Haise (Lens)	0.309
47	Stefano Pioli (AC Milan)	0.596	47	Alexander Zorniger (Stuttgart)	0.286
48	Marco Rose (Borussia M.Gladbach)	0.596	48	Jose Mourinho (Tottenham)	0.273
49	Guus Hiddink (Chelsea)	0.595	49	Roberto Mancini (Inter)	0.272
50	Oliver Glasner (Wolfsburg)	0.588	50	Imanol Alguacil (Real Sociedad)	0.267
66	Marcelo Bielsa (Leeds)	0.543	62	Graham Potter (Brighton)	0.147
173	Graham Potter (Brighton)	0.421	92	Marcelo Bielsa (Leeds)	0.016
189	Eddie Howe (Bournemouth)	0.410	216	Eddie Howe (Bournemouth)	-0.347
246	Gary Neville (Valencia)	0.344	258	Gary Neville (Valencia)	-0.539

Notes: Club with which a manager appeared most recently in the sample is in parentheses.

4.5.2 League coefficients and league adjusted player ranking

Following Section 4.4.2, we first estimate model (4.2) to obtain league coefficients. Table 4.6 presents the resulting ranking of the league strengths, i.e. the negative value of the estimated league-specific effects.

Table 4.6: Estimated league coefficients

	League	$-\hat{\rho}_l$
1	Premier League (England)	0.378
2	LaLiga (Spain)	0.348
3	Serie A (Italy)	0.292
4	Bundesliga (Germany)	0.224
5	Ligue 1 (France)	0.212
6	Brasileirão (Brazil)	0.183
7	Championship (England)	0.159
8	Super Lig (Turkey)	0.123
9	Premier League (Russia)	0.109
10	Liga NOS (Portugal)	0.089
11	League One (England)	-0.007
12	Major League Soccer (USA)	-0.010
13	Bundesliga II (Germany)	-0.024
14	Jupiler Pro League (Belgium)	-0.024
15	Eredivisie (Netherlands)	-0.060
16	Premiership (Scotland)	-0.091
17	League Two (England)	-0.147

Notes: Reference league is Superliga (Argentina).

The Table suggests that prominent differences exist across leagues. Other things being equal, being in the English Premier Leagues lowers an individual player's ratings by 0.378 points relative to the reference category of the Argentinian Superliga. It is also worth noting that the Big Five are indeed ranked in the top five.

Table 4.7 shows the highest values of the weighted average of ratings (\overline{WS}_t) and League Adjusted WhoScored ratings (\overline{LAWS}_t) achieved by a unique individual. Applied to both cases are time discounting based on the weighting function (4.4) and shrinkage towards the league average as the initial value described in Section 4.4.2. The only difference between the left-hand side and the right-hand side of the table, therefore, is whether the ratings are adjusted for the league strengths. Lionel Messi and Neymar attained the highest ratings in both cases. As expected, however, without the league adjustment (\overline{WS}_t) some players in the lower tier leagues achieved higher ratings, for instance, James Tavernier in Premiership

(Scotland), who is ranked higher than Kevin De Bruyne in Premier League (England). The left-hand side of the table produces the ranking, which is probably more in line with what one would expect; the well-known players from the Big Five, for instance, Kevin De Bruyne, Harry Kane, and Sergio Agüero are included in the top ten.

To measure the strength of a squad at time t , therefore, we take the average value of $\overline{\text{LAWS}}_t$ for the players available on a match day t .

Table 4.7: Weighted average of WS and LAWS ratings

	Player	Date	League	\overline{WS}_t		Player	Date	League	\overline{LAWS}_t
1	Lionel Messi	2021/04/25	LaLiga (Spain)	8.025	1	Lionel Messi	2021/04/25	LaLiga (Spain)	8.370
2	Neymar	2021/02/07	Ligue 1 (France)	7.767	2	Neymar	2020/11/20	Ligue 1 (France)	8.053
3	Hakim Ziyech	2020/02/16	Eredivisie (Netherlands)	7.548	3	Cristiano Ronaldo	2021/01/17	Serie A (Italy)	7.822
4	Cristiano Ronaldo	2021/04/21	Serie A (Italy)	7.504	4	Kevin De Bruyne	2021/01/20	Premier League (England)	7.726
5	Robert Lewandowski	2021/05/15	Bundesliga (Germany)	7.474	5	Robert Lewandowski	2021/05/15	Bundesliga (Germany)	7.713
6	Zlatan Ibrahimovic	2021/02/13	Serie A (Italy)	7.439	6	Eden Hazard	2020/06/21	LaLiga (Spain)	7.712
7	James Tavernier	2020/12/30	Premiership (Scotland)	7.429	7	Luis Suárez	2020/06/13	LaLiga (Spain)	7.698
8	Hulk	2016/05/15	Premier League (Russia)	7.379	8	Harry Kane	2021/05/19	Premier League (England)	7.672
9	Steven Berghuis	2021/04/25	Eredivisie (Netherlands)	7.366	9	Zlatan Ibrahimovic	2021/02/13	Serie A (Italy)	7.627
10	Kevin De Bruyne	2021/01/20	Premier League (England)	7.355	10	Sergio Agüero	2019/09/28	Premier League (England)	7.626

Notes: Table shows the top 10 weighted average of WhoScored ratings (\overline{WS}_t) and League Adjusted WhoScored ratings (\overline{LAWS}_t) achieved by a unique player during the period from 2013/07/14 to 2021/05/23. Date and League are the date when the latest rating used to obtain the weighted average is realised, and the league the player belonged to at that time.

4.5.3 Estimated production function

Using expected goals and adjusted ratings presented above, together with transfer budgets to control for a club's financial strength, we estimate the model (4.9). The estimation results are reported in Table 4.8. The dependent variable is the log-difference of expected goals from a home club's perspective (d_log_xG). Other inputs (*Player ratings* and *Transfer budgets*) are also in the log-difference form according to the model (4.9). In columns (1) and (2), *Player ratings* is based on the mean value of average adjusted ratings of a squad, whilst (3) and (4) include separately that of the starting lineup and substitutions, *Player ratings (starting)* and *Player ratings (substitutions)*, respectively. Manager fixed effects are included in columns (2) and (4) but not in columns (1) and (3). The coefficient estimates on the manager dummies, i.e. the estimates of manager fixed effects, are not reported in Table 4.8 since they will be presented and discussed further in the next subsection.

All the coefficients on the relative input variables have expected signs; the relative quality of players and financial strengths positively affect the relative performance measured by expected goals. In addition, home advantage (Constant) is positive and significant in all specifications, as expected. The comparisons between columns (1) and (2) as well as (3) and (4) imply that managerial input also influences field performance. The inclusions of manager fixed effects reduce the size of coefficients for labour and capital inputs, whilst that of home advantage is not affected. This implies that players' effects in columns (1) and (3) partly capture the manager's input. Considering that the production function models the match-day contribution of a manager given the resources available on the day, it can be argued that managerial input adds to the players' impacts on the field performance through the decisions on tactics and substitution, as well as psychological supports.

The estimated parameters show statistically significant effects of each input, except for the strength of substitutions in column (4), where such effects are not significant at the 5% significance level. The possible explanation for this is that the seemingly positive effects of substitution quality in column (3) are due to a manager's decision on substitutions during the match, rather than the strengths of substitutions themselves. In other words, whilst the robust set of substitutions can marginally add to the performance, it is possibly the quality of decisions on substitutions made by a manager that matters more.

The joint significance of manager fixed effects is tested by means of the F-test in Table 4.9. The model numbers in Table 4.9 correspond to the column numbers in Table 4.8. The comparisons between models (1) and (2) as well as (3) and (4) confirm that manager fixed effects are jointly significant. This provides evidence that managers overall affect output level, after explicitly controlling for players' strength on a particular match day and financial resources.

Table 4.8: Estimated parameters for production functions

	<i>Dependent variable:</i>			
	d.log_xG			
	(1)	(2)	(3)	(4)
Player ratings	27.441*** (0.949)	18.236*** (1.377)		
Player ratings (starting)			23.662*** (0.906)	16.244*** (1.197)
Player ratings (substitutions)			2.807*** (0.806)	1.543* (0.855)
Transfer budgets	0.092*** (0.009)	0.065*** (0.013)	0.085*** (0.009)	0.059*** (0.013)
Constant	0.266*** (0.007)	0.266*** (0.007)	0.266*** (0.007)	0.266*** (0.007)
Manager fixed effects	No	Yes (N > 10)	No	Yes (N > 10)
Observations	12,668	12,668	12,668	12,668
R ²	0.227	0.317	0.229	0.318
Adjusted R ²	0.227	0.299	0.229	0.300

Notes: *p<0.1; **p<0.05; ***p<0.01. All variables are in the log-difference form from a home club's perspective. In columns (1) and (2), *Player ratings* is based on the mean value of average adjusted ratings of a squad, whilst (3) and (4) include separately that of the starting lineup and substitutions. Manager fixed effects are included in columns (2) and (4), but not in columns (1) and (3). The individual coefficients on manager fixed effects are not reported.

Table 4.9: F-test for significance of manager effects

A. Model (1) v. (2)						
Model	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
(1)	12,665	8,657.173				
(2)	12,343	7,651.564	322	1,005.609	5.038	0
B. Model (3) v. (4)						
Model	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
(3)	12,664	8,628.030				
(4)	12,342	7,638.022	322	990.008	4.968	0

4.5.4 Estimated manager coefficients

In Table 4.10, we report the estimated coefficients for individual managers with the 50 largest coefficients and selected managers. The estimation is based on the specification in column (4) in Table 4.8. As explained in Section 4.3, we include dummy variables for the 322 managers who appeared in more than 10 matches in our sample. Therefore, our reference group is effectively the pool of managers who do not satisfy this criterion. That is, the estimated coefficients presented in Table 4.10 are relative to the average performance of the 83 managers with less than or equal to 10 appearances. As described in Section 4.3, the latter group is mainly represented by caretaker managers.

Table 4.10 provides evidence of heterogeneity among managers in European football, controlling for home advantage, players' strengths, and financial resources. Our estimation suggests that Josep Guardiola, who managed FC Bayern Munich (German *Bundesliga*) and Manchester City F.C. (English *Premier League*) during our sample period, made the largest contribution to clubs' production. Specifically, the relative size of his influence to home advantage is 3.443 ($\approx 0.916/0.266$). With similar comparisons, the effects of the top 50 managers presented in Table 4.10 are all larger than the size of home advantage. The impact of an average manager (one with an average coefficient of 0.1163) is less than half the size of home advantage. For instance, Peter Stöger, the former manager of the German *Bundesliga* clubs, FC Cologne and Borussia Dortmund, has an estimated coefficient of 0.1189, which is equal to 0.447 relative to the size of home advantage.

Table 4.10: Estimated manager coefficients

ranking	manager	Estimate	Clubs (Seasons)
1	Josep Guardiola	0.916	Bayern ('14-'16), Man City ('16-'21)
2	Frank Lampard	0.839	Chelsea ('19-'21)
3	Luis Enrique	0.782	Barcelona ('14-'17)
4	Laurent Blanc	0.716	PSG ('14-'16)
5	Javier Aguirre	0.701	Leganes ('19-'20)
6	Jose Rojo Martin	0.692	SD Huesca ('20-'21)
7	Jurgen Klopp	0.640	Borussia Dortmund ('14-'15), Liverpool ('15-'21)
8	Josef Heynckes	0.619	Bayern ('17-'18)
9	Gian Piero Gasperini	0.605	Genoa ('14-'16), Atalanta ('16-'21)
10	Maurizio Sarri	0.577	Empoli ('14-'15), Napoli ('15-'18), Chelsea ('18-'19), Juventus ('19-'20)
11	Franck Haise	0.570	Lorient ('16-'17), Lens ('20-'21)
12	Graham Potter	0.561	Brighton ('19-'21)
13	Thomas Tuchel	0.561	Borussia Dortmund ('15-'17), PSG ('18-'20), Chelsea ('20-'21)
14	Frank Kramer	0.555	Arminia Bielefeld ('20-'21)
15	Massimiliano Allegri	0.547	Juventus ('14-'19)
16	Ralf Rangnick	0.536	RBL ('18-'19)
17	Stefano Pioli	0.520	Lazio ('14-'16), Inter ('16-'17), Fiorentina ('17-'19), AC Milan ('19-'21)
18	Sabri Lamouchi	0.504	Rennes ('17-'19)
19	Julen Lopetegui	0.496	Real Madrid ('18-'19), Sevilla ('19-'21)
20	Pascal Plancque	0.484	Nimes ('20-'21)
21	Jose Bordalas	0.478	Getafe ('17-'21)
22	Nuno Espirito Santo	0.466	Valencia ('14-'16), Wolves ('18-'21)
23	Vahid Halilhodzic	0.459	Nantes ('18-'19)
24	Antonio Conte	0.458	Chelsea ('16-'18), Inter ('19-'21)
25	Julian Nagelsmann	0.455	Hoffenheim ('15-'19), RBL ('19-'21)
26	Clarence Seedorf	0.451	Deportivo ('17-'18)
27	Brendan Rodgers	0.447	Liverpool ('14-'16), Leicester ('18-'21)
28	Bo Svensson	0.447	Mainz ('20-'21)
29	Imanol Alguacil	0.445	Real Sociedad ('17-'21)
30	Sergio Conceicao	0.443	Nantes ('16-'17)
31	Ole Gunnar Solskjaer	0.428	Man Utd ('18-'21)
32	Bruno Genesio	0.415	Lyon ('15-'19), Rennes ('20-'21)
33	Niko Kovac	0.414	Eintracht Frankfurt ('15-'18), Bayern ('18-'20), Monaco ('20-'21)
34	Jose Luis Mendilibar	0.414	Levante ('14-'15), Eibar ('15-'21)
35	Pablo Machin	0.408	Girona ('17-'18), Sevilla ('18-'19), Espanyol ('19-'20), Deportivo Alaves ('20-'21)
36	Giampiero Ventura	0.402	Torino ('14-'16), Chievo ('18-'19)
37	Rudi Garcia	0.400	Roma ('14-'16), Marseille ('16-'19), Lyon ('19-'21)
38	Carlo Ancelotti	0.398	Real Madrid ('14-'15), Bayern ('16-'18), Napoli ('18-'20), Everton ('19-'21)
39	Stephane Moulin	0.394	Angers ('15-'21)
40	Urs Fischer	0.386	Union Berlin ('19-'21)
41	Paulo Fonseca	0.385	Roma ('19-'21)
42	Luciano Spalletti	0.384	Roma ('15-'17), Inter ('17-'19)
43	Marcelo Bielsa	0.383	Marseille ('14-'16), Lille ('17-'18), Leeds ('20-'21)
44	Mikel Arteta	0.382	Arsenal ('19-'21)
45	Scott Parker	0.380	Fulham ('18-'21)
46	Hubert Fournier	0.376	Lyon ('14-'16)
47	Julien Stephan	0.375	Rennes ('18-'21)
48	Cristobal Parralo	0.373	Deportivo ('17-'18)
49	Bernard Blaquart	0.371	Nimes ('18-'20)
50	Edin Terzic	0.367	Borussia Dortmund ('20-'21)
150	Jose Mourinho	0.144	Chelsea ('14-'16), Man Utd ('16-'19), Tottenham ('19-'21)
154	Eddie Howe	0.136	Bournemouth ('15-'20)
308	Gary Neville	-0.333	Valencia ('15-'16)

Notes: Table reports the estimated manager coefficients for those with the 50 largest values and selected managers. The reference group is a set of managers who appeared in less than 10 matches during the sample period. Also presented are clubs and seasons where a manager was observed during the sample period.

Depicted in Figure 4.3 are the coefficients of the top 10 and selected managers with 95% confidence intervals. These effects of the top 10 managers are individually different from zero, and even among this very top group, the heterogeneity is evident.

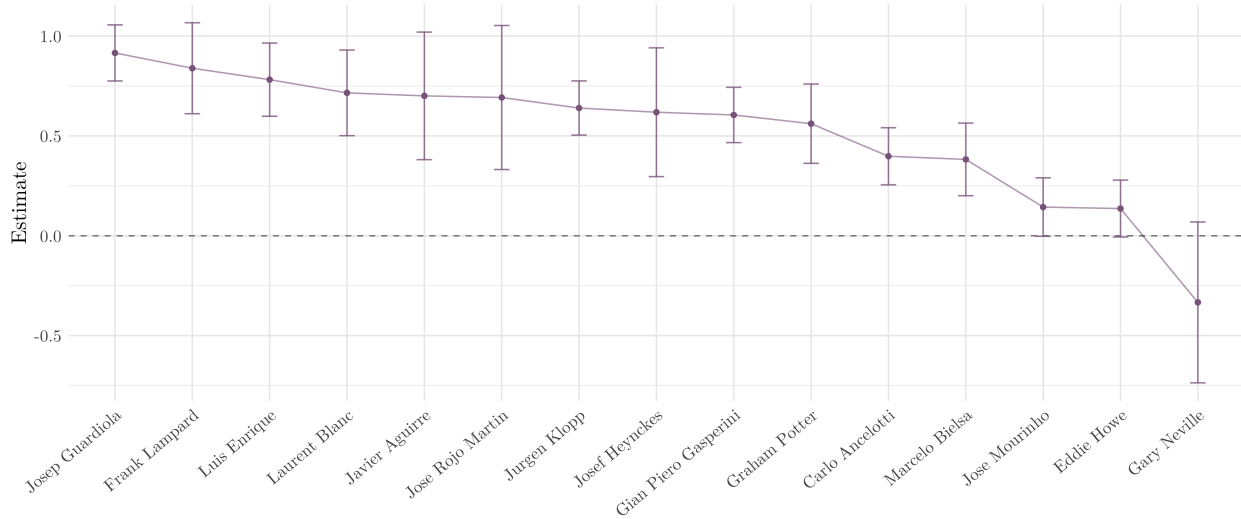


Figure 4.3: Manager coefficients with 95% confidence intervals

Recalling the naive rankings in Table 4.5, we compare our ranking based on the manager coefficients (MFE) with the two naive rankings in Figure 4.4. In Panel (I), the 322 managers are ordered with respect to the naive ranking based on the win percentage (WP) on the x-axis in descending order, which is plotted against the ranking based on the MFE, together with a 45-degree line. Panel (II) is similar, except that the naive ranking is based on the average net expected goals (d_xG).

It is evident from Figure 4.4 (I) that the two rankings produce quite different pictures. The correlation between the two rankings is 0.589. The sources of discrepancy between the two rankings are that the naive ranking is based on a more crude measure of output (WP) and that it also does not take into account the contributions of other inputs. In Figure 4.4 (II), the naive ranking is based on the average net expected goals, the same output measure used to obtain MFEs. Nevertheless, there is still some discrepancy between the two rankings (the correlation between the two is 0.8320), suggesting that taking into account the input levels is also important, as well as measuring the output that is more robust to randomness.¹⁰

In both cases, managers above (below) the 45-degree lines are potentially under (over) rated relative to our ranking, which takes into account the other inputs and is more robust to randomness. The further the distance from the 45-degree line, the greater the disparity

¹⁰The correlation between WP and MFE and that between d_xG and MFE in actual values (as opposed to rankings) are 0.6130 and 0.8164, respectively.

between the respective naive ranking and the ranking based on MFEs. One of the potentially most under-rated managers is Clarence Seedorf, a former Deportivo (Spanish *LaLiga*) manager. He is ranked 270th and 79th out of the 322 managers in the naive rankings in terms of WP and xG differentials but is ranked 26th in the MFE ranking. The disparity between his WP and MFE rankings is more prominent than that between xG differential and MFE rankings, therefore, he can be considered as one of the “unlucky” managers, whose performance was not necessarily reflected in match outcomes. Similarly, Graham Potter (Brighton & Hove Albion F.C., England) is ranked a lot higher in the MFE ranking (12th), than the WP and xG differential rankings (173rd and 62nd, respectively). Given the resources available, his performance is notable, being just below the top ten managers.

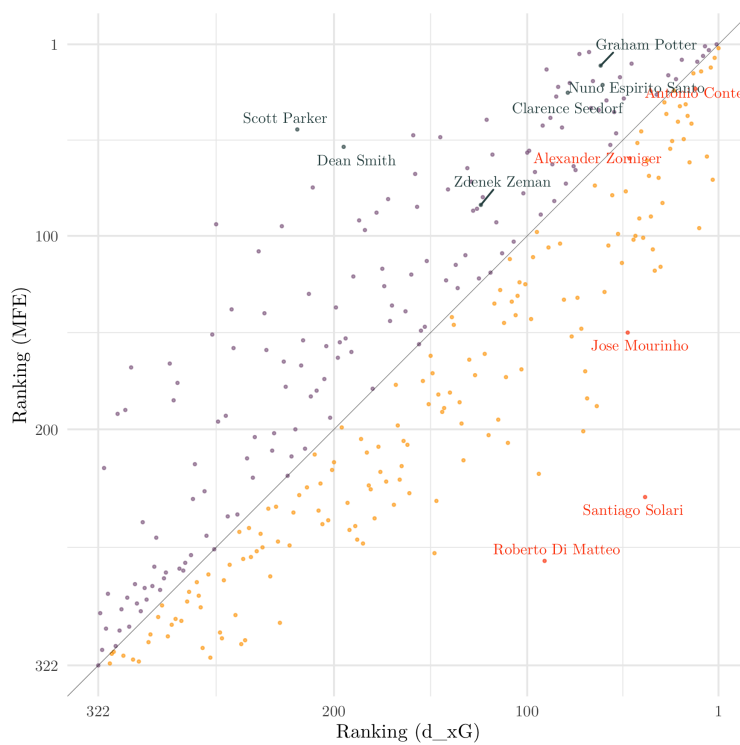
On the other hand, our estimation suggests that some managers are over-rated when relying on naive measures of a manager’s performance. For instance, Santiago Solari (Real Madrid, Spain) and Roberto Di Matteo (Schalke, Germany) are ranked in the top quartile in terms of the winning percentage but in the 3rd and the bottom quartile, respectively, in the MFE ranking.

As seen in the examples above, when there is an apparent disparity in rankings between one naive measure and MFE, it is usually the case also between the other naive measure and MFE. For instance, if a manager is ranked much higher in the WP table than they are in the MFE table, they also tend to be ranked higher in the net xG table than they are in the MFE table. One exception is Alexander Zorniger (VfB Stuttgart, ’15-’16), who is ranked 291st and 27th in the WP and net xG tables, whilst his ranking is 60th in terms of MFE. In this case, therefore, he may be under-rated in the crude measure of match outcome, but the chances created with the given squad and financial source are rather over-rated.

These examples show that consideration of performance that is more robust to randomness and disentangling a manager’s contribution from that of other inputs are both relevant in comparing managers.



(I) WP v. manager FE



(II) d_xG v. manager FE

Figure 4.4: Naive rankings and manager fixed effects

4.5.5 Case study

To put the results into perspective, we compare expected points (xP) obtained by a club under different managers. In doing so, we first obtain fitted values for d_log_xG with an actual manager and alternative managers based on our estimation results presented in Table 4.8 (column (4)) and Table 4.10. Then, we convert these values into expected points and observe how they are accumulated throughout a certain season.

To obtain xP, we first establish the relationship between the probabilities of each outcome (home win, draw, and away win) and the log-difference of xGs (d_log_xG). Therefore, we estimate the ordered logit model with outcome variable y where $y = 0$ for away win, $y = 1$ for draw, and $y = 2$ for home win and a predictor d_log_xG . Specifically, we estimate the parameters (μ_1, μ_2, β) in the following model:

$$y = \begin{cases} 0 \text{ (away win)} & \text{if } y^* \leq \mu_1, \\ 1 \text{ (draw)} & \text{if } \mu_1 < y^* \leq \mu_2, \\ 2 \text{ (home win)} & \text{if } \mu_2 < y^*, \end{cases} \quad \text{where } y^* = \beta d_log_xG + \varepsilon. \quad (4.10)$$

Our estimation results are presented in Table 4.11.

Table 4.11: Estimated parameters for ordered logit model

	<i>Dependent variable:</i>
	y
d_log_xG	0.981*** (0.022)
0 1	-0.700*** (0.021)
1 2	0.528*** (0.020)
Observations	12,668
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Given this, we discuss a case study of Everton F.C. (English *Premier League*) in Season 2020/2021. In the preceding season, former manager Marco Silva was replaced by Carlo Ancelotti in the first half of the season, and the club finished the season with accumulated points of 49 and is ranked 12th in the league. Carlo Ancelotti was in charge for the following

season, where the club obtained 59 points overall and went up the table by two places. At the end of the season, however, he left the club, resulting in another managerial change for the club. A new manager Rafael Benítez was introduced to the club prior to Season 2021/2022.

To see the impacts of different managers on expected points as well as season outcome, we compare the predicted performance of Everton F.C. for Season 2020/2021 fixtures for the following managers: Carlo Ancelotti, Gary Neville, Graham Potter, Josep Guardiola, Marco Silva, and Rafael Benitez. For each of these managers we obtain the fitted values for the log-difference of xGs (d_log_xG), given the quality of players and financial strengths, those of opponents, and home advantage, if any. Then, convert this into expected points using the estimated ordered logit model parameters in Table 4.11. The accumulated expected points predicted to be obtained throughout the season under different managers are depicted in Figure 4.5.

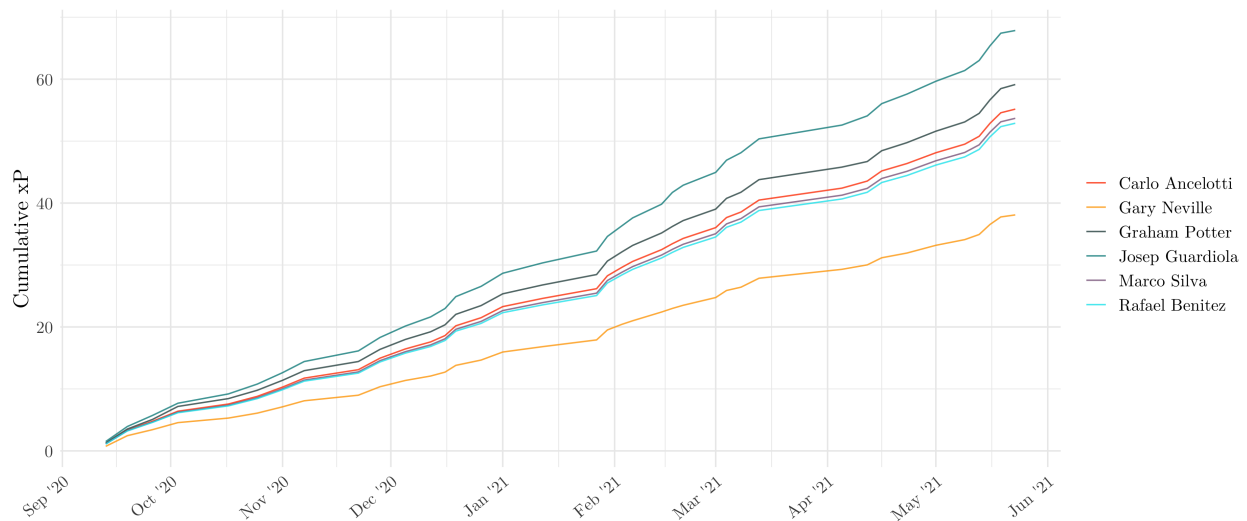


Figure 4.5: Cumulative expected points (xP) for Everton F.C. (2020/2021)

The Figure provides further insight into the implications of different managers, who are ranked quite differently as in Table 4.10, on the seasonal outcome. In terms of expected points, as opposed to actual points, Carlo Ancelotti, who was ranked 38th in our manager ranking, gained 55.16 points, which places the club the 6th in the league. Recall that the club actually finished with 12th place in Season 2020/2021, implying that the club might have had a relatively more unlucky season than the other clubs might have.

How well would the club have performed if Josep Guardiola, who is identified as the best manager in terms of contribution, had been in charge instead of Carlo Ancelotti, all other things being equal? The club would have accumulated the expected points of 67.85, and it would have ended up in the top three in the league, resulting in qualification for

the UEFA Champions League. With Graham Potter, the club would have qualified for the UEFA Europa League and performed slightly better than what would have been expected under Carlo Ancelotti. Marco Silva and Rafael Benitez, who were ranked 58th and 74th in the manager ranking, respectively, would have led to the club just the top half of the league table (8th and 9th, respectively). The club could have been relegated with managers found towards the bottom of our ranking. For instance, under Gary Neville, its expected points would have been 38.09, which would have resulted in finishing the season second to the bottom of the league table. Overall, our analysis shows that the seasonal outcome could have been drastically different under a different manager, all other things being equal.

4.6 Conclusion

Consistent with previous studies, our analysis has shown that managers do matter in explaining firm production. The evidence we present here is arguably even more compelling since the overall significance of managers is still present even after explicitly controlling for other inputs by utilising data from professional football leagues. In addition, our analysis highlights the importance of taking into account the randomness of outcome and resources available at the hand of a manager to evaluate managerial contributions to a club's performance. In particular, we demonstrate how the ranking of football managers can differ with and without these considerations. Therefore, relying on a "naive" measure of managerial performance, such as a winning percentage, can be costly to a club since it may overestimate or underestimate a manager's contributions.

An individual manager's coefficient estimated in our model can be interpreted as the manager's match day contribution to a club's performance, given the quality of a squad and financial strengths on the day. This is useful information since the level of players and financial strength can vary over time and reflect on a manager's performance, and we, therefore, believe that this is a well-founded way of comparing different managers. In addition, this paper expands the previous studies by comparing managers observed in different leagues, which can be particularly useful for clubs seeking a manager.

Nevertheless, there are some limitations to note. Firstly, individual manager effects are effectively held constant at the average level over the past seven seasons (2014/2015-2020/2021). Therefore, this does not provide a full picture of how a manager's ability may have changed over time. It is also likely that the more recent observations carry higher predictive power, hence weighting these observations equally may not be optimal.

Secondly, our measure of managerial contribution reflects their ability to produce outcomes, given the quality of players on the day. Therefore, this captures their ability related

to tactics, substitution decisions, and motivational role on the day. Although we argue that this is a valid way to evaluate a manager, it does not capture other important roles as a manager. For instance, good managers would contribute to players' growth over their career, not necessarily just bringing the best out of the player on a particular match day. Therefore, one could evaluate managers in terms of how they influenced individual players' performance over time. This could be done, for instance, using the historical performance measure employed in this study, which is allowed to vary over time.

Chapter 5

Conclusion

This thesis presented a collection of papers that addressed various issues in leadership research and the economics of managers. Our original empirical studies incorporated in each paper employed data from professional football leagues. As argued throughout the thesis, professional sports produce valuable data that can be seen as an instrument to test economic and management theories. Leadership research can particularly benefit from such data since the role played by a manager in professional sports clubs are similar to a leader's role in a more general context, such as that of a CEO in corporations.

Chapter 2 focused on methodological issues that could arise when estimating causal effects in observational studies. Social science often relies on observational data, as opposed to data generated through randomised control trials, since the latter could be unfeasible or even non-natural. Estimation of leadership succession effects is a precise example. Such events are unlikely to occur randomly but tend to happen, for instance, when a firm's performance is not up to standard. One of the remedies for this problem is a propensity score analysis, and the Chapter offered a description of the method and a review of the use of this method in leadership research and related fields.

In addition, the Chapter demonstrated how to conduct a propensity score analysis (inverse propensity score weighting in particular) using the real-world example of head coach changes in Italian professional football. The study highlighted how useful yet straightforward it is to employ such a method whilst offering some insights into leadership succession. Through this case study, we showed how to extend the analysis of managerial succession effects by viewing a leadership change as simultaneous changes in leader characteristics, such as their age, experience, and association with the organisation. We found that differences between new and dismissed managers in certain characteristics can enhance the chances of successful turnover.

The causes and consequences of managerial dismissal decisions were further examined in

Chapter 3. In the Chapter, we extended the analysis by separately identifying the factors affecting the first and second managerial dismissals that occur in a given season and estimating the impact of such events on subsequent performance. We also incorporated a machine learning technique as an alternative to a more conventional method of logistic regression in order to identify the determinants of each type of dismissal and estimate propensity scores. As was the case in our analysis, the use of a machine learning algorithm can improve the performance in terms of balancing covariates between the treated and control observations.

Our findings suggest that the two decisions are motivated by different factors. The most noticeable difference between the two is that the likelihood of first within-season dismissals is negatively related to favourable field performance measured both in absolute and relative terms, whereas that of second dismissals is predominantly affected by relative performance against expectations. As such, we argued that given that a club has dismissed a manager since the beginning of the respective season, a decision to replace yet another manager is made with greater caution. Furthermore, this implies that different sets of covariates need to be taken into account when estimating the causal effects of the two types of dismissal.

Given this, we separately estimated the average treatment effects of the first and second within-season replacements using inverse propensity score weighting. The analysis found that when a club replaces a manager for the first time in a given season, its performance can improve to a certain extent. On the other hand, if a club has already replaced a manager in a particular season, another managerial change makes no difference to subsequent performance. These findings, together with the findings from the case study in Chapter 2, contribute to understanding certain circumstances under which the decision to dismiss a manager could be beneficial.

Chapter 4 diverted from the specific issues around the causes and consequences of leadership succession and examined the role of leaders themselves in firm production. The fact that a football club's output and inputs can be clearly and regularly measured allowed us to estimate a firm's production function and disentangle the contribution of managerial inputs from that of other inputs. For instance, exploiting the worker (player)-level data of historical performance, we measured individual players' strengths on a particular match day and used this as a proxy for labour input. Therefore, we could evaluate how much an individual manager adds to a club's performance, given the resources at his disposal.

The Chapter provided clear evidence for the overall significance of managerial inputs in explaining heterogeneity in outputs across firms after controlling for labour and capital inputs. In fact, as our case study showed, having different managers can have both economically and statistically significant impacts on a club's outcomes. We then presented the ranking of the managers who appeared in the five European top-flight football leagues in our

sample period based on their individual coefficients. Our ranking was not quite consistent with the “naive” ranking based on winning percentage, which does not take into account playing talent or financial strengths of working clubs. Therefore, comparisons across managers based on the latter method could be misleading. It was also highlighted that taking into account the randomness of the outcome is important to evaluate an individual manager’s contributions. We overcame this issue by employing expected goals as a performance measure, which reflects the quality of chances created within a match.

The papers presented in this thesis, therefore, collectively contribute to leadership research and the economics of managers in the following ways. First, we provide practical guidance about propensity score weighting, which researchers in these fields and beyond can benefit from. Second, our studies offer further insights into leadership succession effects by identifying some conditions under which the decision to replace a manager can be beneficial. Finally, we add to the empirical evidence for the significance of leaders whilst proposing an alternative way to evaluate individual managers. We believe that future studies can also benefit from employing additional techniques and metrics that are used throughout the thesis. For instance, as mentioned above, the use of machine learning could perform better in some aspects of empirical analysis relative to more traditional methods. Additionally, whilst it is well-established that data from the sports industry offer valuable resources for economics and management research, future studies can take further advantage of such data. For example, they can make use of the individual players’ historical performance data or event data that allows us to compute the advanced measure of club performance used in our analyses.

To conclude, we outline potential avenues for future research. First, one can assess the risk associated with managerial decisions. In Chapter 3, for example, it remains unclear why clubs still sack managers for the second time, even though it does not improve performance. One possibility is that they are clubs facing a desperate situation and may be willing to gamble on managerial change. Depending on a club’s risk-preferences, volatility in an individual manager’s performance may also be of an important concern. Whilst Chapter 4 identified managers who perform well on average, some of these managers may have a performance with a very high variance. If a club sees them as unreliable, they may prefer another manager who delivers slightly worse results but with lower variability.

Second, further analysis can be conducted to identify what determines the heterogeneity among managers. One can follow Bertrand and Schoar (2003)’s approach, where they linked the estimated leader coefficients with their characteristics, such as age and MBA qualification. In particular, one can examine whether the differences among individual managers in terms of our managerial contribution estimates can be explained by their background, as a player and in terms of coaching experience and any previous association with the particular

club. In light of the recent finding by Tur et al. (2021), one can also analyse the role of charisma in this context, as well as leadership succession. Given that football managers are public figures and they tend to appear on social media, it could be the case that such a leadership trait could potentially be of importance.

Third, complementarity between a leader and his/her subordinates and their match quality, i.e. the degree to which workers cooperate effectively across layers to impact firm success, may merit further research. Peeters et al. (2020) show that, in addition to the separate contributions of upper and middle managers, their match quality can significantly affect firm performance. In the context of professional football, Bridgewater et al. (2011) find that some managers are good at dealing with superstars whilst others are better at teaching less skilled players. A future study could adopt a similar approach to study complementarity between managers and players and identify the factors that determine their match quality.

Lastly, we proposed one way to rate managers based on their ability to produce favourable outcomes given the quality of players and financial strengths measured on a particular match day. Therefore, this essentially reflects their ability related to tactics, substitution decisions, and motivational role on the pitch. However, it could be argued that another crucial role of a leader is to improve their subordinates' ability. Therefore, one could alternatively evaluate a football manager based on their contributions to the growth of playing talent. Again, this type of analysis is feasible in the context of professional sports since the historical performance of individual players is publicly available, and it has the potential to advance leadership research.

Bibliography

- 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. (1999). Organizational Performance and Managerial Turnover. *Managerial and Decision Economics*, 20(6):305–318.
- 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.
- Bertrand, M. and Schoar, A. (2003). Managing with style: The effect of managers on firm policies. *Quarterly Journal of Economics*, 118(4):1169–1208.
- Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., and Roberts, J. (2013). Does management matter? Evidence from India. *Quarterly Journal of Economics*, 128(1):1–51.
- Bloom, N., Lemos, R., Sadun, R., Scur, D., and Reenen, J. V. (2014). The new empirical economics of management. *Journal of the European Economic Association*, 12(4):835–876.
- 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.
- Boshnakov, G., Kharrat, T., and McHale, I. G. (2017). A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, 33(2):458–466.
- Boyne, G. and Dahya, J. (2002). Executive succession and the performance of public organizations. *Public Administration*, 80(1):179–200.
- Branco, P., Torgo, L., and Ribeiro, R. P. (2017). A survey of predictive modelling under imbalanced distributions. *ACM Computing Surveys*, 49(2):1–50.
- Brechot, M. and Flepp, R. (2020). Dealing With Randomness in Match Outcomes: How to Rethink Performance Evaluation in European Club Football Using Expected Goals. *Journal of Sports Economics*, 21(4):335–362.
- Breiman, L. (2001). Random Forests. *Machine Learning 2001 45:1*, 45(1):5–32.
- Brian, G. (2013). Contributions of Managerial Levels: Comparing MLB and NFL. *Managerial and Decision Economics*, 34(6):428–436.
- 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.
- Bruinshoofd, A. and Ter Weel, B. (2003). Manager to go? Performance dips reconsidered with evidence from Dutch football. *European Journal of Operational Research*, 148(2):233–246.
- Bryson, A., Buraimo, B., Farnell, A., and Simmons, R. (2021a). Special Ones? The Effect of Head Coaches on Football Team Performance. *IZA Discussion Paper*, No. 14104.
- Bryson, A., Buraimo, B., Farnell, A., and Simmons, R. (2021b). 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.

- Buzzacchi, L., Caviggioli, F., Milone, F. L., and Scotti, D. (2021). Impact and Efficiency Ranking of Football Managers in the Italian Serie A: Sport and Financial Performance. *Journal of Sports Economics*, 22(7):744–776.
- 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.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357.
- 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.
- D’Addona, S. and Kind, A. (2014). Forced Manager Turnovers in English Soccer Leagues: A Long-Term Perspective. *Journal of Sports Economics*, 15(2):150–179.
- 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.

- Dawson, P., Dobson, S., and Gerrard, B. (2000). Stochastic Frontiers and the Temporal Structure of Managerial Efficiency in English Soccer. *Journal of Sports Economics*, 1(4):341–362.
- Day, D. V., Gordon, S., and Fink, C. (2012). The sporting life: Exploring organizations through the lens of sport. *Academy of Management Annals*, 6(1):397–433.
- de Jong, A. and Naumovska, I. (2016). A note on event studies in finance and management research. *Review of Finance*, 20(4):1659–1672.
- Decroos, T., Bransen, L., Haaren, J. V., and Davis, J. (2020). VAEP: An Objective Approach to Valuing On-the-Ball Actions in Soccer. *In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 4696–4700.
- 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.
- Dehejia, R. H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1):151–161.
- 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.
- Desai, M. N., Lockett, A., and Paton, D. (2018). Information Asymmetries in the Hiring Process and the Risk of New Leader Dismissal: Insights from English Premier League Soccer Organizations. *British Journal of Management*, 29(1):26–42.
- 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.
- Flepp, R. and Franck, E. (2021). The performance effects of wise and unwise managerial dismissals. *Economic Inquiry*, 59(1):186–198.
- Flores, R., Forrest, D., and Tena, J. D. (2012). Decision taking under pressure: Evidence on football manager dismissals in Argentina and their consequences. *European Journal of Operational Research*, 222(3):653–662.
- 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.
- Frick, B., Barros, C. P., and Prinz, J. (2010). Analysing head coach dismissals in the German “Bundesliga” with a mixed logit approach. *European Journal of Operational Research*, 200(1):151–159.
- Frick, B. and Simmons, R. (2008). The impact of managerial quality on organizational performance: Evidence from German soccer. *Managerial and Decision Economics*, 29(7):593–600.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.

- 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.
- Gamson, W. A. and Scotch, N. A. (1964). Scapegoating in baseball. *American Journal of Sociology*, 70(1):69–72.
- 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.
- Gilfix, Z., Meyerson, J., and Addona, V. (2020). Longevity differences in the tenures of American and foreign Major League Soccer managers. *Journal of Quantitative Analysis in Sports*, 16(1):17–26.
- Gordon, G. E. and Rosen, N. (1981). Critical factors in leadership succession. *Organizational Behavior and Human Performance*, 27(2):227–254.
- Graham, J. R., Li, S., and Qiu, J. (2012). Managerial attributes and executive compensation. *Review of Financial Studies*, 25(1):144–186.
- Grusky, O. (1963). Managerial Succession and Organizational Effectiveness. *American Journal of Sociology*, 69(1):21–31.
- 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.

- Hendricks, K. B., Hora, M., and Singhal, V. R. (2015). An empirical investigation on the appointments of supply chain and operations management executives. *Management Science*, 61(7):1562–1583.
- Hill, G. C. (2009). The effect of frequent managerial turnover on organizational performance: A study of professional baseball managers. *Social Science Journal*, 46(3):557–570.
- 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.
- Hofler, R. A. and Payne, J. E. (2006). Efficiency in the National Basketball Association: A Stochastic Frontier Approach with Panel Data. *Managerial and Decision Economics*, 27(4):279–285.
- 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.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to Statistical Learning with applications in R*. Springer, New York.
- Jensen, M. C. and Murphy, K. J. (1990a). CEO Incentives : It’s Not How Much You Pay, But How. *Harvard Business Review*, 3(May-June):138–153.
- Jensen, M. C. and Murphy, K. J. (1990b). Performance Pay and Top-Management Incentives. *Journal of Political Economy*, 98(2):225–264.
- Kahn, L. M. (2000). The sports business as a labor market laboratory. *Journal of Economic Perspectives*, 14(3):75–94.

- Kaplan, S. N., Klebanov, M. M., and Sorensen, M. (2012). Which CEO Characteristics and Abilities Matter? *Journal of Finance*, 67(3):973–1007.
- Khalik, A. A., Thompson, D. M., and Walston, S. L. (2006). Perceptions of hospital CEOs about the effects of CEO turnover. *Hospital topics*, 84(4):21–27.
- Kharrrat, T., McHale, I. G., and Peña, J. L. (2020). Plus-minus player ratings for soccer. *European Journal of Operational Research*, 283(2):726–736.
- Kim, Y., Jeong, S. S., Yiu, D. W., and Moon, J. (2021). Frequent CEO Turnover and Firm Performance: The Resilience Effect of Workforce Diversity. *Journal of Business Ethics*, 173(1):185–203.
- Kiss, A. N., Cortes, A. F., and Herrmann, P. (2021). CEO proactiveness, innovation, and firm performance. *The Leadership Quarterly*, 101545, In press.
- Kwon, I. (2005). Threat of Dismissal: Incentive or Sorting? *Journal of Labor Economics*, 23(4).
- Lazear, E. P., Shaw, K. L., and Stanton, C. T. (2015). The value of bosses. *Journal of Labor Economics*, 33(4).
- Lefgren, L., Platt, B., and Price, J. (2015). Sticking with what (Barely) worked: A test of outcome bias. *Management Science*, 61(5):1121–1136.
- 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.
- Muehlheusser, G., Schneemann, S., Sliwka, D., and Wallmeier, N. (2018). The Contribution of Managers to Organizational Success: Evidence from German Soccer. *Journal of Sports Economics*, 19(6):786–819.
- Olmos, A. and Govindasamy, P. (2015). A Practical Guide for Using Propensity Score Weighting in R. *Practical Assessment, Research & Evaluation*, 20(13).
- Ong, W. J. (2021). Gender-contingent effects of leadership on loneliness. *Journal of Applied Psychology*, Advance online publication.
- Papadopoulos, A. (2021). The Effects of Management on Production: A Survey of Empirical Studies. In *Handbook of Production Economics*, pages 1–47. Springer, Singapore.
- Peeters, T., Szymanski, S., Fumagalli, C., and Thomas, C. (2014). Financial fair play in European football. *Economic policy*, 29(78):343–390.
- Peeters, T. L., Salaga, S., and Juravich, M. (2020). Matching and Winning? The Impact of Upper and Middle Managers on Firm Performance in Major League Baseball. *Management Science*, 66(6):2735–2751.
- Peeters, T. L., Szymanski, S., and Tervii, M. (2017). The Inefficient Advantage of Experience in the Market for Football Managers. *SSRN Electronic Journal*.
- Pieper, J., Nüesch, S., and Franck, E. (2014). How performance expectations affect managerial replacement decisions. *Schmalenbach Business Review*, 66:5–26.

- 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.
- Scelles, N., Llorca, M., et al. (2020). Head coach change and team performance in the French men’s football Ligue 1, 2000-2016. *Economics Bulletin*, 40(2):920–937020.
- Schepker, D. J., Kim, Y., Patel, P. C., Thatcher, S. M., and Campion, M. C. (2017). CEO succession, strategic change, and post-succession performance: A meta-analysis. *The Leadership Quarterly*, 28(6):701–720.

- 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.
- Siebert, W. S. and Zubanov, N. (2010). Management economics in a large retail company. *Management Science*, 56(8):1398–1414.
- Sparks, R. (1986). A Model of Involuntary Unemployment and Wage Rigidity: Worker Incentives and the Threat of Dismissal. *Journal of Labor Economics*, 4(4):560–581.
- 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.
- ter Weel, B. (2011). Does Manager Turnover Improve Firm Performance? Evidence from Dutch Soccer, 1986–2004. *De Economist*, 159(3):279–303.
- 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.

- Wang, H., Zhao, S., and Chen, G. (2017). Firm-specific knowledge assets and employment arrangements: Evidence from CEO compensation design and CEO dismissal. *Strategic Management Journal*, 38(9):1875–1894.
- Westreich, D., Lessler, J., and Funk, M. J. (2011). Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8):826–833.
- 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.

University of Liverpool Management School PhD Thesis – PhD Structured as Papers

AUTHORSHIP DECLARATION – JOINT AUTHORED PAPERS - APPENDIX B

1. CANDIDATE

Name of the Candidate	Student number
Kaori Narita	200966587
Thesis Title	
Essays in Economics of Managers. Insights from Professional Football Leagues.	

2. FORMAT OF THE THESIS

Is the candidate intending to structure their thesis as papers?	Yes	If Yes, please complete Section 3 (sole authored paper) OR 4 (joint paper) If No, you do not need to complete this form
---	-----	---

3. PAPER INCLUDED IN THE THESIS – JOINT AUTHORED PAPER

Title of the paper	Has this paper been published, presented at a conference or under review with a journal	If Yes, please complete the boxes below. If No, go to section 4
Causal Inference with Observational Data: A Tutorial on Propensity Score Analysis	Yes	
If the paper has already been published please refer to the University guidelines on presentation of publications within a PGR Thesis - https://www.liverpool.ac.uk/media/livacuk/tqsd/code-of-practice-on-assessment/annex-7.1-PGR-CoP.pdf		
If the paper is under review with a journal, give details of the journal, including submission dates and the review stage		
The paper was submitted to <i>The Leadership Quarterly</i> on 05/11/2020 and has gone through two rounds of major revisions. The latest decision received was a minor revision, which has completed and submitted on 15/07/2022.		
If the paper is presented at a conference, give details of the conference		
The earlier version of the paper was presented at the 11 th European Sports Economics Association conference and the 27 th EASM European Sport Management Conference in 2019.		


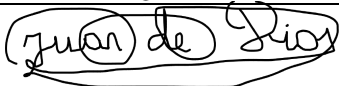

4. DESCRIPTION OF ALL AUTHOR CONTRIBUTIONS (including the PhD candidate)

Name and affiliation of author	Contribution(s) (for example, conception of the project, design of methodology, data collection, analysis, drafting the manuscript, revising it critically for important intellectual content, etc.)
Kaori Narita	Development of the idea, design of methodology, data collection, data analysis, drafting the manuscript, responding to the reviewers' comments.
Juan De Dios Tena Horrillo (University of Liverpool)	Intellectual guidance and suggestions for the entire paper production process, drafting the manuscript, responding to the reviewers' comments.
Claudio Detotto (University of Corsica)	Data collection, responding to the reviewers' comments.

5. AUTHOR DECLARATIONS (including the PhD candidate)

I agree to be named as one of the authors of this work, and confirm:

- i. that the description in Section 4 of my contribution(s) to this publication is accurate,*
- ii. that there are no other authors in this paper,*
- iii. that I give consent to the incorporation of this paper/publication in the candidate's PhD thesis submitted to the University of Liverpool*

Name of author	Signature*	Date
Kaori Narita		14/09/2022
Juan De Dios Tena Horrillo		14/09/2022
Claudio Detotto		14/09/2022

6. OTHER CONTRIBUTOR DECLARATION

I agree to be named as a non-author contributor to this work.

Name and affiliation of contributor	Contribution	Signature* and date

This consent form (Appendix B) or the sole author consent form (Appendix A) for each paper must be completed and kept with the PhD candidate once the paper is finalised. If the paper is to be included as part of the thesis, a copy of this form must be included in the thesis with each publication.

University of Liverpool Management School PhD Thesis – PhD Structured as Papers

AUTHORSHIP DECLARATION – JOINT AUTHORED PAPERS - APPENDIX B

1. CANDIDATE

Name of the Candidate	Student number
Kaori Narita	200966587
Thesis Title	
Essays in Economics of Managers. Insights from Professional Football Leagues.	

2. FORMAT OF THE THESIS

Is the candidate intending to structure their thesis as papers?	Yes	If Yes, please complete Section 3 (sole authored paper) OR 4 (joint paper) If No, you do not need to complete this form
---	-----	---

3. PAPER INCLUDED IN THE THESIS – JOINT AUTHORED PAPER

Title of the paper	Has this paper been published, presented at a conference or under review with a journal	If Yes, please complete the boxes below. If No, go to section 4
Causes and Consequences of Recurrent Managerial Changes	No	
If the paper has already been published please refer to the University guidelines on presentation of publications within a PGR Thesis - https://www.liverpool.ac.uk/media/livacuk/tqsd/code-of-practice-on-assessment/annex-7.1-PGR-CoP.pdf		
If the paper is under review with a journal, give details of the journal, including submission dates and the review stage		
If the paper is presented at a conference, give details of the conference		


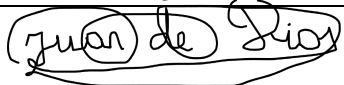
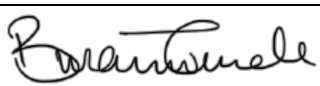
4. DESCRIPTION OF ALL AUTHOR CONTRIBUTIONS (including the PhD candidate)

Name and affiliation of author	Contribution(s) (for example, conception of the project, design of methodology, data collection, analysis, drafting the manuscript, revising it critically for important intellectual content, etc.)
Kaori Narita	Development of the idea, design of methodology, data collection, data analysis, drafting the manuscript.
Juan De Dios Tena Horrillo (University of Liverpool)	Intellectual guidance and suggestions for the entire paper production process.
Babatunde Buraimo (University of Liverpool)	Intellectual guidance and suggestions for the entire paper production process.

5. AUTHOR DECLARATIONS (including the PhD candidate)

I agree to be named as one of the authors of this work, and confirm:

- i. that the description in Section 4 of my contribution(s) to this publication is accurate,*
- ii. that there are no other authors in this paper,*
- iii. that I give consent to the incorporation of this paper/publication in the candidate's PhD thesis submitted to the University of Liverpool*

Name of author	Signature*	Date
Kaori Narita		14/09/2022
Juan De Dios Tena Horrillo		14/09/2022
Babatunde Buraimo		14/09/2022

6. OTHER CONTRIBUTOR DECLARATION

I agree to be named as a non-author contributor to this work.

Name and affiliation of contributor	Contribution	Signature* and date

This consent form (Appendix B) or the sole author consent form (Appendix A) for each paper must be completed and kept with the PhD candidate once the paper is finalised. If the paper is to be included as part of the thesis, a copy of this form must be included in the thesis with each publication.

University of Liverpool Management School PhD Thesis – PhD Structured as Papers

AUTHORSHIP DECLARATION – JOINT AUTHORED PAPERS - APPENDIX B

1. CANDIDATE

Name of the Candidate	Student number
Kaori Narita	200966587
Thesis Title	
Essays in Economics of Managers. Insights from Professional Football Leagues.	

2. FORMAT OF THE THESIS

Is the candidate intending to structure their thesis as papers?	Yes	If Yes, please complete Section 3 (sole authored paper) OR 4 (joint paper) If No, you do not need to complete this form
---	-----	---

3. PAPER INCLUDED IN THE THESIS – JOINT AUTHORED PAPER

Title of the paper	Has this paper been published, presented at a conference or under review with a journal	If Yes, please complete the boxes below. If No, go to section 4
Rating Football Managers with Match-Day Contribution to Performance	No	
If the paper has already been published please refer to the University guidelines on presentation of publications within a PGR Thesis - https://www.liverpool.ac.uk/media/livacuk/tqsd/code-of-practice-on-assessment/annex-7.1-PGR-CoP.pdf		
If the paper is under review with a journal, give details of the journal, including submission dates and the review stage		
If the paper is presented at a conference, give details of the conference		




4. DESCRIPTION OF ALL AUTHOR CONTRIBUTIONS (including the PhD candidate)

Name and affiliation of author	Contribution(s) (for example, conception of the project, design of methodology, data collection, analysis, drafting the manuscript, revising it critically for important intellectual content, etc.)
Kaori Narita	Development of the idea, design of methodology, data collection, main data analysis, drafting the manuscript.
Benjamin Holmes (University of Liverpool)	Part of data collection and analysis.
Ian McHale (University of Liverpool)	Intellectual guidance and suggestions, drafting the manuscript.

5. AUTHOR DECLARATIONS (including the PhD candidate)

I agree to be named as one of the authors of this work, and confirm:

- i. that the description in Section 4 of my contribution(s) to this publication is accurate,*
- ii. that there are no other authors in this paper,*
- iii. that I give consent to the incorporation of this paper/publication in the candidate's PhD thesis submitted to the University of Liverpool*

Name of author	Signature*	Date
Kaori Narita		14/09/2022
Benjamin Holmes		14/09/2022
Ian McHale		14/09/2022

6. OTHER CONTRIBUTOR DECLARATION

I agree to be named as a non-author contributor to this work.

Name and affiliation of contributor	Contribution	Signature* and date

This consent form (Appendix B) or the sole author consent form (Appendix A) for each paper must be completed and kept with the PhD candidate once the paper is finalised. If the paper is to be included as part of the thesis, a copy of this form must be included in the thesis with each publication.