Title
Measuring the Robustness of the National Basketball Association Home Court Advantage Effect

Author
Cheng, Wesley

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Measuring the Robustness of the National Basketball Association Home Court Advantage Effect

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Wesley Cheng

2019

ABSTRACT OF THE THESIS

Measuring the Robustness of the National Basketball Association
Home Court Advantage Effect

by

Wesley Cheng
Master of Science in Statistics
University of California, Los Angeles, 2019
Professor Chad J. Hazlett, Chair

Home court advantage is an often confused term in sports literature that encompasses a variety of factors. Our goal is to identify and quantify home court advantage in the National Basketball Association within a causal framework, and to conduct an extended sensitivity analysis to measure the robustness of our estimated effect. The traditional omitted variable bias framework succeeds in providing an interpretable way of approaching sensitivity analysis, however it also introduces subjectivity in interpreting sensitivity plots and does not extend to multiple confounders very well. By reparameterizing the omitted variable bias framework in terms of R^2 we can use expert knowledge to formally bound the strength of unobserved confounders and provide a more systematic way to conduct sensitivity analyses. We use the imbalance in number of rest days between home and away teams to gauge the robustness of our estimated home court advantage effect.

The thesis of Wesley Cheng is approved.

Frederic R. Paik Schoenberg

Yingnian Wu

Chad J. Hazlett, Committee Chair

University of California, Los Angeles

2019

To my parents,

for their unconditional love and support

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Home court advantage in professional sports is a topic that has been well debated within sports literature and media. A FiveThirtyEight article suggests that after a Cleveland Cavaliers loss in a critical Game 5 to the Boston Celtics even LeBron James, one of the greatest NBA players of all time, could not overcome Boston's home court advantage (Paine, 2018). The Cavaliers would go on to win that playoff series anyways, but this exemplifies the common belief that the home court advantage in sports is very real. In three of the top professional sports leagues in America, the home team wins about 63% (NBA), 54% (MLB), and 58% (NFL) of the time (Moskowitz and Wertheim, 2012). A common guideline in sports betting for the NBA and NFL is that simply playing at home adds about three points to the point spread in favor of the home team. However, while the effect's existence seems be certain, determinations of its constituents and to what extent the effect actually has, is much more unclear.

There has been many studies done to determine what causes home court advantage to exist. Referee bias has been commonly hypothesized as one of the main contributors to home court advantage. Observed goal differentials between home and away teams in the English Premier League were found to be partly due to subjective refereering (Boyko, R Boyko, and G Boyko, 2007) and referees in the NBA have been found to typically make calls in favor of the home team (Price, Remer, and Stone, 2019). Crowd effects have also frequently been pointed to when discussing home court advantage, as it might be natural to conjecture players' performance is enhanced when given positive visual and verbal support. Increased game efficiency (Yi, 2017) and likelihood of wins (Boudreaux, Sanders, and Walia, 2017) by the home team have been found to be attributed to crowd effects.

For the NBA in particular, there has also been much debate regarding the evolution and trend of the home court advantage. An ESPN article in 2015 pointed to the downward trend in home win percentage since 1975 as evidence that the effect was gradually dwindling over time (Haberstroh, 2015). One possible explanation was that an increased reliance on the 3-pointer, which inherently leads to less contact, leaves less room for referee influence. Technology is also pointed to as a potential culprit, as it becomes more common for fans to be immersed with their phones rather than the game itself. Shortly after however, an article from The Economist quickly supplied a retort, suggesting that this purported trend misinterpreted statistical noise for a true signal, going as far to claim that home court advantage in the NBA was actually 'as sweet as ever' (economist.com, 2015).

While there has been much work in the area of determining what home court advantage entails, potential confounders to the effect have been much less studied. Once two teams begin a game, factors such as the crowd engagement and referee subjectivity can help the the home team perform better, but perhaps at a baseline level, away teams are already at a disadvantage when compared to home teams. Travel fatigue (Stefani, 2008) and lack of rest (Moskowitz and Wertheim, 2012) have often been mentioned as competitive obstacles that put away teams at a disadvantage. The imbalance in number of rest days between home and away teams has been quantitatively examined in an attempt to determine what proportion of the home court advantage in the NBA can actually be attributed to rest (Entine and Small, 2008). It was shown that visiting teams play a disproportionate percentage of games on zero days rest compared to home teams (33% to 15%) and that the overall imbalance in rest accounts for about 0.3 points of the 3.24 point home court advantage that home teams enjoyed in the 2004-05 and 2005-06 seasons. This finding serves as the primary motivation for this thesis. For if rest can account for a significant portion of the home court advantage in the NBA, then is it plausible that there potentially exists some unmeasurable confounder that could problematically change our estimated effect?

The goal of this thesis is twofold. First, we estimate the NBA home court advantage effect within a causal framework. Second, we apply an extended sensitivity analysis to measure the robustness of our estimated effect. This thesis is then organized as follows. Chapter

2 provides an overview of the causal model and sensitivity analysis which will serve as the theoretical backbone for the application portion of this thesis. Chapter 3 takes a dive into our NBA data and demonstrates how it fits within the causal and sensitivity frameworks. Chapter 4 examines and discusses the results. Concluding remarks are made in the final chapter.

CHAPTER 2

Sensitivity Analysis for Causal Inference

2.1 Potential Outcome Model

The potential outcome model, also known as the Neyman-Rubin model, is a framework for causal inference that relies on the idea of observed outcomes and unobserved counterfactuals. The concept of potential outcomes was first introduced within the context of randomized experiments (Neyman, 1923). It was further popularized by its application to the estimation of casual effects (Rubin, 1974) and extensions to observational studies when randomization is not possible (Rubin, 1977). We first go over potential outcomes and its general notation, then move into a discussion of when causal effects can be identified in experimental and observational settings.

2.1.1 Potential Outcomes

Potential outcomes can be thought of from a factual and counterfactual point of view. Let D_i denote the treatment assignment for unit i, with $D_i = 1$ when a unit is assigned to treatment, and $D_i = 0$ when a unit is assigned to control. Let Y_i denote the observed outcome and X_i denote the pre-treatment covariates for unit i. We then define the potential outcomes for unit i as follows.

$$Y_i(T_i) = \begin{cases} Y_i(1) & \text{Potential outcome for unit } i \text{ when they receive treatment} \\ Y_i(0) & \text{Potential outcome for unit } i \text{ when they receive control} \end{cases}$$

We can then define the unit treatment effect as

$$\tau_i = Y_i(1) - Y_i(0)$$

With the potential outcomes framework we are outlining, we are also implying a set of assumptions known as SUTVA (Stable Unit Treatment Value Assignment). We first assume there is a single version of the treatment. In addition, we also assume there is no interference between the units. This means that the potential outcomes for one unit is not affected by the treatment assignment for other units. Take for example the case where our sample consists of two units. If interference exists between the two units, each unit now has four potential outcomes $(Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1))$. The result is that the causal effect for each unit now depends on the treatment assignment of the other unit (Rubin, 2003), which can be problematic especially as our sample size increases.

Given this, the primary quantity that we are interested in measuring is the sample average treatment effect

$$\begin{aligned}\tau &= \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0)) \\ &= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]\end{aligned}$$

This average treatment effect is unidentifiable without further assumptions as we cannot calculate these quantities directly given our inability to observe both potential outcomes for any given unit i.

2.1.2 Randomization and Observational Data

The usage of randomization as the foundation for causal inference was popularized by Fisher's work in the early 20th century (Fisher, 1925; Fisher, 1937). The idea is that if the treatment assignment mechanism is random, then any underlying characteristics that could also explain variation in the outcome variable are uncorrelated with the treatment. With potential outcomes, this implies $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i$, which is known as the ignorability assumption. This means our units' potential outcomes are independent of their treatment assignment. A result of proper randomization is that we can recover our average treatment effect. Consider Neyman's difference in means estimator (Neyman, 1923)

$$\begin{aligned}
\hat{\tau} &= \bar{Y}_t^{obs} - \bar{Y}_c^{obs} \\
&= \frac{1}{n_t} \sum_{i:D_i=1} Y_i - \frac{1}{n_c} \sum_{i:D_i=0} Y_i \\
&= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]
\end{aligned}$$

Under randomization, $\mathbb{E}[Y_i(1)|D_i = 1] = \mathbb{E}[Y_i(1)|D_i = 0] = \mathbb{E}[Y_i(1)]$ and similarly for $\mathbb{E}[Y_i(0)|D_i = 0]$. This implies

$$\hat{\tau} = \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] = \tau$$

In most real world data however, we have no control over the treatment assignment mechanism, with the primarily implication being that our ignorability assumption no longer holds. A particularly common problem with observational data is that our treatment variable of interest is now correlated with other variables which also account for a proportion of the variance in the outcome variable. This leads to unreliable biased estimates of our average treatment effect. Because of this we need additional assumptions to draw causal inferences from observational data. Imagine if we looked within a specific slice of our sample by fixing a set of covariates. Upon controlling for these covariates, perhaps it could be argued that our original ignorability assumption holds, in that the treatments are randomly assigned. Then we could treat this particular strata as a randomized experiment and simply weight these strata effects by the probability density of our covariates to estimate our average treatment effect. This new assumption states $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i \mid X_i$ and is known as the conditional ignorability assumption. An additional and often overlooked assumption needed is the common support assumption which states that within any strata, there is a nonzero probability of receiving treatment, i.e., $P(D_i = 1|X_i = x) > 0 \; \forall \; x \in X$. This is needed to prevent improper comparisons between treatment and control groups that have little to no overlap.

We can show that our average treatment effect within each strata is identified in a similar fashion above by using the conditional ignorability assumption. That is, $\tau(X) = \mathbb{E}[Y_i(1) -$

$Y_i(0)|X_i = x] = \mathbb{E}[Y_i(1)|X_i = x, D_i = 1] - \mathbb{E}[Y_i(0)|X_i = x, D_i = 0]$. Using common support and the law of iterated expectations allows us to recover the overall average treatment effect.

$$
\begin{aligned}
\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1) - Y_i(0)|X_i]] \\
&= \int \Big(\mathbb{E}[Y_i(1)|X_i = x, D_i = 1] - \mathbb{E}[Y_i(0)|X_i = X, D_i = 0]\Big)p(x)dX
\end{aligned}
$$

2.1.3 Regression Adjustment

To estimate the average treatment effect in the observational setting, we need a method to perform the conditioning and weighting as described above. Matching is a commonly used technique particularly with propensity scores, which rely only on covariates and not the outcomes (Rosenbaum and Rubin, 1983), however in this thesis we will use regression adjustment as a model-based approach because it dovetails nicely with the sensitivity framework that will be applied. With this model we can define our outcome as

$$Y_i = \beta_0 + \beta_1 D_i + \boldsymbol{\alpha}^T X_i + \epsilon_i$$

We can show that β_1 is an unbiased estimator of the average treatment effect if we make two additional assumptions. If we assume the specification of our model is correct, then in conjunction with the conditional ignorability assumption,

$$
\begin{aligned}
\mathbb{E}[\beta_1] &= \mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i] \\
&= \mathbb{E}[Y_i(1)|X_i] - \mathbb{E}[Y_i(0)|X_i] \\
&= \tau(X)
\end{aligned}
$$

From here, to recover the average treatment effect, the second assumption needed is that the treatment effect across the various strata is constant, i.e. $\tau(X) = \tau$.

One caveat with using a regression based method is that since it is a parametric method, it imposes a functional form on the potential outcomes. This can cause problems if overlap between treatment and control groups is poor as the model will extrapolate counterfactuals for treatment/control observations if the other does not exist (Gelman and Hill, 2006).

Therefore if the model is poorly specified, misspecification bias can render the treatment effect estimate unreliable.

2.2 An Extended Sensitivity Analysis Framework

Sensitivity analyses are used to better understand the relationship between outcome and input variables. This is particularly important in the case where we may have immeasurable confounding variables that could change the estimate of our quantity of interest. In a causal setting, the main concern is how confounding variables may change our treatment effect estimate, which relies on the absence of unobserved confounders. Traditional sensitivity analysis typically involves the investigation of omitted variable bias. This framework can be extended by reparameterizing the omitted variable bias formula in terms of partial R^2 and introducing bounds to the strength of potential confounders by utilizing the researcher's knowledge of observed covariates (Cinelli and Hazlett, 2018). In particular, this thesis is focused on applying Cinelli and Hazlett's work to NBA home court data. For this portion, a brief introduction to omitted variable bias is given, then key highlights and results from Cinelli and Hazlett's extended framework are noted.

2.2.1 Omitted Variable Bias

Omitted variable bias allows the researcher to investigate the effect of unaccounted for variables in the context of their existing model. More concretely, suppose the following restricted regression was run

$$Y = \hat{\tau}_{\text{res}} D + \mathbf{X}\hat{\beta}_{\text{res}} + \hat{\epsilon}_{\text{res}}$$

where $\hat{\tau}_{\text{res}}$ is our estimated constant treatment effect across all the strata in our data and D is a vector of treatment indicator variables. However, while we have controlled for the covariate set $\mathbf{X}$, suppose there is also a confounding variable Z (or set of confounding variables) that we are not able to measure, whether it be for ethical, monetary, or feasibility reasons. The fully specified model that we did not run can be written as

$$Y = \hat{\tau}_{\text{full}} D + \mathbf{X}\hat{\beta}_{\text{full}} + \hat{\gamma} Z + \hat{\epsilon}_{\text{full}}$$

If we can make the argument that upon controlling for $\mathbf{X}$ and Z, the treatment assignment is as-if random, our conditional ignorability assumption would be satisfied, i.e $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i \mid X_i, Z_i$, and if our regression assumptions hold, then $\hat{\tau}_{\text{full}}$ is an unbiased estimate of our sample average treatment effect. Conversely, this implies that our realized estimate $\hat{\tau}_{\text{res}}$ is a biased estimate. While it would be impossible to recover $\hat{\tau}_{\text{full}}$ unless we explicitly measure Z, omitted variable bias allows us to precisely quantify how much our estimate of τ would change depending on the hypothetical influence of Z. Formally we can write the bias of our estimate as

$$\widehat{\text{bias}} = \hat{\gamma}\hat{\delta} \quad \text{where} \quad \hat{\delta} = \frac{\text{cov}(D^{\perp \mathbf{X}}, Z^{\perp \mathbf{X}})}{\text{var}(D^{\perp \mathbf{X}})}$$

The $\perp\mathbf{X}$ notation denotes removing the proportion linearly explained by X from the aforementioned variable. Note $\hat{\delta}$ is equivalent to the coefficient for our treatment indicator if we regressed Z on D and X,

$$Z = \hat{\delta} D + \mathbf{X}\hat{\eta} + \hat{\epsilon}_Z$$

Both coefficients $\hat{\gamma}$ and $\hat{\delta}$ then have a traditional regression interpretation. $\hat{\gamma}$ represents the marginal effect Z has on Y holding D and Z constant, and $\hat{\delta}$ represents the marginal effect D has on Z holding X constant. This means bias can be interpreted as the product of the confounder's various relationships with both the treatment and outcome variables. Sensitivity plots for our causal quantity of interest could then easily be made by plugging in hypothetical values for $\hat{\gamma}$ and $\hat{\delta}$ and adding or subtracting this bias from our original estimate.

However, Cinelli and Hazlett note multiple limitations of the omitted variable bias framework. From the current formulation it is difficult to measure the changes multiple confounders would have on the treatment coefficient. The current formulation also does not allow us to gauge the sensitivity of the treatment coefficient's standard errors, which is important in determining the statistical significance of any treatment effect estimate we may find. In addition, when interpreting the sensitivity plots, the intuition is clear, but determining plausible hypothetical values needed to calculate the bias is challenging and can be subjective.

2.2.2 Extension of Omitted Variable Bias

Cinelli and Hazlett primarily the extend the preexisting sensitivity analysis framework by introducing two new sensitivity measures and a method to bound the strength of confounders with observed covariates. Altogether this allows for a more complete illustration of our quantity of interest's robustness. To do so, the omitted variable bias formula is reparemetrized in terms of partial R^2. For detailed derivations for some of the following results, please refer to Cinelli and Hazlett's appendix section (Cinelli and Hazlett, 2018).

2.2.2.1 Relevant Quantities and Measures

By using the Frisch-Waugh-Lovell theorem to partial out the effects of X and D, and useful identities of the R^2 measure, it can be shown that the bias of our estimate can be rewritten as

$$|\hat{\text{bias}}| = \text{se}(\hat{\tau}_{\text{res}})\sqrt{\frac{R^2_{Y\sim Z|\mathbf{X},D}R^2_{D\sim Z|\mathbf{X}}}{1-R^2_{D\sim Z|\mathbf{X}}}(df)}$$

Here, df denotes the number of degrees of freedom in the restricted regression model without the confounder. The estimated standard error of $\hat{\tau}$ can be expressed as

$$\text{se}(\hat{\tau}) = \text{se}(\hat{\tau}_{\text{res}})\sqrt{\frac{1-R^2_{Y\sim Z|\mathbf{X},D}}{1-R^2_{D\sim Z|\mathbf{X}}}\left(\frac{df}{df-1}\right)}$$

By rearranging terms for our standard error expression, the sensitivity of the variance can be written as the relative change

$$\frac{\text{var}(\hat{\tau})}{\text{var}(\hat{\tau}_{\text{res}})} = \left(1-R^2_{Y\sim Z|\mathbf{X},D}\right)\left(\frac{1}{1-R^2_{D\sim Z|\mathbf{X}}}\right)\left(\frac{df}{df-1}\right)$$

Two new sensitivity measures are introduced to provide a systematic way to quickly asses the overall sensitivity of our estimate. The first is a robustness value which considers a confounder that has equal partial R^2 values with both the treatment and outcome. Define this partial R^2 as the robustness value RV_q. This robustness value can be written as

$$RV_q = \frac{1}{2}\left(\sqrt{f_q^4 + 4f_q^2} - f_q^2\right) \quad \text{where } f_q := q|f_{Y\sim D|\mathbf{X}}|$$

and tells the researcher how strong the partial R^2 would have to be between the confounder and both the treatment and outcome. Confounders that have partial R^2 values that exceed this robustness value would be strong enough to change the estimated effect by $(100 \times q)\%$.

Hypothesis testing can also be incorporated into this robustness value by specifying a significance level α. For the null hypothesis $H_0 : \tau = (1-q)|\hat{\tau}_{\text{res}}|$ to not be rejected, the partial R^2 between the confounder and both the treatment and outcome would need to exceed $RV_{q,\alpha}$ which is defined as

$$RV_{q,\alpha} = \frac{1}{2}\left(\sqrt{f_{q,\alpha}^4 + 4f_{q,\alpha}^2} - f_{q,\alpha}^2\right) \quad \text{where } f_{q,\alpha} := q|f_{Y\sim D|\mathbf{X}}| - \frac{|t_{\text{df}-1}^{\alpha}|}{\sqrt{\text{df}-1}}$$

and $t_{\text{df}-1}^{\alpha}$ denotes the critical value at the α significance level and $df - 1$ degrees of freedom.

The second sensitivity measure proposed is $R^2_{Y\sim D|\mathbf{X}}$, which conveniently also describes how strong the association between confounder and treatment would have to be in the case where the confounder explained all the remaining residual variance of Y. By including this measure it is easy to use this value as a quick benchmark. If $R^2_{Y\sim D|\mathbf{X}}$ is very low, it is possible that a confounder having weak association with treatment, would be enough to change our conclusions regarding the estimated treatment effect. A more formal sensitivity analysis would need to be done, however, this measure can help guide the direction of such an analysis.

2.2.2.2 Bounding Confounder Strength

With bias reparameterized in terms of partial R^2's of the confounder, sensitivity contour plots could be made, however it would be difficult to hypothesize plausible partial R^2 values without further information. A key result from Cinelli and Hazlett's paper is the usage of observed covariates to bound these partial R^2 values. Denote X_j as one particular covariate (although these results also extend to multiple covariates) that will be used to bound the confounders' strength. Then define the following quantities

$$k_D := \frac{R^2_{D\sim Z|\mathbf{X}_{-j}}}{R^2_{D\sim X_j|\mathbf{X}_{-j}}} \qquad k_Y := \frac{R^2_{Y\sim Z|\mathbf{X}_{-j},D}}{R^2_{Y\sim X_j|\mathbf{X}_{-j},D}}$$

These quantities describe the relative strength of the confounder with respect to the treatment and outcome when compared to X_j. Then the strength of the confounders can be expressed as

$$R^2_{D\sim Z|\mathbf{X}} = k_D f^2_{D\sim X_j|\mathbf{X}_{-j}} \qquad R^2_{Y\sim Z|\mathbf{X},D]} \leq \eta^2 f^2_{Y\sim X_j|\mathbf{X}_{-j},D}$$

$$\text{where} \quad \eta = \frac{\sqrt{k_Y} + |f_{K_D} \times f_{D\sim X_j|\mathbf{X}_{-j}}|}{\sqrt{1 - f^2_{K_D} \times f^2_{D\sim X_j|\mathbf{X}_{-j}}}} \quad \text{and} \quad f_{K_D} := \frac{\sqrt{k_D} R_{D\sim X_j|\mathbf{X}_{-j}}}{\sqrt{1 - k_D R^2_{D\sim X_j|\mathbf{X}_{-j}}}}$$

Together this allows the researcher to bound the strength of unobserved confounders by using their knowledge of the data and subject matter at hand. If the researcher can make strong educated guesses regarding the strength of the confouder relative to the strength of an observed covariate, these formal bounds can then be used to calculate the bias and relative change in variance of our estimate. This allows for the investigation of whether the conclusions regarding the estimated treatment effect would be changed significantly.

CHAPTER 3

Research Design

3.1 Data Collection

Individual NBA player game log data for the 2013-14 through 2017-18 seasons is collected from `https://www.basketball-reference.com/`. These years were selected as they represent the most recently completed NBA seasons. The Python module `basketball_reference_web_scraper` is used to scrape the data.[1] Player data is aggregated up to the team and day level to obtain team matchup results.

There were two potential choices for the outcome variable. One option was to leave the grain of the dataset at the matchup level and have the outcome variable Y_{ij} represent the point margin for the home team i against visiting team j.

home_team	away_team	year	home_days_rest	home_travel	away_days_rest	away_travel	point_margin
CHICAGO BULLS	NEW YORK KNICKS	2013	1	-1	0	-1	1
LOS ANGELES CLIPPERS	GOLDEN STATE WARRIORS	2013	1	0	0	0	11

Table 3.1: Data with Point Margin as Outcome

The alternative option was to use each individual team's points scored as the outcome Y_i and split the matchup data into two individual rows. This latter option allows for the usage of a treatment variable to indicate whether a team was home or away which fits within our particular causal and sensitivity framework favorably. Keeping the data at the matchup level means the home court advantage estimate would be returned as an intercept rather than coefficient.

[1] https://github.com/jaebradley/basketball_reference_web_scraper

team	opponent	year	team_days_rest	team_travel	opp_days_rest	opp_travel	treatment	points
CHICAGO BULLS	NEW YORK KNICKS	2013	1	-1	0	-1	1	82
LOS ANGELES CLIPPERS	GOLDEN STATE WARRIORS	2013	1	0	0	0	1	126
NEW YORK KNICKS	CHICAGO BULLS	2013	0	-1	1	-1	0	81
GOLDEN STATE WARRIORS	LOS ANGELES CLIPPERS	2013	0	0	1	0	0	115

Table 3.2: Data with Points Scored as Outcome

In addition to a home/away treatment indicator variable, the number of rest days and time zones traveled are recorded for the team and opponent. These variables serve as covariates that will be included in the model. The number of rest days and time zones traveled requires looking at teams' prior games in relation to their current game. The direction of travel is also taken into account with + corresponding to eastwardly travel and - denoting westwardly travel. The season start year and team names are included in the dataset to allow for the incorporation of fixed effects into our model. Playoff games and games where at least one of the teams were playing their first game of the season (no rest data) are removed from the dataset. Altogether the final dataset contains 6067 games and 12134 observations.

3.2 A Look Into NBA Matchup Data

Creation of the NBA schedule is a sophisticated and opaque process that takes into account various factors. Aside from scheduling 82 games within a restricted time window, other factors must also be taken into consideration. For example, some NBA teams share an arena with NHL team and if a NHL team is not using the arena, the venue is often being used for concerts or other non-sports related events, raising arena availability constraints. Teams are also required to play other teams a certain number of times depending on the year, division, and conference. Schedules with national TV channels like ESPN and TNT must also be taken into account. Given these multitude of circumstances, it would not be surprising to see certain systematic patterns arise with the NBA schedule.

	0	1	2	3+
Away	1763	3237	808	259
Home	822	3943	943	359

Table 3.3: Imbalance in Rest Days between Home and Away Teams

Back-to-back games refer to situations where a NBA team is required to play games on consecutive days. The impact of rest on NBA team performances has been previously studied (Steenland and Deddens, 1997; Entine and Small, 2008) and naturally this variable seems to call for further investigation given its potential impact of team performance. Table 3.3 shows the imbalance in rest days between home and away teams. It is immediately clear that away teams play almost twice the number of back-to-backs compared to home teams. When considering the time associated with air travel, it is not necessarily surprising to see visiting teams play on a more condensed schedule. Allowing for more rest with away teams would prolong already taxing road trips that often consist of multiple consecutive road games. There has been evidence that implies that the NBA league office schedules back-to-back games in order to minimize travel time and costs (Kelly et al., 2008). Regardless of such motives, the result is that on average home teams are entering matchups more rested than their visiting counterparts.

Home teams certainly seem to exhibit some type of advantage over visiting team as seen in Figure 3.1. A five year boxplot of the distributions in points scored for home and away teams show an uptick in points scored for home teams and an overall upward trend in points scored over the past couple of NBA seasons. In an era where there is an increasingly heavier focus on 'space and pace' in the modern NBA offense, an overall increase in total number of points scored is expected. As with most panel data it makes sense to model these time effects by incorporating fixed effects to produce time-invariant estimates.

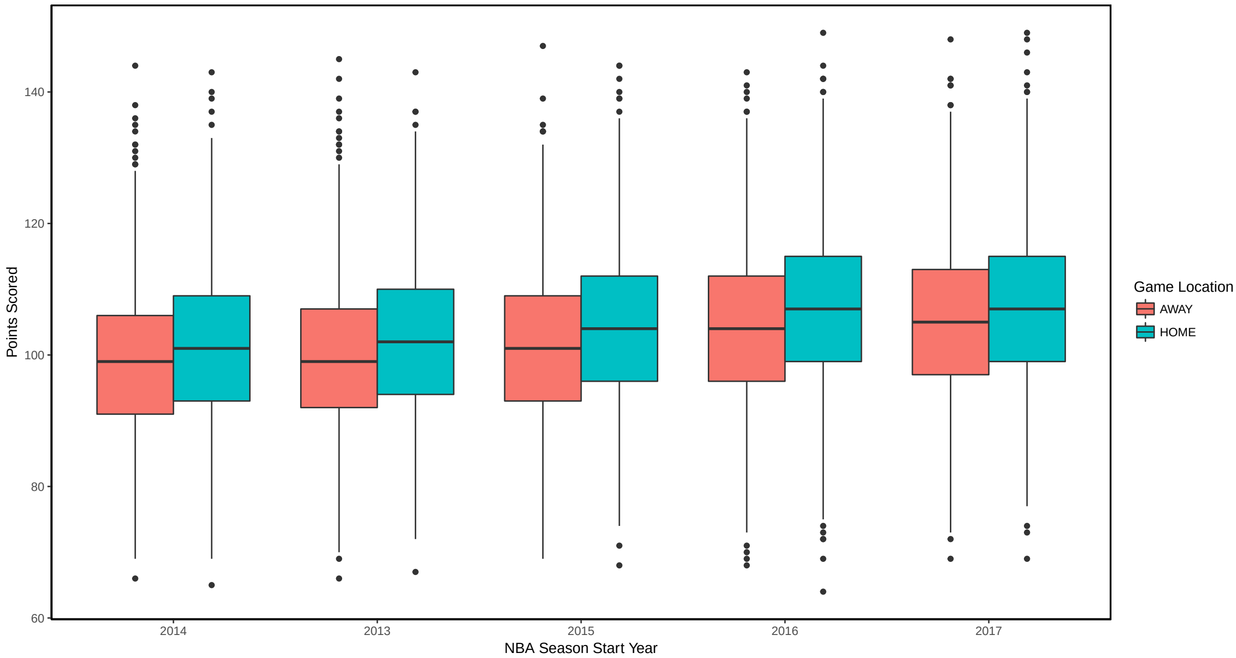

Figure 3.1: Home and away team point discrepancy

3.3 Methodology

In order to incorporate the NBA data within the theoretical framework explained in Chapter 2, three key components must be established. We first define a causal quantity of interest and associated identification strategy. A model specification is then proposed to generate an unbiased estimate of the quantity of interest. Finally we evaluate the robustness of our estimate by conducting a sensitivity analysis. The following sections outline these components.

3.3.1 Identification Strategy

The causal quantity of interest is the effect of playing at home on points scored. Define T_i as the binary indicator variable *treatment*, as illustrated in Table 3.2. For teams playing at home we only see $Y_i(1)$, that is their potential outcome when $T_i = 1$. A quick glance at the data might suggest that this resembles a type of natural experiment in that whether a team plays at home versus away seems to be essentially random. However, Table 3.3 highlights the lack of balance in rest days between home and away teams and it is reasonable to hypothesize that this imbalance might be correlated with our observed outcomes, given the strenuous nature of professional sports. This suggests that any estimate from a simple difference in means estimator would be biased as our potential outcomes are not independent of the treatment assignment. A rest variable then seems to serve as a sensible control. Another factor that could be associated with the number of points scored is a travel variable. Travel is similar to rest in that teams that travel across a greater number of time zones could be at a disadvantage in terms of fatigue. Deviations in points scored due to travel differences are not a primary quantity of interest, however controlling for travel prevents their effects from potentially confounding with our treatment effect estimate.

		Days of Rest			
		0	**1**	**2**	**3+**
Time Zones Traveled	**-3**	0 (0)	31 (40)	7 (16)	9 (13)
	-2	1 (1)	130 (128)	38 (57)	12 (24)
	-1	151 (319)	354 (608)	88 (176)	25 (53)
	0	548 (1140)	3004 (1543)	640 (330)	260 (90)
	+1	127 (303)	319 (686)	103 (170)	33 (61)
	+2	0 (0)	88 (185)	47 (39)	14 (13)
	+3	0 (0)	17 (47)	20 (20)	6 (5)

Table 3.4: Number of home (away) teams by travel and rest

The conditional ignorability and common support assumptions are relied upon as key elements of our identification strategy. For any given season (with the exception of lockout years), teams in the league are required to play 82 games split evenly between at home and on the road. Teams also play every team in the league at least once as both a host and visitor. While we cannot definitively prove that conditional ignorability holds, controlling for both rest and travel would seem to encapsulate most key confounders, given the relatively random nature of the NBA schedule. Table 3.4 shows the number of home and away games per rest/travel combination. The common support assumption does not seem to pose a serious problem in this case as there are an adequate number of treated and control units in each cell that contains units. Note the NBA essentially never schedules a back-to-back if a team is required to travel more than one time zone.

3.3.2 Modeling

A least squares regression model is used to estimate the treatment effect of playing at home, while controlling for confounders like rest and travel. The number of points scored for team i against team j during season t, can be modeled as

$$\begin{aligned}
Y_{ijt} = \mu &+ \alpha_{it} + \gamma_{jt} + \lambda_t + \tau D_i \\
&+ \beta_1 \cdot \mathbb{1}[rest_i = 0] - \beta_1 \cdot \mathbb{1}[rest_j = 0] \\
&+ \beta_2 \cdot \mathbb{1}[rest_i = 1] - \beta_2 \cdot \mathbb{1}[rest_j = 1] + \beta_3 \cdot \mathbb{1}[rest_i = 2] - \beta_3 \cdot \mathbb{1}[rest_j = 2] \\
&+ \beta_4 \cdot \mathbb{1}[travel_i = -3] - \beta_4 \cdot \mathbb{1}[travel_j = -3] \\
&+ \cdots + \beta_9 \cdot \mathbb{1}[travel_j = +3] - \beta_9 \cdot \mathbb{1}[travel_j = +3] + \epsilon_{ijt}
\end{aligned}$$

where α_{it} and γ_{jt} represent team-year fixed effects for teams i and j and λ_t represents league wide season fixed effects. Including these fixed effects allows for the estimation of τ to be based off the within-team variation in points scored rather than the league-wide variation (Mummolo and Peterson, 2018). Rest and travel variables are converted into dummy variables and these coefficients for team i and team j are constrained to be equal. This structure is motivated by Entine and Small (2008) who show that this constraint on the rest coefficient does not impact the error sum of squares greatly compared to when the coefficients are left unconstrained. By constraining these coefficients, this model is also more intuitive to interpret. Teams that enter games on the same number of days of rest and time zones traveled, would mean the model for Y_{ijt} is reduced to

$$Y_{ijt} = \mu + \alpha_{it} + \gamma_{jt} + \lambda_t + \tau D_i + \epsilon_{ijt}$$

which also means the only way team j would enter the equation is through its fixed effects.

3.3.3 Sensitivity Analysis

While factors like rest and travel are controlled for in the model, it is possible there exists unmeasured confounders that could still affect our home court treatment effect estimate and violate our conditional ignorability assumption. Such a scenario would imply that our original causal effect estimate is biased.

One possible confounder is the quality of practice leading up to a game. Due to the limited access nature of the NBA it is difficult to collect data and information on these closed door practices. Some coaches like Clippers coach Doc Rivers prefer lighter practices[2] while others, like former Bulls coach Tom Thibodeau[3], are notorious their high intensity practices. For teams on the road, fitting team activities let alone full practices, into the schedule could be challenging. Given the nature of back-to-backs it simply seems not feasible for teams to run consistent comprehensive practices aside from film studies and shootarounds. Perhaps visiting teams are at a disadvantage compared to home teams because they are unable to practice certain plays tailored for their upcoming opponent, or only have time to break down a limited number of strategies and fixes during their film studies due to time constraints.

A metric measuring the quality of practice might then be a valuable variable to control for. Unfortunately there is no publicly available historical data on teams' past practices. We can then conduct our sensitivity analysis with *practice* as a hypothetical confounder and evaluate how it might affect our original results, if its exclusion were to indeed violate our absence of unobserved confounders assumption. The desired model to be run then becomes

---

[2]https://www.nba.com/clippers/clippers-dont-practice-often-save-energy

[3]https://sports.yahoo.com/richard-hamilton-on-playing-for-tom-thibodeau—practice-was-tough-181655127.html?y20=1

$$
\begin{aligned}
Y_{ijt} = \mu &+ \alpha_{it} + \gamma_{jt} + \lambda_t + \tau D_i \\
&+ \beta_1 \cdot \mathbb{1}[rest_i = 0] - \beta_1 \cdot \mathbb{1}[rest_j = 0] \\
&+ \beta_2 \cdot \mathbb{1}[rest_i = 1] - \beta_2 \cdot \mathbb{1}[rest_j = 1] + \beta_3 \cdot \mathbb{1}[rest_i = 2] - \beta_3 \cdot \mathbb{1}[rest_j = 2] \\
&+ \beta_4 \cdot \mathbb{1}[travel_i = -3] - \beta_4 \cdot \mathbb{1}[travel_j = -3] \\
&+ \cdots + \beta_9 \cdot \mathbb{1}[travel_j = +3] - \beta_9 \cdot \mathbb{1}[travel_j = +3] \\
&+ \delta \cdot practice_i - \delta \cdot practice_j + \epsilon_{ijt}
\end{aligned}
$$

We can use our *rest* variable, which we do have measurements for, to help us bound the strength of the unobservable *practice* variable. Although this is a potential model that includes *practice* as a confounder, *practice* does not need to be specified in this particular linear functional form. The extended sensitivity analysis that will be carried out also allows for confounders that act non-linearly (Cinelli and Hazlett, 2018).

CHAPTER 4

Results

4.1 Model Results

Table 4.1 shows results for the adjusted and unadjusted regression models. The estimated average treatment effect is 2.33 points scored and is significant at the $\alpha = .05$ level. This means that home teams entering a game on equal rest and equal travel, on average, score an additional 2.33 points than if they would have played as a visiting team. Compared to the model that does not take into account the rest and travel control variables, the unadjusted model has an upwardly biased estimate of 2.59 points due in part to the difference in distributions for rest between home and visiting teams as described in the previous section. This 10% difference is inline with Entine and Small's findings, even though they use point margin as the dependent variable rather than points scored.

Of the various categorical dummy variables, only the binary variable indicating if a team was on zero days rest (playing a back-to-back) was found to be significant. The reference level for the rest variable was set at three days of rest, which suggests that holding all other factors equal, a team that plays a back-to-back, on average, scores 1.10 points less than if they had three or more days of rest. Coefficients for variables corresponding to one and two days of rest were also negative, suggesting an overall disadvantageous position when teams do not get three or more days of rest, although these coefficients were not statistically significant, so any meaningful interpretation for these particular coefficients is limited.

Coefficients for the travel dummy variables were relative to no time zones traveled. However, none of these of these coefficients were found to be statistically significant.

Table 4.1: Regression Adjustment Results

	Dependent variable:	
	points	
	(1)	(2)
treatment	2.332***	2.591***
	(0.215)	(0.194)
rest0	−1.096***	
	(0.395)	
rest1	−0.363	
	(0.373)	
rest2	−0.245	
	(0.405)	
travelwest3	−0.536	
	(0.739)	
travelwest2	0.434	
	(0.430)	
travelwest1	0.181	
	(0.225)	
traveleast3	−0.921	
	(0.724)	
traveleast2	−0.161	
	(0.405)	
traveleast1	−0.159	
	(0.216)	
Observations	12,134	12,134
R^2	0.249	0.247
Adjusted R^2	0.230	0.229
Residual Std. Error	10.687 (df = 11829)	10.693 (df = 11838)
F Statistic	12.892*** (df = 304; 11829)	13.194*** (df = 295; 11838)

Note: made with stargazer Hlavac, 2018 package in R — *p<0.1; **p<0.05; ***p<0.01

4.2 Sensitivity Analysis Results

An initial minimal reporting table that summarizes and highlights some simple robustness measures is shown in Table 4.2. Commonly reported regression results are shown alongside quantities that allow the researcher to quickly gain a rough understanding of the sensitivity of their quantity of interest. Looking at the estimate, standard error, and t-value would quickly lead most to suspect that the existence home court effect is indeed evident, though the additional measures provide a more granular view. The robustness value (RV) tells us that a confounder explaining 9.44% of the residual variance in treatment and outcome would be strong enough to bring our home court effect estimate to zero. The quantity $R^2_{Y\sim D|\mathbf{X}}$ considers the worst case scenario where some unmeasured confounder explains all the residual variance in the outcome variable. In this case, $R^2_{D\sim Z|\mathbf{X}}$ would need to be at least as large as $R^2_{Y\sim D|\mathbf{X}}$, or 0.98% in order to bring our estimated effect down to zero.

Treatment:	Est.	SE	t-value	$R^2_{Y\sim D\mid\mathbf{X}}$	RV	$RV_{\alpha=0.05}$
home	2.332	0.215	10.83	0.98%	9.47%	9.46%
df = 11829, Bound (Z as strong as *rest*): $R^2_{Y\sim Z\mid\mathbf{X},D} = 1.22\%$, $R^2_{D\sim Z\mid\mathbf{X}} = 3.03\%$						

Table 4.2: Minimal Reporting Table for Sensitivity Analysis

If we consider a confounder as strong as our observed covariate *rest*, we can bound the strength of such a confounder by using the bounds described at the end of Chapter 2. It can be easily seen that our bounds for $R^2_{Y\sim Z|\mathbf{X},D}$ and $R^2_{D\sim Z|\mathbf{X}}$ both fall short of the robustness value, meaning such a confounder would not be strong enough to reduce our point estimate to zero.

To visualize the strengths of these hypothetical confounders, a contour plot similar to those used in traditional sensitivity analyses can be made by using the partial R^2 values of our confounder on the axes and representing the adjusted estimated treatment effect on the contour lines. Figure 4.1 shows the adjusted treatment effect estimate at varying combinations of partial R^2 values. By bounding the partial R^2 of hypothetical unobserved

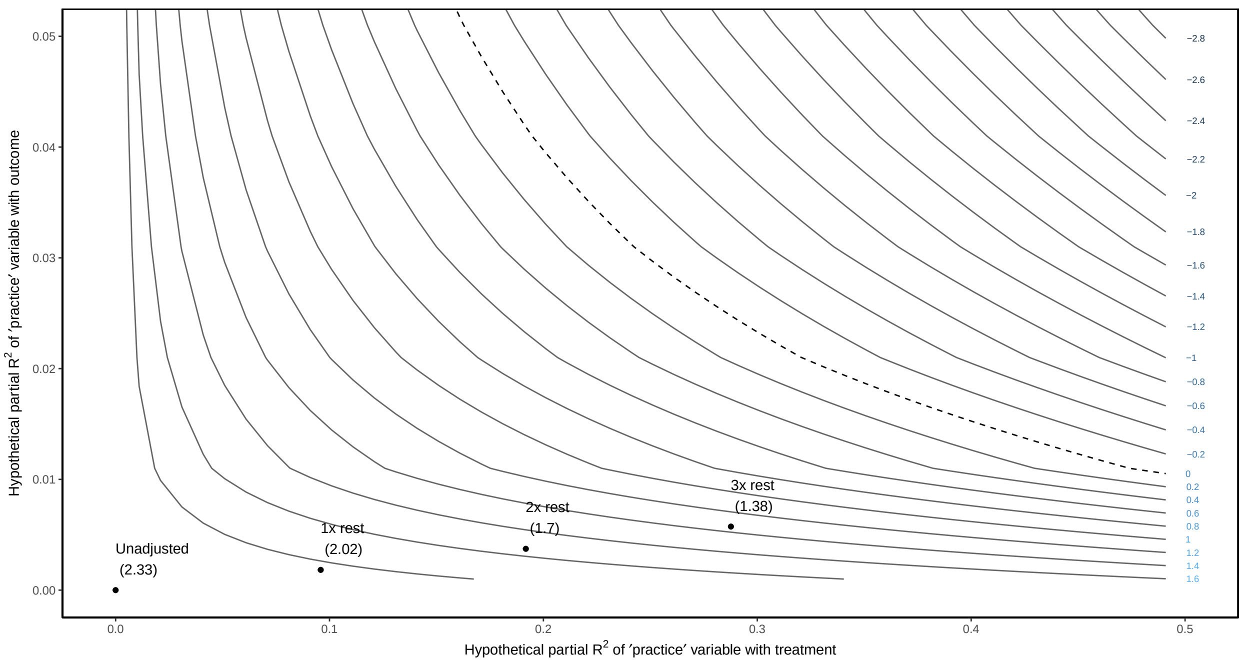

Figure 4.1: Sensitivity contour plot of *home* treatment variable point estimate

confounders k times stronger than our rest variable, we can see that even though a confounder three times stronger would reduce our estimate by about 40%, it would not be nearly enough to completely eliminate our point estimate.

It is also possible to conduct a similar type of sensitivity analysis on the t-statistic of our treatment variable to see if unmeasured confounders would cause inferential concerns. In this particular case, the contour line of interest is the one that corresponds to the critical value for our particular t-distribution. Given our large sample, this critical value roughly corresponds to a value of 2. The t-statistic is sensitive to changes to both the point estimate and variance and we can see this enlarged sensitivity in Figure 4.2 where a confounder three times stronger reduces the t-statistic by over 50% from 10.83 to 4.73. While this percentage reduction is larger than that of the point estimate's and the new value comes closer to our contour line of interest, the statistical significance of estimate is still clearly robust to a confounder three times as strong as our *rest* variable. In fact, it would take a confounder a little more than four times as strong as *rest* before statistical significance is threatened.

Recall the quantity $R^2_{Y\sim D|\mathbf{X}}$, which corresponds to the value that $R^2_{D\sim Z|\mathbf{X}}$ would need

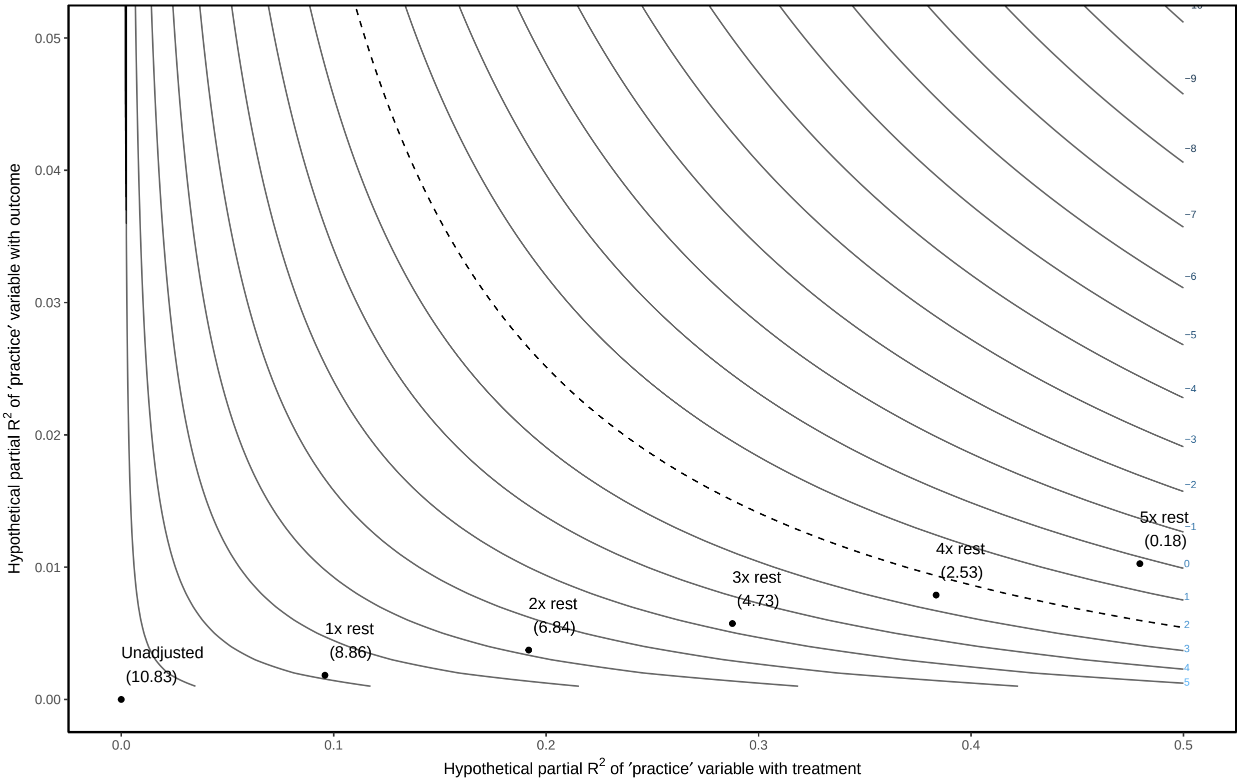

Figure 4.2: Sensitivity contour plot of *home* treatment variable t-statistic

to be in order to eliminate our point estimate if a confounder explained all the residual variance in the outcome variable. The extreme scenario plot in Figure 4.3 helps visualize how our treatment effect estimate is changed in various scenarios where the confounder explains a large proportion of the residual variance in the outcome. The vertical dotted line corresponds precisely with $R^2_{Y\sim D|\mathbf{X}}$. As we consider less extreme scenarios, higher values of $R^2_{D\sim Z|\mathbf{X}}$ are needed to bring the treatment effect estimate down to the same value when compared to the most extreme scenario. This plot is particularly useful when we do not have a good barometer of the potential strength of unobserved confounders relative to our observed covariates. By considering these extreme scenarios, we can gauge the plausibility of the corresponding $R^2_{D\sim Z|\mathbf{X}}$ to see how robust our treatment effect estimate is. Determining the plausibility of the events discussed in this section depends on the context of the research problem at hand.

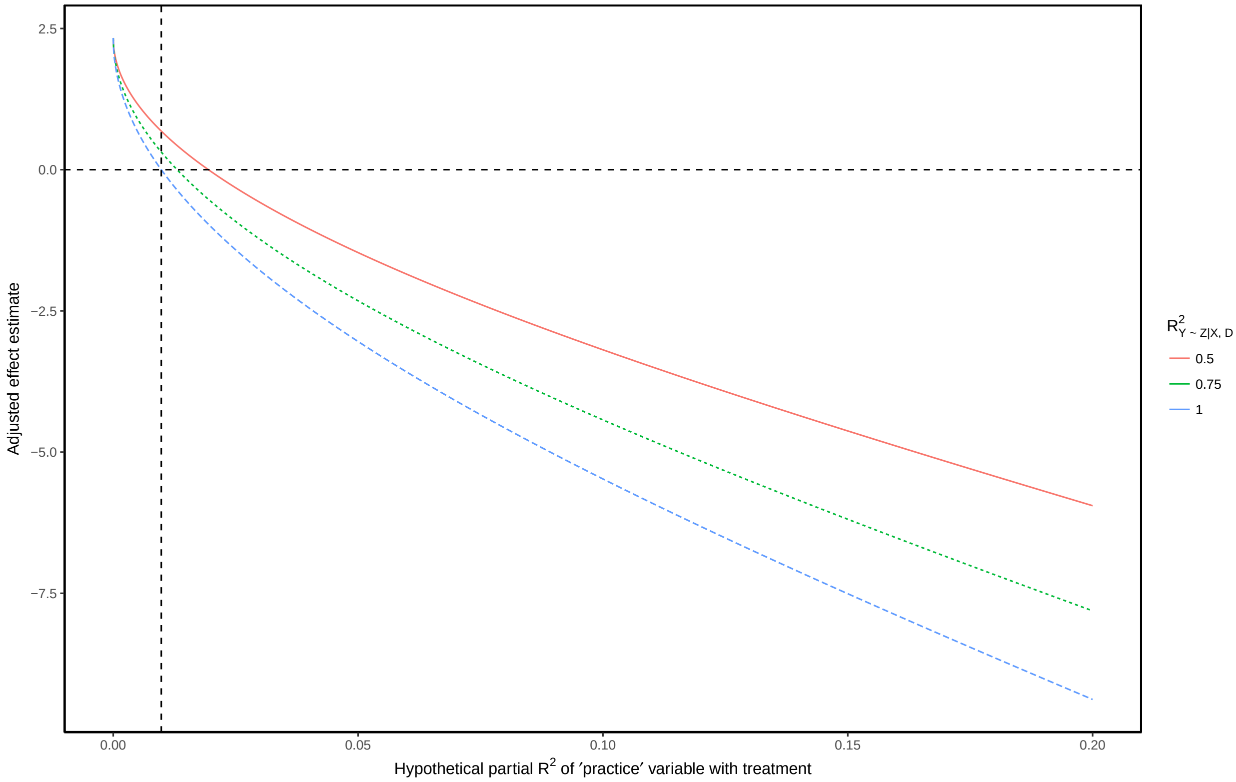

Figure 4.3: Extreme scenario plot of an unobserved confounder

4.3 Discussion

From the results, there indeed seems to be evidence of a significant home court advantage effect that is robust to strong confounders. It would take a confounder four times as strong as *rest* to change our inferential conclusions. One important question to consider is how likely it would be for such a confounder to exist. With sports literature to date, there has been little evidence to suggest that there exists other strong systematic biases built into the NBA's scheduling formula. On the surface, the highlighted imbalance of rest in favor of the home team would seem to play a significant role in contributing to home court advantage, yet its exclusion only contributes a slight upwardly bias in the home court effect.

We previously mentioned a *practice* variable measuring the quality of practice proceeding a particular game as a potential unobserved confounder. The discussion that remains is if such a hypothetical confounder could possibly be four to five times stronger than our baseline pre-treatment covariate *rest*. While it might be strategically optimal for teams to fit in complete practice sessions before each game in order to best prepare themselves for their upcoming opponent's strategy and key players, it simply seems physically unlikely that either team would be running full practices on a consistent basis given the rigors of the NBA's 82 game regular season. Even if home teams were to run more complete practices, with the preexisting high skill level of these professional athletes, the utility to be gained from these sessions over simply watching game film and going over key game plan strategies would seem to be limited. Using *rest* as a comparison, which has more tangible physical impact on athletes' bodies and a more direct connection to their performance during a game, it seems unreasonable to claim that a *practice* variable confounder would be over four times as strong as *rest*.

Examining extreme case scenarios for this particular analysis is not as insightful when compared to other potential sensitivity analyses for a couple of reasons. Considering these scenarios are important when the researcher is not able to make clear comparisons to observed covariates. In our particular case, there is reasonable evidence to suggest that unobserved confounders could not be much stronger than the *rest* variable. In addition, there is a significant amount of noise associated with the outcome variable. The number of points scored by a team in a particular game is affected by a wide range of variables, some being very difficult to measure such as the day-to-day mentality of a team, and is in general fairly difficult to predict. Our treatment variable *home* has a very low partial R^2 of less than 1%, so considering potential confounders that explain a significant proportion of the residual variance in outcome, let alone 100%, would not be very practical for this study.

CHAPTER 5

Concluding Remarks

This thesis discusses the concept of home court advantage in the NBA within a causal framework. In conjunction with assuming the correct specification of our causal model, conditional ignorability is used as an identification technique. This provides an initial unbiased estimate of the casual effect that playing at home has on the number of points scored, given that the absence of unobserved confounders assumption holds true. A detailed sensitivity analysis allows us to investigate how hypothetical confounders, such as *practice*, would affect our original home court advantage estimate if their exclusions did indeed threaten the validity of our prior assumption.

Overall, we find that playing at home is worth about an extra 2.3 points scored, upon controlling for factors such the teams' strength, rest, and travel. Our sensitivity analysis reveals that our treatment effect estimate is robust to confounders that are up to four times as strong as our *rest* covariate. When considering other possible unmeasured variables such as the quality of practice leading into a matchup, it does not seem likely that there exists a confounder strong enough to come close to eliminating our point estimate for the home court advantage effect. Confounders one to two times as strong as our *rest* variable would slightly decrease our treatment effect estimate, although the presence of such confounders seems unlikely given the physical importance of rest in competitive basketball.

There are numerous potential next steps to build on this study. To better understand what leads to the existence of such an effect, even after controlling for the aforementioned variables, it would be interesting to also control for the referees as they are indeed observable for each game. The addition of such a covariate would control for any variations and inconsistencies in home bias by the referees. It also isolates the home court advantage effect estimate to other factors such as those relating to the psychological aspect of playing on the road, which is often what one thinks of when considering home court advantage.

A detailed study regarding the factors considered when creating the NBA schedule would be extremely beneficial in identifying potential factors that could affect our estimated effect. Intricate information regarding the construction of the NBA schedule remains scarce. Extending this study to other professional sports leagues would help paint a more complete image of the competitive landscape in professional sports. Applications from this study include its effect within the sports betting sphere. Modeling the impact that findings from this study have on long term betting strategies could be a potentially rewarding and exciting investigation.

Bibliography

Boudreaux, Christopher J, Shane D Sanders, and Bhavneet Walia (2017). "A natural experiment to determine the crowd effect upon home court advantage". In: *Journal of Sports Economics* 18.7, pp. 737–749.

Boyko, Ryan, Adam R Boyko, and Mark G Boyko (2007). "Referee bias contributes to home advantage in English Premiership football". In: *Journal of sports sciences* 25, pp. 1185–94. DOI: 10.1080/02640410601038576.

Cinelli, Carlos and Chad Hazlett (2018). *Making sense of sensitivity: Extending omitted variable bias.* Tech. rep. Working Paper.

economist.com, ed. (2015). *As Sweet as Ever.* URL: www.economist.com/game-theory/2015/05/24/as-sweet-as-ever.

Entine, Oliver A and Dylan S Small (2008). "The role of rest in the NBA home-court advantage". In: *Journal of Quantitative Analysis in Sports* 4.2.

Fisher, RA (1925). "Statistical methods for research workers. Edinburgh: Oliver and Boyd, 1925". In: *FisherStatistical Methods for Research Workers1925.*

— (1937). *The design of experiments.* Oliver and Boyd; Edinburgh; London.

Gelman, Andrew and Jennifer Hill (2006). *Data analysis using regression and multilevel/hierarchical models.* Cambridge university press.

Haberstroh, Tom (2015). *Home-Court Advantage? Not so Much.* Ed. by ESPN.com. URL: espn.com/nba/story/_/id/12241619/home-court-advantage-decline.

Hlavac, Marek (2018). *stargazer: Well-Formatted Regression and Summary Statistics Tables.* R package version 5.2.2. Central European Labour Studies Institute (CELSI). Bratislava, Slovakia. URL: https://CRAN.R-project.org/package=stargazer.

Kelly, Yvan J et al. (2008). *Cost Reduction, Competitive Balance, and the Scheduling of Back-to-Back Games in the NBA.* Tech. rep.

Moskowitz, Tobias and L Jon Wertheim (2012). *Scorecasting: The hidden influences behind how sports are played and games are won.* Three Rivers Press (CA).

Mummolo, Jonathan and Erik Peterson (2018). "Improving the interpretation of fixed effects regression results". In: *Political Science Research and Methods* 6.4, pp. 829–835.

Neyman, Jerzy S (1923). "On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480)". In: *Annals of Agricultural Sciences* 10, pp. 1–51.

Paine, Neil (2018). *Not Even LeBron Could Overcome The Celtics' Home-Court Magic*. Ed. by FiveThirtyEight.com. URL: `https://fivethirtyeight.com/features/not-even-lebron-could-overcome-the-celtics-home-court-magic/`.

Price, Joseph, Marc Remer, and Daniel Stone (2019). "Profitable Biases of NBA Referees". In:

Rosenbaum, Paul R and Donald B Rubin (1983). "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1, pp. 41–55.

Rubin, Donald B (1974). "Estimating causal effects of treatments in randomized and non-randomized studies." In: *Journal of educational Psychology* 66.5, p. 688.

— (1977). "Assignment to treatment group on the basis of a covariate". In: *Journal of educational Statistics* 2.1, pp. 1–26.

— (2003). "Basic concepts of statistical inference for causal effects in experiments and observational studies". In: *Course material in Quantitative Reasoning* 33.

Steenland, Kyle and James A Deddens (1997). "Effect of travel and rest on performance of professional basketball players". In: *Sleep* 20.5, pp. 366–369.

Stefani, Ray (2008). "Measurement and interpretation of home advantage". In: *Statistical thinking in sports*, pp. 203–216.

Yi, David T (2017). "The Home Court Advantage: Evidence from Men's College Basketball." In: *Sport Journal*.