UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, CONTABILIDADE E ATUÁRIA
DEPARTAMENTO DE ECONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA

Second Leg Home Advantage: Um estudo sobre a Copa do Brasil

Pedro Acioly Teixeira
Orientador: Prof. Dr. Márcio Issao Nakane

São Paulo
2023

Prof. Dr. Carlos Gilberto Carlotti Junior
Reitor da Universidade de São Paulo

Profa. Dra. Maria Dolores Montoya Diaz
Diretora da Faculdade de Economia, Administração, Contabilidade e Atuária

Prof. Dr. Claudio Ribeiro de Lucinda
Chefe do Departamento de Economia

Prof. Dr. Mauro Rodrigues Junior
Coordenador do Programa de Pós-Graduação em Economia

PEDRO ACIOLY TEIXEIRA

Second Leg Home Advantage: Um estudo sobre a Copa do Brasil

Dissertação apresentada no Departamento de Economia da Faculdade de Economia, Administração, Contabilidade e Atuária da Universidade de São Paulo como requisito da obtenção do título de Mestre em Ciências.

Orientador: Prof. Dr. Márcio Issao Nakane

Versão Final

(versão original disponível na Faculdade de Economia, Administração, Contabilidade e Atuária)

São Paulo

2023

Teixeira, Pedro Acioly
Second Leg Home Advantage: Um estudo sobre a Copa do Brasil/ Pedro Acioly Teixeira. – São Paulo, 2023-
67p.

Orientador: Prof. Dr. Márcio Issao Nakane

Dissertação (Mestrado) – Universidade de São Paulo, 2023.

1. Futebol. 2. Econometria. 2. Microeconomia Aplicada. I. Universidade de São Paulo. II. Faculdade de Economia, Administração, Contabilidade e Atuária. III. Second Leg Home Advantage - Copa do Brasil.

Agradecimentos

Agradeço aos meus pais, Luciana Acioly e Alexandre Teixeira, por todo o apoio que sempre me deram profissionalmente. Sou grato ao professor Márcio Nakane, orientador do meu trabalho, pela confiança depositada em minha pesquisa e aos ensinamentos me passado durante esse período. Menciono também os professores Ricardo Avelino e Mauro Rodrigues, pela ajuda tanto na qualificação quanto na avaliação de progresso. Assim, presto meus agradecimentos à banca examinadora pela atenção e tempo depreendido.

Expresso aqui também minha gratidão ao professor Carlos Azzoni, pelas oportunidades dadas a mim no mundo da pesquisa. Em função do apoio prestado durante grande parte da formulação de minha pesquisa, sou grato aos meus colegas do DEE-CADE, especialmente ao Fernando Franke.

Obrigado aos meus amigos por me manterem motivado durante todo o processo. Nomeadamente, sou extremamente grato ao Carlos Nomura, Zeca, Pedro Moreth, Nathaniel, João Marcos, PV, Gabriela Ribeiro, João Paulo, Guilherme Aranha, Giovanni Castiglioni, Leonardo Merlini e Renato Potter.

Por último, quero agradecer também a Universidade de São Paulo, a CAPES e a FIPE por viabilizarem a pesquisa.

Resumo

A presente dissertação estuda o efeito de decidir um confronto eliminatório em mando de campo próprio por uma equipe de futebol, também conhecido como *Second Leg Home Advantage* (SLHA). A abordagem empírica aponta para a existência dessa vantagem, bem como explora a condicionalidade do efeito ao resultado da primeira partida. É aplicado o método das variáveis instrumentais logo após o desenvolvimento de um controle para a qualidade das equipes utilizando as *odds* dos jogos de ida e de volta. Tais estimações apresentam vantagens comparativamente a outros métodos utilizados na literatura. São discutidos o impacto de cada placar do primeiro jogo na probabilidade de vitória na segunda partida, bem como o impacto do gol qualificado nessa variável dependente. Por fim, investiga-se a hipótese de possíveis fatores estratégicos que poderiam explicar essa influência da ordem dos mandos de campo como determinante dos resultados, e infere-se que tais diferenças estratégicas são possíveis razões para a existência de SLHA.

Palavras-chave: Futebol, Econometria, Micro Aplicada, Home Advantage, Second Leg Home Advantage, Economia do Esporte, Sports Economic

Classificação JEL: C50; Z20; Z29

Abstract

This dissertation studies the effect of deciding a knockout match on a team's own home field, also known as *Second Leg Home Advantage* (SLHA). The empirical approach points to the existence of this advantage and explores the conditionality of the effect on the result of the first leg. The instrumental variables method is applied after creating a control variable for team quality using the odds of the first and second leg matches. These estimations offer advantages compared to other methods used in the literature. The impact of each score in the first leg on the probability of winning in the second leg is discussed, as well as the impact of the away goals rule on this dependent variable. Finally, the hypothesis of possible strategic factors that could imply the influence of the home field advantage in the second leg result is investigated, and it is inferred that such strategic differences are possible reasons for the existence of SLHA.

Keywords: Football, Soccer, Econometrics, Applied Microeconomics, Home Advantage, Second Leg Home Advantage, Sports Economic

JEL Classification: C50; Z20; Z29

Sumário

1 INTRODUÇÃO 21

2 REVISÃO DE LITERATURA 23

3 ESTRATÉGIA EMPÍRICA 29
3.1 Dados 29
3.2 Regulamento 30
3.3 Construção do conjunto de dados 31
3.4 Construção das variáveis de controle 32
3.4.1 Classificação dos times no Campeonato Brasileiro 32
3.4.2 *Odds* 34
3.5 Modelo 35

4 ESTATÍSTICAS DESCRITIVAS 37
4.1 Evidências de *Second Leg Home Advantage* 37

5 ESTIMAÇÕES 43

6 EXTENSÃO: UM BOM RESULTADO NA PRIMEIRA PARTIDA . 47

7 ESTRATÉGIAS 53

8 CONCLUSÕES 59

REFERÊNCIAS 60

APÊNDICES 63

APÊNDICE A – TABELAS ADICIONAIS 64

APÊNDICE B – FIGURAS ADICIONAIS 66

Lista de ilustrações

Figura 1 – Proporção vitórias em cada Edição - Mandantes da segunda partida . . 38
Figura 2 – Distribuição da diferença de gols no agregado - Mandantes da segunda partida 40
Figura 3 – Diferença de gols no agregado em cada edição - Mandantes da segunda partida 40
Figura 4 – Proporção vitórias em cada edição e valores previstos - Mandantes da segunda partida 46
Figura 5 – Proporção amostral das vitórias condicional à diferença de gols na primeira partida 51
Figura 6 – Média de Δc por edição da Copa do Brasil 66

Lista de tabelas

Tabela 1 – Vencedores dos Confrontos por Critério 37
Tabela 2 – Vencedores dos Confrontos por Critério - A partir de 2009 37
Tabela 3 – Vencedores dos Confrontos por Fase 39
Tabela 4 – Vencedores dos Confrontos por Fase - A partir de 2009 39
Tabela 5 – Confrontos: Divisão dos Clubes na amostra 41
Tabela 6 – Confrontos: Divisão dos Clubes na amostra - A partir de 2009 42
Tabela 7 – Estimações 43
Tabela 8 – SLHA predito e diferença para SLHA amostral 44
Tabela 9 – Estimações - Modelos com *odds* 44
Tabela 10 – Diferença entre SLHA predito x SLHA amostral - Modelos com *odds* . 45
Tabela 11 – Home Advantage Condicional 48
Tabela 12 – Chance de vitória condicional ao resultado do primeiro jogo 49
Tabela 13 – Home Advantage Condicional - Especificações alternativas 50
Tabela 14 – Diferenças de média: Mandantes ($\overline{\Delta S^m}$) e Visitantes ($\overline{\Delta S^m}$) 56
Tabela 15 – Primeiro Estágio com instrumento 64
Tabela 16 – Home Advantage: Resultados das equipes em casa na Copa do Brasil . 64
Tabela 17 – Variáveis in-match, Médias por trecho do confronto 65

Lista de abreviaturas e siglas

CBF	Confederação Brasileira de Futebol
HA	*Home Advantage*
PCA	*Principal Component Analysis*
RSSSF	*Rec.Sport.Soccer Statistics Foundation*
SLHA	*Second-Leg Home Advantage*
UEFA	*Union of European Football Associations*

1 Introdução

O futebol é, sem dúvidas, o esporte mais popular do mundo. Além da paixão envolvida, o jogo é, em si, um objeto de estudo das mais variadas áreas de conhecimento, como a área da saúde, até temas relacionados à sociologia, história e à economia. Obviamente, o interesse por esse esporte nas áreas sociais é reflexo direto de seu impacto social e da crescente e grandiosa movimentação financeira envolvida.

Em vista disso, existem diversos fatores que tornam o futebol um objeto de considerável relevância para os economistas. A proximidade da racionalidade do jogo e da racionalidade econômica é notória quando se observa o comportamento dos agentes diante de um problema de maximização de *payoffs*. Na teoria econômica, firmas e consumidores escolhem estratégias visando maximizar lucro e utilidade, respectivamente, e levando em conta suas restrições intrínsecas. Em uma partida de futebol, inserida na circunstância de uma competição, os times escolhem estratégias intencionando obter o melhor resultado possível, e consideram seu nível de força e demais restrições para tal.

Em adição a isso, o contexto futebolístico permite estudar diversas dimensões envolvendo interação entre diferentes agentes em ambientes competitivos, com pressão e na presença de diversos fatores que podem influenciar no *payoff* esperado. Por exemplo, duas equipes adversárias podem obter resultados completamente distintos jogando entre si, a depender de qual time possui o mando de campo. Ou seja, a mudança de uma única variável é fator determinante do jogo. Assim ocorre também nas relações econômicas entre firmas, governo e indivíduos: alterações na taxa de juros, mudanças intrínsecas nas preferências individuais, entre outros, modificam por completo as interações entre os agentes em ambientes econômicos.

Nesse sentido, compreender os fenômenos que transcorrem no ambiente futebolístico colabora para elucidar discussões da ciência econômica. A dissertação em foco visa explorar o tema popularmente conhecido como *Second-Leg Home Advantage*, hipótese na qual o time que possui o mando de campo na partida de volta de um confronto eliminatório desfruta de vantagem em relação ao adversário que disputa a primeira partida em casa.

A existência de SLHA, que será demonstrada mais à frente, é resultado direto do comportamento dos times em cada momento do confronto, e revela um dilema na escolha estratégica das equipes. Por conseguinte, o texto tem a intenção de examinar e detalhar o tema aqui referido, utilizando como base as edições da Copa do Brasil. Como será visto mais adiante, os resultados empíricos indicam um efeito positivo do mando de campo na segunda partida de um confronto eliminatório, bem como um empate como no primeiro trecho mostra-se um ótimo resultado para o clube que decide em casa, em termos

probabilísticos.

Diante do exposto, além dessa seção de introdução, essa dissertação será dividida da seguinte maneira: no capítulo 2 será feita a revisão de literatura. No capítulo 3, estarão descritas a metodologia e criação da base de dados, bem como estará exposto o desenvolvimento das variáveis de controle, inclusive utilizando variável instrumental. Tal contribuição metodológica apresenta vantagens em relação à adoção de variáveis de controle comumente empregadas na literatura, as quais são desenvolvidas com certo grau de arbitrariedade, conforme será tratado posteriormente. No capítulo 4, serão abordados as primeiras estatísticas descritivas. O capitulo 5 apresentará as estimções econométricas, enquando o capítulo 6 contará com as extensões dos modelos. O capítulo 7 abordará as discussões em relação as estratégias do jogo. Finalmente, o capítulo 8 apresenta a conclusão desta dissertação.

2 Revisão de Literatura

A existência de *Home Field Advantage* (HA) é reconhecida na literatura sobre vários esportes, inclusive o futebol (DOBSON; GODDARD, 2011). De maneira concisa, pode-se definir tal fenômeno como a vantagem relativa usufruída por uma equipe ao jogar uma partida em seu próprio estádio comparativamente a disputar esse mesmo jogo em campo adversário.

Frequentemente, de acordo com Dobson e Goddard (2011), os determinantes de HA podem ser divididos em 4 diferentes categorias: familiaridade, distância entre os clubes, efeitos da torcida e regras que favoreçam o time da casa. Familiaridade refere-se ao conhecimento natural que o time mandante possui sobre seu próprio estádio - ou desconhecimento, por parte dos visitantes. A distância, por sua vez, pode interferir na preparação e cansaço do time visitante para a partida, dada a necessidade de viajar. Em relação aos efeitos de torcida, estes podem ocorrer por 3 diferentes canais, via encorajamento do time da casa, intimidação dos visitantes e influência sobre a arbitragem (*referee bias*). Por fim, regras como a de gol qualificado fora de casa em confrontos eliminatórios também podem interferir em HA.

Ademais, outros possíveis efeitos também foram explorados como explicação para a vantagem de jogar em mando próprio. Táticas específicas usadas pelos times, com os visitantes jogando mais defensivamente em geral, poderiam implicar em uma vantagem psicológica do time da casa (POLLARD, 1986). Uma evidência apontada por Pollard e Pollard (2005) é que Home Advantage é consideravelmente maior em duelos eliminatórios do que em partidas da liga nacional, considerando o cenário das competições europeias. Isso poderia ser atribuido ao fato de que uma derrota fora de casa por pequena diferença de gols é em geral encarada como reversível no jogo de volta, o que levaria aos times visitantes obedecerem a táticas ainda mais defensivas.

Fatores psicológicos também são discutidos como determinantes de HA. Os autores argumentam que desde o início do futebol inglês - ainda no século XIX - existe notável vantagem de se jogar em casa. As próprias crenças das equipes (fundamentadas ou não) sobre familiaridade e outros fatores reforçariam essa vantagem. Por fim, Neave e Wolfson (2003) argumentam para a territorialidade como razão adicional de HA. De acordo com os autores, territorialidade é definida como o senso de proteger seu território de invasores. Curiosamente, os níveis de testosterona dos jogadores, medidos antes das partidas, foram significativamente maiores antes de jogos em casa, comparativamente a partidas como visitantes.

Assim, o fenômeno de Home Advantage e suas causas são bem reconhecidos na

literatura sobre futebol. No entanto, uma parte importante da discussão ainda não apresenta consenso: a existência de Second-Leg Home Advantage (SLHA).

SLHA pode ser definido como a vantagem de se jogar o segundo jogo de um confronto eliminatório como mandante em relação a disputar essa mesma partida como visitante. Em outras palavras, SLHA implica em maior efeito de HA no segundo jogo de um torneio, tudo o mais constante.

Dessa maneira, Page e Page (2007) observam a existência de SLHA ao analisarem as copas europeias ao longo de 51 anos, concluindo que o time mandante no jogo de volta possui em média mais de 50% de chance de avançar para a segunda rodada. Lidor et al. (2010) ao examinarem também as competições europeias chegaram a concluisões similares. No entanto, Eugster, Gertheiss e Kaiser (2011) e Amez et al. (2020) apontam para a inexistência de SLHA e que toda a diferença notada pode ser explicada pelo desempenho dos clubes na fase de grupos da *Champions League*. Tal conclusão é corroborada por Abad et al. (2017) utilizando os dados das partidas da Copa Libertadores da América.

Dessa forma, inexiste consenso quanto ao assunto. É necessário, portanto, investigar quais as diferenças metodológicas que levaram às diferenças nos resultados encontrados.

Page e Page (2007) iniciam a investigação desse fenômeno, utilizando dados das 3 maiores competições europeias entre 1955 e 2006. Os torneios analisados foram a *Champions League* (*Champions Cup* até 1998), a *UEFA Cup* (*Inter-Cities Fairs Cup*), e a *Cup Winners Cup*. Vale ressaltar que em geral o chaveamento dos confrontos eliminatórios não era randomizado ao longo do período estudado, com os times mais fortes sendo alocados para receber o jogo de volta em seu estádio. Dessa maneira, os autores decidem por controlar pela diferença de habilidade entre os times, utilizando o índice de habilidade dos times da UEFA - que a própria utiliza para o chaveamento dos times.

$$habilidade_{it} = \sum_{k=1}^{5} coef_{it-k} \tag{2.1}$$

Em que t representa o ano da edição em questão e k representa a defasagem. Portanto, o índice acima define a habilidade do time i no ano t, resultante do somatórios dos coeficientes ($coef$) da UEFA nos últimos 5 anos. Vale ressaltar que o método de cálculo do coeficiente $coef$ mudou duas vezes ao longo do período. Dessa maneira, os autores estimam a seguinte regressão logística para calcular a probabilidade p para o time mandante no jogo de volta avançar à próxima fase:

$$log\left(\frac{p}{1-p}\right) = \alpha + \beta \Delta habilidade \tag{2.2}$$

$$
\begin{aligned}
log\left(\frac{p}{1-p}\right) = \alpha &+ \beta_1 I_{61-99}\Delta habilidade \\
&+ \beta_2 I_{00-04}\Delta habilidade \\
&+ \beta_3 I_{05-06}\Delta habilidade
\end{aligned}
\tag{2.3}
$$

Em que $\Delta habilidade = habilidade_i - habilidade_j$ representa a diferença de força entre duas equipes i e j em determinada edição. A equação (2.3) é equivalente à equação (2.2), porém exibindo o termos I, os quais assumem o valor 1 se a partida aconteceu em data entre o período sinalizado em seus subscritos e 0 caso contrário. O termo I é importante pois como os coeficientes da UEFA são calculados de maneiras diferentes em cada um desses períodos, estamos na verdade interessados em três valores para β, isto é, β_1, β_2 e β_3.

Repare que, se não há diferença entre habilidade dos times, p é uma função de α, de tal forma que se $\alpha = 0$, não há Second Leg Home Advantage.

Os resultados estimados pelos autores indicam que há SLHA nas três competições analisadas. Além disso, encontram que apesar de existir essa vantagem, ela vem diminuindo com o passar do tempo.

Adicionalmente, o estudo indica que parte dessa vantagem vem de fato da prorrogação, onde os times mandantes do segundo jogo usufruem de 30 minutos extras de HA. Nas disputas de pênaltis, também existe essa vantagem.

No entanto, cabe ressaltar que esses dois fatores não são suficientes para explicar o fenômeno por completo. Mesmo removendo as partidas decididas no tempo extra e nas disputas de penalidades, a equipe da casa na segunda partida ainda possui cerca de 54% de chance de avançar, tudo o mais constante.

Lidor et al. (2010) enunciam a mesma questão relacionando SLHA com o número de gols marcados em cada partida do confronto. Ou seja, se esse efeito está ligado ao número de gols marcados pela equipe mandante do segundo jogo. A análise da variância do número de gols marcados em cada um dos dois jogos revelou que o número médio de tentos marcados pelo clube da casa é maior no segundo jogo. Portanto, os autores chegam a conclusão similar a Page e Page (2007). Todavia, não foram diretamente empregados controles para a força de cada um dos times do confronto.

De maneira geral, outros autores exploram os dados de maneira semelhante ao modelo proposto por Page e Page (2007), com algumas poucas alterações, mas que levam a resultados diferentes. Por exemplo, Eugster, Gertheiss e Kaiser (2011) adicionam outros controles para a diferença de força entre os times. Primeiramente, os autores consideram a performance das equipes ao longo da fase de grupos do torneio em questão, juntamente

com os coeficientes da UEFA. Ademais, a forma de calcular a diferença C de habilidade das equipes é ponderada pela habilidade do time mais forte da competição:

$$C = \frac{CoeficienteUEFA(T_2) - CoeficienteUEFA(T_1)}{max(CoeficienteUEFA)} \tag{2.4}$$

Em que C é diferença de habilidade entre as equipes T_2 e T_1, o mandante e o visitante na segunda partida, respectivamente.

Ademais, ao adicionar um controle relativo ao desempenho dos clubes na fase de grupos da competição, tem-se uma medida mais atualizada do desempenho das equipes. Utilizando tais controles, os autores concluem que não há vantagem de jogar o segundo jogo de um confronto eliminatório em casa, e propõem que o time melhor classificado na primeira fase deveria, então, poder livremente escolher a ordem dos mandos de campo.

Waquil, Horta e Moraes (2020) adotam uma análise semelhante para a Copa do Brasil, porém constroem uma proxy para a habilidade dos times baseada no desempenho das equipes no Campeonato Brasileiro e no próprio torneio. O modelo econométrico estimado, como os próprios autores argumentam, é uma extensão de Page e Page (2007). A conclusão do artigo aponta para a existência de SLHA no torneio, entretanto, os autores selecionam uma amostra de partidas diferente do presente texto, na qual consideraram também confrontos onde se aplicavam a regra de eliminação por dois ou mais gols fora de casa na primeira partida - que será detalhada mais adiante.

Todavia, em contraposição à maior parte dos estudos, Geenens e Cuddihy (2018) optam por uma abordagem não-paramétrica da questão. Também utilizando as informações da *Champions League* e *Europa League*, os autores indicam que há de fato um benefício de se disputar o segundo jogo com o mando de campo. Porém, a contribuição mais importante desse artigo é em relação a sua metodologia.

Os autores argumentam que, embora muito difundido, o uso de modelos de regressação logística podem não ser adequados em certas ocasiões. Além disso, estudos como Page e Page (2007) e Eugster, Gertheiss e Kaiser (2011) partiriam do pressuposto de que os modelos especificados se adequam aos dados sem observarem previamente como esses dados se comportam. Portanto, Geenens e Cuddihy (2018) decidem adotar uma abordagem não-paramétrica, utilizando o estimador NW (Nadaraya–Watson) e diferentes intervalos de confiança (Wilson e Agresti-Coull).

Em todo o caso, é notável que grande parte do debate acerca da presença SLHA insere-se em um contexto controverso: a escolha da medida de força entre os times. Em algum grau, as variáveis de controle escolhidas serão sempre arbitrárias, uma vez que a diferença de força entre os times é não observável de maneira direta e não há consenso de como definí-la.

Além desse ponto, as possíveis causas de SLHA ainda não são claras. Quais fatores levariam uma equipe a usufruir de maior efeito de Home Advantage na partida de volta comparado com a partida de ida? Algumas possíveis explicações são rapidamente apresentadas na literatura, dentre as quais estão as diferentes estratégias utilizadas pelos times em cada partida do confronto eliminatório, a regra do gol qualificado fora de casa e a possível prorrogação no segundo jogo.

Em relação à primeira explanação acima, argumenta-se que no primeiro jogo as equipes optam em geral pela estratégia de jogar da melhor maneira possível, de modo a buscarem maximizar seus pay-offs (LIDOR et al., 2010). No entanto, na segunda parte do confronto, as equipes lançam a estratégia de buscar o mínimo resultado necessário para avançarem à próxima fase, escolhendo, portanto, uma meta definida de gols.

Por sua vez, a regra do gol qualificado pode também implicar em uma maior vantagem de se jogar a última partida em território próprio. Se no trecho de ida a equipe visitante conseguir marcar gols, ao jogar a partida seguinte ela susufruirá da vantagem de perder (ganhar) pela mesma diferença de gols que ganhou (perdeu) anteriormente para avançar a próxima fase, desde que não sofra um numero de tentos maior em casa.

Além desses fatores, a existência do tempo extra na partida de volta pode também exercer influência no *payoff* final do confronto. Caso ao final dos 180 minutos totais o confronto termine empatado (inclusive em relação aos critérios de desempate), o time mandante ainda terá 30 minutos adicionais em seu campo para definir o confronto, ou seja, maior tempo sob o efeito de HA. Inclusive, se a partida for definida em última instância nas cobranças de pênalti, é possível que a equipe mandante também disfrute de alguma vantagem adicional.

Como consequência, a presente dissertação buscou explorar a Copa do Brasil para contribuir com o debate sobre SLHA. A ideia é que a Copa do Brasil é possivelmente o torneio mais próximo do ideal para se avaliar a presença, os efeitos e as causas de SLHA.

Em primeiro lugar, o formato do torneio desde seu início, em 1989, é parcialmente homogêneo. Embora ocorreram diversas mudanças tanto na definição dos classificados para a disputa do torneio quanto na forma de disputa das primeiras fases da competição, sempre foi adotado o sistema de confronto em dois jogos obrigatórios pelo menos a partir das oitavas de final. Ademais, nunca houve a implementação de prorrogação como desempate, o que exclui a vantagem artificial natural desse critério. Por fim, entre 1989 e 2017, era adotado o critério do gol qualificado fora de casa, o que caiu em desuso a partir da edição de 2018. Além disso, a determinação de qual equipe possuirá o mando de campo na volta é de maneira aleatória há várias edições.

Desse modo, a Copa do Brasil pode ser considerada como uma competição que permite observar o fenômeno de SLHA em um cenário empírico próximo a um ensaio

em laboratório. Torna-se relativamente direto e simples testar a existência do efeito de SLHA. Consequentemente, permite-se testar qual (ou quais) seriam boas medidas de força e habilidade das equipes. Por fim, também tal cenário permite examinar as causas, ou pelo menos as singularidades, que justificam a presença ou ausência de SLHA.

Logo, utilizando os dados do torneio em questão, serão abordados os seguintes pontos: i) A existência de SLHA; ii) Qualidade dos controles de força; iii) Diferenças estratégicas em cada partida do confronto.

3 Estratégia Empírica

3.1 Dados

Os dados foram extraídos das seguintes fontes: do Bola na Área (*bolanaarea.com*), *oddsportal.com*, *betexplorer.com*, dos sites da CBF (*cbf.com.br*) e RSSSF Brasil (*rsssfbrasil.com*) e, por fim, do Sofascore (*sofascore.com.*)

As informações correspondentes aos resultados das partidas das edições entre 1989 a 2008 da Copa do Brasil foram extraídas do primeiro site citado acima, enquanto as observações correspondentes aos anos de 2009 a 2020, foram extraídas do *oddsportal*, site agregador dos *odds* das partidas. Os dados relacionados às edições de 2021 e 2022 foram obtidos do *betexplorer*, que também apresenta o agregado das *odds* das casas de apostas. Tais *odds* representam a média dos valores definidos pelas casas de apostas, e foram utilizados nas regressões apresentadas como medidas de força relativa entre os times, o que será mais bem detalhado na seção sobre estratégia empírica.

A partir de 2012, também foram utilizadas as informações da CBF para compor a base de dados. O site da RSSSF Brasil foi utilizado em especial como consulta aos regulamentos das edições e eventuais dados faltantes. Também foram obtidas as informações específicas de cada partida, normalmente chamados de dados *in-match*. Tais dados foram obtidos do *Sofascore*, e correspondem às edições de 2015 em diante. Cabe ressaltar, no entanto, que tal base possui várias lacunas (*missings*), que serão discutidos mais adiante. Na tabela 17, consta a lista das variáveis extraídas.

Conforme será abordado mais adiante, as informações *in-match* foram utilizadas para testar a hipótese de que os jogos de volta são disputados com estratégias distintas do primeiro. Em outras palavras: os mandantes do jogo de volta se comportam diferentemente dos mandantes do jogo de ida, o que também vale para as equipes visitantes. Consequentemente, a existência de SLHA poderia estar parcialmente relacionada a uma (má) escolha estratégica das equipes.

Por fim, também do *Sofascore*, foram obtidas as classificações das equipes no campeonato brasileiro, desde 2008 para a série A e série B, e desde 2013 para série C. As edições de 2009 a 2012 da terceira divisão foram ignoradas na construção do conjunto de dados, por se entender que o regulamento da época - dividindo em 4 grupos de 5 equipes - não possibilitaria a criação de um bom critério para controle. Adicionalmente, do Bola na Área extraiu-se a classificação das equipes entre as edições de 1988 a 2007 do campeonato nacional, para as duas divisões superiores do torneio. O Bola na Área utiliza um critério próprio para calcular os postos finais de cada clube durante as edições anteriores ao sistema

de pontos corridos, levando em conta tanto o desempenho nas primeiras fases quanto os confrontos da fase eliminatória do campeonato brasileiro, anteriormente à adoção do sistema de pontos corridos. Na seção seguinte será detalhada o uso dessas informações para a parte empírica do texto.

3.2 Regulamento

Historicamente, diversos regulamentos foram utilizados na competição, com critérios diferentes tanto para a classificação ao torneio quanto em relação ao formato do torneio em si. A grande vantagem de utilizar a Copa do Brasil como objeto de estudo, no entanto, é que o sistema de chaveamento aleatório foi adotado há muito tempo para os confrontos em que dois jogos eram mandatórios. Dessa forma, os melhores e piores times tiveram seus mandos de campo nas partidas de ida e volta decididos de modo randomizado, o que torna a presente análise mais direta.

Deve-se destacar, ainda, que na Copa do Brasil nunca houve a adoção de prorrogação como critério de desempate. Conforme abordado na revisão de literatura, a adoção da prorrogação cria um viés (artificial) para o estudo de SLHA: o time mandante do segundo jogo estará por 30 minutos a mais sob o efeito de HA, após empate no tempo regulamentar. Além disso, o visitante dessa partida poderá usufruir de vantagem extra caso consiga um tento durante a prorrogação. Como não se sabe a magnitude exata desses efeitos, torna-se complicado calcular o real efeito de SLHA.

Outro ponto histórico importante foi a regra do gol qualificado fora de casa, utilizada em todas as edições entre 1989 até 2017. Em 2018, tal regra cai como um todo, porém ressalta-se que desde 2015 as finais já não possuíam esse critério de desempate. Em adição, dois outros confrontos também foram disputados sem a regra do gol qualificado: Atlético Mineiro e Cruzeiro, em 2014, e Flamengo e Vasco, em 2006. Neles, tanto o primeiro jogo quanto a segunda partida foram disputados no mesmo estádio (Mineirão e Maracanã, respectivamente) e houve o entendimento que o mando de campo deveria ser considerado neutro. Dessa forma, de antemão optou-se por não se aplicar tal parte do regulamento . De qualquer modo, tem-se um cenário em que ambas as equipes estão expostas pelo mesmo período de tempo sob esse critério na disputa.

No que se refere às mudanças históricas no torneio em questão, entre 1989 a 1994, os classificados para o torneio eram os campeões estaduais (embora alguns estados ainda não possuíssem representantes, cada vez mais estados aderiram à competição) junto a algumas outras equipes, classificadas por outros critérios, como exemplo, o ranking da CBF. Todas as fases eram disputadas em dois jogos obrigatórios: ida e volta.

Ao longo dos anos, o número de participantes na competição aumentou significativamente, com a modificação das regras de classificação e também via convite a equipes. O

detalhamento dessa parte dos regulamentos, embora curioso, foge do escopo do texto em questão. Como resultado, mais fases foram estabelecidas no torneio, de modo a acomodar a maior quantidade de times.

A partir de 1995 até a edição de 2016, havia, nas duas primeiras fases, do torneio um critério de desempate adicional. O time visitante que vencesse por dois ou mais gols de diferença no primeiro jogo (três ou mais na edição de 1995) se classificaria automaticamente para a próxima fase, sem a necessidade de uma segunda partida. Adicionalmente, em tais encontros, o chaveamento não era aleatório no período em questão: os times com melhor ranking na CBF disputavam o primeiro jogo fora de casa. Aliás, esse foi o único critério de chaveamento das equipes para a determinar os mandos dos jogos de um confronto durante esse período da Copa do Brasil.

Do ano de 2017 em diante, as duas primeiras rodadas foram disputadas em jogos únicos, sendo que na primeira rodada o visitante (time com melhor ranking da CBF) avançaria até mesmo em caso de empate e, na segunda rodada, o desempate seria via cobrança das penalidades.

3.3 Construção do conjunto de dados

Defina-se "confronto"como sendo o conjunto de jogos de um time i contra um time j em uma mesma fase da Copa do Brasil para determinado ano em questão. Ou seja, um confronto pode ser composto de dois jogos (um de ida e um de volta), ou de uma única partida, como explicitado anteriormente.

Para a criação do conjunto de dados final foram feitas uma série de filtragens em relação às partidas. Primeiramente, removeram-se os jogos nos quais ocorreram algum empecilho para sua realização ou se houve qualquer decisão extracampo. Destaca-se que, se o problema ocorreu no segundo jogo, ambas as partidas foram desconsideradas. No total, 18 observações foram removidas da amostra.

Em concordância com o descrito na seção acima, para a construção da base de dados foram excluídas da amostra todas as observações referentes às duas primeiras rodadas de 2017 até 2022, uma vez que tais confrontos foram disputados em jogos únicos.

Além disso, considerando os regulamentos expostos, para as duas primeiras rodadas das edições de 1995 a 2016, é necessário lidar com o critério eliminatório do jogo de ida. Como o interesse da pesquisa está em descobrir se a equipe mandante do jogo de volta possui uma vantagem em comparação com os visitantes (considerando tudo o mais constante, incluindo o nível de força), poderia-se pensar em duas soluções plausíveis.

A primeira solução viável seria excluir da amostra os confrontos em que essa regra se aplicou. Ou seja, eliminar aqueles em que não foram diputados o segundo jogo, uma vez

decididos na primeira partida por dois ou mais gols de diferença pelo time visitante. Porém, vale ressaltar que, pela existência dessa regra, o incentivo é completamente diferente para os times nesse tipo de confronto. É factível pensar que a equipe mais forte entre com a mentalidade de definir a disputa logo, enquanto os mandantes, mais fracos, busquem apenas sobreviver na primeira partida, a título de exemplo. Como o interesse do estudo é avaliar o comportamento das equipes em confrontos em que há necessariamente a disputa de dois jogos, optou-se por uma segunda solução: apagar essas fases do torneio por completo da amostra.

Com as exclusões acima, o conjunto de dados final consta com 1494 partidas, ou 746 confrontos disputados em dois jogos, entre 1989 e 2022. Estão disponibilizadas informações para: i) Placar da partida; ii) Time mandante; iii) Time visitante; iv) Fase do jogo; v) Se o jogo disputado era referente à ida ou a volta do confronto eliminatório; vi) Vencedor da disputa de pênaltis; vii) A partir de 2009, as odds médios das casas de apostas para vitória de cada time e empate viii) estatísticas *in-match*, a partir de 2015; ix) classificação das equipes no campeonato nacional (a partir de 1989 para série A e série B e a partir de 2013 para série C).

3.4 Construção das variáveis de controle

3.4.1 Classificação dos times no Campeonato Brasileiro

Assim como abordado anteriormente, foram empregados como controle de força de cada equipe a classificação no campeonato brasileiro. No que se refere aos anos entre 2006 e 2022, a construção da variável se deu da seguinte maneira: primeiro, a classificação da série A seguiu o ordenamento padrão do primeiro ao vigésimo colocado, enquanto para a série B somaram-se 20 colocações para cada posição. Em outras palavras, o líder da segunda divisão seria o 21º colocado e o último, o 40º, por exemplo.

No tocante à série C, nas edições sem observações, considerou-se os times como o 41º na classificação. A partir de 2013, somaram-se 40 colocações para cada posição. Como o torneio é organizado em 2 grupos de 10 times, temos para cada colocação entre 41 e 50, duas equipes. As equipes que disputam a série D ou que não participam de nenhuma divisão foram consideradas como últimas colocadas.

Para os torneios compreendidos entre 1989 e 2005 a construção da variável se deu de maneira diversa, devido às peculiaridades do torneio nacional. Como os regulamentos do campeonato brasileiro variaram frequentemente em relação ao formato, número de participantes e quantidade de rodadas disputadas por cada equipe, foram definidos alguns critérios para ordenar as equipes.

Primeiro, foram considerados todos os times que não participavam de nenhuma

das duas divisões superiores como últimos colocados. Em outras palavras, as equipes sem divisão ou que participavam da série C foram classificadas como de mesma força. Essa escolha se deu devido ao fato de que em vários anos a terceira divisão não foi disputada e, ademais, grande parte dos times disputavam uma quantidade ínfima de jogos o que, para critérios de ordenamento em uma classificação geral, torna um ranking pouco explicativo. Por exemplo, na edição de 1995 da terceira divisão, do 41º colocado ao 107º tais times disputaram seis ou menos partidas.

Já para os times da Série A, assim como nas edições a partir de 2008, foram consideradas as posições ordenadas naturalmente no torneio disputado. Para a Série B, somou-se o número de postos disponíveis da Série A mais a colocação de cada equipe na segunda divisão. Cabe ressaltar que o número de participantes na primeira divisão variou muito ao longo do tempo, chegando a um máximo de 32 participantes na edição de 1993. Tal observação também é aplicável para a Série B, onde - a título de exemplo - a edição de 1989 contou com 96 participantes.

Dessa maneira, com a intenção de deixar o *ranking* proposto de forma mais homogênea, definiu-se que a pior posição possível para uma equipe seria a 51ª. Isso ocorre pelo mesmo motivo do que fora argumentado anteriormente sobre não inserção da classificação da Série C na construção da variável para os anos apontados acima. A pequena quantidade de jogos nas edições com muitas equipes (em especial para os piores colocados) não atribui muito valor explicativo para a variável. Por exemplo, na própria edição de 1989 da segunda divisão, apenas 3 pontos separam o 54º colocado (Sobradinho Esporte Clube) e o 90º colocado (ACEC Baraúnas). Portanto, incluir a posição exata desses clubes provavelmente gera mais viés para a relação linear dessa variável do que considerar ambas equipes com o mesmo nível de força.

Depois, para cada confronto, obteve-se a diferença de posição entre o time mandante e o time visitante, obtendo, assim, a variável:

$$\Delta c = c_m - c_v \tag{3.1}$$

Em que Δc é a variável de controle para classificação no campeonato nacional utilizada nas regressões, c_m é a posição do mandante e c_v a posição do visitante na tabela.

Também se definiu a mesma variável defasada em um período:

$$\Delta c_{-1} = c_{m-1} - c_{v-1} \tag{3.2}$$

Em que o subscrito -1 indica a defasagem para a edição anterior. É possível mostrar que, $-50 \leq \Delta c \leq 50$ e, quanto menor seu valor, mais forte o time mandante será em relação ao time visitante. Em outras palavras, o valor da variável quando negativo

significa que o mandante possui maior nível técnico que o visitante. Caso o valor seja positivo, o contrário ocorre. Portanto, quanto maior os valores de Δc e Δc_{-1}, espera-se que menor seja a probabilidade de um mandante vencer um confronto.

3.4.2 *Odds*

Empregando as informações sobre as odds das partidas, foi desenvolvida uma medida de força relativa dos times. Converteu-se, primeiramente, os *odds* em probabilidades: inverteram-se os *odds* de vitória, empate e derrota e normalizou-se a soma desses valores de forma que somassem 1.

No entanto, essa abordagem gera um problema: as probabilidades do primeiro jogo estão condicionadas ao fato de o time jogar tal partida em casa ou não. Dessa maneira, se utilizarmos essa variável estaríamos superestimando a força do time visitante no segundo jogo. O oposto ocorre se utilizarmos as probabilidades da segunda partida, com a agravante de as probabilidades também estarem condicionadas ao efeito de SLHA e, certamente, ao resultado da primeira partida. Tal problema ocorre no caso de as casas de apostas e os apostadores acreditarem na existência de SLHA.

Combinações entre as probabilidades dos jogos de ida e de volta podem resolver parte do problema. É razoável assumir que a média dos *odds* mitigue os efeitos de HA de cada time no confronto, mas ainda é preciso lidar com a endogeneidade gerada por se considerar SLHA nas *odds*. Amez et. al (2019) não consideram esse ponto na medida de força relativa calculada, o que pode eliminar o efeito explicativo da variável de interesse.

Levando em conta os pontos expostos acima, optou-se por adotar uma abordagem por variáveis instrumentais. Abaixo, estão explicitadas a criação da variável de força e do instrumento. Deste ponto em diante do texto, os termos "mandante"e "visitante"farão referência à situação dos times na partida de volta, a menos quando explicitado.

Seja $\Delta f_i = Pm_i - Pv_i$, em que Pm_i indica a probabilidade de vitória do mandante na partida i e Pv_i indica a probabilidade de vitória do visitante em tal partida. Por sua vez, i indica se encontro em questão refere-se à ida ou volta do confronto ($i = \{ida, volta\}$). Adicionalmente, considere $P\mu_i$ a probabilidade de empate.

i) Primeiro, seja a variável de força como definida abaixo:

$$\Delta f = \Delta f_{ida} + \Delta f_{volta} \tag{3.3}$$

Em que $-2 \leq \Delta f \leq 2$. Assim como comentado acima, assume-se que os efeitos de HA para ambos os times são iguais, em média. Portanto, no agregado de duas partidas, ambos os efeitos seriam anulados na amostra devido ao elevado número de observações. Observe, também, que tal medida de força ainda é endógena.

ii) Considere a seguinte regressão linear:

$$\Delta f_{volta} = \beta_0 + \beta_1 \Delta f_{ida} + \beta_2 P\mu_{ida} + \epsilon \tag{3.4}$$

Note que o termo de erro ϵ da equação (3.4) será composto, entre todos os outros fatores, pelo componente relativo a SLHA.

iii) Repare que os valores previstos da equação acima serão exógenos ao efeito de SLHA:

$$\hat{\Gamma_j} = \hat{\beta}_0 + \hat{\beta}_1 \Delta f_{idaj} + \hat{\beta}_2 P\mu_{ida} \tag{3.5}$$

Em que Γ é o instrumento definido para Δf e j é o índice que representa o confronto.

Nota-se que o instrumento cumpre os dois requisitos de um bom instrumento (WOOLDRIDGE, 2010): i) cumpre a restrição de exclusão, isto é, afeta a variável dependente apenas através de Δf e ii) é fortemente correlacionado com a variável endógena. No apêndice, encontra-se a estimação do primeiro estágio. Por fim, perceba que quanto maior Δf, maior o nível técnico do mandante em relação ao visitante.

É possível argumentar que há uma grande vantagem de se utilizar as *odds* como controle. As casas de aposta buscam de forma mais atual e exata possível representar a probabilidade de cada time vencer um jogo, levando em consideração uma grande quantidade de variáveis relevantes, como o desempenho recente dos times, qualidade dos jogadores disponíveis para o jogo, entre outras informações que captam um maior número de explicações comparativamente com a elaboração de um ranking baseado na posição das equipes no campeonato brasileiro, como desenvolvido na subseção anterior. Ademais, por essas mesmas razões, as *odds* estão menos sujeitas ao problema de arbitrariedade dos rankings, em especial quando se utiliza a média das *odds* de diferentes casas de aposta, como no caso dessa dissertação.

Além disso, as *odds* são frequentemente utilizadas na literatura sobre esportes, inclusive no futebol. Por exemplo, Xu (2011) explora as eficiências de se utilizar *odds* como preditores dos resultados das partidas na *Premier League*. O resultado do artigo mostra que as probabilidades das casas de apostas para a temporada 2006-2007 do torneio são previsões eficazes dos resultados das partidas de futebol.

3.5 Modelo

De maneira geral, os artigos que abordam SLHA estimam a existência do efeito usando modelos de resposta binária - *logit, probit* ou modelo de probabilidade linear. Nessa

dissertação, escolheu-se reportar os resultados estimados pelos modelos utilizando *probit*, embora os resultados nas regressões logísticas tenham sido muito semelhantes.

Seja p a probabilidade do mandante vencer o confronto. Considere o probit a seguir:

$$p = \Phi(\alpha + \beta \Delta força) \tag{3.6}$$

Em que $\Phi(.)$ representa a função de densidade acumulada da distribuição normal padrão. Para dois times com o mesmo nível de força ($\Delta força = 0$), existe SLHA se $p > 0.5$, indicando que o mandante do segundo jogo possui mais de 50% de chance avançar embora possua o mesmo nível de habilidade do seu adversário. Isso é equivalente a obter $\alpha > 0$ no modelo estimado. Consequentemente, $\alpha \leq 0$ indica ausência de SLHA.

Claramente, $\Delta força$ refere-se às variáveis de controle criadas na seção 3.4 do texto.

4 Estatísticas Descritivas

4.1 Evidências de *Second Leg Home Advantage*

As estatísticas descritivas apontam para a existência de SLHA na Copa do Brasil. Compreendendo o período completo de realização do torneio, os mandantes do segundo jogo avançaram em 56,43% das oportunidades. Considerando apenas os confrontos vencidos por saldo de gols, as equipes da casa derrotaram seus adversários em 57,1% das vezes. Ambos os resultados são significantes a 0,1%, sobrevivendo tanto ao teste t unilateral quanto ao bicaudal para diferença de média, contra a hipótese nula de 50% . A tabela abaixo resume tais estatísticas:

Tabela 1 – Vencedores dos Confrontos por Critério

	Mandante	Visitante	Total	Mandante %	Visitante %	Variável
1	421	325	746	56.43%	43.57%	Total
2	350	263	613	57.1%	42.9%	Saldo
3	34	35	69	49.28%	50.72%	Gol Fora
4	37	28	65	56.92%	43.08%	Pênaltis

Elaborado pelo autor.

Observa-se que, considerando as partidas decididas pelo critério do gol qualificado fora de casa, aparentemente tal regra não beneficiou qualquer grupo. Por outro lado, há vantagem dos mandantes do segundo jogo nas penalidades. É possível, entretanto, que tal diferença seja explicada pelo desbalanceamento nos anos iniciais da amostra, conforme observado anteriormente. Assim, é possível que as equipes com maior nível técnico tenham disputado as penalidades em seu mando de campo mais frequentemente. No entanto, tais diferenças não são significantes, talvez devido ao baixo número de observações.

Não obstante, a tabela abaixo reporta os mesmos resultados acima, porém considerando a subamostra a partir de 2009, visto que nesse período os dados estão livres de possíveis vieses de chaveamentos não-aleatórios.

Tabela 2 – Vencedores dos Confrontos por Critério - A partir de 2009

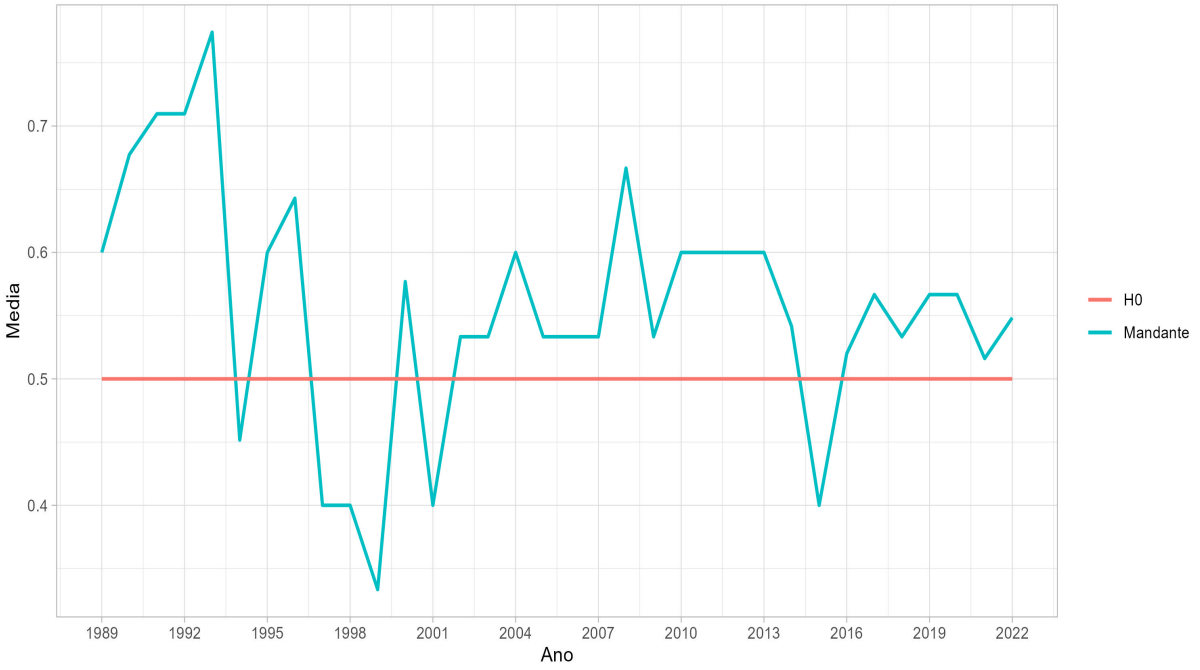
	Mandante	Visitante	Total	Mandante %	Visitante %	Variável
1	185	156	341	54.25%	45.75%	Total
2	155	126	281	55.16%	44.84%	Saldo
3	9	13	22	40.91%	59.09%	Gol Fora
4	21	18	39	53.85%	46.15%	Pênaltis

Elaborado pelo autor.

Adicionalmente, buscou-se avaliar ano a ano a presença deste efeito no torneio. O próximo gráfico (Figura 1) resume com que frequência os times da casa derrotaram seus adversários por edição. SLHA parece um efeito presente ao longo do tempo, porém apresentando diferentes magnitudes. Dos 34 anos de nossa amostra, em 7 os visitantes do jogo de volta avançaram mais vezes em relação aos mandantes: 1994, 1997, 1998, 1999, 2001, 2007 e 2015.

Em uma análise qualitativa mais detalhada dos confrontos nos anos em que não se observou SLHA na Copa do Brasil, notou-se, nas edições dos anos 90, que ocorreram algumas "zebras" e alguns dos chaveamentos aleatoriamente alocou as equipes mais fortes para jogarem a segunda partida fora de casa. Por exemplo, em 1994 o Ceará eliminou Palmeiras e Internacional como visitante do jogo de volta. O time cearense disputava a série B na época, mas avançou até a final da Copa do Brasil no ano. Estes dois confrontos, caso vencidos pelos favoritos à época, por si só levariam à constatação de SLHA naquela edição.

Figura 1 – Proporção vitórias em cada Edição - Mandantes da segunda partida



Elaborado pelo autor.

No triênio de 1997 a 1999, e em 2001, observa-se também a presença de algumas zebras - como a eliminação do Vasco em 1997, que viria a ser campeão brasileiro naquele ano, pelo Atlético Paranaense. Porém, especialmente, destaca-se que o tamanho da amostra entre 1995 e 2012 (exceção do ano 2000) é menor, restrita a 15 confrontos em nossa base para cada um desses anos, devido aos regulamentos da época. Recapitula-se que no período mencionado, as duas primeiras fases eram disputadas sob a regra da eliminação em jogo único caso o visitante vencesse o jogo de ida por dois ou mais gols de vantagem e, portanto,

foram excluídas do conjunto de dados.

Dessa maneira, a amostra fica mais suscetível a grandes variações de resultados ocasionadas pelos chaveamentos aleatórios, no caso de melhores times serem alocados para jogar o segundo jogo fora de casa ou mesmo no próprio caso da existência de zebras. De qualquer forma, não deixa de ser notável a aparente ausência de SHLA nessas edições. Mais à frente, essa questão será novamente abordada, levando em consideração os desbalanceamentos da amostra.

Outrossim, pode-se decompor o efeito por etapa do torneio. Para tal, como as fases foram oficialmente denominadas de maneiras diversas ao longo das edições do torneio, considerou-se a criação de uma nomenclatura relativa à distância da fase para a final da copa. Em outras palavras, tem-se as seguintes fases: final, semifinais, quartas de final, oitavas de final e 1/16 avos de final. Novamente, foram reportados tanto os resultados para a amostra completa quanto para as edições a partir de 2009.

As tabelas a seguir reportam tais resultados. É notável que a presença de SLHA não ocorre durante as finais disputadas, contrastando com o observado nas demais etapas do torneio.

Tabela 3 – Vencedores dos Confrontos por Fase

Fase	Mandante	Visitante	Total	Mandante %	Visitante %
1/16	140	99	239	58.58%	41.42%
Oitavas	149	120	269	55.39%	44.61%
Quartas	74	62	136	54.41%	45.59%
Semi	43	25	68	63.24%	36.76%
Final	15	19	34	44.12%	55.88%

Elaborado pelo autor.

Tabela 4 – Vencedores dos Confrontos por Fase - A partir de 2009

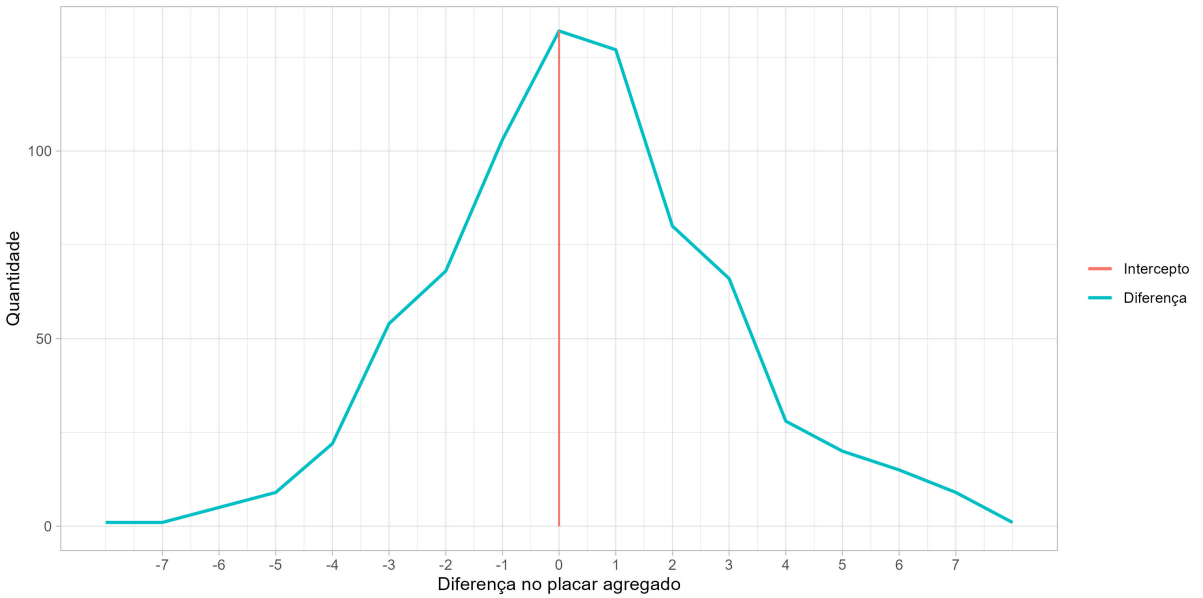
Fase	Mandante	Visitante	Total	Mandante %	Visitante %
1/16	66	66	132	50.00%	50.00%
Oitavas	58	53	111	52.25%	47.75%
Quartas	35	21	56	62.50%	37.50%
Semi	20	8	28	71.43%	28.57%
Final	6	8	14	42.86%	57.14%

Elaborado pelo autor.

Uma forma alternativa de se medir a magnitude do efeito de SLHA é através da diferença na quantidade de gols marcados por cada time no placar agregado das duas partes do confronto. Do total de gols marcados no confronto, pode-se obter a diferença de gols feitos por mandantes e visitantes. Nessa medida alternativa, se a diferença no agregado for positiva, há indícios de SLHA. Caso contrário, inexiste o efeito.

Assim como consta na figura a seguir, é notável que a distribuição da variável tem maior massa à direita do ponto 0, indicando a existência de SLHA na janela de tempo entre 1989 e 2022.

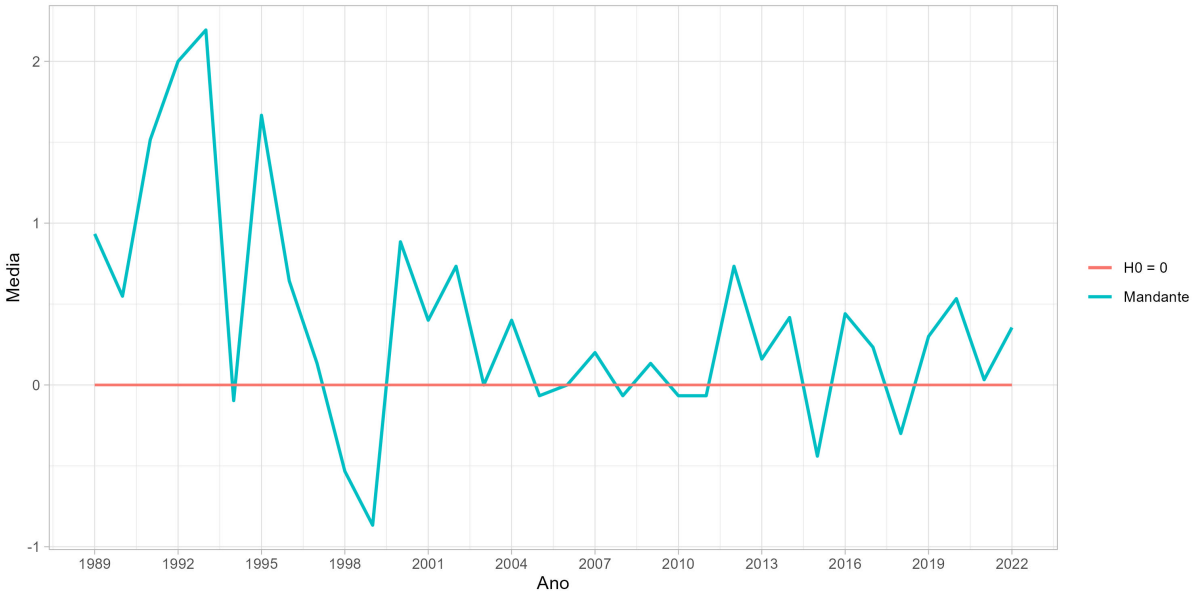
Figura 2 – Distribuição da diferença de gols no agregado - Mandantes da segunda partida



Elaborado pelo autor.

O gráfico abaixo (Figura 3) desagrega, por ano, a média dessa diferença de tentos no agregado. Ressalta-se que os mandantes foram responsáveis por 60,44% dos tentos durante o período analisado. No entanto, em algumas edições os visitantes levam vantagem nessa variável, conforme mostrado a seguir.

Figura 3 – Diferença de gols no agregado em cada edição - Mandantes da segunda partida



Elaborado pelo autor.

A média dessa diferença para o período compreendido é de 0,46, e estatisticamente diferente de 0 a 0,1% de significância, para o teste t realizado. Portanto, há evidência da presença de SLHA no período de nossa amostra. Contudo, destaca-se que, dos 34 anos analisados, tal valor foi negativo em 9 edições, a saber: 1994, 1998, 1999, 2005, 2008, 2010, 2011, 2015 e 2018.

Adicionalmente, antes de adentrar nos resultados das estimações, é importante qualificar os confrontos descritos no conjunto de dados a ser utilizado nas regressões. A próxima tabela caracteriza os confrontos de acordo com a divisão em que ambas as equipes se encontram. Espera-se que, devido ao chaveamento aleatório, exista um número próximo de confrontos de cada grupo, isto é, considerando a divisão do mandante e do visitante.

Tabela 5 – Confrontos: Divisão dos Clubes na amostra

	Mandante X Visitante	Total	% Amostral
1	Série A X Série A	337	45.17%
2	Série A X Série B	111	14.88%
3	Série B X Série A	76	10.19%
4	Série A X Sem Divisão	72	9.65%
5	Sem Divisão X Série A	42	5.63%
6	Série B X Série B	27	3.62%
7	Sem Divisão X Sem Divisão	16	2.14%
8	Série C X Série A	16	2.14%
9	Série B X Sem Divisão	14	1.88%
10	Sem Divisão X Série B	13	1.74%
11	Série A X Série C	7	0.94%
12	Série B X Série C	7	0.94%
13	Série C X Série C	3	0.4%
14	Série C X Série B	3	0.4%
15	Sem Divisão X Série C	1	0.13%
16	Série C X Sem Divisão	1	0.13%

Elaborado pelo autor.

O grupo "Sem Divisão"inclui os times da Série D e os de fato sem divisão entre os anos 2013 a 2022. Entre os anos 1989 e 2012, esse grupo engloba também os times da Série C, pelos motivos já discutidos anteriormente. Adicionalmente, pode-se checar o balanceamento das observações utilizando-se a média da variável de diferença de classificação do time mandante comparativamente com a média dos adversários (definida anteriormente como Δc).

A média amostral dessa variável é de -2,14. Esse resultado indica um leve desbalanceamento na alocação das equipes, com os times da casa do segundo jogo ocupando uma colocação um pouco melhor. Por ano, a média da diferença (Δc) dessas variáveis está exposta no apêndice. Para o recorte da amostra a partir de 2009, a variável possui média de 0,3871, indicando uma diferença ínfima no chaveamento.

A tabela 6 abaixo replica os resultados da tabela anterior, considerando apenas a

subamostra a partir de 2009.

Tabela 6 – Confrontos: Divisão dos Clubes na amostra - A partir de 2009

	Mandante X Visitante	Total	% Amostral
1	Série A X Série A	155	45.45%
2	Série A X Série B	45	13.2%
3	Série B X Série A	36	10.56%
4	Série A X Sem Divisão	23	6.74%
5	Sem Divisão X Série A	21	6.16%
6	Série C X Série A	16	4.69%
7	Série B X Série B	9	2.64%
8	Sem Divisão X Série B	7	2.05%
9	Série A X Série C	7	2.05%
10	Série B X Série C	7	2.05%
11	Série B X Sem Divisão	5	1.47%
12	Série C X Série C	3	0.88%
13	Série C X Série B	3	0.88%
14	Sem Divisão X Sem Divisão	2	0.59%
15	Sem Divisão X Série C	1	0.29%
16	Série C X Sem Divisão	1	0.29%

Elaborado pelo autor.

5 Estimações

Nesse capítulo, encontram-se divididos os resultados das estimações em dois grupos. As primeiras tabelas reportam os resultados obtidos utilizando as observações completas do conjunto de dados. Posteriormente, estão reportados os valores encontrados para as observações a partir de 2009, na intenção de possibilitar a comparação das variáveis de controle de classificação com a variável derivada das *odds* das casas de apostas.

Tabela 7 – Estimações

		Vitória	
Modelo:	(1)	(2)	(3)
	Probit	Probit	Probit
Variáveis			
Constante	0.0986*	0.1303***	0.1109**
	(0.0490)	(0.0488)	(0.0495)
Δc_{-1}	-0.0260***		-0.0126***
	(0.0025)		(0.0043)
Δc		-0.0268***	-0.0163***
		(0.0025)	(0.0044)
Estatísticas			
Observações	746	746	746
Corr^2	0.15018	0.16161	0.16961
Pseudo R^2	0.11903	0.12452	0.13282
BIC	912.93	907.33	905.46

Erros-padrões IID entre parênteses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Nota: Os modelos utilizam as observações desde 1989 As colunas de (1) a (3) representam diversas especificações, incluindo os diferentes controles apresentados. A coluna (1) apresenta a especificação defasada da variável de controle apresentada, conforme discutido anteriormente. A coluna (3) adiciona ambas variáveis.

A tabela 7 acima expõe os resultados selecionados das regressões. Como se observa, em todas as especificações a constante é positiva e significante a pelo menos 10%, indicando a existência de SLHA. Ainda, constata-se que o sinal das variáveis de controle possuem o sinal esperado: quanto maior o nível técnico do mandante em relação ao visitante (menor Δc) mais elevada será a probabilidade de triunfo no confronto. Em termos de interpretação, o interesse está na probabilidade de vitória em casa em um confronto, considerando a força das equipes como iguais ($\Delta c = 0$ e $\Delta c_{-1} = 0$).

Logo, para calcular o efeito estimado de SLHA, deseja-se obter $\hat{p_0} = \Phi(\hat{\alpha})$, onde $\hat{\alpha}$ é o valor estimado da constante nos modelos acima. A tabela abaixo compara tais valores com a média de vitórias do mandante encontrado na amostra - os quais avançaram em

56,57% dos confrontos, entre 1989 e 2022.

Tabela 8 – SLHA predito e diferença para SLHA amostral

Modelo	$\hat{p_0}$	Diferença
(1) Probit	0.5393	-0.0264
(2) Probit	0.5518	-0.0138
(3) Probit	0.5442	-0.0215
Observado	0.5657	

Elaborado pelo autor.

Interessantemente, todas as três especificações acima subestimam o efeito de SLHA em relação ao valor observado na amostra. De certa forma, tal fato é esperado, dado que há um desbalanceamento no chaveamento das equipes assim como abordado acima e, portanto, os controles mitigam esse problema encontrado. Por sua vez, as estimações presentes na tabela 9 reportam os resultados com as variáveis de controle derivadas das *odds* - Δf -, incluindo as estimações com variável instrumental, com o instrumento Γ definido anteriormente. A janela temporal engloba as informações a partir de 2009.

Tabela 9 – Estimações - Modelos com *odds*

	Vitória					
Modelo:	(1) Probit	(2) Probit (IV)	(3) Probit	(4) Probit	(5) Probit	(6) Probit (IV)
Variáveis						
Constante	0.1322*	0.1267*	0.1393*	0.1460**	0.1450*	0.1254*
	(0.0761)	(0.0760)	(0.0732)	(0.0738)	(0.0740)	(0.0761)
Δf	1.296***	1.322***				1.3678***
	(0.1423)	(0.1426)				(0.2384)
Δc_{-1}			-0.0311***		-0.0104	0.0015
			(0.0040)		(0.0078)	(0.0070)
Δc				-0.0323***	-0.0235***	
				(0.0040)	(0.0078)	
Estatísticas						
Observações	340	340	341	341	341	340
$Corr^2$	0.27496		0.19132	0.21860	0.22173	
Pseudo R^2	0.22374		0.15331	0.16967	0.17337	
BIC	375.48		409.53	401.84	405.93	

Erros-padrões IID entre parênteses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
Nota: Os modelos utilizam as observações da base de dados composta por 341 confrontos disputados entre 2009 e 2022. As colunas de (1) a (6) representam diversas especificações do modelo apresentado, incluindo os diferentes controles apresentados. A coluna (1) apresenta a especificação endógena do modelo, conforme discutido anteriormente. A coluna (2) especifica o mesmo controle, porém instrumentalizando a variável Δf. O confronto entre Grêmio e Bahia em 2012 não possuia *odds* disponíveis para a primeira partida

As regressões apresentadas acima corroboram a hipótese de existência de SLHA. Em todos os modelos, a constante é positiva e significante a pelo menos 10%. Outrossim, constata-se que o sinal das variáveis de controle possuem o sinal esperado: quanto maior o nível técnico do mandante em relação ao visitante (maior Δf e menor Δc) mais elevada será a probabilidade de triunfo no confronto. Destaca-se dos resultados acima que, mesmo na especificação endógena do primeiro modelo, rejeita-se a hipótese nula de inexistência do efeito em questão, o que - em outras palavras - indica que as casas de aposta (e, portanto, também os apostadores) subestimam a magnitude de SLHA. Nota-se, ademais, uma visível superioridade do modelo derivado das *odds* em termos de ajuste do modelo, o qual possui o Pseudo R^2 mais elevado e acima 0,2, indicando um bom poder preditivo (MCFADDEN, 1974).

Em consonância com a tabela 8, reportam-se abaixo os efeitos estimados, $\hat{p}_0$, encontrados acima. Para os anos em vigência, os donos da casa no jogo de volta avançaram em 54,55% das oportunidades.

Tabela 10 – Diferença entre SLHA predito x SLHA amostral - Modelos com *odds*

Modelo	$\hat{p_0}$	Diferença
(1) Probit	0.5526	0.0071
(2) Probit (IV)	0.5504	0.0050
(3) Probit	0.5553	0.0099
(4) Probit	0.5580	0.0126
(5) Probit	0.5576	0.0122
(6) Probit (IV)	0.5498	0.0044
Observado	0.5455	

Elaborado pelo autor.

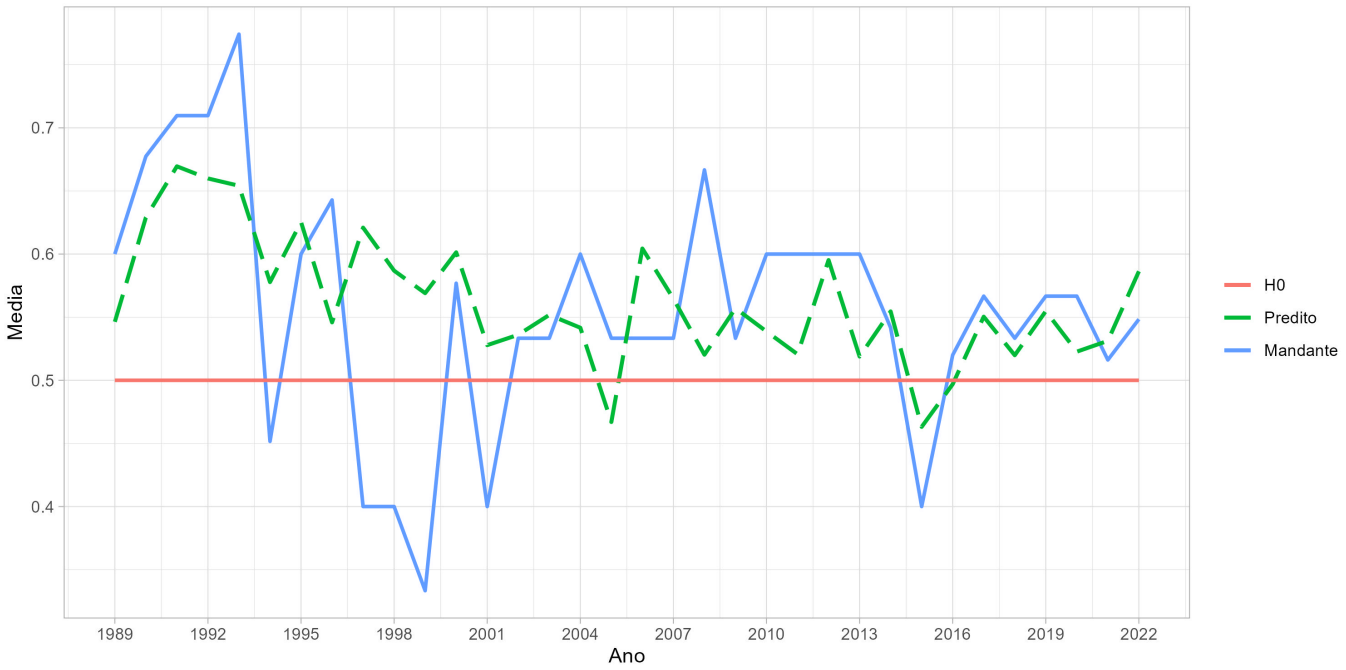
Dos valores reportados imediatamente acima, é notável que os modelos que utilizam o controle Δf apresentam o coeficiente $\hat{p_0}$ numericamente mais próximo do observado na amostra.

Retomando os resultados apresentados na tabela 7, uma das questões importantes a serem respondidas nessa dissertação se refere à evolução do efeito de SLHA ao longo dos anos. Em uma rápida análise da figura 1, além da grande variância em sua magnitude, há um aparente decréscimo no impacto de se disputar a volta de um mata-mata em mando de campo próprio. Como tem-se que as observações são levemente desbalanceadas em especial nas edições iniciais da Copa do Brasil, uma forma de se abordar esse ponto é através dos valores preditos pelas regressões. O gráfico abaixo - figura 4 - compara anualmente os resultados da figura 1 com a média prevista pelo modelo (3) da tabela 7. A escolha dessa especificação foi dada devido à maior capacidade preditiva em termos de Pseudo R^2.

Constata-se da figura 4 que, ao controlar pela diferença de força entre as equipes, não se observa a existência de SLHA em apenas duas edições do torneio. Em outras

palavras, verifica-se que tal efeito é relativamente constante ao longo do tempo na Copa do Brasil, com dois eventuais momentos em que não se pode afirmar sua presença.

Figura 4 – Proporção vitórias em cada edição e valores previstos - Mandantes da segunda partida



Elaborado pelo autor.

Por conseguinte, as evidências empíricas apresentadas aqui revelam que os clubes que possuem o mando de campo na segunda partida usufruem de uma vantagem extra, com as chances de vitória estimadas entre 53,93% e 55,80%, mantendo as demais variáveis constantes. Outrossim, o efeito calculado mostra-se históricamente presente ao longo das edições do torneio.

6 Extensão: Um bom resultado na primeira partida

Essa seção do texto procurou explorar o que acontece na segunda parte do confronto condicional ao resultado da primeira partida. A ideia é que, dada a diferença de gols marcados entre ambas as equipes no primeiro jogo da disputa, se observe a probabilidade de vitória. O modelo básico estimado, portanto, está definido a seguir.

$$p_{|ida} = \Phi(\alpha + \beta_1 \Delta força + \beta_2 \Delta g_{ida}) \tag{6.1}$$

Em que $p_{|ida}$ é a probabilidade de vitória no confronto condicional ao resultado da partida de ida e $\Delta g_{ida} = Gm_{ida} - Gv_{ida}$, onde Gm_{ida} e Gv_{ida} são a quantidade de gols marcados no primeiro jogo pelo mandante e pelo visitante, respectivamente. É imediato perceber que se $\Delta g_{ida} < 0$, o visitante saiu vencedor no primeiro jogo.

É possível argumentar que, além da variável de diferença de placar na primeira partida disputada (Δg_{ida}), a quantidade de gols marcados fora de casa nesse mesmo jogo poderia influenciar na chance de vitória, nas edições em que houve a adoção de tal critério. Assim, defina Fm_{ida} como:

i) $Fm_{ida} = Gm_{ida}$, se o confronto possui o critério de desempate pela regra do gol qualificado;

ii) $Fm_{ida} = 0$, caso contrário.

Dessa maneira, Fm_{ida} captura o efeito da quantidade de gols marcados como visitante na probabilidade de vitória do confronto na equação abaixo:

$$p_{|ida} = \Phi(\alpha + \beta_1 \Delta força + \beta_2 \Delta g_{ida} + \beta_3 Fm_{ida}) \tag{6.2}$$

No entanto, a estimação da especificação (6.2) acima não retornou significância estatística no coeficiente de Fm_{ida}, além de seu valor ser muito próximo de zero. Em outras palavras, mesmo se for assumido que o coeficiente exerce impacto nas chances de vitória, o efeito marginal da variável é extremamente pequeno. Ademais, os coeficientes das demais variáveis permaneceram em magnitude muito próxima aos valores estimados pela equação (6.1). Diante do exposto, optou-se por reportar o primeiro modelo e suas interpretações na presente seção. De qualquer modo, as estimações para a especificação (6.2) encontram-se computadas e discutidas no texto posteriormente[1].

[1] Outras especificações para Fm_{ida} também foram testadas: utilizando seu logaritmo natural, *dummies*

Tabela 11 – Home Advantage Condicional

	Vitória					
Modelo:	(1) Probit	(2) Probit (IV)	(3) Probit	(4) Probit	(5) Probit	(6) Probit (IV)
Variáveis						
Constante	0.3639***	0.3687***	0.3783***	0.3961***	0.3811***	0.3709***
	(0.0895)	(0.0892)	(0.0599)	(0.0595)	(0.0602)	(0.0894)
Δg_{ida}	0.4966***	0.5024***	0.5895***	0.5892***	0.5834***	0.5034***
	(0.0705)	(0.0717)	(0.0498)	(0.0500)	(0.0501)	(0.0718)
Δf	0.9477***	0.9293***				0.8172***
	(0.1579)	(0.1605)				(0.2664)
Δc_{-1}			-0.0175***		-0.0094*	-0.0041
			(0.0029)		(0.0048)	(0.5823)
Δc				-0.0176***	-0.0100***	
				(0.0029)	(0.0048)	
Estatísticas						
Observações	340	340	746	746	746	340
$Corr^2$	0.43861		0.36731	0.37015	0.37430	
Pseudo R^2	0.35990		0.31432	0.31484	0.31861	
BIC	317.49		720.73	716.37	719.17	

Erros-padrões IID entre parênteses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
Nota: Os modelos compreendem as observações da base de dados utilizando as informações disponíveis das variáveis. As colunas de (1) a (6) representam diversas especificações do modelo apresentado, incluindo os diferentes controles apresentados. A coluna (1) apresenta a especificação endógena do modelo, conforme discutido anteriormente. A coluna (2) especifica o mesmo controle, porém instrumentalizando a variável Δf.

Considere $\hat{p}_{|ida=x}$ representando a probabilidade condicional de vitória do mandante no confronto, dado $x = \Delta g_{ida}$ antes da partida de volta. Analogamente à interpretação dos coeficientes das tabelas 7 e 9, pode-se obter a chance de vitória do mandante condicional à diferença de gols na primeira partida através dos coeficientes estimados acima. Considerando a diferença de força igual a 0, em caso de 0 a 0 (ou qualquer outro empate, $\Delta g_{ida} = 0$) tem-se $\hat{p}_{|ida=0} = \Phi(\hat{\alpha})$. Já em caso de vitória do mandante por um gol de diferença na ida, tal probabilidade é dada por: $\hat{p}_{|ida=1} = \Phi(\hat{\alpha} + \hat{\beta}_2)$. A tabela abaixo calcula essas informações, utilizando como base a especificação do modelo (2) acima.

para cada gol marcado fora de casa, uma única *dummy* indicando se ocorreu pelo menos um gol fora de casa e a interação dessas variáveis com Δg_{ida}. Nenhum coeficiente calculado mostrou significância estatística.

Tabela 12 – Chance de vitória condicional ao resultado do primeiro jogo

Δg_{ida}	$\hat{p}_{\|ida=x}$	Efeito(%)	$\hat{p}_{v\|ida=x}$
-5	0.0160		0.9840
-4	0.0504	3.44%	0.9496
-3	0.1275	7.71%	0.8725
-2	0.2624	13.49%	0.7376
-1	0.4468	18.44%	0.5532
0	0.6438	19.70%	0.3562
1	0.8082	16.44%	0.1918
2	0.9152	10.70%	0.0848
3	0.9697	5.45%	0.0303
4	0.9913	2.16%	0.0086
5	0.9980	0.67%	0.0020

Elaborado pelo autor.

A coluna "Efeito(%)"mede o impacto de um gol marcado a mais - tudo o mais constante - na probabilidade de vitória calculada para o mandante. Tal medida está expressa em termos percentuais. Por sua vez, $\hat{p}_{v|ida=x} = 1 - \hat{p}_{|ida=x}$ é a probabilidade de vitória do visitante.

Ainda considerando as equipes com mesmo nível técnico, é notável que, para o mandante, um empate na primeira partida eleva as chances de classificação para uma próxima fase para 64,38%. Uma derrota por um gol de diferença na primeira partida, porém, já reverte esse cenário de vantagem. Em adição, a especificação do modelo (2) incondicional apresentada na tabela 9 do capítulo 5 aponta a presença de SLHA em uma magnitude de 55,04%. Em outras palavras, o resultado do modelo condicional obtido indica também que, em média, os visitantes são derrotados no primeiro jogo por uma diferença de 0.4817 gols:

$$0.5504 = \Phi(0.3687 + \hat{\beta}_2\overline{\Delta g_{ida}})$$

$$\Phi^{-1}(0.5504) = 0.3687 + 0.5024\overline{\Delta g_{ida}}$$

$$\overline{\Delta g_{ida}} = -0.4817$$

Alternativamente, dos resultados acima, pode-se calcular qual a diferença de gols necessária no primeiro trecho disputado para que inexista SLHA. Se, os mandantes fossem derrotados, em média, por 0.7339 gols de diferença, a probabilidade dos visitantes avançarem seria de 50% Qualquer resultado do mandante superior a tal valor na ida, proporciona a visualização do efeito do mando de campo na volta.

Além do exposto acima, considera-se na tabela abaixo as estimações das especificações advindas do modelo populacional (6.2), para fins de comparação com os resultados até agora relatados.

Tabela 13 – Home Advantage Condicional - Especificações alternativas

				Vitória		
Modelo:	(1) Probit	(2) Probit (IV)	(3) Probit	(4) Probit	(5) Probit	(6) Probit (IV)
Variáveis						
Constante	0.3829***	0.3881***	0.3752***	0.3946***	0.3804***	0.3869***
	(0.1096)	(0.1095)	(0.0820)	(0.0821)	(0.0825)	(0.1096)
Δg_{ida}	0.5043***	0.5115***	0.5884***	0.5886***	0.5832***	0.5109***
	(0.0745)	(0.0764)	(0.0533)	(0.0537)	(0.0537)	(0.0763)
Fm_{ida}	-0.0342	-0.0349	0.0038	0.0019	0.0009	-0.0291
	(0.1150)	(0.1141)	(0.0709)	(0.0707)	(0.0710)	(0.1150)
Δf	0.9490***	0.9305***				0.8218***
	(0.1584)	(0.1600)				(0.2633)
Δc_{-1}			-0.0175***		-0.0094**	-0.0040
			(0.0029)		(0.0048)	(0.0075)
Δc				-0.0176***	-0.0100**	
				(0.0029)	(0.0048)	
Estatísticas						
Observações	340	340	746	746	746	340
$Corr^2$	0.43768		0.36737	0.37020	0.37432	
Pseudo R^2	0.36010		0.31432	0.31484	0.31861	
BIC	323.23		726.72	726.19	728.95	

Erros-padrões IID entre parênteses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
Nota: Os modelos compreendem as observações da base de dados utilizando as informações disponíveis das variáveis. As colunas de (1) a (6) representam diversas especificações do modelo apresentado, incluindo os diferentes controles apresentados. A coluna (1) apresenta a especificação endógena do modelo, conforme discutido anteriormente. A coluna (2) especifica o mesmo controle, porém instrumentalizando a variável Δf.

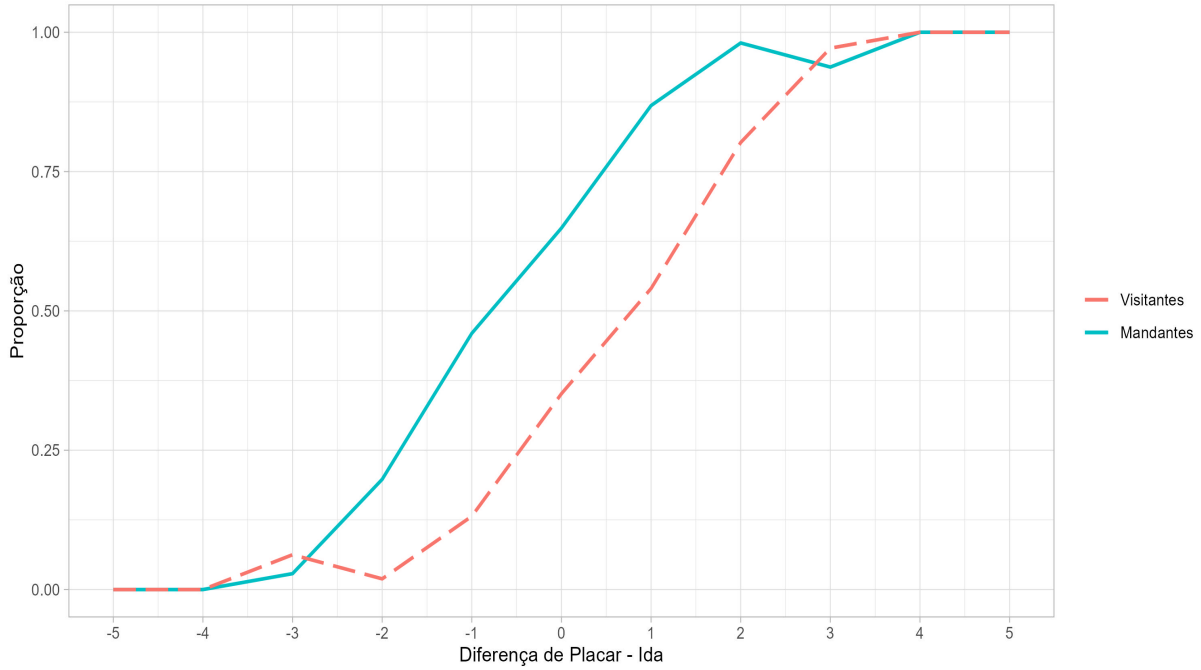
Conforme abordado previamente, em nenhuma estimação o coeficiente de Fm_{ida} é significante. Ademais, tendo em vista o mandante da volta, nota-se que o efeito adicional de um gol marcado fora de casa no primeiro jogo seria muito pequeno (mantendo a diferença de gols entre as equipes constante), ainda que desconsiderada a insignificância estatística. Por exemplo, a diferença entre um empate sem gols e um empate por 1 a 1 no modelo (3), o qual possui o maior coeficiente calculado para Fm_{ida}, é mínima. Um 0 a 0 implicaria em uma probabilidade de vitória do mandante de 64,62%, enquanto o empate por 1 a 1 elevaria essa chance para 64,77%, uma diferença de cerca de 0,15%. Com intuito de facilitar a interpretação, essas análises consideram adversários com mesmo nível de força, porém, a lógica apresentada não muda.

Por sua vez, uma derrota por um gol de diferença na ida, sem gols marcados fora de casa, comparada com uma derrota por 2 a 1, apresenta uma diferença inferior a 0,15%

nas chances de avanço, pelo mesmo modelo (e novamente desconsiderando a significância estatística). Tais resultados encontrados são notáveis em especial por contrariarem o senso comum, sintetizado pela frase "gol fora vale por dois", quando na verdade o efeito da regra na probabilidade condicional de vitória do confronto é estatisticamente nula.

Como maneira de ilustrar o que fora discutido nessa seção, pode-se obter graficamente a frequência de vitória dos mandantes e visitantes após o placar da primeira partida na amostra. A figura 5 abaixo reporta essas informações:

Figura 5 – Proporção amostral das vitórias condicional à diferença de gols na primeira partida



Elaborado pelo autor.

Resumidamente, o presente capítulo expôs uma abordagem empírica para estimar a probabilidade de vitória em um confronto condicional ao resultado do primeiro jogo. Um empate na partida de ida, por exemplo, mostra-se um bom resultado para o mandante e eleva a probabilidade de sucesso para acima de 64%, considerando dois adversários com mesmo nível técnico. Ainda, a regra do gol qualificado não aparenta exercer impactos significativos nessas probabilidades, contrariando a ideia de que "o gol fora vale por dois", frequentemente argumentada pelo senso comum.

7 Estratégias

O capítulo em foco procura iniciar a discussão sobre as causas do efeito encontrado na dissertação e nos demais trabalhos sobre o tema. Aqui, argumenta-se que tais diferenças no desempenho de mandantes e visitantes ocorrem por escolhas estratégicas equivocadas das equipes que decidem fora de casa. A hipótese principal levantada nessa dissertação é a de que, uma vez que no primeiro jogo os times não observam o resutado da segunda partida, as equipes que possuem o mando de campo na ida não aproveitam da melhor maneira sua vantagem de mandante - i.e, seu *Home Advantage* (HA). Dessa maneira, não há maximização do *payoff*, dada tal escolha estratégica.

Romer (2006) aborda uma questão semelhante para futebol americano. O autor encontra evidências empíricas de que as equipes não maximizam o retorno esperado na escolha estratégica do chamado *fourth down*. Resumidamente, o objetivo principal do jogo de futebol americano é chegar à linha de fundo (*end zone*) do adversário. Para isso, o time busca avançar pelo menos dez jardas a cada quatro tentativas. Quando o time consegue as dez jardas, conquista o chamado (*first down*) e possui mais quatro tentativas para ganhar outras dez jardas.

Caso não alcance o objetivo durante as tentativas, a equipe devolve a bola para o adversário no ponto em que a última jogada terminou. Por conta disso, comumente os times escolhem, na última descida, chutar um *punt*, em vez de buscar um *first down*, na intenção de forçar o oponente a começar em uma posição pior. O *punt* basicamente é um chute para frente. No entanto, o autor encontra que essa decisão estratégica frequentemente não maximiza as chances de vencer.

No cenário em questão nessa dissertação, em uma partida de futebol existe uma larga quantidade de estratégias possíveis, que variam ao longo dos minutos disputados em uma partida. Santos (2014) utiliza um PCA (*Principal Component Analysis*) para determinar e resumir as estratégias das equipes em defensiva ou ofensiva em cada tempo de uma partida. O autor avalia os jogos da fase de grupos da *Champions League* das edições entre 1997-1998 até 2009-2010 e sugere que os times adotam estratégias mais defensivas que o ótimo recomendado.

Ressalta-se, porém, que a análise de jogos em fase de grupos diferem totalmente de confrontos eliminatórios em termos de objetivo. Pode-se argumentar, sem grande rigor, que em fases de grupos os competidores procuram uma sequência de resultados perante seus oponentes intencionando alcançar a colocação necessária para passar de fase. Ainda, os times consideram sua restrição de força, *Home Advantage*, e os resultados obtidos até a rodada de disputa. Claramente, os *payoffs* possíveis de cada partida são também avaliados

individualmente pelos clubes (por exemplo, uma vitória concede três pontos à equipe, enquanto o empate premia apenas um ponto e não há pontuação para classificação em caso de derrota).

Por outro lado, em um confronto eliminatório, o objetivo final de ambos os clubes é vencer o confronto de 180 minutos. É factível assumir que as equipes não olham o resultado de cada um dos dois jogos de maneira isolada e, por sua vez, o saldo de gols em cada uma das partidas possui um peso muito maior na decisão estratégica das equipes em relação aos jogos em fase de grupos. Portanto, para estabelecer uma estratégia ótima para cada um dos adversários em confrontos eliminatórios é uma tarefa distinta, que necessita uma formalização do problema de maximização de cada time.

Ainda assim, esse texto buscou levantar evidências de que as escolhas estratégicas dos clubes visitantes não são ótimas e, por isso, ocorre SLHA. A ideia é que, caso as equipes adotassem estratégias iguais quando são mandantes (visitantes) independentemente se no primeiro ou no segundo jogo, tal efeito seria mitigado. Possivelmente, tais times não o fazem justamente por não observarem o resultado da segunda partida e, portanto, não usufruem do efeito de *Home Advantage* em sua totalidade na primeira parte do confronto.

A hipótese de comportamento das equipes pode ser inicialmente testada de modo simples. Primeiramente, supõe-se que existe vantagem de se jogar em mando próprio independentemente do trecho do confronto e, por sua vez, tal vantagem é traduzida em diferenças no comportamento médio da equipe da casa em relação a seu adversário. Em outras palavras HA é o efeito predominante e as equipes adaptam suas estratégias cientes dese fato. Para corroborar essa suposição, além da literatura citada no capítulo 2, identifica-se que ocorre a vantagem de jogar em mando próprio tanto no primeiro quanto no segundo jogo dos confrontos: a tabela 16, no apêndice, expõe a existência de HA na Copa do Brasil, tanto na ida quanto na volta.

Ademais, a tabela 17, no apêndice, apresenta as médias das variáveis *in-match* para os times que possuem o mando de campo e para seus adversários tanto para as partidas de ida quanto para o segundo jogo, colaborando para elucidar como cada equipe joga em cada trecho do confronto, em média.

Além disso, defina $S^m_{ik} = (x^1_{ik}, x^2_{ik}, ..., x^n_{ik})$ o vetor das n variáveis *in-match* observadas em uma partida $i = \{ida, volta\}$ em um confronto k qualquer para uma equipe mandante. Por analogia, seja $S^v_{ik} = (y^1_{ik}, y^2_{ik}, ..., y^n_{ik})$ o vetor de tais variáveis para um visitante. Como as variáveis *in-match* são correlacionadas com a estratégia escolhida pelas equipes, S^m e S^v podem ser interpretados como vetores que representam a estratégia do clube condicional à estratégia de seu oponente.

A variação no comportamento das equipes na situação em que possuem o mando de campo relativamente a quando visitam seu adversário pode ser expressa da seguinte

maneira:

i) Para os mandantes (do jogo de volta):

$$\Delta S_k^m = (x^1_{volta_k}, x^2_{volta_k}, ..., x^n_{volta_k}) - (x^1_{ida_k}, x^2_{ida_k}, ..., x^n_{ida_k}) \tag{7.1}$$

$$\Delta S_k^m = (\Delta x^1_k, \Delta x^2_k, ..., \Delta x^n_k) \tag{7.2}$$

ii) Para os visitantes (do jogo de volta):

$$\Delta S_k^v = (y^1_{ida_k}, y^2_{ida_k}, ..., y^n_{ida_k}) - (y^1_{volta_k}, y^2_{volta_k}, ..., y^n_{volta_k}) \tag{7.3}$$

$$\Delta S_k^v = (\Delta y^1_k, \Delta y^2_k, ..., \Delta y^n_k) \tag{7.4}$$

Logo, dados os k confrontos presentes na amostra, pode-se obter as médias amostrais $\overline{\Delta S^m}$ e $\overline{\Delta S^v}$, onde estão implicitas as médias amostrais da variação das n variáveis *in-match* (denotadas por $\overline{\Delta x^j}$ e $\overline{\Delta y^j}$, respectivamente, com $j = 1, ..., n$). Assim, em média, $\overline{\Delta S^m}$ e $\overline{\Delta S^v}$ representam como as equipes variaram sua forma de jogar em casa em relação a fora de casa, dado a condição de mando de campo do segundo jogo. Tal resultado claramente é condicional à forma dos oponentes atuarem e, consequentemente, não é possível inferir qual foi a escolha estratégica pura.

Caso $\overline{\Delta S^m} = \overline{\Delta S^v}$, as estratégias escolhidas pelas equipes nos confrontos independem do mando da segunda partida e, desse modo, a explicação para a existência de SLHA não está relacionada a uma questão estratégica. Por outro lado, se $\overline{\Delta S^m} \neq \overline{\Delta S^v}$ há evidências de que as escolhas estratégicas contribuem, pelo menos em parte, para o efeito.

Nota-se que, para $\overline{\Delta S^m} \neq \overline{\Delta S^v}$ basta que $\overline{\Delta x^j} \neq \overline{\Delta y^j}$ para algum $j = 1, ..., n$. Todavia, é razoável pensar que quanto mais variáveis diferirem maior o indício de diferenças estratégicas em razão do mando do segundo jogo, bem como maior a contribuição desse fator para a extensão de SLHA. A tabela 14 na próxima página revela os resultados calculados utilizando os dados das edições da Copa do Brasil desde 2015, ordenados por grau de significância estatística.

Das 33 variáveis utilizadas, 11 delas possuem médias estatísticamente diferentes entre os grupos para o teste t realizado, a um nível de significância de 10%. Se forem desconsideradas as variáveis com menos de 30 observações, tem-se 11 de 29 variáveis com diferenças significativas. O resultado é um forte indício de que os clubes visitantes não escolhem sua estratégia de maneira ótima quando possuem o mando na partida de ida.

Tabela 14 – Diferenças de média: Mandantes ($\overline{\Delta S^m}$) e Visitantes ($\overline{\Delta S^m}$)

Variável	Mandantes	Visitantes	Diferença	p_valor	Observações
Posse perdida	6.813	-3.780	10.593	0.000	123
Lançamentos	1.774	-4.726	6.500	0.001	106
Cartões Amarelos	0.039	-0.645	0.684	0.002	152
Passes certos	0.013	0.039	-0.025	0.002	143
Duelos aéreos ganhos	2.077	-0.329	2.406	0.003	143
Duelos ganhos	3.406	-0.552	3.958	0.008	143
Faltas	0.747	-0.772	1.519	0.015	162
Passes	25.385	56.657	-31.273	0.018	143
Dribles certos	-0.012	0.052	-0.064	0.057	106
Gols	0.588	0.358	0.230	0.082	257
Chutões	-3.547	-5.962	2.415	0.088	106
Cruzamentos	6.283	4.557	1.726	0.207	106
Cruzamentos certos	-0.007	0.019	-0.026	0.239	106
Dribles	0.057	1.132	-1.075	0.276	106
Finalizações certas	1.525	1.173	0.352	0.305	162
Impedimento	0.203	0.406	-0.203	0.356	138
Cartões Vermelhos	-0.167	-0.333	0.167	0.375	18
Defesas	-0.722	-0.957	0.235	0.428	162
Finalizações (B)	0.822	1.053	-0.230	0.459	152
Interceptações	-0.387	-0.953	0.566	0.505	106
Finalizações (T)	0.182	0.000	0.182	0.514	22
Contra-ataques	0.381	0.143	0.238	0.524	21
Finalizações (FA)	1.622	1.434	0.189	0.670	143
Lançamentos certos	0.056	0.064	-0.008	0.683	106
Desarmes	0.745	0.396	0.349	0.712	106
Grandes Chances (P)	0.240	0.293	-0.053	0.798	75
Finalizações (CA)	0.182	0.091	0.091	0.817	11
Escanteios	1.615	1.704	-0.089	0.828	169
Finalizações (E)	1.309	1.370	-0.062	0.877	162
Grandes Chances	0.441	0.468	-0.027	0.906	111
Finalizações (DA)	2.392	2.350	0.042	0.931	143
Finalizações	3.572	3.579	-0.006	0.992	159
Posse	0.045	0.045	0.000	1.000	155

Nota: Em parênteses estão a qualificação das variáveis, explicadas a seguir. B: Bloquedo; T: na trave; FA: Fora da Área; P: Perdidas; CA: Contra-ataque; E: Errados; DA: Dentro da Área. Posse, Passes certos, Dribles certos e Cruzamentos certos, são apresentados em porcentagem. As demais variáveis estão expressas em valor total.

Elaborado pelo autor.

Perceba que, se o sinal das variáveis Δx_k^j ou Δy_k^j é positivo, a equipe em questão manifestou maior número de tal estatística durante seu mando de campo do que fora dele. Se o sinal for negativo, a interpretação é contrária.

Notavelmente, a quantidade de gols, o número maior de cartões amarelos, faltas e duelos vencidos, a menor quantidade de passes certos e maior quantidade de lançamentos

e posse de bola perdida que os mandantes apresentam em relação a quando jogam fora de seu mando sugerem uma maior adesão ao risco de tais equipes quando passam a jogar em casa, comparadas com o grupo dos visitantes.

Essas diferenças sinalizam que esses clubes, na condição de mandantes, passam a perder mais a posse de bola, mas também a recuperá-la com mais frequência (observa-se que a média de posse de bola é a mesma) e se transformam em mais violentos. Do contrário, os visitantes são mais agressivos quando jogam fora de casa do que quando possuem o mando na ida. Desse modo, é razoável hipotetizar que os visitantes, caso aderissem a maior risco no jogo de ida em seu mando, talvez pudessem obter um melhor *payoff*. Em outras palavras, quando disputam o primeiro trecho em casa, os visitantes poderiam tentar uma vitória por um placar mais elástico. Ressalta-se, no entanto, a necessidade de maiores investigações acerca deste ponto, conforme abordado no início deste capítulo.

8 Conclusões

A dissertação em foco procurou explorar a hipótese da vantagem de decidir um confronto eliminatório em próprio mando de campo, utilizando as informações da Copa do Brasil. A presente análise identificou a existência de tal efeito, com extensão calculada pelos modelos variando entre 53,93% e 55,80%, levando em conta a diferença na qualidade das equipes. Os valores preditos pelos modelos também indicam que SLHA é observado desde o início do período estudado, não sendo observada em apenas dois anos da amostra. Nota-se, também, que os modelos que utilizam as *odds* para o cálculo da variável de força mostram um melhor ajuste aos dados. Combinando esse controle para o nível técnico das equipes com a aplicação do instrumento escolhido nas estimações com variável instrumental, retornaram-se as estimativas mais próximas do valor observado na amostra.

Além do exposto acima, o texto investigou o que ocorre na segunda partida de um confronto, condicional ao resultado observado no primeiro encontro. Um resultado relevante é que, considerando adversários com o mesmo nível técnico, um empate fora de casa no primeiro jogo implica em uma chance de vitória superior a 64% (no modelo mais conservador) para o time mandante na volta.

Além disso, a diferença de gols entre os clubes no primeiro jogo é importante determinante na probabilidade de vitória do confronto, mas a quantidade de gols marcados fora de casa não se mostra tão relevante em nenhum dos modelos estimados. Esse último ponto é um resultado forte, visto que comumente se assume que um empate por 0 a 0 fora de casa no primeiro jogo é um cenário bem melhor para o visitante da segunda partida que um empate com gols. Mesmo considerando tal afirmativa como verdadeira e ignorando a ausência de significância estatística, o efeito aparente é pequeno e, na melhor das hipóteses, pouco muda a chance de classificação das equipes.

Por fim, a hipótese de diferenças estratégicas entre as equipes como fator explicativo para a existência de SLHA foi testada. Utilizando a média da diferença de cada variável in-match de quando os times jogam em casa subtraída dos valores de quando jogam sem o mando de campo, comparou-se o que acontece nas partidas do ponto de vista do mandante e do visitante. A diferença calculada entre os dois grupos sugere inicialmente que os mandantes possuem maior variação de adesão ao risco quando jogam em casa. Tal hipótese sugere que talvez os visitantes do segundo jogo não aproveitem seu HA quando disputam a primeira partida em casa. No entanto, recomenda-se uma investigação mais aprofundada em relação a este tópico.

Referências

ABAD, C. C. C. et al. Having the second leg at home: advantage in the UEFA champions league knockout phase? *Motriz. Revista de Educação Física*, v. 23, n. 3, p. 1–8, 2017. Citado na página 24.

AMEZ, S. et al. No evidence for second leg home advantage in recent seasons of European soccer cups. *Applied Economics Letters*, v. 27, n. 2, p. 156–160, 2020. Disponível em: <https://EconPapers.repec.org/RePEc:taf:apeclt:v:27:y:2020:i:2:p:156-160>. Citado na página 24.

DOBSON, S.; GODDARD, J. *The Economics of Football*. [S.l.]: Cambridge University Press, 2011. ISBN 0521517141. Citado na página 23.

EUGSTER, M. J.; GERTHEISS, J.; KAISER, S. Having the second leg at home - advantage in the UEFA champions league knockout phase? *Journal of Quantitative Analysis in Sports*, v. 7, n. 1, p. 1–11, 2011. Disponível em: <https://EconPapers.repec.org/RePEc:bpj:jqsprt:v:7:y:2011:i:1:n:6>. Citado 3 vezes nas páginas 24, 25 e 26.

GEENENS, G.; CUDDIHY, T. Non-parametric evidence of second-leg home advantage in European football. *Journal of the Royal Statistical Society Series A*, v. 181, n. 4, p. 1009–1031, 2018. Disponível em: <https://EconPapers.repec.org/RePEc:bla:jorssa:v:181:y:2018:i:4:p:1009-1031>. Citado na página 26.

LIDOR, R. et al. On the advantage of playing the second game at home in the knockout stages of European soccer cup competitions.". *International Journal of Sport and Exercise Psychology*, v. 8, n. 3, p. 312–325, 2010. Citado 3 vezes nas páginas 24, 25 e 27.

MCFADDEN, D. Conditional logit analysis of qualitative choice behavior. *Frontiers in econometrics*, p. 105–142, June 1974. Citado na página 45.

NEAVE, N.; WOLFSON, S. Testosterone, territoriality, and the "home advantage". *Physiol Behav*, v. 78, 2003. Citado na página 23.

PAGE, L.; PAGE, K. The second leg home advantage: evidence from European football cup competitions. *Journal of Sports Sciences*, v. 14, 2007. Citado 3 vezes nas páginas 24, 25 e 26.

POLLARD, R. Home advantage in soccer: a retrospective analysis. *Journal of Sports Sciences*, v. 4, 1986. Citado na página 23.

POLLARD, R.; POLLARD, G. Long-term trends in home advantage in professional team sports in North America and England (1876-2003). *Journal of Sports Sciences*, v. 4, 2005. Citado na página 23.

ROMER, D. Do firms maximize? evidence from professional football. *Journal of Political Economy*, v. 114, n. 2, p. 340–365, 2006. Disponível em: <https://EconPapers.repec.org/RePEc:ucp:jpolec:v:114:y:2006:i:2:p:340-365>. Citado na página 53.

SANTOS, R. Optimal soccer strategies. *Economic Inquiry*, v. 52, n. 1, p. 183–200, 2014. Disponível em: <https://EconPapers.repec.org/RePEc:bla:ecinqu:v:52:y:2014:i:1:p:183-200>. Citado na página 53.

WAQUIL, A. P.; HORTA, E.; MORAES, J. C. Home advantage and away goals rule: An analysis from Brazil cup. *Journal of Sports Analytics*, v. 6, n. 1, p. 13–24, 2020. Citado na página 26.

WOOLDRIDGE, J. *Econometric analysis of cross section and panel data*. Cambridge, Mass: MIT Press, 2010. ISBN 0262232588. Citado na página 35.

XU, J. S. *Online sports gambling: a look into the efficiency of bookmakers' odds as forecasts in the case of English Premier League*. Tese (Doutorado) — University of California, 2011. Citado na página 35.

Apêndices

APÊNDICE A – Tabelas Adicionais

Tabela 15 – Primeiro Estágio com instrumento

	Δf
Modelo:	(1)
Variáveis	
Constante	-0.4258***
	(0.0092)
Γ	2.132***
	(0.0259)
Estatísticas	
Observações	340
R^2	0.95236
R^2 Ajustado	0.95222

Erros-padrões IID entre parênteses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Nota: A regressão utiliza as observações da base de dados composta por 341 confrontos disputados entre 2009 e 2022. O confronto entre Grêmio e Bahia em 2012 não possuia *odds* disponíveis para a primeira partida.

Tabela 16 – Home Advantage: Resultados das equipes em casa na Copa do Brasil

		Ida	Volta
1	Vitória	43.97%	54.83%
2	Empate	29.76%	24.53%
3	Derrota	26.27%	20.64%

Elaborado pelo autor.

Tabela 17 – Variáveis in-match, Médias por trecho do confronto

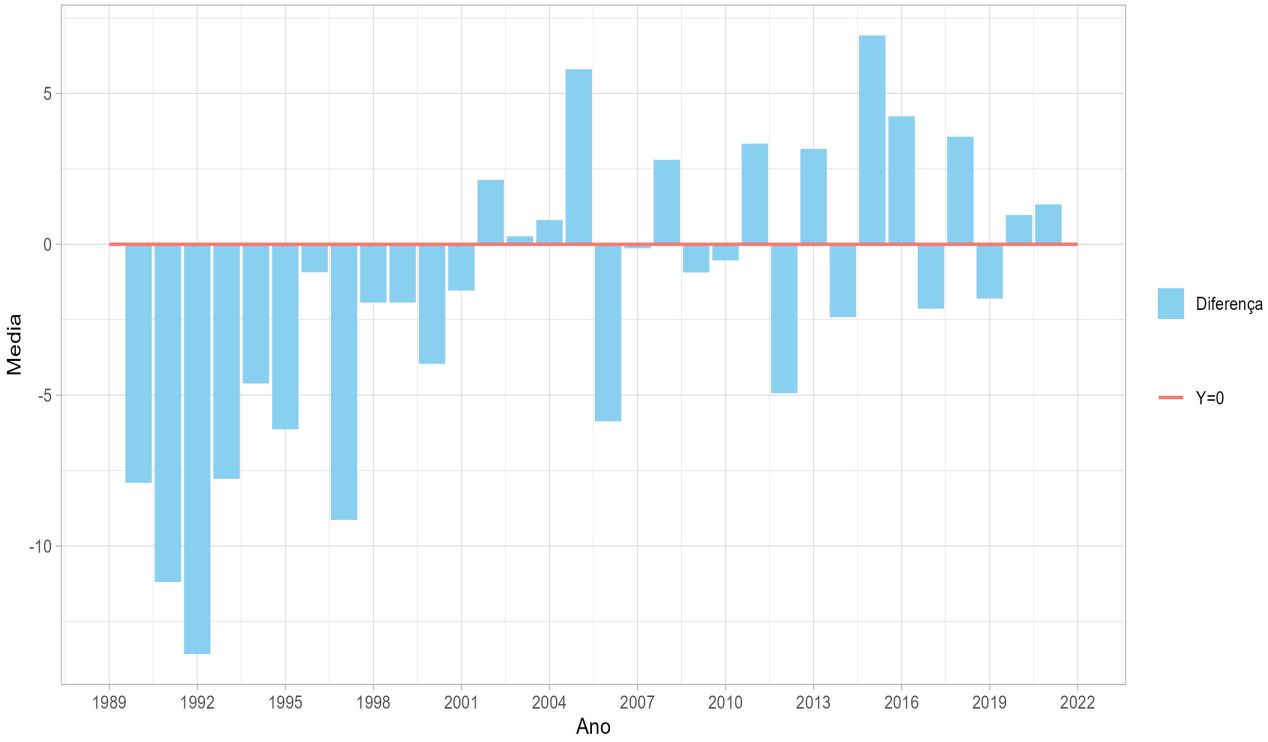
	Variavel	ida_home	ida_away	volta_home	volta_away
1	Posse	0.53	0.47	0.51	0.49
2	Finalizações	14.11	9.92	13.42	10.63
3	Finalizações certas	4.65	3.27	4.75	3.52
4	Finalizações (E)	6.12	4.39	5.69	4.76
5	Finalizações (B)	3.44	2.31	3.11	2.44
6	Escanteios	5.88	4.10	5.70	4.24
7	Impedimentos	1.68	1.70	1.79	1.35
8	Faltas	14.52	14.36	15.09	15.23
9	Cartões amarelos	2.13	2.24	2.25	2.76
10	Grandes Chances	1.65	1.42	1.88	1.13
11	Grandes Chances (P)	1.09	0.94	1.21	0.81
12	Finalizações (DA)	7.63	5.56	7.90	5.36
13	Finalizações (FA)	6.85	4.62	6.23	5.46
14	Defesas	2.45	3.42	2.73	3.38
15	Passes	461.62	405.97	431.01	406.32
16	Passes certos	0.82	0.79	0.80	0.78
17	Lançamentos	53.98	53.96	55.78	58.86
18	Cruzamentos	20.36	14.75	21.30	15.77
19	Dribles	15.14	15.36	15.51	14.07
20	Posse perdida	128.40	125.45	132.37	132.21
21	Duelos ganhos	50.92	50.46	53.86	51.65
22	Duelos aéreos ganhos	14.39	13.83	15.94	14.94
23	Desarmes	15.20	15.09	15.76	14.84
24	Interceptações	10.55	11.83	11.34	11.46
25	Chutões	13.95	19.60	15.95	20.27
26	Finalizações (T)	0.73	0.53	0.64	0.64
27	Contra-ataques	0.78	0.76	1.12	0.85
28	Finalizações (CA)	0.48	0.87	0.78	0.59
29	Gols (CA)	0.38	0.62	0.55	0.55
30	Cartões Vermelhos	0.38	0.42	0.42	0.54
31	Lançamentos certos	0.54	0.47	0.53	0.48
32	Cruzamentos certos	0.22	0.24	0.23	0.20
33	Dribles certos	0.60	0.58	0.57	0.55

Nota: Em parênteses estão a qualificação das variáveis, explicadas a seguir. B: Bloquedo; T: na trave; FA: Fora da Área; P: Perdidas; CA: Contra-ataque; E: Errados; DA: Dentro da Área. Posse, Passes certos, Dribles certos e Cruzamentos certos, são apresentados em porcentagem. As demais variáveis estão expressas em valor total.

Elaborado pelo autor.

APÊNDICE B – Figuras Adicionais

Figura 6 – Média de Δc por edição da Copa do Brasil



Elaborado pelo autor.