Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

Mando de Campo e Gol Qualificado - uma Análise da Vantagem na Copa do Brasil

Alice Paul Waquil

26 de janeiro de 2018

Alice Paul Waquil

Mando de Campo e Gol Qualificado - uma Análise da Vantagem na Copa do Brasil

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientadores:
Prof. Dr. Eduardo de Oliveira Horta
Prof. Dr. Jean Carlo Pech de Moraes

Porto Alegre
Janeiro de 2018

Alice Paul Waquil

Mando de Campo e Gol Qualificado - uma Análise da Vantagem na Copa do Brasil

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pelos Orientadores e pela Banca Examinadora.

Orientadores:

Prof. Dr. Eduardo de Oliveira Horta, UFRGS
Doutor pela Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil

---

Prof. Dr. Jean Carlo Pech de Moraes, UFRGS
Doutor pela University of New Mexico, Albuquerque, Estados Unidos

Banca Examinadora:

Prof. Dr. Álvaro Vigo, UFRGS
Doutor pela Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Prof. Filipe Pereira Gamba, PUCRS
Pós-Graduado pela Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Porto Alegre
Janeiro de 2018

Resumo

No futebol, muitas pessoas defendem que em confrontos de mata-mata – isto é, disputas eliminatórias com jogos de ida e volta – o time que faz o segundo jogo em seu estádio teria uma vantagem. Essa crença vem do fato, amplamente reconhecido na literatura científica, de que o fator local é uma vantagem em uma partida de futebol, bem como em outros esportes. Popularmente, se pensa que fazer o segundo jogo com essa vantagem traria uma maior probabilidade de vitória no resultado final de uma disputa com o sistema de mata-mata.

A proposta deste estudo é verificar a veracidade dessa afirmação utilizando os dados da Copa do Brasil, torneio que segue as características descritas anteriormente. Encontra-se evidências de que a diferença de qualidade entre os times participantes de um confronto é o principal fator que explica a vitória de uma das equipes. Para mensurar esse aspecto, foi criada uma *proxy* que mede as qualidades dos times em todos os anos, com base nos critérios definidos pela CBF.

A amostra é constituída por 1093 confrontos da Copa do Brasil. Estima-se que 36% das disputas terminam empatadas e precisam de um critério para determinar o vencedor. Desses, 51% utilizam a decisão por saldo de gols, 29% o gol qualificado e 20% a disputa de pênaltis. Ao se considerar o campeonato em geral tem-se evidências de que o mandante vence o confronto em aproximadamente 63% das disputas, uma vantagem significativa. Entretanto, nos confrontos que foram decididos pela regra do gol qualificado ou pela disputa de pênaltis, o percentual de classificação do mandante é 20% menor, indicando que esses critérios equiparam as probabilidades dos dois times.

Palavras-Chave: Vantagem em casa, Copa do Brasil, Futebol, Mata-mata, Gol qualificado, Estatística esportiva, Métodos estatísticos aplicados à análise futebolística.

Abstract

In football, many people argue that in knockout matches – that is, knockout competitions with two-leg matches – the team that plays the second game in their stadium would have an advantage. This belief comes from the fact, widely acknowledged in the scientific literature, that the local factor is an advantage in a football match, as well as in other sports. Popularly, it is thought that making the second game with this advantage would bring a greater probability of victory in the final outcome of a knockout match.

The purpose of this study is to verify the veracity of this statement using data from the Brazilian Cup, a tournament that follows the characteristics described above. There is evidence that the quality difference between the teams participating in a showdown is the main factor that explains the victory of one of the teams. To measure this aspect, a proxy was created that measures the qualities of the teams in each year, based on the criteria defined by the CBF.

The sample is constituted by 1093 confrontations of the Brazilian Cup. It is estimated that 36% of the matches end tied and need a criterion to determine the winner. Of these, 51% use the goal balance decision, 29% qualified goal and 20% penalty shootout. When considering the championship in general there is evidence that the home team wins the match in approximately 63% of the matches, a significant advantage. However, in the confrontations that were decided by the away goal rule or the penalty shootout, the home team wins percentage is 20% lower, indicating that these criteria match the odds of both teams.

Keywords: Home advantage, Brazilian Cup, Football, Soccer, Knock-out competitions, two-leg matches, Away goals rule, Sports statistics, Statistical methods applied to soccer analysis.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

O futebol é um esporte de alcance mundial: estima-se que essa modalidade movimente anualmente mais de R$ 550 bilhões (Nobre, 2016), valor maior do que o PIB de vários países (TradingEconomics, 2017). Com tanto dinheiro envolvido, o futebol se aliou à tecnologia e à ciência para, cada vez mais, entregar um produto de qualidade para seus espectadores. Os times investem em medicina avançada para prevenir e combater lesões; métodos estatísticos são usados para, por exemplo, decidir escalações e esquemas táticos. Entretanto, apesar desses avanços, há ainda algumas crenças no futebol que não encontram suporte na literatura científica, muitas vezes pela simples ausência de estudos que busquem avaliar tais questões.

Uma crença antiga, compartilhada por grande parte dos amantes de esportes, é que no futebol, assim como em outros esportes, um time jogar uma partida em casa, isto é, em seu estádio, representa uma vantagem. Muitos estudos, sobre diversos campeonatos do mundo inteiro, evidenciam que essa vantagem de fato existe. O fenômeno tem origem histórica, ocorrendo há mais de 100 anos na Inglaterra (Pollard, 1986). No entanto, está ocorrendo um declínio nessa vantagem, possivelmente devido a mudanças nas regras, como o aumento do número de pontos ganhos por vitória, entre outras que minimizam o posicionamento defensivo do time visitante (Sánchez et al., 2009).

Julga-se que a existência de vantagem associada ao mando de campo já está comprovada e, portanto, os artigos atualmente buscam, em geral, compreender suas possíveis causas. Os principais fatores considerados são: torcida, fadiga de viagens, familiaridade com o local, viés do árbitro, territorialidade, táticas especiais, regras e fatores psicológicos (Pollard, 2008).

Estima-se que a vantagem é de, em média, 61,5% no mundo e 64% na América do Sul, sendo medida como percentual de pontos ganhos em casa sobre o total de pontos ganhos (Pollard, 2006). Entretanto, para clássicos locais, conclui-se que a vantagem de jogar em seu estádio é significativamente menor (Pollard, 1986; Seckin e Pollard, 2008). Em alguns países, como Turquia e Espanha, evidencia-se que não há diferença significativa entre a vantagem de jogar em seus domínios para campeonatos da primeira e da segunda divisão (Seckin e Pollard, 2008; Sánchez et al., 2009). Contudo, no Brasil, constata-se percentuais médios de aproximadamente 65% na Série A e 69% na Série B do Campeonato Brasileiro, indicando que, no país, a vantagem na segunda divisão é significativamente maior do que na primeira (de Almeida et al., 2011). Também são encontrados resultados que indicam um aumento linear significativo entre a primeira e a quarta divisão da Inglaterra, sendo a quarta divisão a que tem a maior vantagem (Pollard, 1986). No entanto, é possível

que isso não ocorra atualmente, uma vez que isso foi concluído por um estudo antigo e, conforme descrito anteriormente, está ocorrendo um declínio na vantagem, contudo não é sabido se há diferenças entre essa diminuição para as diversas divisões ou se o crescimento linear se mantém.

Analisando a existência de discrepâncias regionais na vantagem do mando de campo dentre os países da Europa e da América do Sul, onde as principais ligas estão próximas à média mundial, relata-se que essa variação se deve, principalmente, às localizações geográficas (Pollard, 2006). Conclui-se que a vantagem é maior em locais remotos e etnicamente distintos (Seckin e Pollard, 2008). No Brasil, isso ocorre principalmente nas regiões Norte e Sul por causa das grandes diferenças climáticas e culturais, que também geram um maior sentimento de territorialidade nos times locais (Pollard et al., 2008).

No Campeonato Brasileiro a segunda divisão é mais nacionalizada do que a primeira, e por isso é considerada mais semelhante à Copa do Brasil, já que ambas incluem times de todas as regiões do país. Essa inclusão faz com que as distâncias percorridas e as (muitas vezes) más condições de viagens enfrentadas também sejam mais parecidas entre esses dois campeonatos. O tamanho e a diversidade climática do Brasil fazem com que o efeito das viagens seja, possivelmente, mais importante do que em outros países. Obtêm-se resultados significativos para a Série A indicando que se espera 0,115 gols a mais para o time da casa a cada 1.000 km viajados pelo visitante (Pollard et al., 2008). Como na Série B a vantagem de jogar em casa e as distâncias percorridas são maiores, é razoável supor que seriam esperados ainda mais gols para o time da casa, o mesmo ocorrendo para a Copa do Brasil.

Além disso, a Copa do Brasil, assim como a Série B, inclui times nacionalmente menos expressivos, que geralmente jogam em estádios de pequeno porte, o que pode dar a sensação de maior presença da torcida, consequentemente aumentando a pressão sobre os jogadores, especialmente sobre os visitantes. Outro ponto relevante são as condições do campo, que costumam ser piores, fazendo com que a familiaridade com o local seja um fator ainda mais decisivo na vantagem de jogar em casa.

A crença de que existem vantagens associadas ao mando de campo também ocorre quando se leva em conta confrontos de eliminatórias simples, sistema utilizado na Copa do Brasil, em que são disputados dois jogos, ocorrendo um no estádio de cada time. Neste caso, acredita-se que cada time terá uma vantagem quando jogar na sua casa, mas que o time mandante da segunda partida terá uma vantagem maior no total do confronto. Existem diversos estudos sobre vantagem de jogar em casa, porém poucos têm como objetivo identificar a existência de vantagem em confrontos de eliminatórias simples; em particular inexistem na literatura artigos que abordem[1] os efeitos da regra do gol qualificado.

Para esse sistema, identifica-se uma vantagem significativa: o percentual de classificação é 54,98% para o mandante do segundo jogo, na média de três campeonatos analisados (todos europeus); controlando a análise pela qualidade dos times, a vantagem é de 54,33%. Estima-se, a partir de uma regressão logística, que a probabilidade do mandante do segundo jogo vencer o confronto, ajustada pela qualidade, é de 53,77%, significativamente favorável ao mandante. Além disso, conclui-se que a vantagem permaneceu existente ao longo do tempo, mas tem uma significativa tendência decrescente, assim como foi identificado nas ligas comuns, de pontos corridos (Page e Page, 2007).

[1] Até o conhecimento presente da autora.

Quando confrontos de eliminatórias simples estão empatados em número de pontos, é necessário um critério para definir o vencedor; os três mais usados são:

- *saldo de gols*: vence o time que marcar mais gols;
- *gol qualificado*: vence o time que marcar mais gols ao jogar como visitante;
- *disputa de pênaltis*: vence o time que marcar mais gols na disputa alternada de pênaltis.

A regra do gol qualificado foi criada em 1965, época em que vencer um jogo fora de casa era considerado uma grande façanha. Isso ocorria, principalmente, por causa da condição física dos jogadores: primeiramente porque não havia o mesmo nível de preparo físico que existe atualmente; ademais, as viagens eram muito mais desgastantes, de forma que os visitantes já entravam em campo em uma clara desvantagem. Originalmente a intenção era estimular o time visitante, para que buscasse marcar mais gols, ao invés de manter uma postura totalmente defensiva (Peron, 2017). Contudo, o resultado não foi o esperado: o que aconteceu foi que os times mandantes se tornaram mais receosos, e o jogo não passou a ter dois times ofensivos, como se pretendia. Pelo contrário, as duas equipes passaram a montar estratégias cautelosas, assim reduzindo o dinamismo das partidas. O jornalista Linekher de Andrade opina que, muitas vezes, o segundo jogo acaba por ter um clima de amistoso, visto que é muito difícil um time compensar uma vantagem criada pelos gols qualificados marcados pelo adversário (de Andrade, 2017).

Esse critério de desempate é atualmente utilizado nos principais campeonatos com sistema de eliminatórias no mundo inteiro. Todavia, cada vez mais, a regra do gol qualificado é contestada pelos amantes do esporte, inclusive por membros da FIFA (Federação Internacional de Futebol), como o ex-presidente Joseph Blatter (Uol, 2014). Na Copa do Brasil, na Copa Libertadores da América e na Copa Sul-Americana, essa norma não é mais utilizada na final do campeonato, muitos então levantam indagações, por exemplo, visto que na final esse critério não é aplicado a fim de evitar algum tipo de injustiça, tem sentido permanecer válido no resto da competição? A CBF anunciou que, pela primeira vez, em 2018, após vinte e nove anos de competição e três anos em que a regra não foi válida na final do campeonato, o gol qualificado não será utilizado como critério de desempate em nenhuma fase da Copa do Brasil (CBF, 2017b).

Ao se considerar que, ao término do primeiro jogo, o resultado está fixo (ou seja, o time que jogou como visitante não pode mais alterar o número de gols marcados fora de casa), então no segundo jogo apenas um time pode modificar esse critério. Ou seja, no segundo jogo, ambos os times podem tentar melhorar/mudar os dois primeiros critérios de classificação (pontuação e saldo de gols), mas só o visitante tem controle sobre o terceiro. Por essa razão, considera-se que o gol qualificado se manifesta como vantagem (ou ao menos uma equiparação) para o time visitante do segundo jogo de duas maneiras: minimizando a postura ofensiva do mandante e pela possibilidade de mudar esse critério quando o outro time já não pode mais fazê-lo.

Posto isso, o presente estudo visa investigar se possuir o mando de campo na segunda partida de um confronto mata-mata representa uma vantagem; em particular, auferir de que maneira certas características (como as qualidades dos times participantes e o uso do gol qualificado, por exemplo) influenciam essa suposta vantagem.

1.1 Copa do Brasil

A Copa do Brasil foi criada pela CBF no final da década de 1980 após a diminuição do número de participantes no Campeonato Brasileiro, que fez com que as regiões menos tradicionais (Norte, Centro-Oeste e Nordeste) perdessem sua representatividade naquele campeonato (da Silva, 2016). O objetivo era que clubes de todos os estados da federação ainda tivessem oportunidade de jogar contra os times popularmente considerados "grandes", atraindo público e renda, além de poderem disputar um título nacional e a vaga para a Copa Libertadores da América. Outra meta era a valorização dos campeonatos estaduais, os quais passariam a ser portas de acesso para a nova competição.

Criada nos moldes europeus, a Copa do Brasil teve sua primeira edição em 1989 e vem sendo disputada todos os anos desde então. Embora tenha como objetivo abranger todo o país, em suas 29 realizações, apenas 15 clubes distintos foram campeões, entre os 327 que participaram. A região Sudeste conquistou 20 títulos, 8 foram para o Sul e 1 para o Nordeste. A região Centro-Oeste tem 2 vice-campeonatos e em apenas uma ocasião um time do Norte chegou à semifinal. Portanto, mesmo tendo um campeonato nacionalizado, os principais times, que se concentram nas regiões mais ricas, Sul e Sudeste, permanecem sendo hegemônicos.

Entre 2001 e 2012, os times que participavam da Copa Libertadores da América não competiam na Copa do Brasil devido a conflitos de calendário. Porém, desde 2013 os times que se classificaram para a competição continental passaram a entrar diretamente nas oitavas de final da Copa do Brasil. Antes de 2001 não havia relação entre os dois torneios.

O sistema de disputa adotado ao longo de toda a competição é o de eliminatórias simples, popularmente conhecido como mata-mata, em que os confrontos são definidos agrupando os clubes dois a dois. As disputas acontecem em dois jogos, sendo que cada partida tem um dos times como mandante. Em 1995 foi estabelecido que, nas duas primeiras fases, se o time visitante vencesse o primeiro jogo por uma diferença de 3 gols ou mais, ele estaria imediatamente classificado e não teria necessidade de realizar o segundo jogo. A partir de 1996, essa regra se modificou para permitir a classificação imediata mediante uma diferença de 2 gols ou mais (Wikipédia, 2017a).

Nos confrontos em que acontecem os dois jogos, o time que obtiver mais pontos (vitória representa 3 pontos, empate 1 ponto e derrota 0) é o vencedor e se classifica para a fase seguinte. Em caso de empate no número de pontos obtidos, os critérios utilizados para se determinar qual time obterá a classificação são, respectivamente: saldo de gols; gol qualificado; disputa de pênaltis. A regra do gol qualificado não é válida quando dois clubes da mesma cidade mandam as duas partidas no mesmo estádio, pois o campo é considerado neutro. Nesse caso, se o confronto estiver empatado em número de pontos e saldo de gols, a decisão será disputada nos pênaltis. Desde 2015, o critério do gol qualificado deixou de ser válido na final do campeonato. Em 2018, esse critério não será utilizado em nenhuma fase da Copa do Brasil (CBF, 2017b).

1.2 Definições e Abreviações

De agora em diante, serão utilizadas algumas definições para facilitar a compreensão do texto, são elas:

- IC = intervalo de confiança: intervalo de prováveis estimativas de um parâmetro de interesse, usados para indicar a confiança de uma estimativa pontual; todos os IC aqui calculados utilizam 95% de confiança, portanto cada IC terá 95% de confiança de conter o verdadeiro valor do parâmetro;
- Mata-mata: sistema de eliminatórias simples;
- Jogo: uma partida disputada (90 minutos);
- Confronto: disputa global, inclui os dois jogos (180 minutos);
- Classificação: vencer o confronto;
- Fases finais: a partir dos dezesseis avos de final;
- Empate no confronto: no final do confronto os dois times participantes terminaram com o mesmo número de pontos;
- Visitante: time que jogou o segundo jogo do confronto fora de casa;
- Mandante: time que jogou o segundo jogo do confronto em seu estádio;
- DQ = Diferença de qualidade padronizada: diferença entre as pontuações que representam as qualidades dos times participantes de cada confronto; [2]
- CV = Classificação do visitante: time que jogou o segundo jogo do confronto fora de casa venceu o confronto; [3]
- CM = Classificação do mandante: time que jogou o segundo jogo do confronto em seu estádio venceu o confronto;
- PT = Classificação por pontos: o time que obteve mais pontos venceu o confronto, portanto não precisou utilizar critério de desempate;
- SG = Classificação por saldo: o confronto terminou empatado em número de pontos, então o time que marcou mais gols venceu o confronto;
- GQ = Classificação por gol qualificado: o confronto terminou empatado em número de pontos e no saldo de gols, então o time que marcou mais gols ao jogar no campo adversário venceu o confronto;
- PN = Classificação por pênaltis: o confronto terminou empatado em número de pontos, em saldo de gols e no número de gols qualificados marcados, então o time que marcou mais gols na disputa de pênaltis venceu o confronto.

[2] Esse item trata-se de uma variável contínua, utilizada nas análises posteriores.

[3] Esse e os itens subsequentemente definidos, tratam-se de variáveis dicotômicas, utilizadas nas análises posteriores, indicando a ocorrência do evento descrito por cada item.

2 Objetivos e Hipóteses

2.1 Objetivos

De forma geral, o principal objetivo é analisar, no Brasil, a vantagem de decidir confrontos mata-mata em casa e, especialmente, medir a influência da regra do gol qualificado sobre isso. Para atender essa meta, tem-se o intuito de estimar as seguintes probabilidades:

- $\mathbb{P}$(empate no confronto | DQ)
- $\mathbb{P}$(GQ | DQ)
- $\mathbb{P}$(CM | resultado do primeiro jogo)
- $\mathbb{P}$(CM | DQ, PT)
- $\mathbb{P}$(CM | DQ, SG)
- $\mathbb{P}$(CM | DQ, GQ)
- $\mathbb{P}$(CM | DQ, PN)

Num cenário ideal, teria-se acesso à dados de duas competições em tudo o mais semelhantes, exceto pelo uso do gol qualificado como critério de desempate – uma delas utilizando-o e a outra não. O tipo de inferência que se poderia fazer seria distinto daquele que nosso conjunto de dados ora permite. Neste estudo, portanto, deve-se levar em conta que, na Copa do Brasil, os times participantes estão conscientes do eventual uso da regra e, muitas vezes, consideram isso ao determinar as táticas que serão utilizadas no confronto que será disputado.

Além disso, tem-se como objetivo utilizar os resultados obtidos pelos modelos estimados para predição. Isto é, a partir das probabilidades estimadas, predizer qual seria o resultado de determinado confronto.

2.2 Hipóteses

As hipóteses que serão analisadas no estudo, pensadas com base na revisão bibliográfica e no conhecimento pessoal, por intermédio de crenças populares, são:

- O principal fator determinante para a classificação de um time é a diferença de qualidade entre os participantes do confronto.
- A probabilidade de classificação do mandante é maior do que 0,5, ou seja, o mandante se classifica em mais da metade dos confrontos, representando a vantagem de jogar o segundo jogo de um confronto mata-mata em seu estádio.

- Ao aumentar a diferença de qualidade (quanto melhor o mandante e pior o visitante), as probabilidades de classificação do mandante, dadas as diferenças de qualidade, aumentam, independente do critério de classificação.
- Ao utilizar gol qualificado ou pênaltis a probabilidade de classificação do mandante é aproximadamente 0,5, independente da diferença de qualidade. Isto é, quando a definição do confronto se dá por algum desses dois critérios, os dois times têm iguais probabilidades de classificação.

3 Metodologia

Os dados, contendo os resultados dos confrontos já disputados pela Copa do Brasil, de 1989 a 2017, foram coletados nos sites Wikipédia (2017a) e Bolan@Área (2017), nos quais a informação estava disponível para todos os anos. As análises foram feitas por meio de estatísticas descritivas, testes de hipóteses, regressões não paramétrica e logística utilizando o software R Core Team (2017). Também foram coletados dados do Campeonato Brasileiro das Séries A e B, entre os anos 2012 e 2017, do site TabeladoBrasileirão (2017).

3.1 Confrontos Excluídos

A Copa do Brasil, desde seu início em 1989 até 2017, teve a ocorrência de 1662 confrontos. No entanto, nem todos se enquadram nas características desejadas para este estudo. A Tabela (3.1) apresenta os motivos e os números de confrontos excluídos da análise. Alguns confrontos estão em mais de uma situação (por exemplo, um confronto de fase preliminar que não teve segundo jogo), por isso, a soma das frequências da tabela (663) é maior do que o número de confrontos que foram excluídos (574). Dessa forma, após eliminar todos os casos descritos, o número final de confrontos que serão analisados é 1093.

Para a análise, foram excluídas 155 observações, correspondentes aos anos 1989 a 1993 em que não foi possível determinar adequadamente o índice de classificação dos times, o que será discutido mais detalhadamente na próxima seção. Excluíram-se também 288 observações nas quais o confronto foi decidido em apenas um jogo, pois não é de interesse da presente pesquisa. Optou-se por excluir os 203 confrontos determinados como preliminares pois estes também não apresentam as características em foco na pesquisa (geralmente há grandes discrepâncias de qualidade, além de aproximadamente 45% deles serem definidos em apenas um jogo).

Os 12 confrontos que foram definidos por punições, ou seja, decisões jurídicas que não tem relação com a prática do esporte, também foram excluídos da amostra. Além disso, nos anos 2006 e 2014 houve confrontos entre times que utilizam o mesmo estádio, a saber, Flamengo e Vasco (Maracanã) e Atlético Mineiro e Cruzeiro (Mineirão). O regulamento da competição prevê que nessas situações o mando de campo será considerado neutro e, portanto, a regra do gol qualificado não poderá ser aplicada. Em 2015, 2016 e 2017 também foi determinado que na final do campeonato não seria aplicada a regra do gol qualificado. Uma vez que considera-se que a simples possibilidade de utilizar essa norma modifica as características do confronto, optou-se por excluir esse 5 confrontos na análise.

Tabela 3.1: Confrontos excluídos

Confrontos	n
Sem segundo jogo	288
Preliminares	203
Sem índice	155
Definidos por punições	12
Sem regra do gol qualificado	5

3.2 Formulação do Índice de Qualidade

Uma das hipóteses desta pesquisa é que o principal componente que explica o desfecho de um confronto em duas partidas seja a diferença entre as qualidades dos times participantes do confronto. Foi criada, portanto, com base nos critérios da CBF (CBF, 2014), utilizados em 2016, uma *proxy*, isto é, uma variável que não é observável, que mede a qualidade dos times em cada ano.

Os times recebem uma pontuação de acordo com sua classificação final no Campeonato Brasileiro e na Copa do Brasil. Além disso, recebem uma pontuação bônus como compensação caso tenham sido impedidos de competir a Copa do Brasil por conflito de datas com suas participações nas copas Sul-Americana e Libertadores. Então, a pontuação de um ano é calculada como uma média ponderada dos cinco anos anteriores. Ressalta-se que a pontuação anual contemporânea ao confronto não entra no cômputo do índice.

O número de times participantes no Campeonato Brasileiro costumava variar entre os anos. Por isso, a fim de uniformizar a atribuição de pontos, a convenção da CBF prevê que, a partir do vigésimo-terceiro colocado, todos os times devem receber a mesma pontuação. Dessa forma, se mantém o critério de que todos os participantes de uma série têm a pontuação sempre superior a do primeiro colocado da série imediatamente inferior. Ademais, os anos anteriores a 1994 não possuem o índice completo, uma vez que não havia Copa do Brasil antes de 1989, de forma que a pontuação dos times é mais baixa nos referidos anos. Então, para manter o padrão na pontuação, foi decidido que estes dados não seriam utilizados nas análises, conforme descrito na Seção 3.1.

Nos modelos ajustados, foi utilizada a *proxy* para a qualidade como sendo a padronização da diferença entre as pontuações. Percebe-se que essa variável, além de representar a diferença de qualidade, também capta a variação entre as fases disputadas na competição, já que, geralmente, quanto mais final for a fase, menor é a diferença na qualidade. Da mesma forma, a variação entre os anos é devida, principalmente, a mudanças na qualidade de cada time. A diferença de qualidade, para cada confronto $i = 1, ..., n$, é dada por:

$$d_i = \text{Qualidade do Mandante}_i - \text{Qualidade do Visitante}_i$$

$$z_i = \frac{d_i}{s_d}$$

onde s_d é o desvio padrão amostral da diferença de qualidade.

Isto é, a padronização foi feita dividindo-se o valor da diferença pelo desvio padrão – optou-se por não subtrair a média na padronização para que os valores ficas-

sem centralizados em zero[1]. Dessa forma, valores negativos representam confrontos em que a qualidade do mandante é inferior a do visitante, enquanto valores positivos correspondem a confrontos onde o mandante tem qualidade superior. A Tabela (3.2) e a Figura (3.1) descrevem a diferença de qualidade padronizada. Observa-se que a assimetria é visível no gráfico e confirmada pela mediana igual à 0,67.

Tabela 3.2: Medidas resumo da diferença de qualidade por time classificado

Mínimo	Mediana	Máximo	Média
-2,6260	0,6720	2,5940	0,6466

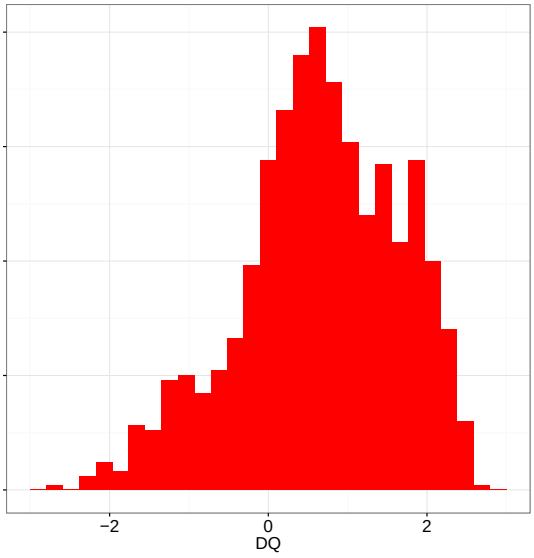


Figura 3.1: Distribuição da diferença de qualidade padronizada

Uma diferença de qualidade igual a um desvio padrão representa uma diferença de 5694,89 pontos no índice de qualidade. Alguns exemplos dessa discrepância, em confrontos que ocorreram na Copa do Brasil, são apresentados na Tabela (3.3). Observa-se que essa situação aconteceu em fases variadas, mas principalmente fases iniciais, incluindo diferenças entre times considerados de "grande" e "médio" porte, bem como entre times "medianos" e "pequenos", que não tem relevante expressão nacional, mas são os principais times de seus respectivos estados e, por isso, acabam sempre participando da Copa do Brasil, o que aumenta sua pontuação no índice.

Tabela 3.3: Exemplos de diferença de qualidade – 1 desvio padrão

Ano	Mandante	Visitante	Diferença	Diferença Padronizada	Fase
2016	Chapecoense(24º)	Princesa de Solimões(93º)	5783	1,0155	64
2015	Bahia(16º)	Luverdense(38º)	5658	0,9935	32
2014	Bahia(17º)	Corinthians(2º)	-5642	-0,9907	16
2014	Figuerense(25º)	Plácido de Castro(104º)	5756	1,0107	64

[1]A média amostral da diferença de qualidade é igual a $0,65$, um valor positivo, indicando que há um ligeiro desequilíbrio no sentido de que times melhores tendem a ter o mando de campo. Esse fato é um indício de que, ao menos em uma parte dos confrontos que compõem nossa amostra, não há aleatorização na atribuição de mando de campo.

3.3 Regressão

Modelos de regressão são utilizados para, a partir de um conjunto de variáveis independentes, explicar o comportamento de uma variável dependente ou predizer uma resposta ou resultado. Essa relação pode ter diversos formatos, entre eles os modelos não paramétrico e logístico, descritos a seguir.

3.3.1 Não Paramétrica

A regressão não paramétrica é adequada quando não há conhecimento *a priori* a respeito da forma da função que será estimada. Isso é, sabe-se que as variáveis x e y são relacionadas através de uma função f, de forma que $y_i := f(x_i) + \epsilon_i, \quad i \in \{1, \ldots, n\}$, porém não se tem informação sobre a natureza dessa função (Aitchison e Aitken, 1976).

Além do objetivo principal, estimar probabilidades de classificação do mandante, tem-se interesse em estimar as probabilidades de um confronto terminar empatado e de utilizar o critério de gol qualificado, visto que pretende-se estudar a influência dessa regra na vantagem de decidir o confronto em casa. As duas regressões consideradas têm como respostas variáveis dicotômicas: a primeira assume valor 1 caso o confronto tenha terminado empatado, enquanto para segunda o *sucesso* acontece quando a disputa foi decidida pelo gol qualificado. Ambos os modelos têm apenas uma variável explicativa (x), que representa a diferença de qualidade entre os times participantes de cada confronto. Para o ajuste, nos dois casos, foram utilizados modelos de regressão não paramétrica pois não se tem conhecimento das funções que modelam as relações entre as variáveis. Computacionalmente, foi utilizado o pacote `np` – ver Hayfield e Racine (2008).

3.3.2 Logística

Modelos de regressão logística são utilizados com o objetivo de relacionar a probabilidade de determinado evento acontecer ($\pi(x)$) com um conjunto de variáveis independentes que explicam esta ocorrência. Nesses modelos, a variável resposta (y) indica a ocorrência, ou não, do evento. O modelo estima a média de y, condicionada aos valores x, que representa a probabilidade de ocorrência do evento (PortalAction, 2017d). Considerando apenas uma variável explicativa (x), o modelo é caracterizado pela seguinte equação:

$$\mathbb{P}(y_i = 1|x_i) = \pi(x_i) = \mathbb{E}(y_i|x_i) := \frac{\exp(g(x_i))}{1 + \exp(g(x_i))}, \quad i \in \{1, \ldots, n\}$$

onde a função de ligação logito se dá por:

$$g(x_i) := \ln\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = \beta_0 + \beta_1 x_{1i}, \quad i \in \{1, \ldots, n\}$$

Neste estudo, a variável resposta é a classificação do mandante, uma variável dicotômica, assumindo valor 1 se o time mandante obteve a classificação (*sucesso*) e 0 caso contrário. Dessa forma, optou-se por utilizar modelos de regressão logística, que são adequados para descrever o tipo de problema abordado pois fazem um ajuste de modo a estimar a probabilidade de *sucesso* da resposta com base nas variáveis explicativas.

3.4 Independência

A independência entre as observações é uma característica importante porque a estimação da regressão logística é feita pelo método de máxima verossimilhança, o qual tem como base a suposição de independência. No caso da Copa do Brasil, os times participantes se repetem ao longo dos anos e, mais especificamente, se repetem ao longo das fases disputadas em um mesmo ano, à medida que vão avançando na competição, de forma que se poderia pensar que existe correlação entre determinadas características. Entretanto, no presente estudo as unidades observacionais são os confrontos disputados, e não os particulares participantes de cada confronto. Ademais, a variável resposta aqui considerada é a classificação, ou não, do time mandante. Entende-se que as características que são correlacionadas entre os confrontos, como os times participantes, não acarretam em problemas pois não são utilizadas nos modelos ajustados. Mesmo que o time se repete, a diferença de qualidade muda nos confrontos que ele participa e é isso que de fato influencia na classificação.

Considera-se que a independência entre as respostas está garantida nos confrontos em que ocorreu sorteio do mando de campo, pois a probabilidade de classificação está ligada ao time ser, ou não, o mandante. Então, mesmo que se tenha informação prévia de quem são os participantes do confronto não se sabe qual será o visitante e, portanto, não há indicativos sobre a probabilidade da resposta.

A fim de certificar-se de que há independência, via o argumento do sorteio dos confrontos e dos mandos de campo, buscou-se os regulamentos da Copa do Brasil, que são modificados a cada ano. No entanto, os regulamentos são de difícil acesso e por isso não foi possível utilizar essa argumentação. Foram encontrados apenas os arquivos dos últimos quatro anos, disponibilizados por notícias antigas do site da CBF (CBF, 2017a). Além disso, obteve-se algumas informações pontuais, na imprensa, sobre os regulamentos mais antigos.

Mesmo sem a confirmação de quais confrontos foram sorteados, optou-se por não excluir nenhum confronto da análise por esse motivo. Acredita-se que não existam problemas com relação a independência, uma vez que as unidades observacionais são os confrontos. Para cada ano, dentro de cada fase, os confrontos são independentes, pois não há nenhuma relação entre as disputas. Também não há dependência entre os diferentes anos, uma vez que os confrontos são diferentes, com times diferentes, em contextos diferentes.

Resta considerar confrontos em um mesmo ano, mas em fases diferentes da competição. Nesse caso, de fato é plausível que haja correlação entre os desfechos de dois ou mais confrontos de um mesmo time em fases distintas. Por exemplo, a informação de que certo time já venceu um confronto como visitante certamente altera a probabilidade de que esse mesmo time venha a vencer um novo confronto, em uma fase posterior, no qual ele joga como visitante: há indícios de que esse time tem competência para classificar-se quando não tem o mando de campo. Porém, em cada fase, novos confrontos são formados, muitas vezes por sorteio, de forma que os mandos de campo são em grande parte aleatorizados. Além disso, a qualidade de cada time se mantém, mas as diferenças de qualidade entre os participantes são outras, de modo que as características no novo confronto, na pior das hipóteses, têm dependência fraca das características do confronto anterior, não sendo suficiente para violar as hipóteses que legitimam os procedimentos de estimação ora adotados.

Por fim, optou-se por utilizar a regressão logística, apesar do conhecimento de que existem limitações intransponíveis por causa da possível existência de correlação entre as unidades observacionais, ainda que as respostas sejam independentes. Uma alternativa é o uso de modelos mistos, em que seria utilizado um termo aleatório para o ano dos confrontos, de forma que a correlação entre confrontos de um mesmo time seria controlada pelo ano da competição. Contudo, esse modelo é mais complexo e levou a resultados semelhantes aos da regressão logística, portanto optou-se pelo modelo mais simples, afim de facilitar a compreensão para o público alvo do estudo, que são os amantes do futebol.

3.5 Avaliação do Ajuste do Modelo

3.5.1 Testes de Ajuste

Após a estimação dos parâmetros do modelo é necessário averiguar a qualidade do ajuste do modelo. Para tanto, são aplicados testes estatísticos que testam a hipótese nula de que o modelo está bem ajustado.

No contexto de modelos de regressão logística, o principal teste de adequação é o de Hosmer e Lemeshow. Este teste faz a avaliação através das diferenças entre as probabilidades ajustadas e observadas. Os valores ajustados são ordenados e então separados em grupos de tamanho aproximadamente igual; os autores do teste propõem o uso de 10 grupos. Porém, quando as frequências esperadas são pequenas para alguns dos grupos é sugerido que se utilize um número menor, respeitando as restrições de que o número de grupos deve ser maior do que três e também maior do que o número de covariáveis do modelo mais um (Hosmer e Lemeshow, 2000).

Apesar de ser amplamente utilizado, o teste de Hosmer e Lemeshow possui um problema de robustez: em aplicações, não raro o nível descritivo amostral do teste pode apresentar diferenças drásticas dependendo do número de grupos utilizado, sendo que a escolha do número de grupos é arbitrária. Outra conhecida adversidade desse teste é que não há sensibilidade à interações e não-linearidade (Allison, 2014a).

Como alternativa, foram propostos diversos testes na literatura. Entre eles, alguns evitam a necessidade de agrupamento, porém ainda não há consenso no uso de apenas um desses testes. Entre os principais estão o Teste Padronizado de Pearson e o Teste de Stukel (Allison, 2014c). Foi demostrado que o teste padronizado tem mais poder do que o teste de Hosmer e Lemeshow para detectar interações e problemas de linearidade. Por outro lado, o teste de Stukel tem menos poder para detectar desvios de linearidade, mas é melhor para constatar interações e desvios da função logística (Hosmer e Hjort, 2002). Por isso, recomenda-se que os dois testes sejam utilizados. A seguir tem-se uma breve descrição dos testes ora mencionados, computacionalmente calculados com a sintaxe disponibilizada por Loughin e Bilder (2013).

Teste Padronizado de Pearson

Proposto por Osius e Rojek, o teste utiliza uma estatística que segue distribuição assintoticamente normal com média e desvio padrão derivados pelos autores (Osius,

1994), calculada como:

$$X^2 := \sum_i \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1 - \hat{\pi}_i)}, \quad i \in \{1, \ldots, n\}$$

onde y_i é a variável resposta admitindo valor 0 ou 1, enquanto $\hat{\pi}_i \in [0, 1]$ é a probabilidade estimada pelo modelo.

Teste de Stukel

Stukel propõe uma generalização da regressão logística com dois parâmetros adicionais, que permitem identificar problemas de assimetria ou desvios na curva logística ao se aproximar de 0 ou 1. Então, o teste verifica a hipótese nula de que os coeficientes desses dois novos parâmetros são iguais a zero, indicando que o modelo está bem ajustado (Allison, 2014b).

Deviance

As observações de um banco de dados podem ser ajustadas por um modelo nulo, que inclui apenas o intercepto, tal que o valor ajustado é sempre igual à média das observações da variável resposta. Em contrapartida, pode-se ajustar um modelo saturado, o mais geral possível, com parâmetros perfeitamente ajustados, pois inclui um parâmetro para cada observação. Qualquer modelo entre esses dois extremos, que são de pouca utilidade prática, é chamado de modelo proposto. Então, os modelos nulo e saturado servem como base de comparação em relação aos modelos propostos. A *deviance* é uma medida da distância entre os modelos saturado e proposto, calculada como a razão de verossimilhança entre ambos. Esse valor é sempre positivo: quanto menor, melhor é o ajuste do modelo proposto.

Pode-se usar a *deviance* do modelo para testar a qualidade do ajuste do modelo, sob certas condições, esse valor segue distribuição assintótica χ^2 com $n - p$ graus de liberdade. Porém, para evitar problemas com os resultados assintóticos, é possível utilizar simulações. A simulação é feita pelo método *bootstrap* de re-amostragem, que consiste em simular re-amostras a partir do modelo original. Em cada simulação, é obtida uma estimativa de *deviance* e então, o nível descritivo amostral é calculado como a proporção das *deviances* que excedem o valor calculado na amostra original (Paula, 2013).

3.5.2 Curva ROC

Para avaliar o ajuste, além dos testes descritos anteriormente, também são utilizadas medidas que avaliam a capacidade preditiva do modelo.

Como mencionado anteriormente, na Seção 3.3.2, na técnica de regressão logística a variável resposta é dicotômica, ou seja assume valor 0 ou 1. No entanto, o modelo estima a *probabilidade* desta variável ser um *sucesso*, e portanto os valores ajustados (preditos) pelo modelo podem assumir qualquer valor entre 0 e 1. Nesse sentido, é usual definir um ponto de corte $\alpha \in (0, 1)$ que determina a partir de qual probabilidade predita o classificador $\widehat{y}_i^{\alpha}$ será definido como *sucesso*. Isto é, após o ajuste do modelo, que estima $\mathbb{P}(y_i = 1|x_i)$, são calculados os valores preditos para

cada observação, da seguinte forma:

$$\widehat{y}_i^{\alpha} := \begin{cases} 1, & \widehat{\mathbb{P}}(y_i = 1|x_i) > \alpha; \\ 0, & \text{caso contrário.} \end{cases}, \qquad \alpha \in (0,1), \quad i \in \{1, \ldots, n\}$$

Além disso, para cada $\alpha \in (0,1)$, a sensibilidade, que mede a capacidade de classificar corretamente os *sucessos*, é definida por:

$$s^{\alpha} := \mathbb{P}(\widehat{y}_i^{\alpha} = 1 \,|\, y_i = 1),$$

enquanto a especificidade, que mede a capacidade de classificar corretamente os *fracassos*, é calculada como:

$$e^{\alpha} := \mathbb{P}(\widehat{y}_i^{\alpha} = 0 \,|\, y_i = 0)$$

Busca-se, então, o valor de α que maximiza as duas medidas, sensibilidade e especificidade, conjuntamente, a fim de obter a maior capacidade possível para o modelo final. Também costuma-se calcular e encontrar o α que maximiza a acurácia do modelo, onde a acurácia mensura a probabilidade incondicional de classificar corretamente (tanto *sucessos* quanto *fracassos*).

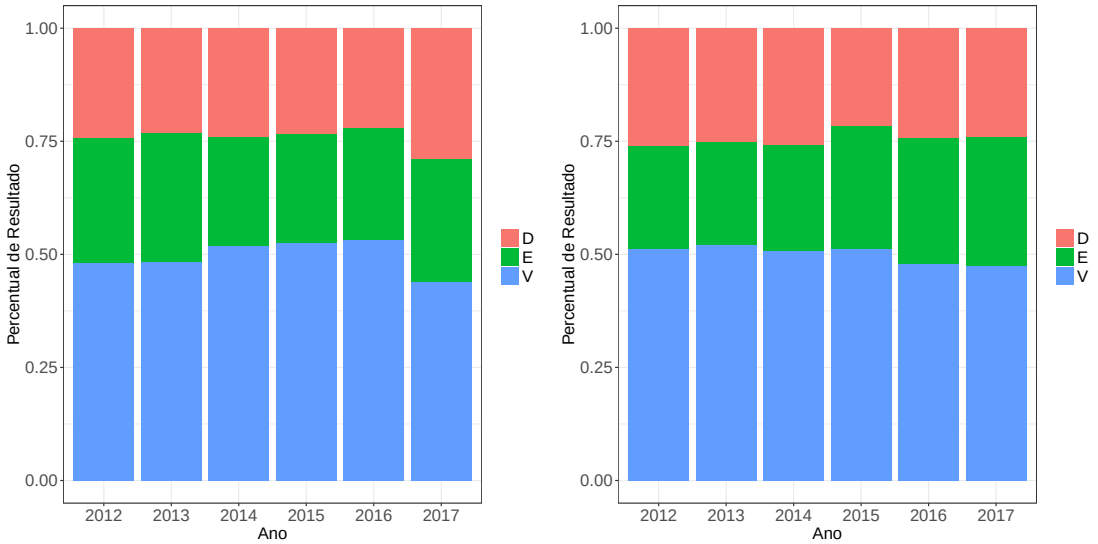
Outra medida de capacidade preditiva é a chamada AUC (do acrônimo em inglês *área abaixo da curva ROC*). A curva ROC (do acrônimo em inglês *Receiver Operating Characteristic Curve*) plota a sensibilidade contra um menos a especificidade, para todos os possíveis pontos de corte α; a área abaixo dessa curva é uma estimativa da capacidade preditiva do modelo ajustado (DeLon et al., 1988).

4 Campeonato Brasileiro

Com a finalidade de ilustrar como a Copa do Brasil, objeto de estudo do presente trabalho, se compara ao Campeonato Brasileiro no que diz respeito à vantagem associada ao mando de campo, são apresentadas nessa seção algumas estatísticas relacionadas a esses campeonatos. Ressalta-se que as duas competições têm características muito distintas, a começar pelo sistema de disputa, por isso, a comparação feita nesse capítulo permite apenas uma compreensão melhor desses torneios e suas diferenças, no entanto eles não devem ser comparados como iguais.

Cada partida de futebol tem três possíveis resultados: empate, vitória ou derrota do time que joga em casa. As proporções em que aconteceram esses possíveis resultados, para o time mandante, em cada ano, no período entre 2012 e 2017, nas séries A e B do Campeonato Brasileiro, são apresentadas na Figura (4.1).

Observa-se que o percentual de vitórias se mantém em torno de 50%, enquanto empates e derrotas ocorrem em cerca de 25% dos casos. Em oposição ao que foi concluído por de Almeida et al. (2011), que calcula a vantagem como percentual de pontos ganhos em casa, não há evidências de que a diferença entre os percentuais médios de vitórias nas duas divisões do campeonato seja significativa (p-valor = 0,7464), 49,69% e 50,13% respectivamente para as séries A e B.



(a) Série A

(b) Série B

Figura 4.1: Resultados do time mandante por ano no Campeonato Brasileiro

Então, agrupando as duas divisões do Campeonato Brasileiro, o percentual médio de vitórias do time mandante é 49,91%, com IC = (0,4701; 0,5281). Portanto, metade das partidas disputados nessa competição são vencidos pela equipe que joga em casa. A outra parte das partidas tem resultados igualmente divididos entre empates e derrotas (p-valor = 0,1283).

A fim de comparar com o Campeonato Brasileiro, o mesmo gráfico foi feito para a Copa do Brasil. A Figura (4.2) dispõe os resultados obtidos pelo mandante de cada jogo na Copa do Brasil, considerando cada jogo separadamente, não como um confronto, diferente do que será feito no resto desse trabalho. A variação ao longo de todo o período aparenta ser maior, porém se comparados apenas os anos a partir de 2012 a Copa do Brasil parece manter um padrão, bem como o Campeonato Brasileiro. Visualmente, os resultados parecem ser levemente mais favoráveis ao mandante do que no Campeonato Brasileiro. O percentual médio de vitórias é de 52,70%, seguido de 27,36% de empates e 19,95% de derrotas. Possivelmente, essa diferença entre os campeonatos se deve às fases iniciais da Copa do Brasil, nas quais as diferenças de qualidade costumam ser mais discrepantes, pois incluem times de todas as divisões.

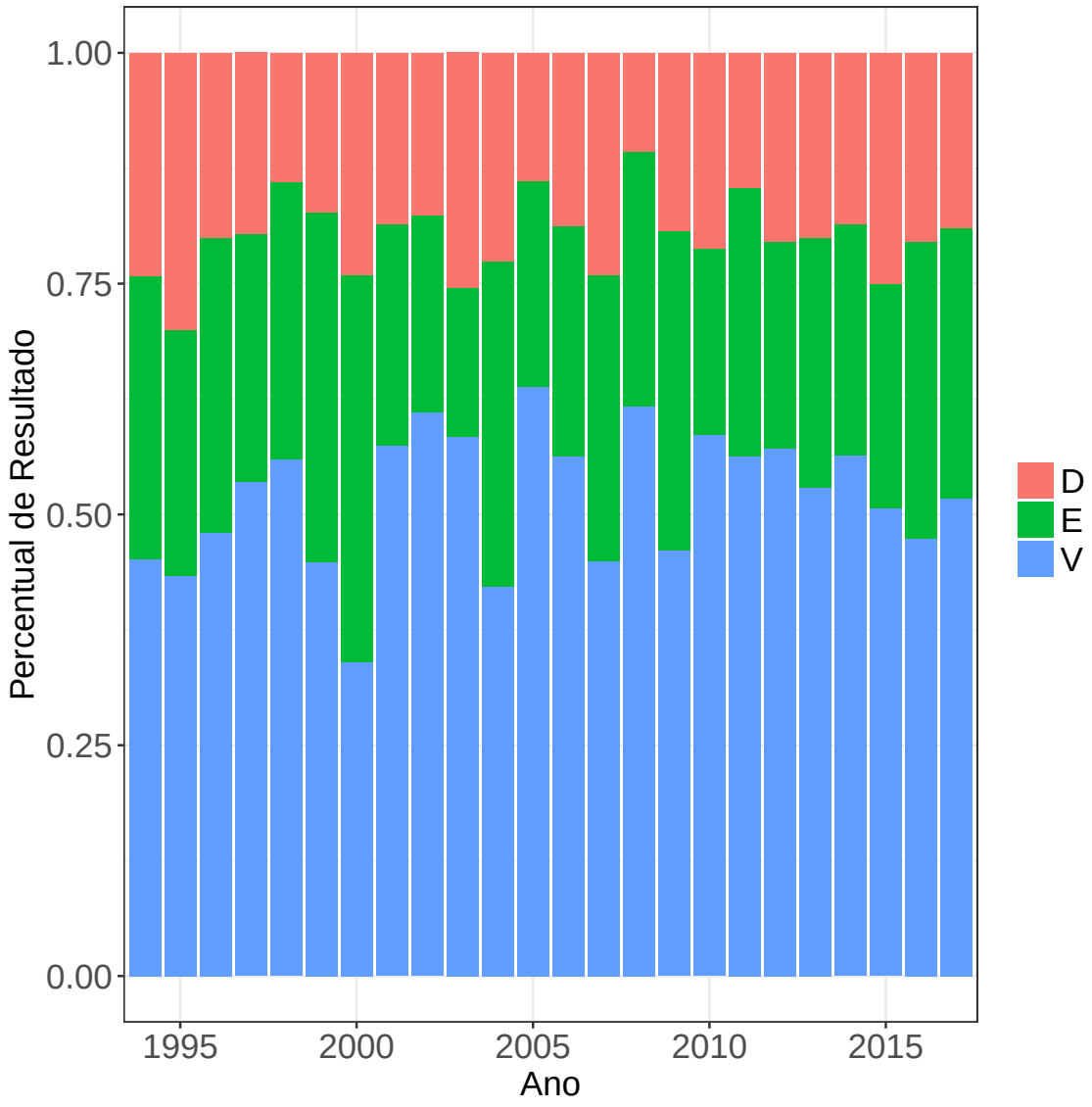


Figura 4.2: Resultados do time mandante por ano na Copa do Brasil

Comparando-se os dois torneios, os percentuais de empate, 27,36% na Copa do Brasil e 25,73% no Campeonato Brasileiro, não são significativamente diferentes (p-valor = 0,1573). Entretanto, os percentuais de vitórias e derrotas têm diferenças significativas entre as duas competições. O percentual de vitórias do mandante na Copa do Brasil é significativamente maior do que no Campeonato Brasileiro (p-valor = 0,0319). Em contrapartida, a porcentagem de derrotas é significativamente menor (p-valor $<$ 0,0001). Conclui-se então, que a vantagem do time mandante, considerando cada partida isoladamente, é maior na Copa do Brasil, uma vez que essa equipe vence mais e perde menos, além de manter o percentual de empates inalterado.

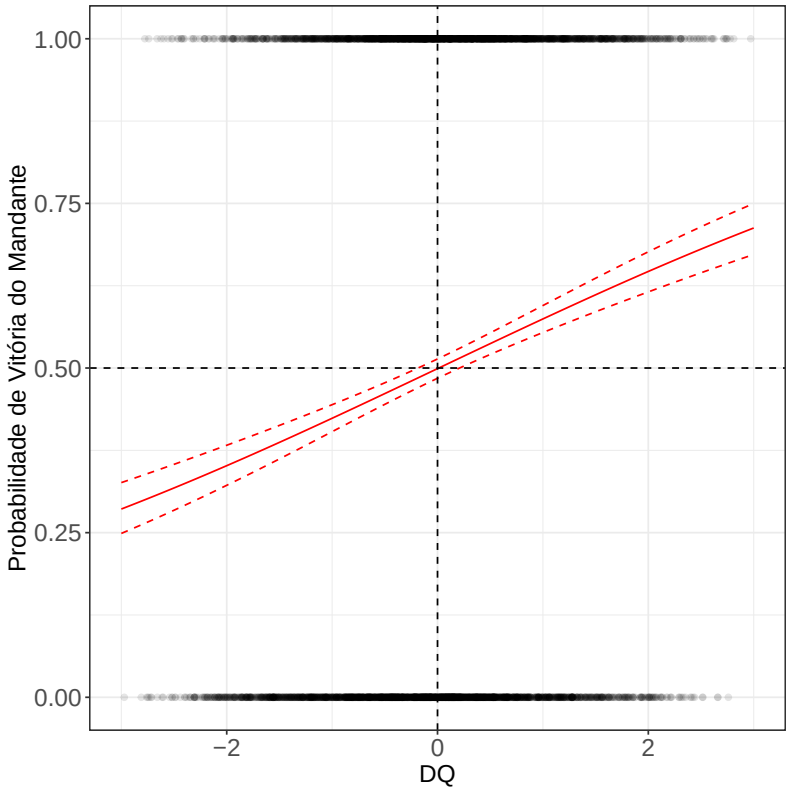
Além de confrontar os resultados obtidos em cada competição, é de interesse comparar as probabilidades do mandante obter a classificação na Copa do Brasil com as probabilidades de vencer um jogo no Campeonato Brasileiro. Para tanto, foram ajustadas duas regressões logísticas, uma para cada torneio, ambas com apenas a diferenças de qualidade como variável explicativa. Deve-se ressaltar que na Copa

do Brasil a resposta é a classificação, ou não, do mandante, ou seja os empates já estão definidos, alguns como *sucesso* outros como *fracasso*. Por outro lado, no Campeonato Brasileiro os empates foram todos classificados como *fracasso*, o que pode gerar a subestimação das probabilidades, em relação às calculadas para a Copa do Brasil.

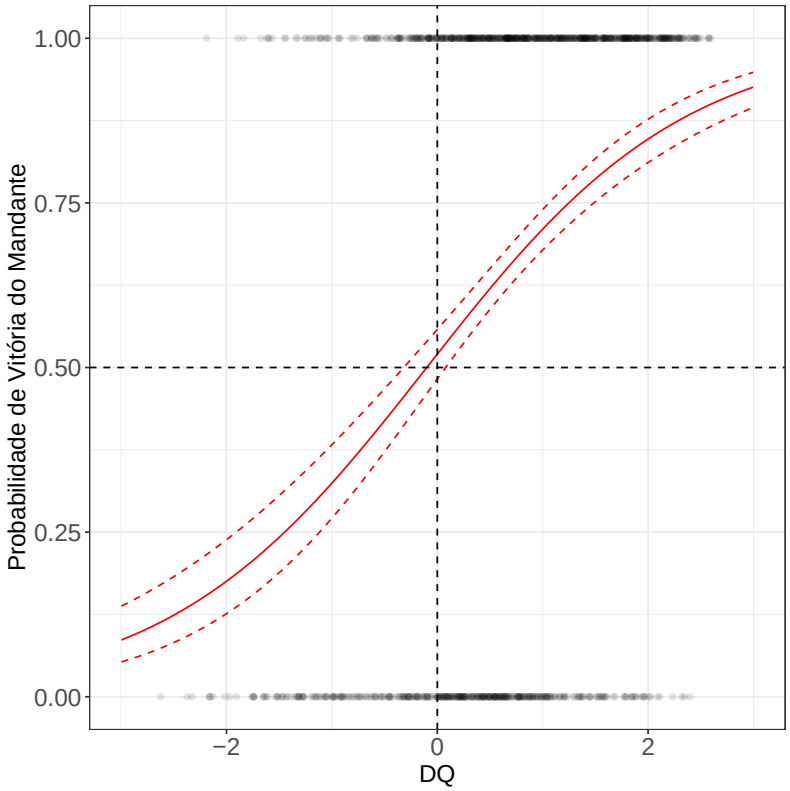
As estimativas obtidas são apresentadas na Figura (4.3), em que o eixo horizontal representa as diferenças de qualidade e o vertical as probabilidades estimadas por cada modelo. Nos dois gráficos, cada ponto representa um confronto, os que estão localizados na linha superior indicam que a determinada disputa foi um *sucesso*, por outro lado, os da linha inferior são *fracasso*. Para cada gráfico, em vermelho, a linha cheia representa as probabilidades estimadas, enquanto as pontilhadas indicam os intervalos de confiança. As linhas pretas pontilhas servem para facilitar a visualização dos gráficos, indicando onde a probabilidade é 0,5 e a diferença de qualidade é zero.

Verifica-se que, nos dois campeonatos, os intervalos de confiança para as probabilidades estimadas incluem o valor 0,5, indicando que os dois times têm iguais probabilidades quando a diferença de qualidade é zero. Além disso, no Campeonato Brasileiro o coeficiente estimado é 0,3040 (IC = (0,2443; 0,3642)), então o aumento de uma unidade na diferença de qualidade não discrimina adequadamente a ocorrência de vitória. Isso acontece porque o impacto da diferença de qualidade é menor nesse campeonato, uma vez que os times tem qualidades mais homogêneas. Por outro lado, na Copa do Brasil, que inclui clubes mais diversificados, o impacto da diferença de qualidade é maior, a estimativa do coeficiente é 0,8146 (IC = (0,6724; 0,9623)). Por isso, as probabilidade estimadas têm o formato mais parecido com um "S", como a curva sigmoide da função logística, e apresenta maior variação, de 0,0400 a 0,9658.

Essa diferença na variação indica que quando o visitante é muito melhor é mais provável que ele vença um jogo no Campeonato Brasileiro do que obtenha a classificação em um confronto da Copa do Brasil. Por exemplo, se a diferença de qualidade for -3 as probabilidades são, respectivamente, 0,2859 e 0,0861. Em contrapartida, se o mandante é muito melhor, diferença de qualidade igual à 3, as probabilidades são 0,7127 e 0,9259. Isso é, o mandante tem maior probabilidade de obter a classificação do que de vencer um jogo.



(a) Campeonato Brasileiro



(b) Copa do Brasil

Figura 4.3: Probabilidades de vitória no Campeonato Brasileiro ou classificação na Copa do Brasil

5 Análise Descritiva

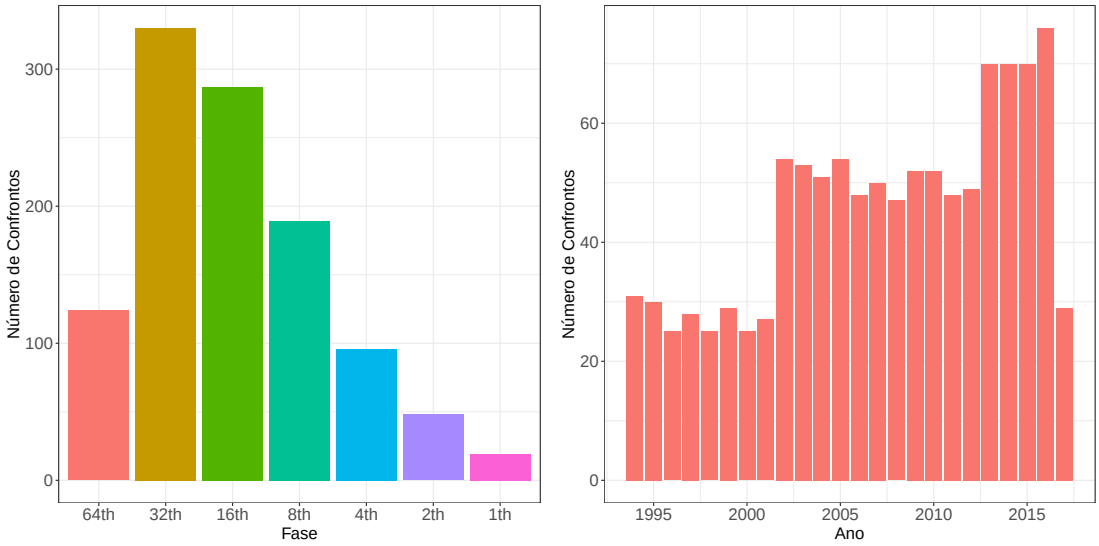
O objetivo deste capítulo é descrever e sumarizar os dados, de modo a contextualizar a Copa do Brasil, proporcionando um melhor entendimento das características desse campeonato. Ressalta-se que, bem como todas análises feitas neste trabalho, a análise descritiva é sobre os dados "limpos", isto é, após a exclusão das observações que não se enquadram nas características necessárias, detalhadamente descritas na Seção 3.1. Inicialmente os confrontos que terminaram empatados e precisaram de um critério de definição serão considerados separadamente de vitórias e derrotas, após a Seção 5.3 que serão considerados somente os desfechos possíveis na Copa do Brasil, classificação do mandante ou do visitante, ou seja, com os confrontos que terminaram empatados já tendo uma definição.

O número de confrontos disputados pela Copa do Brasil sofreu diversas alterações ao longo dos anos. A quantidade de equipes participantes aumentou à medida que o campeonato se consolidou como uma das principais disputas no cenário nacional. Então, a fim de permitir mais participantes, foram adicionadas novas fases à competição. Dentre os dados considerados, uma fase sempre deve ter o dobro de times da fase seguinte para manter o sistema de disputa , logo as fases iniciais são as que têm mais confrontos. A Figura (5.1) dispõe os números de confrontos disputados por fase (Gráfico (a)) e por ano (Gráfico (b)).

Verifica-se, no primeiro gráfico, que a fase sessenta e quatro avos de final tem menos de 150 confrontos acumulados no período total, isso porque ocorreu em apenas 4 dos 24 anos de disputas (entre 2013 e 2016 [1]). Já a fase trinta e dois avos de final aconteceu em 16 edições. Por outro lado, as fases finais representam 639 confrontos, 58,46% do total, e foram disputadas em todos os anos. Observando o gráfico identifica-se que, somente a partir das oitavas de final, o número de confrontos por fase parece respeitar o fato de que uma fase deve ter metade dos confrontos da anterior. Isso se deve ao aumento de fases ao longo dos anos e também aos confronto que foram excluídos da análise, que eram, em grande parte, das fases iniciais.

Conforme o esperado, é possível constatar, no segundo gráfico, um aumento no número de confrontos ao longo dos anos. As variações, no decorrer dos anos, são devidas aos confrontos que foram excluídos da análise. Em 2017 apenas 29 confrontos foram mantidos na análise porque as primeiras fases foram disputadas em partidas únicas, por essa razão o gráfico apresenta uma diminuição neste ano.

[1] Em 2017 houve disputas preliminares de jogo único, o mata-mata só começou na fase trinta e dois avos de final.



(a) Por fase

(b) Por ano

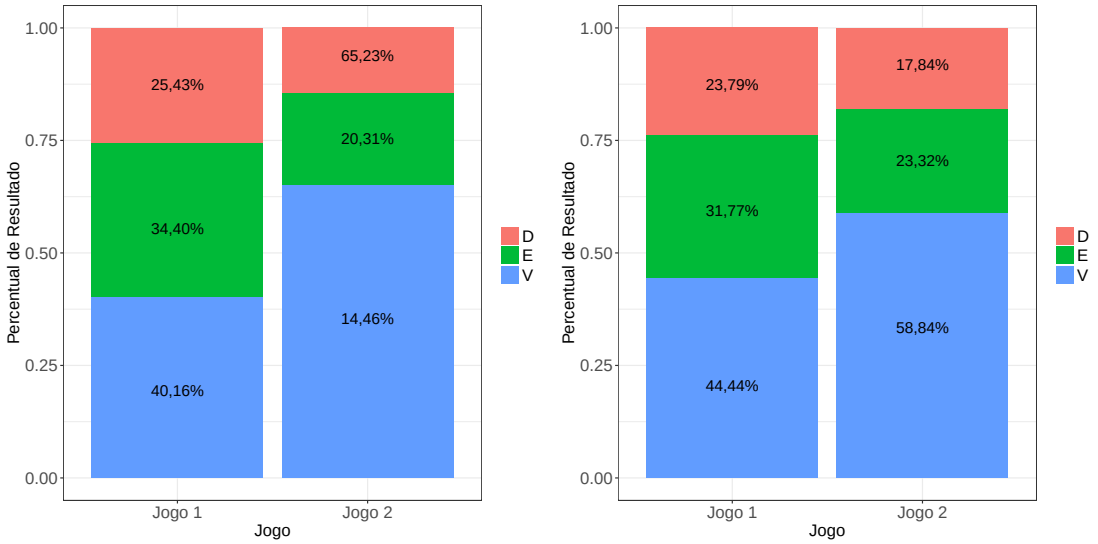
Figura 5.1: Número de confrontos na Copa do Brasil

5.1 Resultados por Jogo

No Capítulo 4, foram descritas as proporções de cada resultado possível nos jogos dos dois campeonatos nacionais, cujos valores eram próximos, apesar de mais favoráveis ao mandante na Copa do Brasil. No entanto, é verossímil supor que essas proporções sejam diferentes quando considerados os primeiros e os segundos jogos do confronto separadamente. A Figura (5.2) exibe um comparativo entre os resultados de cada jogo na Copa do Brasil, considerando todo o campeonato, e considerando apenas as fases finais. A legenda da figura indica “D”, “E” e “V”, que representam, respectivamente, derrota do time que jogou em casa, empate entre as duas equipes e vitória do time que jogou o determinado jogo em seu estádio.

Observa-se que o principal resultado, em todas as situações, é vitória do time que jogou a determinada partida em seu domínio. Isso está de acordo com o esperado, pois indica que existe vantagem para o time mandante, em um jogo isolado. Porém, no primeiro jogo os resultados são mais equilibrados, entre os três possíveis. Além disso, ao considerar o campeonato inteiro, o percentual de vitórias no segundo jogo (65,23%) é significativamente maior do que no primeiro (40,16% com p-valor $<$ 0,0001), concordando com a crença de que jogar o segundo jogo em casa é um benefício. Ressalta-se que ao se considerar todo o campeonato inclui-se as fases iniciais, nas quais, em alguns anos, o mando de campo era concedido ao time melhor – o que poderia explicar essa diferença.

Contudo, ao se considerar apenas para as fases finais, a diferença entre os percentuais de vitórias no primeiro (44,44%) e segundo jogo (58,84%) também é significativa (p-valor $<$ 0,0001). Portanto, leva-se a crer que, mesmo quando há sorteio do mando de campo e os confrontos são disputados por times de qualidades mais próximas, existe uma vantagem em ser o mandante da segunda partida.



(a) Campeonato (b) Fases finais

Figura 5.2: Percentual de resultados por jogo

Após explorar os resultados dos jogos no campeonato (no Capítulo 4) e também separando entre os resultados da primeira e segunda partida, na figura anterior, uma análise complementar é verificar, de forma combinada, os resultados obtidos em cada jogo. Isso é, nas análises já apresentadas, primeiramente foi calculado, no geral, qual é o percentual de vitórias do mandante em um jogo qualquer; em seguida foram computadas as proporções de vitórias do mandante separadamente, no primeiro jogo e no segundo jogo.

A Tabela (5.1) apresenta os percentuais dos resultados combinados, em que "D" indica que o time que jogava em casa perdeu aquela partida, "E" representa empate e "V" a vitória do time que jogou em casa. Então, a tabela informa, por exemplo, a proporção de vezes em que cada time perdeu o jogo que foi disputado em seu domínio (2,84% – coluna D, linha D) ou que o mandante do primeiro jogo perdeu e o segundo jogo foi empate (4,21% – coluna E, linha D).

Identifica-se então, que o resultado mais comum (25,34%) é vitória nos dois jogos, ou seja, cada time venceu ao jogar em seu estádio, situação em que é necessário um critério de desempate. Ressalta-se que não necessariamente ocorreu a reversão do resultado, por exemplo o primeiro jogo pode ter sido 3x0 e o segundo 1x0, os dois times venceram ao jogar em seu estádio. O segundo resultado mais comum é empate no primeiro jogo e vitória no segundo, situação que classifica o mandante, pela pontuação. É possível perceber que, para qualquer resultado da primeira partida, a maior proporção no segundo jogo é de vitória. Isto já era esperado, uma vez que na Figura (5.2) verificou-se que vitória no segundo jogo é o resultado mais comum. Na diagonal principal desta tabela estão os confrontos em que o ocorreu empate, pois a mesma pontuação foi obtida nos dois jogos; a maior parte desses (71,38%) são devidos a duas vitórias.

Tabela 5.1: Resultados combinados

		Jogo 2			Total
		D	E	V	
Jogo 1	D	2,84	4,21	18,39	25,43
	E	5,58	7,32	21,50	34,40
	V	6,04	8,78	25,34	40,16
Total		14,46	20,31	65,23	100

5.2 Resultados Agregados

As possíveis pontuações e suas combinações foram analisas a fim de permitir um melhor entendimento do funcionamento do torneio. Entretanto, o interesse principal desta pesquisa é estudar o resultado agregado do confronto. A Tabela (5.2) apresenta as frequências e percentuais do resultado agregado dos dois jogos, assim como da classificação final.

Verifica-se que, no campeonato como um todo, 44,10% das disputas têm a classificação direta do time mandante, isto é, por pontuação, sem empate no resultado agregado. Porém, ao observar apenas os confrontos das fases finais, esse percentual diminui para 37,25%. Por outro lado, o percentual de confrontos empatados (os que serão definidos por algum dos critérios) se mantém em torno de 36% ao se considerar todo o campeonato ou apenas as fases finais, divididos em cerca de 19% de classificação do mandante e 17% do visitante.

Levando-se em conta todo torneio, verifica-se que, após a definição do confronto, o time mandante obteve a classificação em 63,22% dos confrontos, significativamente mais que a metade (p-valor $< 0{,}0001$), com IC = (0,6036; 0,6608). Por outro lado, ao se considerar apenas os confrontos a partir dos dezesseis avos de final, esse percentual diminui para 56,81%, no entanto também é significativamente mais do que a metade dos confrontos (p-valor = 0,0006), com IC = (0,5297; 0,6065). Apesar de ambos serem maiores do que 0,5, nota-se que as estimativas dos intervalos bastante distintos. Isso ocorre porque, nas fases finais, a disparidade na qualidade dos times é menor, de forma que o visitante consegue vencer mais vezes, ainda que o mandante persista com a vantagem.

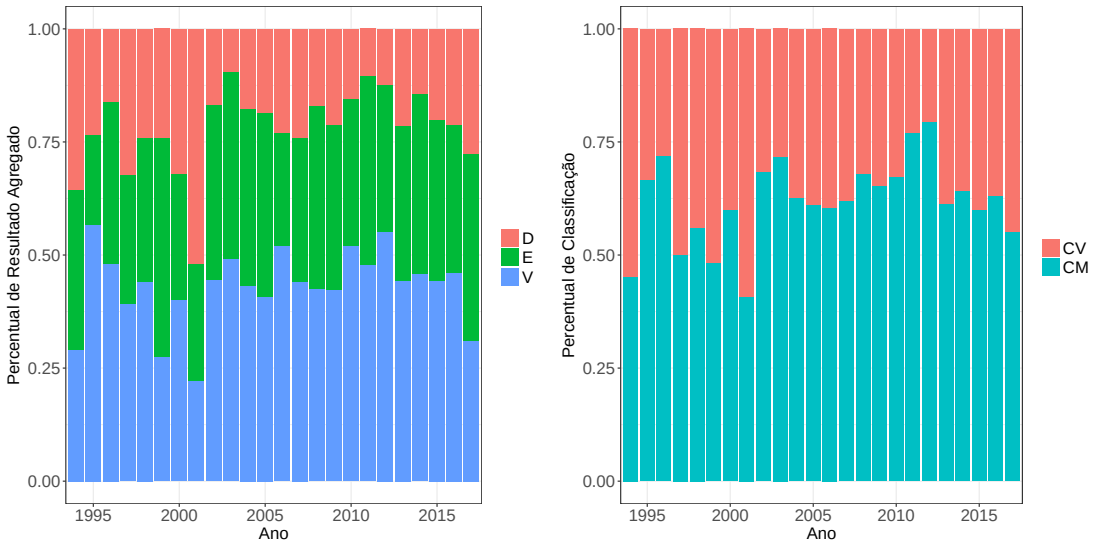
Tabela 5.2: Resultado agregado e classificação

	Campeonato		Fases finais	
Resultado	Frequência	%	Frequência	%
Vitória do mandante	482	44,10	238	37,25
Empate – CM	209	19,12	126	19,72
Empate – CV	179	16,38	110	17,21
Derrota do mandante	223	20,40	165	25,82
Classificação				
Mandante	691	63,22	363	56,81
Visitante	402	36,78	276	43,19
Total	1093	100	639	100

Como foi descrito anteriormente, a Tabela (5.2) apresenta os percentuais dos resultados agregados e da classificação. Porém, no Capítulo 4, observando a Figura

(4.2), que apresenta os resultados do time mandante em cada jogo na Copa do Brasil, identificou-se uma possível mudança ao longo do tempo. Por isso, a Figura (5.3) apresenta os percentuais dos resultados agregados (Gráfico (a)) e dos percentuais de classificação (Gráfico (b)) ao longo dos anos, a fim de analisar se também há diferenças entre os anos, como existe para os resultados quando considerados os jogos separadamente.

Visualmente, identifica-se que existe uma variação entre os anos, nos dois gráficos, mas principalmente nas proporções dos resultados. Acredita-se que isso se deva as fases que foram disputadas. Nos anos em que o campeonato teve mais etapas, ocorreram mais confrontos nas fases iniciais, nas quais nem sempre ocorreu sorteio, além de que, muitas vezes, o time mandante era o de melhor qualidade, consequentemente, aconteciam mais vitórias do mandante. No gráfico da classificação, o comportamento aparenta ser mais constante, especialmente depois de 2002, quando começou a fase trita e dois avos de final. Isso indica que, apesar de o resultado em pontos do confronto ser consideravelmente variável, o desfecho final, a classificação, é relativamente constante no decorrer dos anos.



(a) Resultados (b) Classificação

Figura 5.3: Resultados agregados e classificação do time mandante do confronto, por ano

Ademais da variação ao longo dos anos, sabe-se que pode haver diferença nos percentuais de classificação no decorrer das fases da competição, uma vez que, à medida que a competição avança, a diferença entre as qualidades dos times tende a diminuir, como já foi discutido na Seção 3.2. Nesse sentido, a Figura (5.4) expõe os percentuais de classificação do mandante e do visitante em cada fase da competição.

É possível identificar que, nas duas fases inicias, os mandantes vencem aproximadamente 70% dos confrontos, com IC = (0,6813; 0,7637), significativamente mais que a metade (p-valor $<$ 0,0001). Além disso, essa porcentagem nas fases iniciais também é significativamente maior do que nas fases finais (p-valor $<$ 0,0001), cujo percentual é 56,81%, que, por sua vez, também é significativamente maior do que 0,5 (p-valor = 0,0006), com IC = (0,5297; 0,6065). Isso novamente evidencia que as diferenças de qualidade entre as equipe de um confronto nas fases finais é menor

do que nas inicias. Então, mesmo que o time mandante tenha uma vantagem, o visitante consegue vencer mais do que nas fases iniciais. Entretanto, se considerados apenas os confrontos pós oitavas de final, então o percentual é 51,70%, igualando as porcentagens de classificação do mandante e visitante (p-valor = 0,5235), com IC = (0,4648; 0,5692).

Na final do campeonato o percentual de classificação do mandante é 31,58%, bem menor do que nas demais fases, é a única que demostra que o time visitante tem uma vantagem, é importante ressaltar que esse valor é calculado em apenas 19 confrontos. No entanto, em todos o time visitante venceu ou empatou (respectivamente 10 e 9 vezes) o primeiro jogo, frisando a importância de construir um bom resultado na primeira partida. Também destaca-se a dificuldade que o mandante tem de reverter esse resultado, mesmo com a vantagem de ter o mando de campo no jogo final, das 6 vezes que se classificou, o mandante perdeu o primeiro jogo 2 vezes e empatou 4.

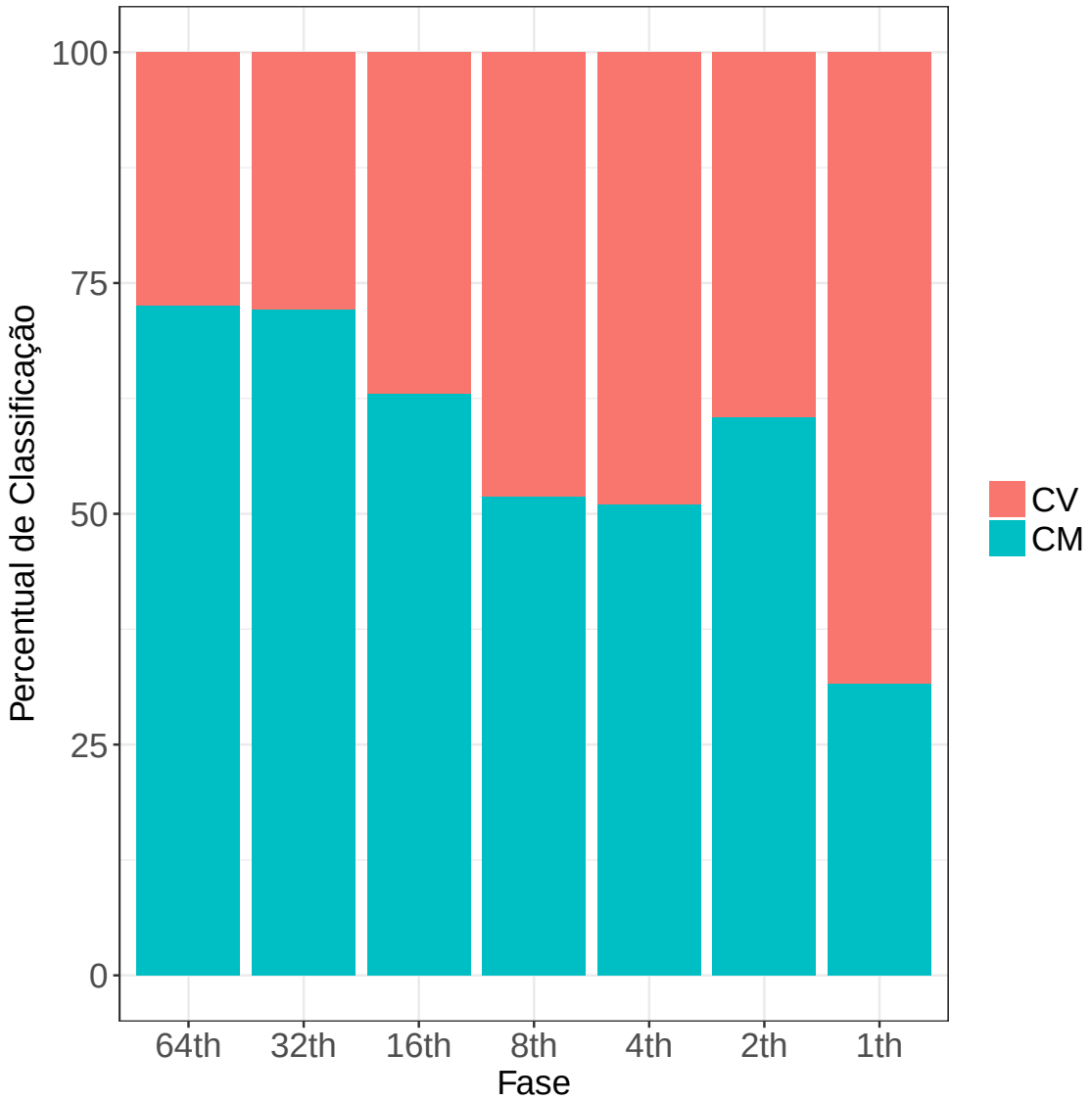


Figura 5.4: Percentual de classificação por fase

5.3 Confrontos Empatados

As análises até agora apresentadas consideram o empate separadamente de vitórias e derrotas. No entanto, esse não é um resultado possível no final do confronto, pois na Copa do Brasil o desfecho final é a classificação do mandante ou do visitante. Então, as próximas análises expõem os resultados para os 388 confrontos que terminaram empatados.

Dos confrontos empatados, 60,82%, com IC = (0,5596; 0,6568) ocorreram nas fases finais, significativamente mais do que a metade (p-valor $<$ 0,0001). Apesar de as fases inicias terem mais confrontos disputados, ocorrem mais empates na fases finais, uma vez que a qualidade dos times é mais equilibrada. A Tabela (5.3) permite analisar os confrontos que terminaram empatados, a fim de verificar as frequências de classificação do mandante e também as frequências de utilização dos critérios, comparando o campeonato inteiro com somente as fases finais.

Percebe-se que, para classificação e critério, os percentuais se mantém nos dois casos, ao se considerar somente os confrontos das fases finais e no campeonato como um todo. Identifica-se que o mandante se classifica em cerca de 53% dos confrontos. Considerando todo o campeonato ou as fases finais, não há evidências de que esse valor seja significativamente diferente de 50% (p-valores 0,1274 e 0,2978), indicando que, nos confrontos empatados, não há vantagem de classificação para nenhum dos times. Além disso, o principal critério de desempate é o saldo de gols, utilizado em cerca de metade dos confrontos, seguido por gol qualificado e pênaltis.

Tabela 5.3: Classificação e critérios utilizados nos confrontos empatados

	Campeonato		Fases finais	
Classificação	Frequência	%	Frequência	%
Mandante	209	53,87	126	53,39
Visitante	179	46,13	110	46,61
Critério				
SG	196	50,52	117	49,58
GQ	113	29,12	73	30,93
PN	79	20,36	46	19,49
Total	388	100	236	100

Os confrontos podem terminar empatados em decorrência de três combinações de resultados: os dois jogos tiveram derrota do time que jogava em casa (DD); os dois jogos terminaram empatados (EE); os dois jogos tiveram vitória do time que jogava em casa (VV). Por isso, além de verificar o uso dos critérios nos confrontos empatados em geral, como feito na Tabela (5.3), considerou-se importante analisar as proporções de utilização dos critérios em cada possível tipo de empate. Essas são descritas na Tabela (5.4).

Verifica-se que, em caso de duas derrotas, o principal critério utilizado é o saldo de gols (41,94%). Porém, essa situação é a de maior equilíbrio de uso dos critérios. Entretanto, quando acontecem dois empates os times têm o mesmo saldo de gols, por isso esse deixa de ser um critério de desempate. Então, o critério mais frequente é o gol qualificado, utilizado em 62,5% dos casos. Já nas situações em que houve duas vitórias o saldo é o critério utilizado na maior parte dos confrontos (66,06%, p-valor $<$ 0,0001).

Tabela 5.4: Percentual do uso de critério por tipo de empate

Critério	DD	EE	VV
SG	41,94	-	66,06
GQ	35,48	62,50	18,77
PN	23,58	37,50	15,16

Também foi considerado interessante verificar as proporções de classificação de mandantes e visitantes ao utilizar cada um dos critérios. Então, a Tabela (5.5) apresenta essa informação. Constata-se que o mandante tem vantagem significativa (p-valor = 0,0006) apenas quando o confronto é decidido no saldo de gols, cujo intervalo de confiança é (0,5345; 0,6903). O uso dos outros critérios não resulta em vantagem significativa, no entanto equiparam as probabilidades dos dois times (p-valores respectivamente 0,2215 e 0,5784 para gol qualificado, que tem IC = (0,3509; 0,5341), e pênaltis, com IC = (0,3584; 0,5784).

Tabela 5.5: Percentual de classificação por critério

Critério	CM	CV
SG	62,24	37,76
GQ	44,25	55,75
PN	46,84	53,16

A fim de combinar as duas informações imediatamente anteriores, apresenta-se a Tabela (5.6). Esta expõe os percentuais de classificação para mandantes e visitantes em cada tipo de critério de desempate, considerando separadamente cada possível combinação de resultados.

Utilizando saldo de gols, se o empate foi causado por duas derrotas, o mandante do confronto tem 23,08% de probabilidade de classificação. De maneira oposta, no caso de duas vitórias a vantagem é de 65,03% para o mandante. Então, se ocorreram duas derrotas a vantagem é do visitante, mas no caso de duas vitórias é do mandante. Pode-se pensar que isso acontece porque o segundo jogo é "mais decisivo", com o resultado do primeiro jogo não podendo mais ser alterado e os times tendo conhecimento do número de gols necessários para utilizar o saldo de gols como critério. Então, o time que vence o segundo jogo se classifica mais frequentemente, sendo ele o mandante ou não. No caso de decisão por gol qualificado o visitante mantém a equiparação da vantagem, independente da pontuação que causou o empate.

Na disputa de pênaltis, o mandante tem vantagem apenas se os dois jogos foram empate, caso contrário a vantagem na probabilidade de classificação é do time visitante. A situação de derrota nos dois jogos, significa que o visitante venceu o segundo jogo, o que tende a fazer com que o time mandante fique abalado, pois tinha uma vantagem, por ter vencido a primeira partida, que foi revertida no seu mando de campo, com isso, o visitante, que está confiante por ter conseguido compensar sua derrota, passa a ter a vantagem. Quando ocorrem duas vitórias, o time mandante entrou para o segundo jogo em desvantagem, mas conseguiu igualar o resultado. Contudo, nessa situação o time visitante costuma jogar retrancado, tentando segurar o resultado construído no primeiro jogo. Então, o mandante precisa de muito esforço durante toda a partida para obter a vitória, por isso os jogadores ficam propensos a estarem mais cansados do que os do time visitante, além de pos-

sivelmente frustrados por não terem conseguido superar o resultado sem precisar da disputa de pênaltis.

Por outro lado, os confrontos que tiveram os dois jogos empatados geralmente acontecem em duas situações. Quando o jogo foi "morno", fazendo com que os dois times estejam descansados, ou porque foram jogos muito disputados e, consequentemente, exigiram muito fisicamente. Das duas formas, entende-se que o desgaste é igual para os dois times, porém o visitante precisa lidar com a pressão da torcida, enquanto o mandante tem o apoio, podendo ser esta a causa para a vantagem do mandante.

Tabela 5.6: Percentual de classificação por critério e por pontos

Critériro	Classificação	DD	EE	VV
SG	CM	23,08	-	65,03
	CV	76,92	-	34,97
GQ	CM	45,45	46,00	42,31
	CV	54,55	54,00	57,69
PN	CM	28,57	63,33	38,10
	CV	71,43	36,67	61,90

5.4 Proporção de Classificação

Com o objetivo de averiguar – globalmente e em cada um dos quatro possíveis tipos de classificação – se a probabilidade de classificação do mandante em confrontos da Copa do Brasil é igual a 0,5, ou seja, se os dois times têm iguais probabilidades de classificação, foram estimadas as respectivas proporções e intervalos de confiança, que são apresentadas na Tabela (5.7).

Verifica-se que, ao se considerar todos os confrontos, a proporção de classificação do mandante é de 63,22%, significativamente maior do que 50%, IC = (0,6036; 0,6608), indicando que o time mandante obtém a classificação em mais da metade das disputas. Dessa forma confirma-se a hipótese inicial de pesquisa de que o mandante tem essa vantagem. O mesmo acontece nos confrontos que foram decididos por pontuação e por saldo de gols, com estimativas de 68,37%, IC = (0,6493; 0,7181), e 62,24%, IC = (0,5540; 0,6909), respectivamente. Entretanto, ao considerarem-se os confrontos decididos por gol qualificado ou que foram para a disputa de pênaltis, as proporções são 44,25% e 46,84%, respectivamente. Portanto, nesses casos, não há evidências de que algum time tenha vantagem de classificação, os dois intervalos de confiança incluem o 0,5.

Tabela 5.7: Estimativas e intervalos de confiança

	Estimativa	IC inferior	IC superior
Geral	0,6322	0,6036	0,6608
PT	0,6837	0,6493	0,7181
SG	0,6224	0,5540	0,6909
GQ	0,4425	0,3495	0,5355
PN	0,4684	0,3559	0,5808

Unificando os casos em que os confrontos foram decididos por gol qualificado ou pênaltis obtém-se 0,4531 como a proporção de classificação do mandante. Por outro lado, juntando as situações em que a decisão foi por pontuação ou saldo, a proporção é de 0,6322. Comparando essas duas estimativas conclui-se que são significativamente diferentes, IC = (-0,2556 ; -0,1026) e p-valor $<$ 0,0001. Portanto, confirmando a análise anterior, a proporção de classificação do mandante por pontuação ou saldo é significativamente maior do que nos outros casos.

A partir das análises feitas, conclui-se que quando o confronto termina empatado e é necessário utilizar gol qualificado ou pênaltis como critério de definição, os dois times têm iguais probabilidades de classificação. Porém, sabendo-se que a decisão deu-se por pontuação ou saldo de gols, então é mais provável que o mandante obtenha a vitória.

6 Diferença de Qualidade

6.1 Relação com o Ano e a Fase

Como foi discutido anteriormente, na Seção 3.2, acredita-se que a variação existente entre os anos deve-se principalmente às mudanças nas qualidades dos times. Além disso, as figuras 4.2 e 5.3 identificam uma variação ao longo dos anos nos resultados dos jogos, individualmente e no confronto, bem como na classificação. Então, a Figura (6.1) permite verificar a relação entre as diferenças de qualidade e os anos dos confrontos disputados.

Observa-se que a variação da diferença de qualidade é similar entre os anos. Entretanto, percebe-se que após 2002 há uma maior concentração de confrontos que têm diferença de qualidade positiva. Isso se deve ao fato de que, em alguns anos, nas fases iniciais, o melhor time recebia o mando de campo do segundo jogo, sem ser feito sorteio. Por outro lado, nota-se que a amplitude dos valores negativos aumentou nos anos mais recentes, indicando que têm acontecido confrontos nos quais times visitantes são muito melhores, o que tinha menos ocorrência antigamente, quando não havia sorteio e o mando de campo era dado ao melhor time, fazendo com que a diferença de qualidade fosse positiva. Entende-se então, que as variações nos resultados e classificação ao longo dos anos podem ser captadas pela diferença de qualidade entre os times que disputam os confrontos.

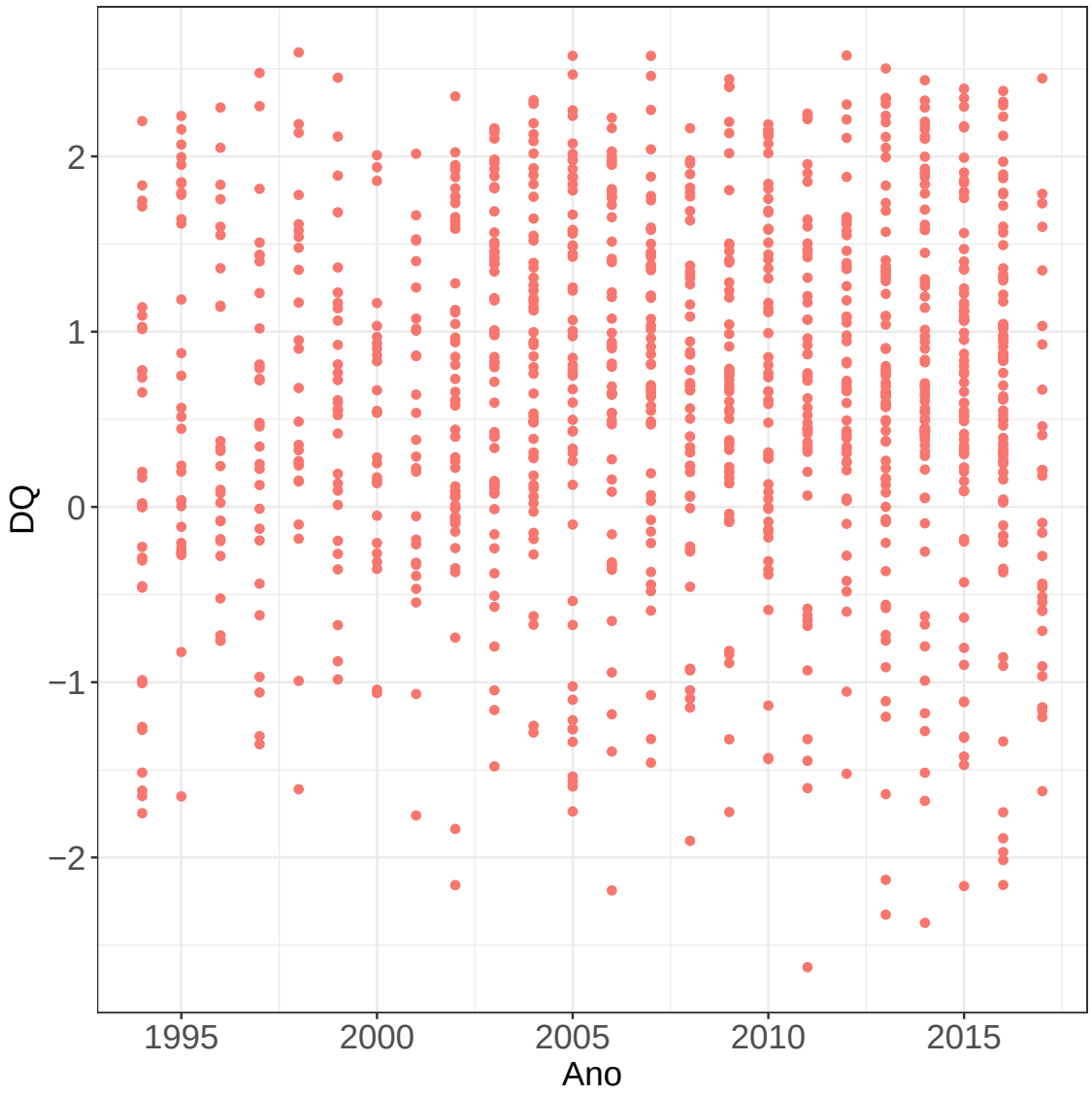


Figura 6.1: Diferença de qualidade x ano

Além disso, a Seção 3.2, bem como a Figura (5.4), trouxeram a discussão sobre a variação dos percentuais de classificação e da diferença de qualidade ao longo das fases da competição. A Figura (6.2) apresenta a relação entre essas variáveis. Observa-se que as fases dezesseis avos e oitavas de final são as de maior variação. Isso ocorre porque são as que apresentam os confrontos mais diversos, alguns entre times de qualidades próximas, outros com grandes discrepâncias, positivas e negativas. Também percebe-se um claro contraste entre as diferenças de qualidade nas diversas fases: as fases dezesseis, trinta e dois e sessenta e quatro avos de final têm as maiores medianas, acima de 0,5. Enquanto as outras fases têm medianas próximas de 0,5. Ressalta-se que a final do campeonato tem apenas 19 observações, podendo levar a conclusões errôneas.

Entende-se que a variação nos percentuais de classificação ao longo das fases da competição se deve às diferenças de qualidades. As fases cujas diferenças de qualidade são mais positivas, os seja os mandantes são melhores, têm os maiores percentuais de classificação do mandante. Por outro lado, nas oitavas, quartas e

semifinais, que têm diferenças de qualidades visualmente simétricas em torno de zero, os percentuais de classificação são cerca de 50%. A final do campeonato não apresenta simetria nas diferenças de qualidade, no entanto, como foi ressaltado anteriormente, esse *boxplot* é composto por apenas 19 observações.

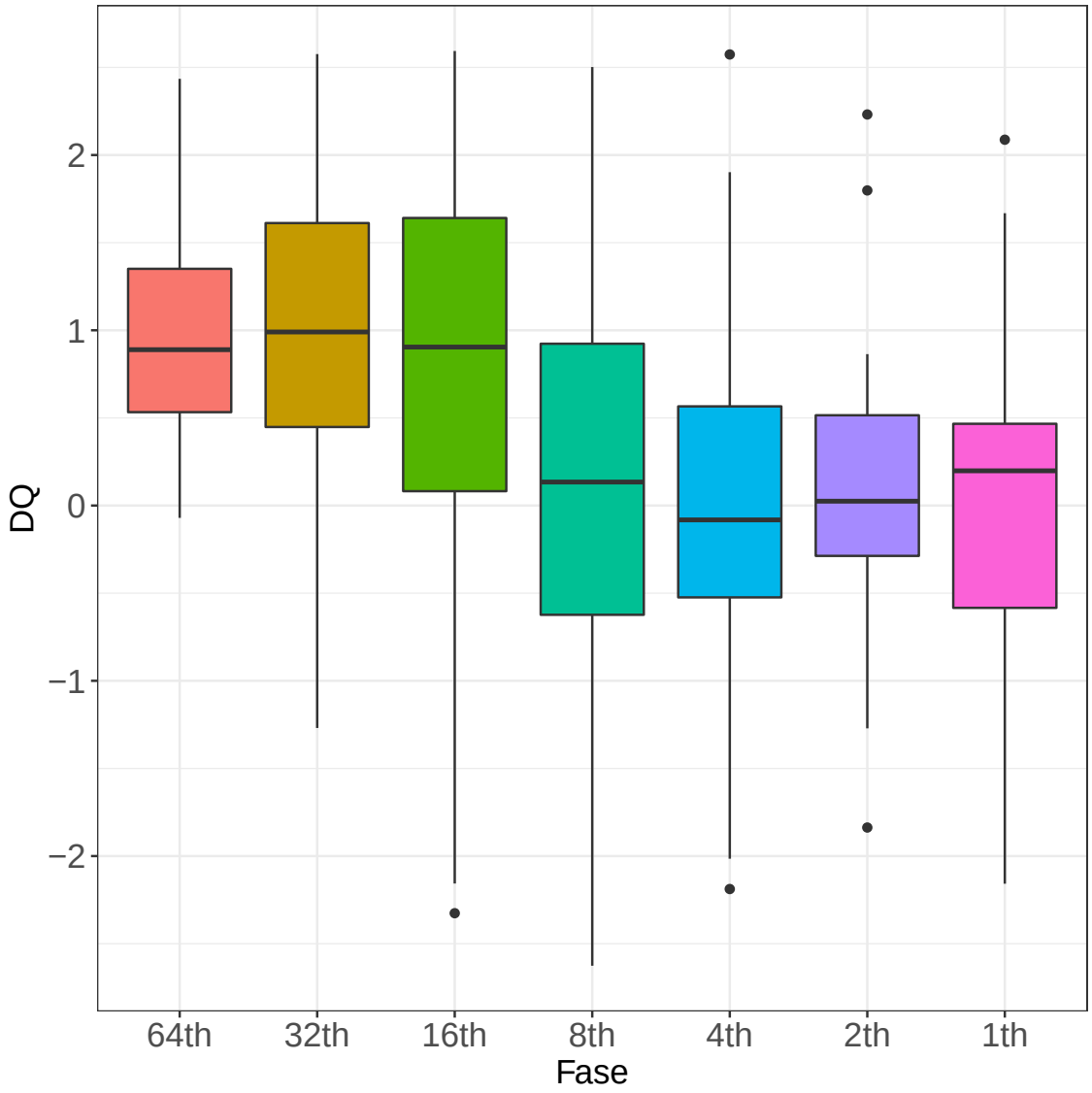


Figura 6.2: Diferença de qualidade x fase

Ao regredir a diferença de qualidade contra ano e fase tem-se todos os coeficientes significativos, bem como o teste de significância global do modelo, evidenciando que, de fato, há uma relação significativa destas variáveis com a diferença de qualidade.

6.2 Relação com a Classificação

A fim de verificar se o padrão da diferença de qualidade é o mesmo nos casos em que o mandante se classificou e naqueles em que o visitante foi o vencedor, a Tabela (6.1) apresenta o resumo estatístico da distribuição da diferença de qualidade

separando essas duas situações. Verifica-se que a variação é maior nos confrontos em que o time visitante obteve a classificação, apesar de o mínimo e o máximo não serem tão discrepantes, as médias e medianas são maiores nos confrontos em que o mandante se classificou. Além disso, as duas medianas são positivas, indicando que o time mandante era melhor, mesmo quando foi o visitante que venceu o confronto.

Tabela 6.1: Medidas resumo da diferença de qualidade por time classificado

Classificação	Mínimo	Mediana	Máximo	Média	Desvio Padrão
Mandante	-2,1879	0,9458	2,5936	0,9188	0,9066
Visitante	-2,6257	0,2641	2,3972	0,1787	0,9808

A mesma ideia apresentada pela tabela anterior pode ser vista na Figura (6.3) que apresenta as distribuições da diferença de qualidade para os dois casos, classificação do mandante ou do visitante, separado por cada tipo de decisão. O principal ponto a ser observado é que, em todas as situações, mesmo nos confrontos em que o visitante se classificou, o time mandante era melhor em mais da metade dos casos. Portanto, há uma assimetria na distribuição da diferença de qualidade, possivelmente influenciada pelos confrontos das fases iniciais, conforme foi discutido na Seção 3.2.

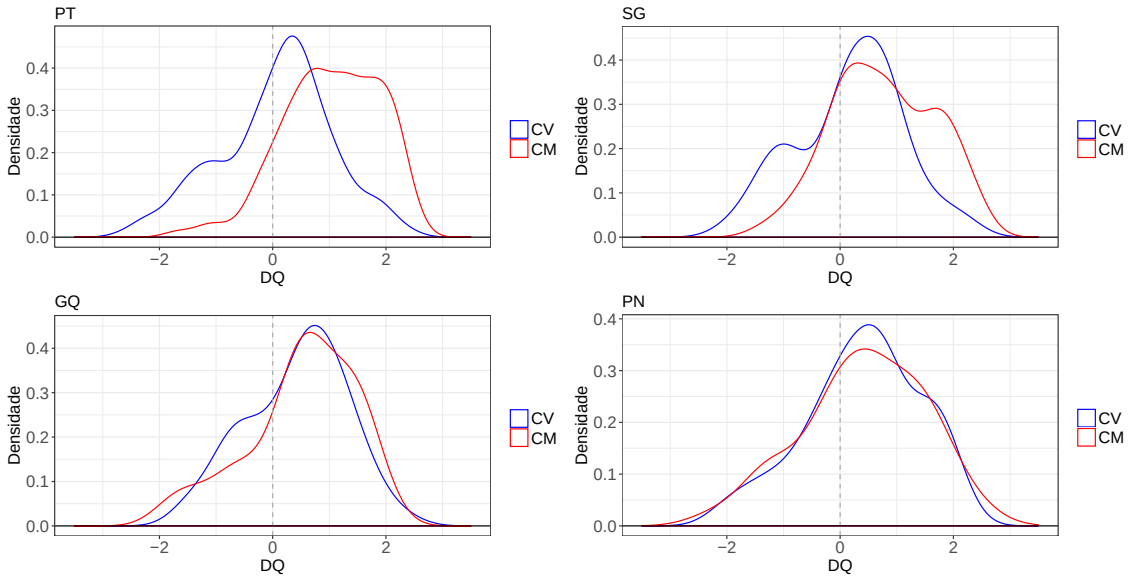


Figura 6.3: Distribuição da diferença de qualidade por time classificado em cada tipo de classificação

6.3 Relação com o Tipo de Classificação

Com o objetivo de identificar possíveis variações nas distribuições da diferença de qualidade nos diversos tipos de classificação são apresentados quatro *boxplots* na Figura (6.4). A variabilidade aparenta ser constante, indicando que não há heterocedasticidade entre as diferentes circunstâncias. Além disso, existem poucos *outliers* (valores atípicos) e estes ocorrem principalmente nos confrontos em que a classificação se deu pela pontuação.

Outra observação relevante, é que todas as medianas são positivas, confirmando que, em todos os casos, o mandante é melhor em mais da metade dos confrontos. Mais uma vez, traz-se a questão da assimetria que é causada pela falta de sorteio em algumas fases. Ademais, as medianas variam de 0,7498 a 0,4494, diminuindo conforme a ordem de utilização dos critérios. É possível que isso aconteça porque, quanto menor a diferença de qualidade entre os times, mais critérios são necessários para definir o vencedor. Assim como as medianas, as médias das diferenças decaem à medida que mais critérios são necessários para definir o vencedor, indicando que há uma relação entre a diferença de qualidade e o tipo de classificação.

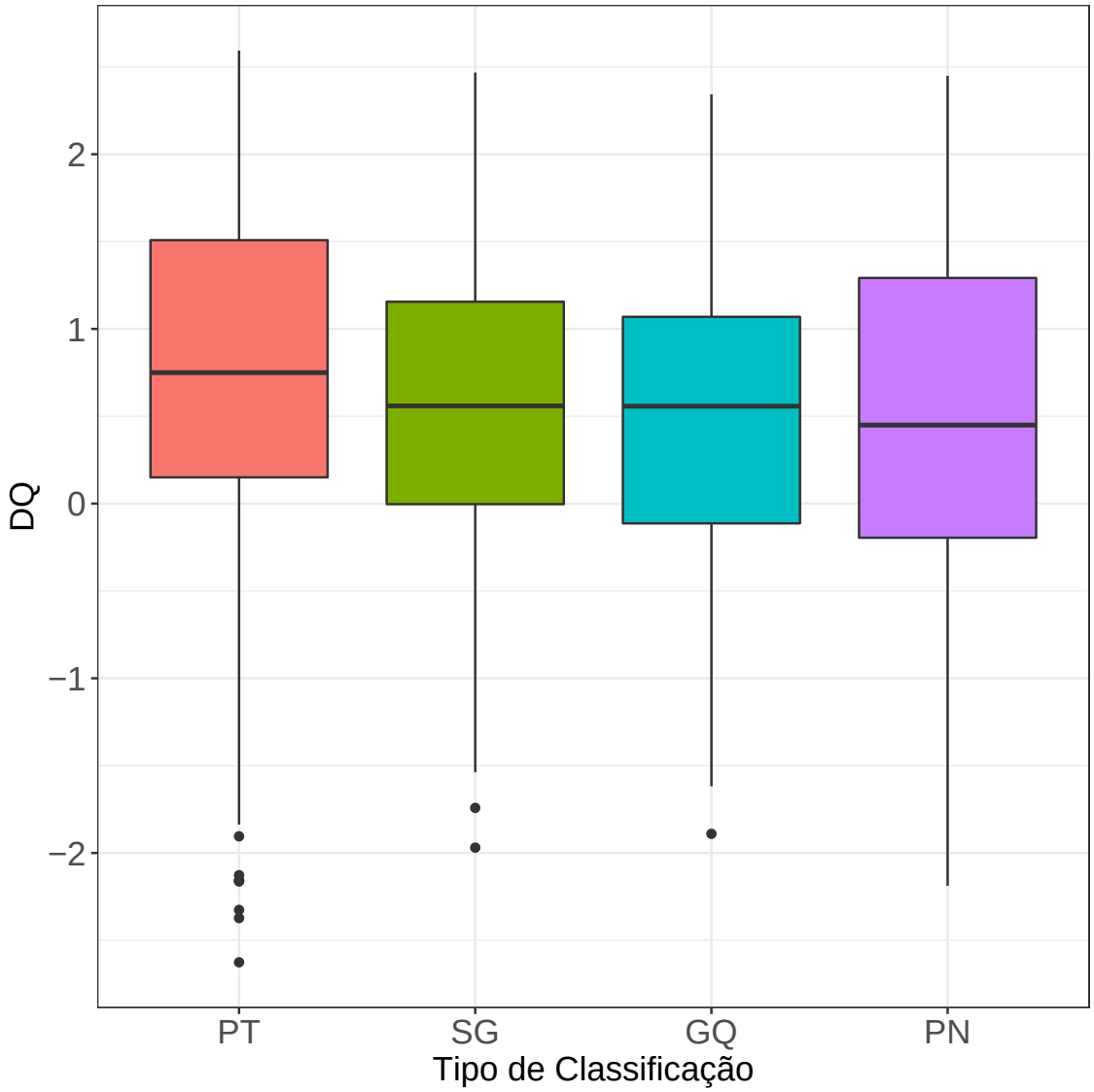


Figura 6.4: Distribuição das diferença de qualidade em cada tipo de classificação

A interação existente entre a diferença de qualidade e o tipo de classificação pode ser melhor observada na Figura (6.5). Verifica-se o decaimento da média da diferença de qualidade (linha azul) à medida que mais critérios de desempate são necessários, confirmando que a ordem de utilização está adequada. Outra informação relevante que o gráfico traz, é que ao longo do uso dos critérios a média da diferença de

qualidade diminui nos confrontos em que o vencedor foi o mandante (linha vermelha) e aumenta naqueles cuja vitória foi do visitante (linha verde). Percebe-se que, nos confrontos definidos pela pontuação, as médias são muito diferentes nos dois casos. Porém, se aproximam e acabam muito próximas quando a decisão é por pênaltis, podendo considerar o saldo de gols e o gol qualificado como estágios de transição entre os critérios.

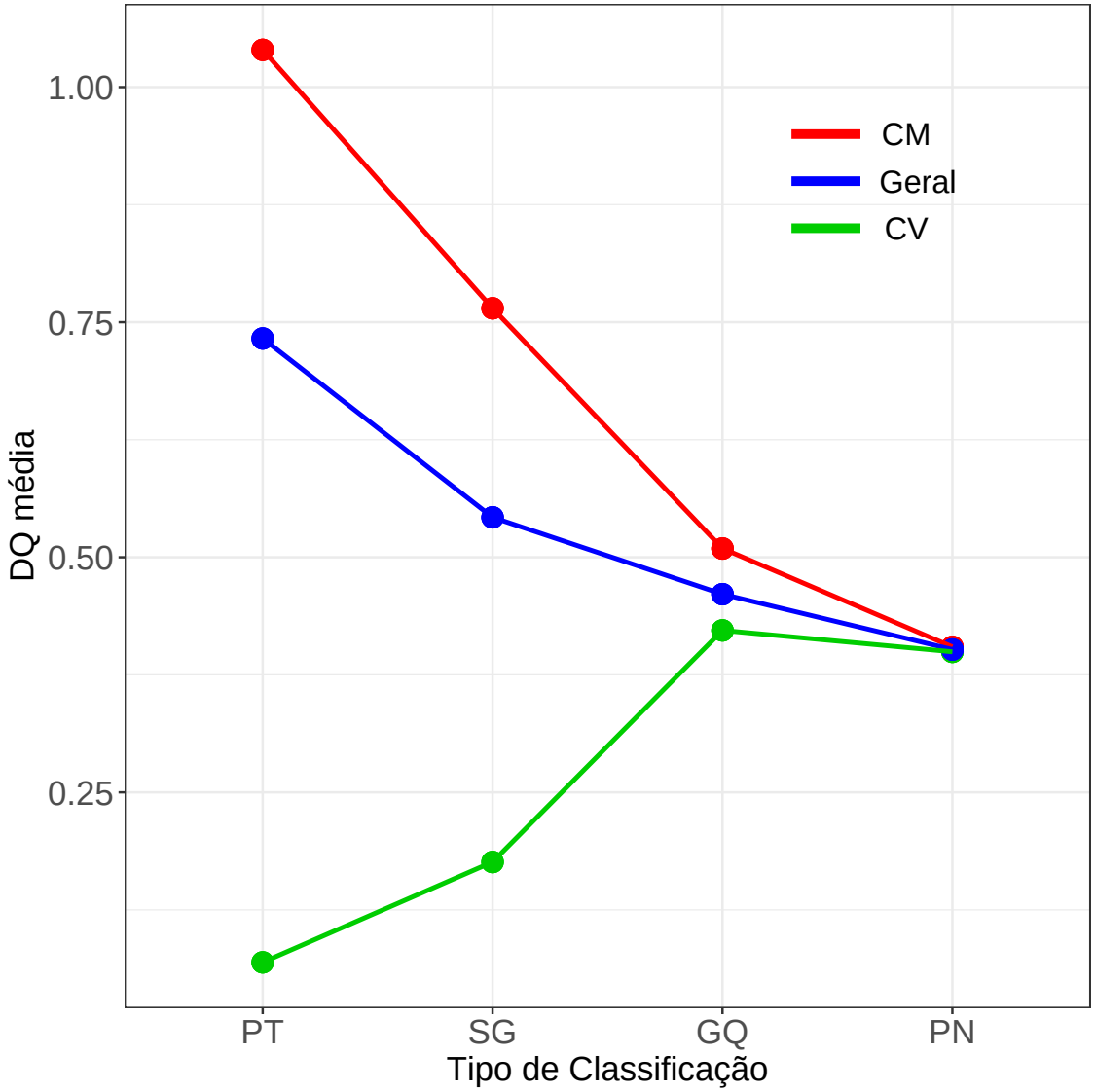


Figura 6.5: Interação entre a diferença de qualidade e o time classificado

A fim de confirmar a relação entre a diferença de qualidade e os tipos de classificação, foi ajustada uma regressão linear. No modelo, a variável resposta é a diferença de qualidade e o tipo de classificação é a variável explicativa, sendo a categoria de referência a classificação por pontos. As estimativas obtidas estão expostas na Tabela (6.2).

Os resultados demonstram que existe uma relação significativa e inversa entre a diferença de qualidade e cada um dos critérios de desempate, indicando que quanto maior a diferença de qualidade, menor a probabilidade de a decisão ser nesses cri-

térios. Por outro lado, o intercepto, que representa a categoria de referência, ou seja, o caso em que os confrontos são definidos pela pontuação, tem uma estimativa positiva. O teste de significância global do modelo indica que de fato as variáveis diferença de qualidade e tipo de classificação têm uma relação significativa (p-valor = 0,0011).

Tabela 6.2: Colinearidade

	Estimativa	Erro Padrão	t	p-valor
Intercepto	0,7327	0,0374	19,572	<0,0001
SG	-0,1902	0,0803	-2,370	0,0180
GQ	-0,2719	0,1007	-2,700	0,0071
PN	-0,3308	0,1179	-2,805	0,0051

Em vista daquilo que foi discutido nessa seção, constata-se que é importante utilizar as interações entre essas variáveis no ajuste do modelo de regressão, além de analisar possíveis problemas de multicolinearidade, que poderiam causar impactos sobre a precisão das estimativas.

7 Análises Exploratórias

7.1 Empate e Uso de Gol Qualificado

Foi discutido que o principal objetivo deste trabalho é analisar a vantagem de jogar como mandante o segundo jogo em uma disputa na Copa do Brasil e, além disso, verificar a influência que o gol qualificado tem sobre isso. Para compreender melhor esse fator é preciso entender quanto e quando ele ocorre. Esse é o objetivo desta seção. Tem-se que independentemente do time favorecido, as proporções estimadas de confrontos que terminam empatados (consequência de três possíveis combinações de resultados: DD, EE ou VV) e de confrontos que são definidos pela regra do gol qualificado são, respectivamente, 35,50%, IC = (0,3266; 0,3834), e 10,34%, IC = (0,0853; 0,1215), do total de disputas. No entanto, acredita-se que isso varie de acordo com a diferença de qualidade entre os times. Isso porque, entende-se que seria mais provável um empate ocorrer em um confronto cujas equipes têm qualidades equiparáveis do que quando um time é muito melhor do que outro.

Então, a fim de estimar as probabilidades de um confronto terminar empatado e de ser decidido pela regra do gol qualificado, dadas as diferenças de qualidade entre os times participantes, foram ajustadas duas regressões não paramétricas, estimadas pelo método Kernell (ver Hayfield e Racine (2008)). As duas variáveis resposta são dicotômicas: para o primeiro modelo, o *sucesso* é empate na disputa; para o segundo, é ter utilizado o gol qualificado – em ambos os casos, independentemente do time que obteve a classificação final. As estimativas são apresentadas na Figura (7.1).

Os pontos do gráfico representam os confrontos que foram disputados: os que aparecem na parte superior, na linha do 1, são os eventos de *sucesso*, isto é, os que terminaram empatados; na linha do zero constam os eventos de *fracasso*. Os confrontos definidos por gol qualificado não estão marcados de forma diferenciada pois são uma parte dos empatados. O posicionamento horizontal de cada ponto representa a diferença de qualidade dos times desse confronto, uma vez que o eixo horizontal dispõe essa variável, onde o 0 indica ausência de diferença. As linhas pontilhadas em preto estão plotadas no gráfico apenas para facilitar a visualização, indicando onde a probabilidade é 0,5 e a diferença de qualidade é igual a 0.

Em vermelho, estão representadas as probabilidades estimadas de um confronto terminar empatado. Identifica-se que as estimativas estão abaixo de 0,5 para todas as diferenças de qualidade; apenas uma pequena parte do intervalo de confiança superior está acima desse valor. Apesar disso, observa-se que o evento é mais provável quando o time visitante é melhor, isto é, para diferenças de qualidade negativas, sendo que o máximo ocorre quando o visitante é cerca de um desvio padrão melhor

e decai à medida que a que diferença de qualidade fica mais positiva. Contrariando o que se esperava, que as maiores probabilidades seriam quando os times têm qualidades próximas.

Todavia, pode-se argumentar que, nos casos em que o time visitante é melhor, sendo o primeiro jogo em sua casa, geralmente o time mandante vai para esse jogo com uma tática defensiva, buscando um empate ou uma derrota simples para que possa "correr atrás do resultado" no segundo jogo. Essa postura, mesmo quando o visitante tem qualidade muito superior, tende a aumentar a probabilidade de o confronto terminar empatado. Por outro lado, quando o mandante é muito melhor essa situação não é comum, e por essa razão a probabilidade de empate no confronto estimada pela regressão é muito mais baixa: ocorre que no primeiro jogo o pior time busca conseguir um bom resultado em sua casa para então mantê-lo no segundo jogo, em que o time mandante, além de ter qualidade superior, irá jogar em seu estádio, já sabendo qual resultado é necessário para obter a classificação.

Entende-se que a probabilidade de empate é maior quando o visitante é melhor do que quando os dois times têm qualidade iguais, por causa da vantagem que o time mandante tem por decidir o confronto em casa. Isso é, se os mandos de campo fossem neutros, seria esperado que confrontos empatados acontecessem com maior probabilidade em disputas de equipes com a mesma qualidade, no entanto, a vantagem de decidir em casa se sobrepõe à essa equiparação.

As probabilidades estimadas de uso do critério do gol qualificado são representadas em azul no gráfico. Essas são sempre menores do que as probabilidades de o confronto empatar, uma vez que esses confrontos são uma parte dos empatados, 29,12% conforme a Tabela (5.3). Como esperado, essa probabilidade é pequena para quaisquer diferenças. Identifica-se que, assim como para a probabilidade de empate nos confrontos, o ponto máximo acontece quando o visitante é aproximadamente um desvio padrão melhor. Acredita-se que as mesmas justificativas dadas para o caso de empate são cabíveis para o gol qualificado. Entre -0,5 e 1 desvio padrão as probabilidades e intervalos de confiança estimados se mantém constantes, com uma probabilidade de aproximadamente 0,10, indicando que entre esses valores de diferença de qualidade as probabilidades de uso do gol qualificado são iguais.

Nas duas situações percebe-se que os intervalos de confiança são maiores para os valores mais extremos da diferença de qualidade. Isso acontece porque ocorrem menos confrontos com tais valores de diferença.

Inicialmente, optou-se por modelos não paramétricos para as duas regressões por causa falta de conhecimento em relação aos formatos das funções. Contudo, percebe-se que este é, de fato, um método adequado para estimar estas probabilidades, uma vez que as curvas não são monótonas, característica de outros métodos de regressão, como logística e linear. Outra alternativa seria a utilização de métodos de *splines*.

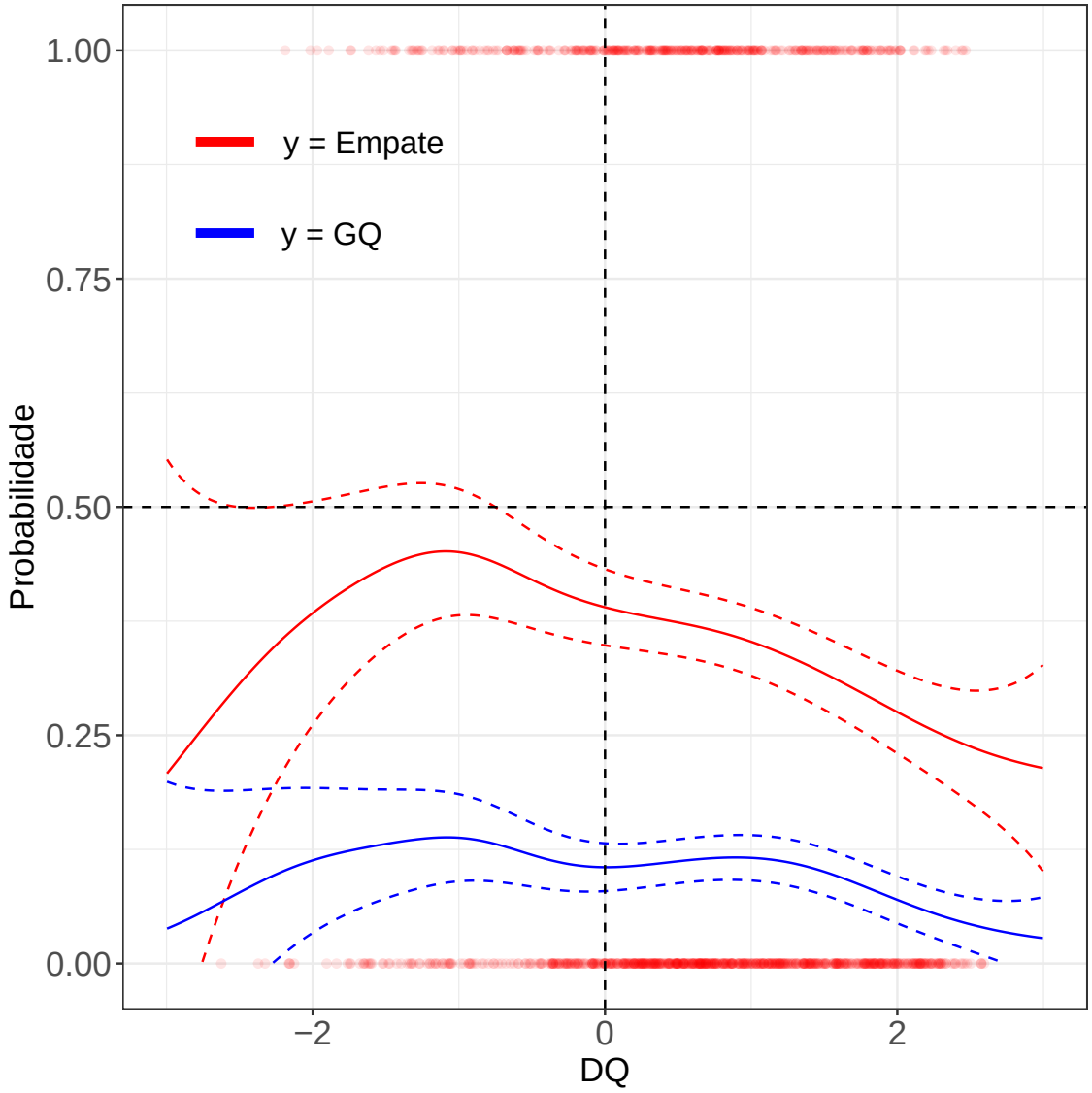


Figura 7.1: Probabilidades estimadas de um confronto terminar empatado ou ser definido por gol qualificado

7.2 Resultado do Primeiro Jogo

Apesar de o segundo jogo ser popularmente considerado o mais decisivo, sabe-se que o resultado da primeira partida tem grande importância para a definição final do confronto, essa seção tem o propósito de avaliar essa influência. A Tabela (7.1) apresenta os percentuais de classificação do mandante para cada possível resultado do primeiro jogo, desconsiderando as diferenças de qualidade. Ressalta-se que, na notação deste trabalho, como foi definido na Seção 1.2, o mandante é o time que tem o mando de campo do segundo jogo, ou seja, no primeiro jogo, objeto de estudo nesta seção, essa equipe joga fora de casa.

Nos casos em que o time visitante, que jogou o primeiro jogo em seu estádio, sofreu uma derrota ou empate, as proporções de classificação dessa equipe (respectivamente 7,91% e 26,06%) são significativamente menores do que 50% (p-valor $<$

0,0001 para as duas situações). Ou seja, o time mandante do segundo jogo construiu uma vantagem, pois fez um bom resultado na partida que disputou fora de seu estádio. A proporção de classificação do mandante em caso de derrota – ou seja, o mandante do segundo jogo venceu a primeira partida –, 92,09%, é significativamente maior do que em caso de empate, 73,94%, (p-valor $<$ 0,0001). Por outro lado, se o visitante venceu o jogo em sua casa, é significativamente mais provável que ele vença o confronto do que o time mandante (p-valor $<$ 0,0001).

Tabela 7.1: Percentual de classificação do mandante pelo resultado do primeiro jogo

Resultado	CM	CV
Derrota	92,09	7,91
Empate	73,94	26,06
Vitória	35,76	64,24

Então, com o objetivo de analisar, de forma exploratória, a probabilidade de classificação em cada possível resultado do primeiro jogo, levando-se em conta a qualidade das equipes, foi estimada uma regressão logística utilizando como resposta a classificação do mandante e como variáveis explicativas o resultado do primeiro jogo do confronto e a diferença da qualidade entre os dois times participantes. Para o resultado do primeiro jogo, que é uma variável categórica politômica, foram criadas duas indicadoras, uma representando empate e outra vitória do time visitante, que jogou este jogo em seu domínio, deixando como referência o resultado de derrota.

Como resultado do modelo tem-se que as duas variáveis são significativas para explicar a resposta (p-valores $<$ 0,0001), porém a interação entre as indicadoras e a diferença de qualidade não (p-valores 0,359 e 0,230). As probabilidades estimadas pelo modelo estão representadas na Figura (7.2), em que o eixo horizontal dispõe as diferenças de qualidade entre os times participantes dos confrontos, sendo que o 0 representa ausência de diferença, ou seja, quanto mais próximo desse valor, maior é a equivalência nas qualidades dos times, conforme descrito na Seção 3.2. Cada ponto no gráfico representa um confronto. Portanto, a posição horizontal de cada ponto representa a diferença de qualidade dos times desse confronto. Os que foram vencidos pelo mandante estão dispostos na parte de cima, pois são os eventos de *sucesso*. Por outro lado, os que estão na linha do 0 são aqueles cuja classificação foi do visitante.

A linhas pontilhadas em preto estão no gráfico apenas para facilitar a visualização, a linha horizontal indica onde a probabilidade é 0,5 enquanto a vertical mostra a diferença de qualidade igual a 0. Então, um ponto que está à direita da linha horizontal (valores positivos) e na parte superior do gráfico, representa um confronto em que o time mandante era melhor e se classificou, por exemplo, Inter x Sampaio Correa em 2017 que o Inter venceu e era 1,60 desvio padrão melhor.

Conforme a legenda do gráfico, a curva azul apresenta as probabilidades de classificação do mandante nos confrontos em que o resultado do primeiro jogo foi derrota do time que jogava em casa naquela partida, ou seja, o time mandante do confronto venceu, situação que aconteceu 278 vezes. As curvas verde e vermelha representam as disputas em que os resultados foram, respectivamente, empate, em 376 confrontos, e vitória do time que jogava em casa o primeiro jogo, em 439 disputas. As linhas centrais, representam as estimativas pontuais das probabilidades, enquanto as linhas pontilhadas representam os intervalos de confiança para as respectivas estimativas.

A amplitude dos intervalos indica a precisão das estimativas para as probabilidades em cada tipo de classificação. Quanto mais observações, maior a precisão e, consequentemente, menor o intervalo de confiança.

O gráfico permite observar que, em todos os casos, a probabilidade de classificação do mandante aumenta à medida que a diferença de qualidade fica mais positiva. Isso acontece porque, quanto melhor for o time mandante e pior o visitante (do confronto), maior é a probabilidade do mandante vencer, para todos possíveis resultados do primeiro jogo. Quando as qualidades são iguais as probabilidades de classificação do mandante são, respectivamente, 0,8373, 0,6869 e 0,2850 para derrota, empate e vitória.

Os três intervalos de confiança têm pequenas amplitudes. No entanto, é visível que, para diferenças muito negativas, o intervalo é mais preciso em caso de vitória do que em empates ou derrotas, indicando que esse é resultado é mais comum. Isso acontece porque são as situações em que o time visitante é muito melhor, então eles costumam vencer o primeiro jogo, que acontece em seus domínios.

As curvas, mesmo os intervalos de confiança, não se cruzam ou sobrepõem, isso indica que, para qualquer diferença de qualidade, a maior probabilidade de classificação do mandante do confronto acontece em caso de derrota, ou seja, quando o mandante venceu o primeiro jogo. Em seguida, tem-se que as probabilidades nos confrontos em que a primeira partida terminou empatada são maiores do que quando o time visitante venceu a primeira partida, que jogou em seu estádio.

Portanto, as análises desta seção ressaltam a importância do primeiro jogo para a classificação dos times na Copa do Brasil. Nos casos em que o mandante do confronto consegue construir um resultado favorável na primeira partida ele leva uma grande vantagem para o segundo jogo. Por outro lado, evidenciou-se que para o time visitante é imprescindível obter um bom resultado ao jogar em seu estádio.

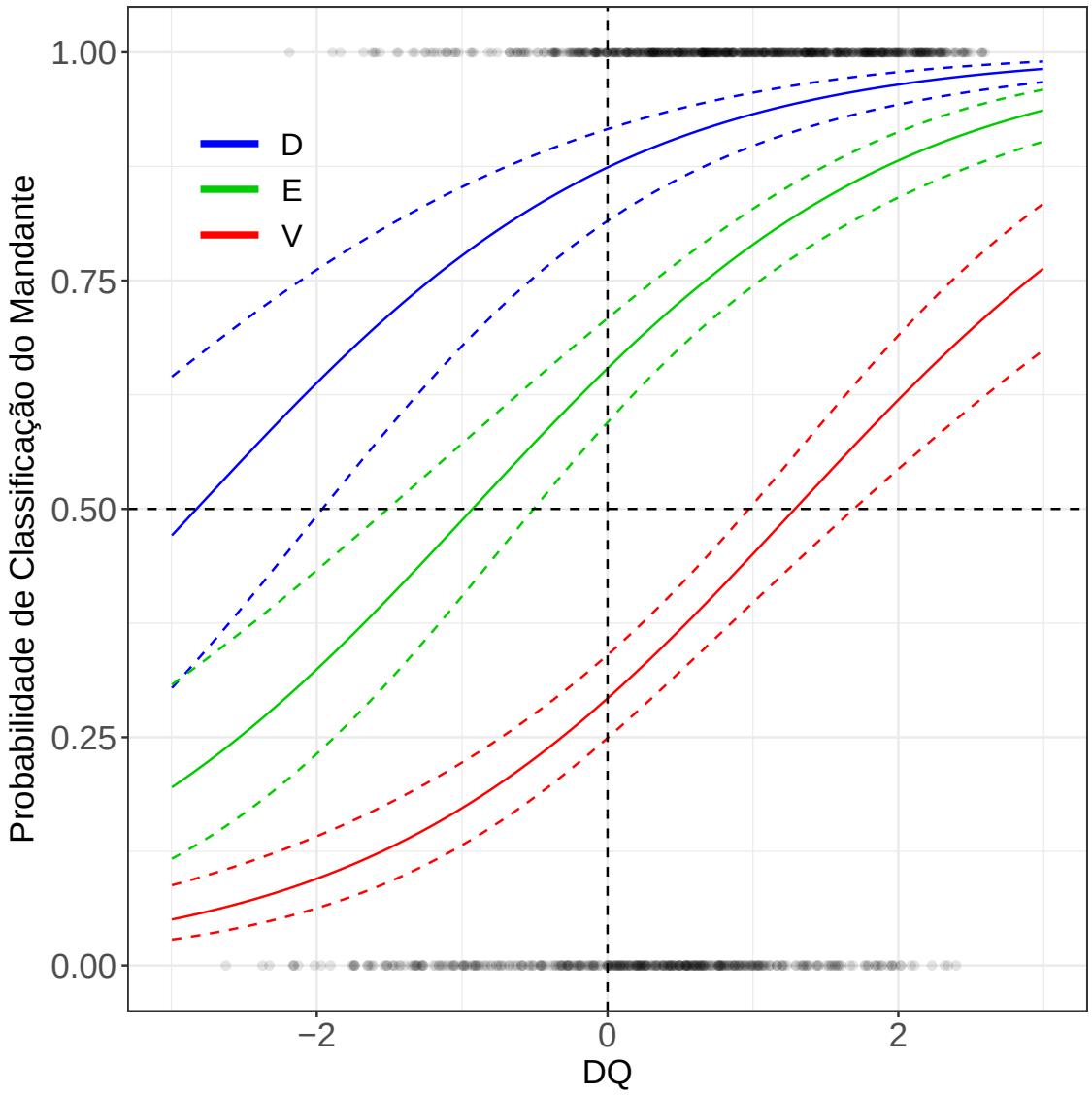


Figura 7.2: Probabilidades estimadas de classificação do mandante dados os resultados do primeiro jogo

8 Classificação na Copa do Brasil

Para realizar o objetivo principal deste trabalho, estimar a probabilidade de classificação do mandante na Copa do Brasil, relacionando com a diferença de qualidade e o critério de decisão utilizado, foi ajustado um modelo de regressão logística. Para tanto, a variável resposta é a classificação do mandante e os componentes explicativos considerados no modelo são a diferença de qualidade (x_1) e o tipo de decisão que é uma variável categórica politômica. Por isso, foram criadas três indicadoras, para saldo de gols, gol qualificado e pênaltis (respectivamente, x_2, x_3 e x_4), mantendo a classificação por pontuação como categoria de referência. Além disso, foram incluídas no ajuste as interações entre a diferença de qualidade e cada uma das indicadoras.

Nas análises descritivas (Capítulo 5), especialmente nas figuras 5.3 e 5.4, foi identificada uma possível relação entre os resultados e a classificação ao longo dos anos e das fases da competição. Contudo, como foi descrito na Seção 3.2 e também discutido na Seção 6.1, acredita-se que a variável diferença de qualidade capta as variações entre os anos e as fases. Por isso, julgou-se desnecessário incluir a fase em disputa, bem como o ano, entre as variáveis explicativas no modelo.

Preliminarmente, foram ajustados dois outros modelos. O primeiro, uma regressão logística com o banco de dados reduzido, contendo apenas os confrontos das fases finais. No entanto, os resultados foram semelhantes aos obtidos pela regressão que será descrita a seguir, portanto concluiu-se que não há necessidade de detalhar ambas, optou-se pelo banco completo porque inclui mais informações. O segundo modelo foi o mesmo estimado pela regressão logística, porém utilizando a abordagem não paramétrica. Os resultados também ficaram próximos aos da regressão logística, indicando que o ajuste pelo método paramétrico é adequado e também descartando a necessidade de descrever essa abordagem. Foi definido que a regressão logística seria utilizada pois facilita a interpretação das análises, além de que o método não paramétrico apresenta dificuldades para estimar o modelo com as interações.

Então, o modelo final ajustado é:

$$\mathbb{P}(y_i = 1|\mathbf{x_i}) = \pi(\mathbf{x_i}) = \frac{\exp(g(\mathbf{x_i}))}{1 + \exp(g(\mathbf{x_i}))}, \quad i \in \{1, \ldots, n\}$$

Onde a função de ligação logito se dá por:

$$g(\mathbf{x_i}) = \ln\left(\frac{\pi(\mathbf{x_i})}{1 - \pi(\mathbf{x_i})}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \\ + \beta_5 x_{1i} x_{2i} + \beta_6 x_{1i} x_{3i} + \beta_7 x_{1i} x_{4i}, \quad i \in \{1, \ldots, n\}$$

Este modelo é útil para compreender o fenômeno da vantagem de jogar em casa a segunda partida de um confronto mata-mata, bem como algumas características que podem ter influência sobre isso. Contudo, esse modelo não permite predizer resultados futuros, uma vez que utiliza a informação sobre qual critério de decisão foi utilizado, que só está disponível após o término do confronto. Para fazer uma predição futura o modelo só pode ser explicado pela diferença de qualidade.

Os coeficientes estimados e suas respectivas significâncias, obtidos pelo teste de Wald, após a estimação do modelo logístico, estão descritos na Tabela (8.1). Verifica-se que apenas a diferença de qualidade é individualmente significativa. Contudo, as interações da diferença de qualidade com cada tipo de classificação são significativas.

Tabela 8.1: Coeficientes da regressão pelo teste de Wald

	Estimativa	Erro Padrão	z	p-valor
Intercepto	0,1229	0,1047	1,1730	0,2408
DQ	1,1087	0,1058	10,4788	<0,0001
SG	0,0476	0,2001	0,2381	0,8118
GQ	-0,4034	0,2382	-1,6940	0,0902
PN	-0,2516	0,2637	-0,9541	0,3400
DQ*SG	-0,4125	0,2034	-2,0282	0,0425
DQ*GQ	-1,0026	0,2349	-4,2675	<0,0001
DQ*PN	-1,1038	0,2429	-4,5450	<0,0001

A Tabela (8.2) apresenta os coeficientes estimados e suas respectivas significâncias, obtidos pelo teste da razão de verossimilhança. Em contraponto com o teste de Wald, verifica-se que, além da diferença de qualidade, as indicadoras de gol qualificado e pênaltis também são individualmente significativas. Entretanto, bem como o teste anterior, as verossimilhanças indicam que as três interações são significativas.

Tabela 8.2: Coeficientes da regressão pelo teste de razão de verossimilhança

	logL	p-valor
DQ	134,965	<0,0001
SG	0,465	0,4955
GQ	17,585	<0,0001
PN	8,328	0,0039
DQ*SG	3,933	0,0473
DQ*GQ	17,028	<0,0001
DQ*PN	19,306	<0,0001

Dessa forma, tem-se evidências de que a probabilidade de classificação do mandante, dado cada tipo de classificação, muda de acordo com a diferença de qualidade entre os times do confronto, conforme discutido na Seção 6.2. Portanto, esses resultados demonstram que todas as variáveis do modelo são importantes para explicar a probabilidade de classificação do time mandante na Copa do Brasil.

As tabelas Tabela (8.1) e Tabela (8.2) possibilitam a verificação das significâncias individuais das variáveis indicadoras do tipo de classificação. Contudo, não é possível identificar se a variável tipo de classificação é significativa como um todo, ou seja, se o tipo de classificação, independente de qual, é significativo para determinar o vencedor de um confronto. A fim de conferir isso, foram feitos dois testes

comparando as *deviances* dos modelos, os resultados estão dispostos na Tabela (8.4). Para tanto, três modelos foram ajustados: o primeiro é o completo, descrito anteriormente; no segundo, foram mantidas todas as variáveis, mas as interações foram retiradas; já o terceiro modelo considera como covariável apenas a diferença de qualidade. As *deviances* e os graus de liberdade dos os três modelos estão descritos na Tabela (8.3).

Tabela 8.3: *Deviances*

Modelo	*Deviance*	Graus de liberdade
Completo (1)	1236,978	1085
Sem interações (2)	1269,148	1088
Apenas DQ (3)	1293,048	1091

Interpretando os resultados da Tabela (8.4), observa-se que o modelo 1 é significativamente mais informativo que o modelo 2 e que este, por sua vez, é significativamente mais informativo que o modelo 3. Portanto, retirar as interações reduz a verossimilhança do modelo, e retirar as indicadoras faz com que diminua ainda mais. Logo, pode-se concluir que a variável tipo de classificação, ao ser considerada globalmente, é de fato significativa para explicar a probabilidade de classificação do mandante. A comparação entre os modelos 1 e 3 informa que, conjuntamente, o tipo de classificação e a interação com a diferença de qualidade melhora significativamente a explicação do modelo.

Tabela 8.4: Testes diferenças entre as *deviances*

Modelos	Diferença	p-valor
1 x 2	32,1702	<0,0001
2 x 3	23,9000	<0,0001
1 x 3	56,0702	<0,0001

Além disso, com a finalidade de averiguar uma das hipóteses iniciais deste trabalho (de que a diferença de qualidade é a variável mais importante para a classificação), verificando se esta informação é a que mais agrega ao modelo, foram feitos testes que comparam os valores das áreas sob a curva ROC, o teste de DeLong, computacionalmente disponível pelo pacote pROC (Robin et al., 2011). Isto é, comparando as capacidades preditivas dos modelos, cujos valores estão descritos na Tabela (8.5). Observa-se que para os três primeiros modelos, que incluem a diferença de qualidade, a capacidade preditiva está acima de 70%. Entretanto, ao retirar essa variável do modelo, a capacidade preditiva tem uma grande redução, diminuindo para 58,33%.

Tabela 8.5: Áreas Abaixo da Curva ROC (AUC)

Modelo	AUC
Completo (1)	0,7439
Sem interações (2)	0,7346
Apenas DQ (3)	0,7094
Apenas o tipo de classificação (4)	0,5833

Os resultados dos testes, que permitem identificar a significância da diminuição na capacidade preditiva do modelo, são apresentados na Tabela (8.6). Verifica-se que as interações, o tipo de classificação e a diferença de qualidade são variáveis importantes, pois sua retirada causa uma diminuição significativa na área abaixo da curva. No entanto, a inclusão que resulta em maior acréscimo na capacidade é a diferença de qualidade, corroborando a hipótese inicial de que essa é a variável mais importante para o modelo.

Tabela 8.6: Testes de AUC

Modelos	z	p-valor
1 x 2	1,9264	0,0541
2 x 3	3,4249	0,0006
2 x 4	8,5427	<0,0001

8.1 Probabilidades Estimadas

Tendo em vista a discussão da seção anterior, considera-se que o modelo mais adequado para capturar as peculiaridades do processo gerador dos dados é o modelo completo. As probabilidades estimadas estão representadas na Figura (8.1). Cada ponto no gráfico representa um confronto: nos casos em que o mandante venceu os pontos estão dispostos na linha do 1, já os que estão na linha do 0 são aqueles cujo resultado foi derrota do mandante. A posição horizontal de cada ponto representa a diferença de qualidade dos times desse confronto, portanto quanto mais próximo de 0 um valor estiver, maior é a equivalência nas qualidades dos times, de acordo com o que foi descrito na Seção 3.2.

Com a finalidade de facilitar a visualização, as duas linhas pretas pontilhaadas estão plotadas no gráfico. A horizontal indicando onde a probabilidade é 0,5 enquanto a vertical representa a diferença igual a 0. Por exemplo, um ponto que está à direita da linha horizontal (valores positivos) e na parte inferior do gráfico, representa um confronto em que o time mandante era melhor, mas não se classificou, como Grêmio x Fluminense, em 2015, no qual o Grêmio era 0,23 desvio padrão melhor, porém não conseguiu a vitória. Como descrito na legenda do gráfico, a curva preta apresenta as probabilidades de classificação do mandante nos confrontos que foram definidos por pontos, ou seja, que não necessitaram de critérios de desempate. As curvas vermelha, azul e rosa representam as disputas definidas, respectivamente, por saldo de gols, gol qualificado e pênaltis.

O gráfico permite observar que, em todos os casos, a probabilidade de classificação do mandante aumenta à medida que a diferença de qualidade fica mais positiva. Em outras palavras, indica a veracidade de uma das hipóteses iniciais de pesquisa, quanto melhor for o time mandante e pior o visitante, maior é a probabilidade do mandante vencer, independente do tipo de decisão. As probabilidades têm maior variação nos casos em que a decisão se deu por pontos ou saldo. A disputa de pênaltis é o tipo de confronto que tem a menor variação, com probabilidade média de 0,4679, ou seja, para esse critério a diferença de qualidade não tem tanta influência, apesar de a interação ser significativa.

Analisando as probabilidades à medida que a diferença de qualidade fica mais positiva, ou seja, quanto mais qualidade tem o mandante e menos tem o visitante,

verifica-se que, para o mandante, o critério que leva às maiores probabilidades de classificação é a disputa de pênaltis até o ponto em que a diferença seja de -0,45 desvio padrão. Então a maior probabilidade é pelo saldo de gols, até que o mandante seja 0,12 desvio padrão melhor. Acima desse valor o critério com maior probabilidade para o mandante é a pontuação.

Em contrapartida, para o visitante, as maiores probabilidades acontecem quando a decisão é por pontos enquanto este time é muito melhor do que o mandante. Quando a diferença é de -0,41 desvios padrão, a menor probabilidade de classificação do mandante é em caso de uso do gol qualificado, isso se mantém até que a diferença de qualidade seja 1,51 desvios padrão. Então a disputa de pênaltis se torna o critério em que o visitante se classifica com maior probabilidade.

Dito isso, percebe-se que o critério do saldo de gols não é o mais favorável, ou seja, não leva as maiores probabilidades de classificação do time visitante para nenhuma diferença de qualidade. Por outro lado, o gol qualificado nunca é o melhor para o mandante, pois as probabilidades são menores. Ressalta-se que isso não significa que esses times não podem utilizar esses critérios, apenas que sempre tem outro que proporciona uma probabilidade de classificação maior. Apesar de os times não escolherem o critério utilizado, no máximo é podem montar uma estratégia de jogo que visa determinado tipo de decisão, porém não depende apenas da equipe.

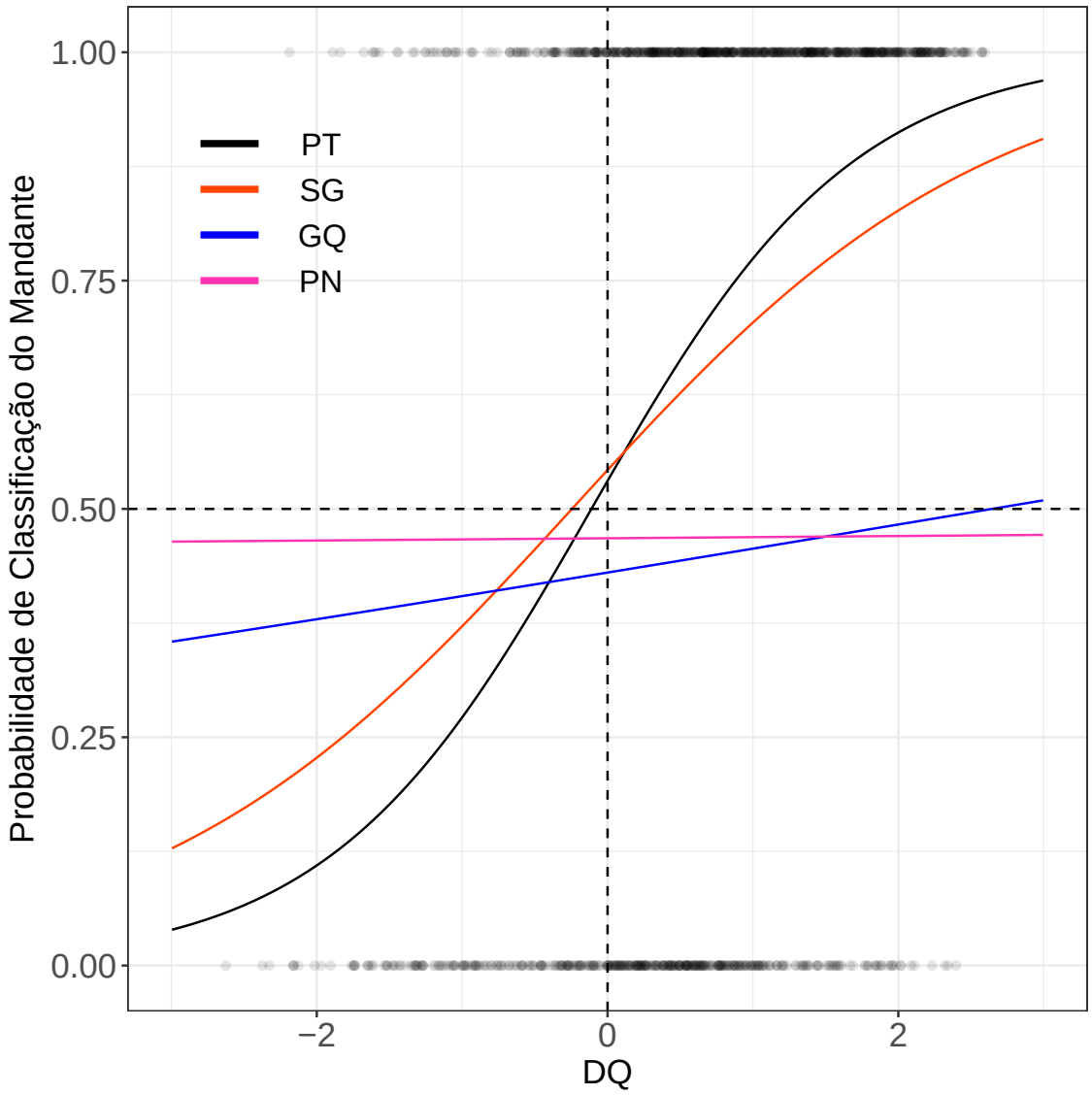


Figura 8.1: Probabilidades estimadas de classificação do mandante

Para complementar a análise, a Figura (8.2) apresenta um quadro com quatro gráficos, um para cada curva exposta na figura anterior, estimadas pelo modelo completo. As curvas foram separadas em quatro gráficos somente para facilitar a visualização. Em cada gráfico, a linha central representa as estimativas pontuais da probabilidade, enquanto as linhas pontilhadas representam os intervalos de confiança correspondentes. Conforme a Tabela (5.3), o saldo de gols foi utilizado em 196 confrontos, o gol qualificado em 113 e a disputa de pênaltis em 79, o que influencia nas precisões dos intervalos de confiança. Os outros 705 confrontos foram decididos pela pontuação e, por isso, não utilizaram nenhum critério de desempate.

Conforme o esperado, verifica-se que as estimativas para as probabilidades de classificação nos confrontos definidos por pontos são as mais precisas, pois o intervalo de confiança tem pequena amplitude ao longo de toda a curva. Para o saldo de gols o intervalo é um pouco mais amplo, pois essa situação tem menos observações. Nesses dois casos, apesar de a estimativa pontual ser cerca de 0,54 quando a diferença de qualidade é nula, o intervalo de confiança inclui o 0,5, portanto não se

pode dizer que o mandante tenha de fato uma vantagem. Por outro lado, para diferenças de qualidade iguais a 0, a probabilidade de classificação do mandante é igual à 0,42 se o confronto foi decidido por gol qualificado e 0,46 se a disputa de pênaltis foi utilizada. Para esses critérios o intervalo de confiança para a probabilidade do mandante vencer o confronto também inclui o 0,5. Portanto, quando os dois participantes do confronto têm qualidades equivalentes, eles têm iguais probabilidades de classificação, independente do critério de decisão.

Nas disputas definidas por gol qualificado ou pênaltis, as estimativas têm menos precisão, por causa do pequeno número de observações. Nota-se que os intervalos de confiança são amplos para todas diferenças de qualidade. Contudo, percebe-se que, para valores da diferença de qualidade próximos de 0, os intervalos de confiança têm as menores amplitudes. Isso acontece porque a maioria dos confrontos que utilizaram esses critérios tinham pequenas diferenças de qualidade, ou seja, quando os times têm qualidades parecidas. Apesar da grande amplitude, essas estimativas trazem informações relevante, pois o intervalo contém o valor igual a 0,5, para todas as diferenças de qualidade. Então, confirmando uma das hipóteses iniciais deste trabalho, pode-se concluir que, para os dois critérios, os times têm as mesmas probabilidades de classificação, independente da diferença de qualidade.

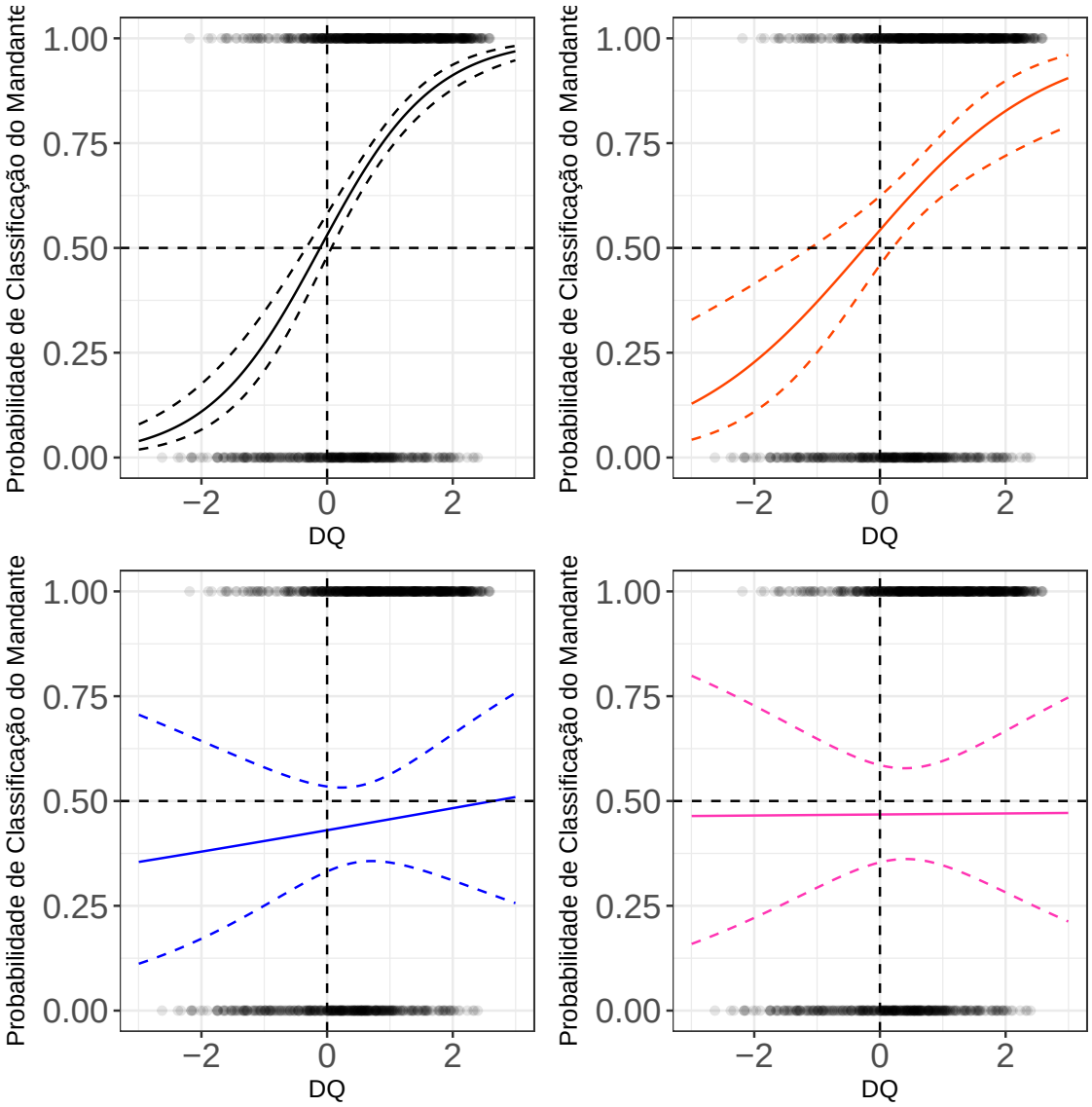


Figura 8.2: Probabilidades estimadas de classificação do mandante e intervalos de confiança

8.1.1 Exemplos

A fim de ilustrar as probabilidades estimadas serão dados alguns exemplos de confrontos que ocorreram na Copa do Brasil. Em cada gráfico, a linha verde apresenta as probabilidades estimadas pela regressão restrita, que considera somente a diferença de qualidade. Por outro lado, a linha rosa representa o modelo completo apresentado anteriormente, ou seja, a regressão que leva em conta também o tipo de classificação e as interações com a diferença de qualidade. Os pontos azuis indicam as probabilidades de classificação estimadas por cada regressão, já o ponto vermelho indica o verdadeiro resultado do confronto.

Exemplo 1

No confronto entre Grêmio e Fluminense, que ocorreu em 2015, o segundo jogo aconteceu em Porto Alegre, então o Grêmio teria a vantagem de ser o mandante, além de ser 0,2255 desvio padrão melhor, de acordo com o índice da CBF. Com isso, esperaria-se uma vitória do time gaúcho, a Figura (8.3) mostra as probabilidades estimadas. Observa-se que, tendo conhecimento somente sobre a diferença de qualidade entre os times, a probabilidade de classificação do Grêmio é de 0,5659. Contudo, ao informar ao modelo o tipo de classificação utilizado (gol qualificado) essa probabilidade passa a ser de 0,4362. De fato, o Fluminense obteve a classificação.

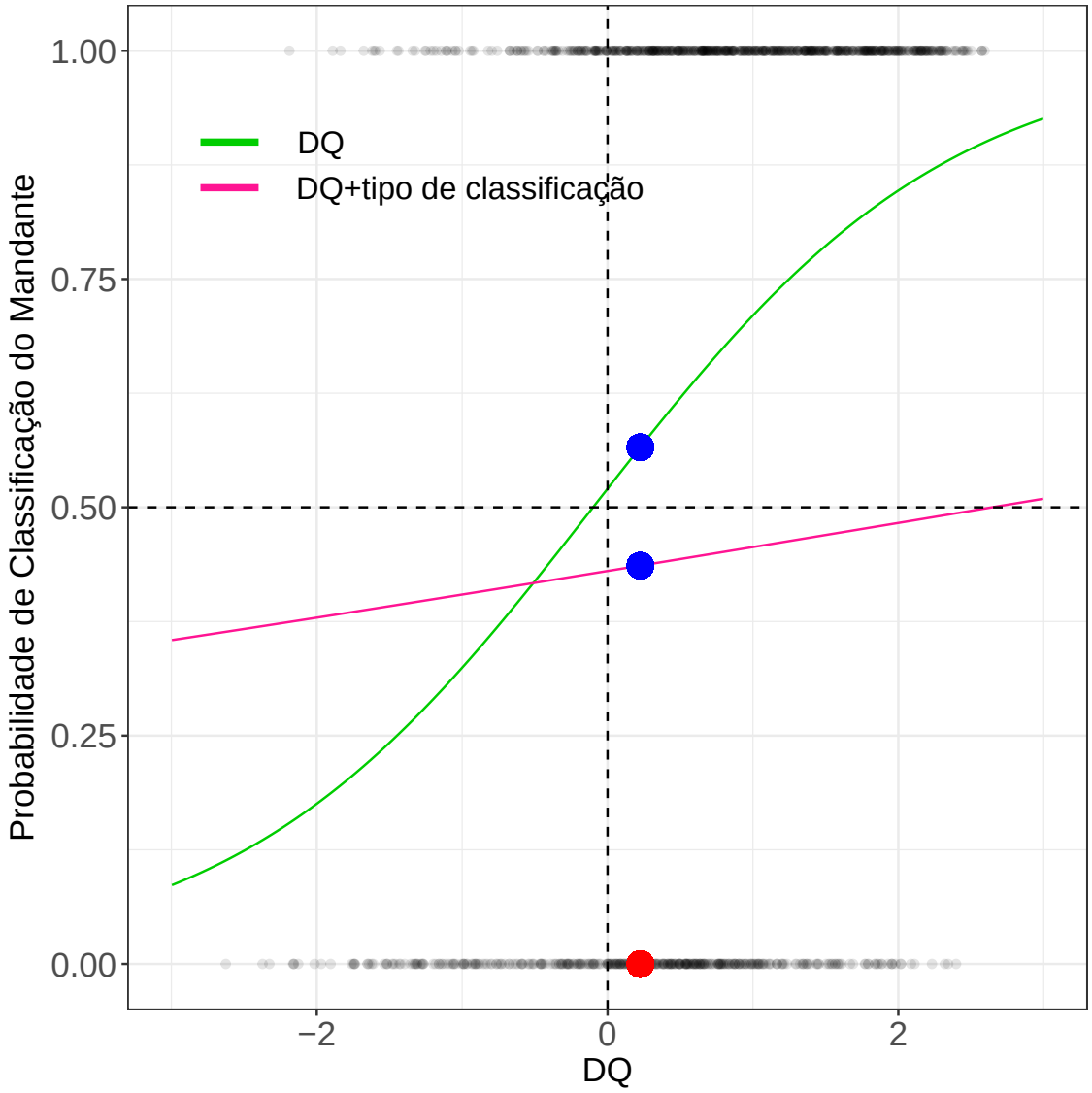


Figura 8.3: Exemplo 1

Exemplo 2

Outro exemplo é o confronto entre Goiás e Ríver-PI, em 2016, nesse caso a diferença de qualidade na época era de 1,5988 desvio padrão. Logo, espera-se que o time goiano tenha uma grande vantagem, uma vez que era muito melhor e decidiu a disputa no seu estádio. Considerando apenas isso, o modelo estima uma probabilidade de 0,7996 de classificação do mandante, como pode ser visto na Figura (8.4). Entretanto, quando o modelo é informado que a decisão foi na disputa de pênaltis a probabilidade passa a ser de 0,4698. O ponto vermelho indica que o Ríver venceu.

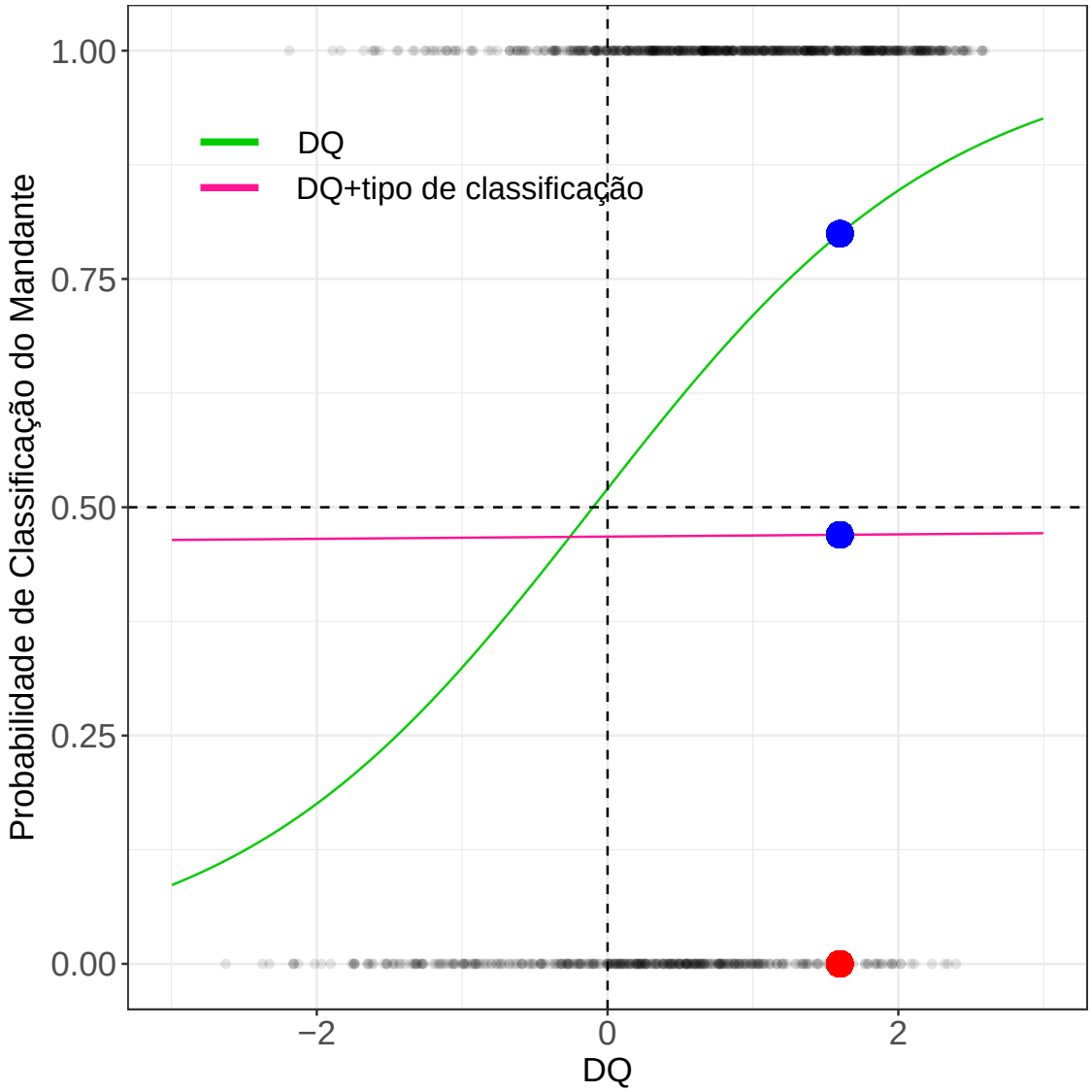


Figura 8.4: Exemplo 2

Exemplo 3

Em 2014, o confronto entre Bahia e Corinthians teve diferença de qualidade igual à -0,9907, valor negativo pois o time paulista, que é o visitante, tinha mais qualidade na época. Na Figura (8.5) é possível visualizar as probabilidades estimadas que são 0,3261 e 0,3730 respectivamente para o modelo sem e com o tipo de classificação. Portanto, percebe-se que, nesse caso, os dois modelos obtiveram estimativas próximas e certeiras, uma vez que o Corinthians de fato venceu o confronto pelo saldo de gols.

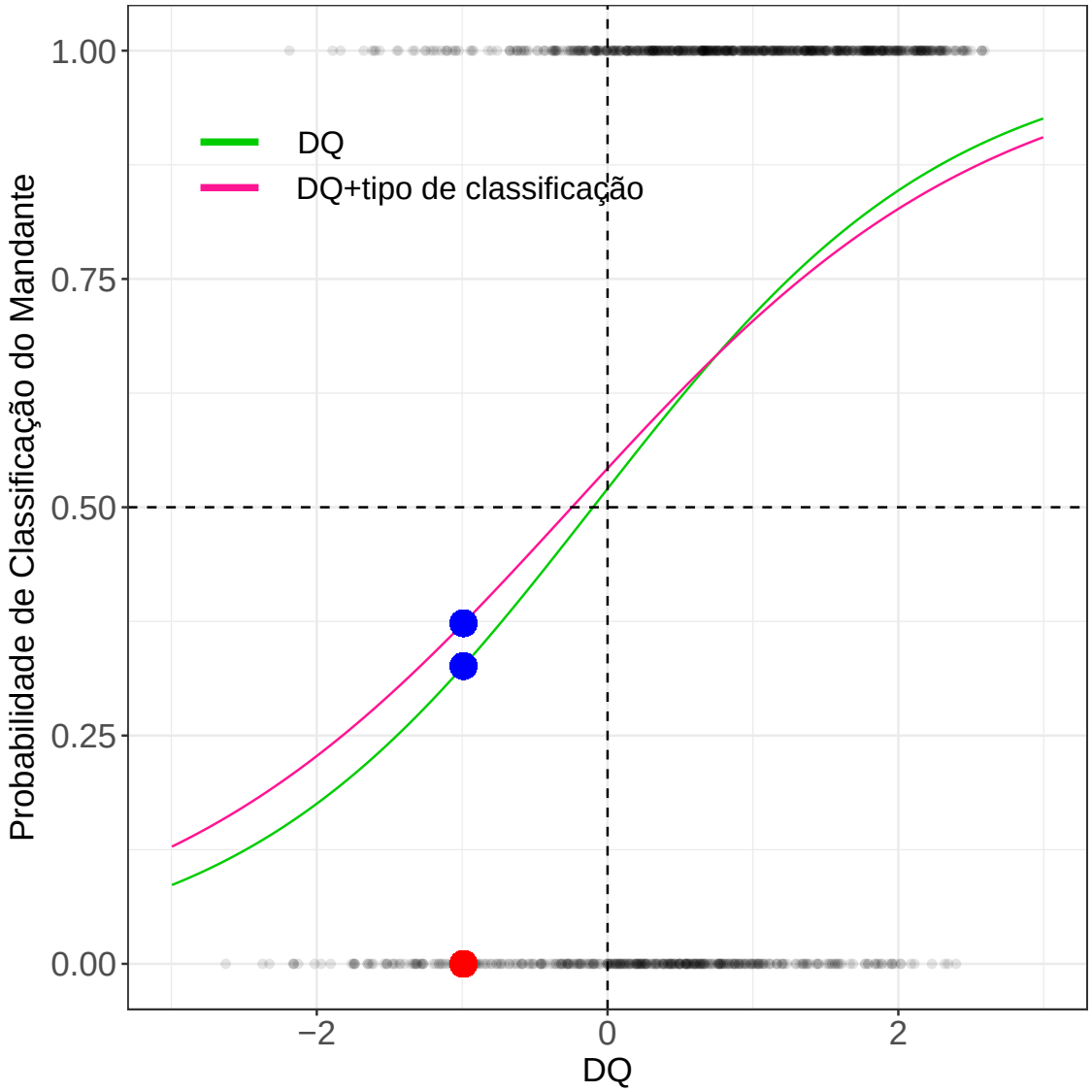


Figura 8.5: Exemplo 3

Exemplo 4

Um exemplo de confronto que foi decidido por pontuação é Internacional x Sampaio Corrêa, ocorrido em 2017. A diferença de qualidade entre esses times era 1,5983 e o segundo jogo foi no Rio Grande de Sul, por isso espera-se uma grande vantagem para o Inter. Como pode ser observado na Figura (8.6), a probabilidade estimada, considerando somente a diferença de qualidade, é de 0,7995; quando informado que a decisão foi por pontos a probabilidade estimada de o Inter se classificar é ainda maior, 0,8693. Conforme o esperado, o Inter venceu.

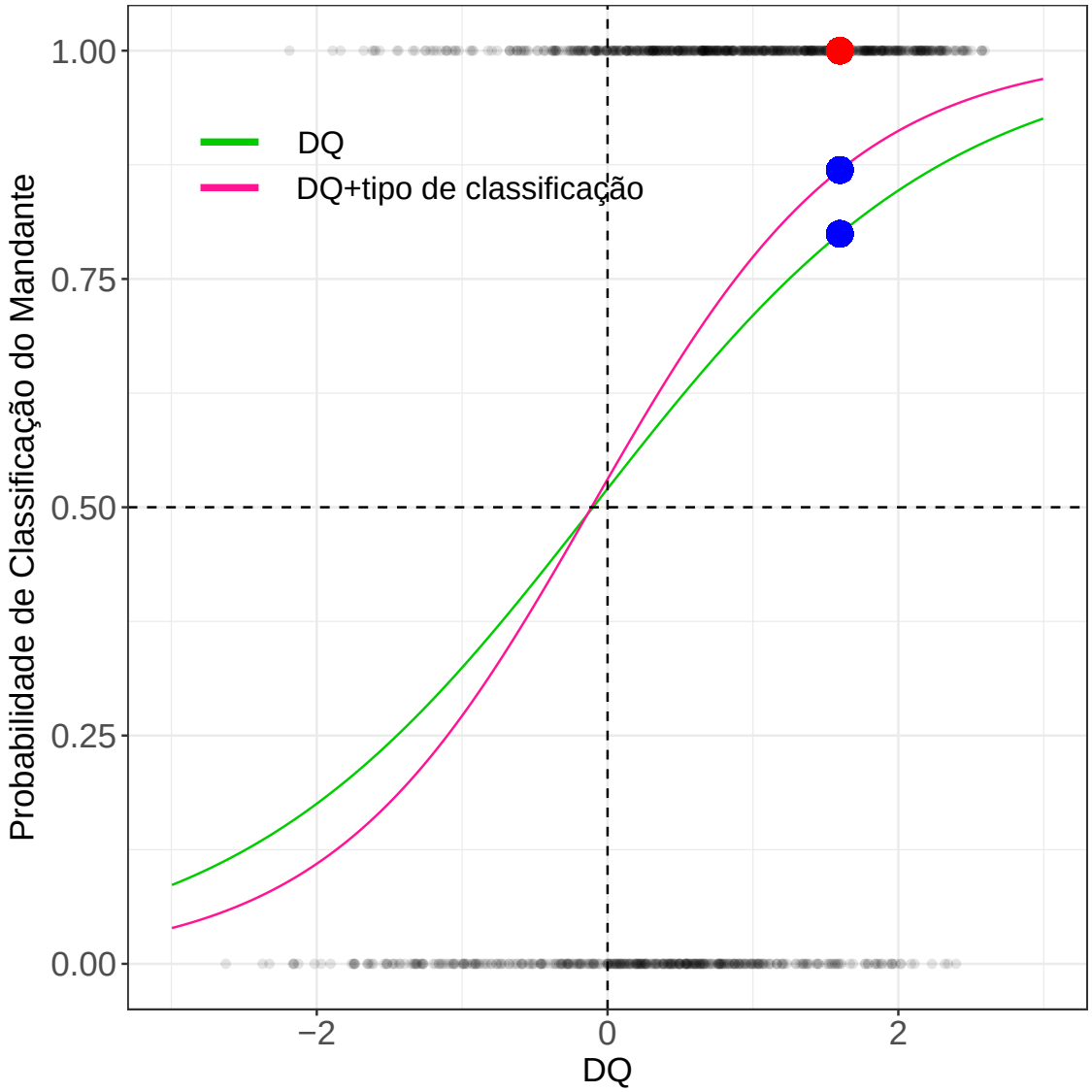


Figura 8.6: Exemplo 4

Conclui-se, então, que o modelo que não considera o tipo de classificação estima probabilidades mais próximas dos casos em que os confrontos são decididos por pontuação. Isso acontece porque a maior parte (64,59%) das disputas na Copa do Brasil utiliza esse critério, influenciando na regressão. Contudo, antes de começar o confronto não se tem conhecimento de qual será o tipo de decisão utilizado. Portanto, o modelo, inicialmente, tem boas estimativas, porém, à medida que mais critérios são necessário para desempatar o confronto, capacidade preditiva do modelo diminui.

8.2 Ajuste do Modelo

8.2.1 Testes de Ajuste

Nesta seção avalia-se a qualidade de ajuste do modelo proposto, como forma de substanciar os resultados obtidos. A Tabela (8.7) apresenta os resultados obtidos pelos testes descritos na Seção 3.5.1. Verifica-se que, pelos testes de Hosmer e Lemeshow e Padronizado de Pearson, o modelo está bem ajustado, indicando que os resultados podem ser utilizados.

No entanto, o teste de Stukel rejeita a hipótese nula, apontando que o ajuste não está adequado. Esse teste é conhecido por ser melhor em identificar problemas de interações ou assimetria, bem como desvios na função logística. Como todas as interações possíveis foram incluídas no modelo e têm coeficientes significativos, entende-se que o ajuste esteja correto nesse sentido. Pode-se testar se o problema é causado por desvios na função logística incluindo um termo no modelo para as probabilidades estimadas elevadas ao quadrado e então faz-se o teste da razão de verossimilhança. Contudo, essa inclusão provocou uma diminuição na *deviance* do modelo, porém, não de forma significativa (p-valor = 0,4905). Além disso, foi descrito anteriormente que, em um estudo preliminar, ajustou-se uma regressão não paramétrica, a qual não faz suposições sobre a função de ligação, e esta obteve resultados muito próximos do modelo logístico. Por isso, conclui-se que a função de ligação utilizada pelo modelo é a correta.

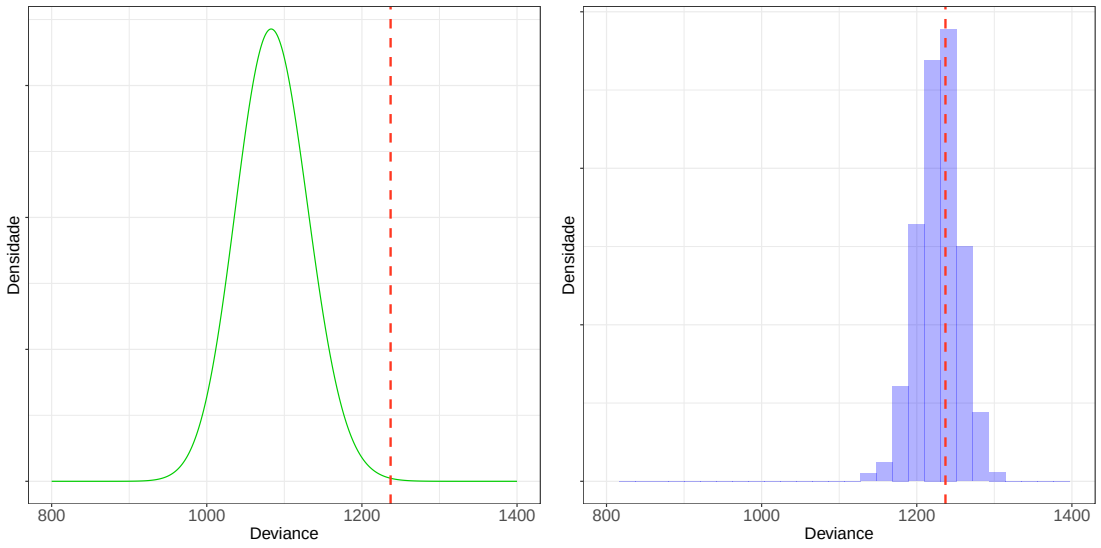
Então, entende-se que o teste de Stukel rejeitou a hipótese de modelo bem ajustado por problemas de assimetria. Havia-se previsto que isso poderia acontecer, uma vez que, na Seção 6.2, foi demonstrado que a diferença de qualidade é positiva na maioria dos casos, independente do time classificado. Possivelmente por causa das fases iniciais, nas quais, em alguns anos, não houve sorteio e o time mandante era o melhor time. Isso pode gerar um viés, aumentando as probabilidades estimadas, em favor do mandante.

Tabela 8.7: Testes de ajuste

Teste	Estatística do teste	p-valor
Hosmer e Lemeshow	10,0006	0,2650
Padronizado de Pearson	-0,1894	0,8498
Stukel	27,8866	<0,0001
Deviance assintótica	1236,9780	0,0009
Deviance bootstrap	1236,9780	0,3910

Além dos descritos anteriormente, a Tabela (8.7) apresenta os testes que avaliam a *deviance* do modelo. Essa é testada de duas formas, com base na distribuição

assintótica, Qui-quadrado (χ^2) e na simulação por *bootstrap*. A Figura (8.7) permite a visualização desses testes, o Gráfico (a) expõe a curva da distribuição χ^2, enquanto o Gráfico (b) apresenta o histograma das estimativas obtidas pela simulação, nos dois, a linha pontilhada em vermelho indica o valor estimado pelo modelo original. Verifica-se que pelo resultado assintótico a hipótese de modelo bem ajustado é rejeitada (p-valor = 0,0009). Porém, de acordo com os resultados simulados, não há evidências de problemas no modelo (p-valor = 0,3910).



(a) Assintótica (b) Bootstrap

Figura 8.7: *Deviance*

Portanto, apesar de o teste assintótico da *deviance* e o teste de Stukel terem encontrado evidências de que o modelo não está bem ajustado, entende-se que isso se deve à falta de sorteio do mando de campo em um grande número de confrontos das fases iniciais, o que causa a assimetria na diferença de qualidade. No entanto, isso não é considerado um problema; apenas ressalta-se que é possível que as probabilidade sejam superestimadas, embora em um estudo preliminar em que foi feito o ajuste do modelo com o banco reduzido (contendo apenas as fases finais e, por consequência, diminuindo o número de confronto sem sorteio), as probabilidades estimadas foram próximas às do modelo apresentado neste trabalho. Na análise preliminar, o teste de Stukel também rejeitou a hipótese de modelo bem ajustado. Contudo, não rejeitou quando utilizados somente os confrontos após a fase oitavas de final, que foram, em maioria, sorteados. Ademais, nessa análise obtiveram-se probabilidades semelhantes às apresentadas neste trabalho. Optou-se por utilizar a análise com o banco que contém todas fases para não perder informações, uma vez os bancos reduzidos têm um número muito menor de confrontos, principalmente de casos em que o gol qualificado e a disputa de pênaltis foram utilizados.

Multicolinearidade

Um possível problema de ajuste, é a multicolinearidade entre as variáveis explicativas, que pode causar impactos sobre as estimativas do modelo. A multicolinearidade é definida como uma relação linear entre uma variável explicativa e as demais

(PortalAction, 2017a). Sabe-se que isso pode ocorrer no modelo ajustado pois, na Seção 6.3, foi identificado que há relação entre as variáveis tipo de classificação e diferença de qualidade. Além disso, como o tipo de classificação é uma variável categórica com quatro categorias, é impossível colocar todas no modelo, justamente pela colinearidade perfeita que existe entre elas. Por isso, são criadas três variáveis indicadoras, deixando uma categoria – no caso, a classificação por pontos – como referência.

Uma forma de verificar a presença dessa alta colinearidade é o cálculo do VIF (Fator de Inflação de Variância). Esse valor indica quanto a variância dos coeficientes estimados pela regressão aumenta por causa de colinearidade. Existem duas regras populares para determinar se há problemas ou não: uma diz que valores $VIF > 10$ causam impactos, outra que $\sqrt{VIF} > 2$. Verifica-se na Tabela (8.8) que não há problemas desse tipo.

Tabela 8.8: Multicolinearidade

Variável	VIF	$\sqrt{VIF}$
DQ	1,8628	1,3649
SG	1,3440	1,1593
GQ	1,3693	1,1702
PN	1,2402	1,1136
DQ*SG	1,6104	1,2690
DQ*GQ	1,5511	1,2454
DQ*PN	1,4045	1,1851

Linearidade

Outra questão importante para o ajuste do modelo é a linearidade dos preditores quantitativos. Isso porque o modelo precisa ajustar corretamente a relação entre as variáveis explicativa e resposta. Caso a relação, a partir da função de ligação, não seja linear, é preciso corrigir esse formato com outros termos no modelo. Então, para o modelo que foi ajustado, é necessário verificar se a relação entre a diferença de qualidade e a classificação do mandante está modelada de forma adequada.

Uma maneira de testar essa relação é pelo Teste do Quartis (Collett, 2003), cujo procedimento é dividir a variável quantitativa em quatro quartis e então ajustar dois modelos, um com a variável como categórica e outro como numérica, que são comparados pelo teste da razão de verossimilhanças. Por esse teste não há diferença significativa (p-valor = 0,3733), indicando que não há problemas na linearidade. A Figura (8.8) plota os coeficientes estimados pela regressão que considera a variável com quatro categorias divididas pelos quartis. Por inspeção visual percebe-se que não há um desvio substancial de uma relação linear.

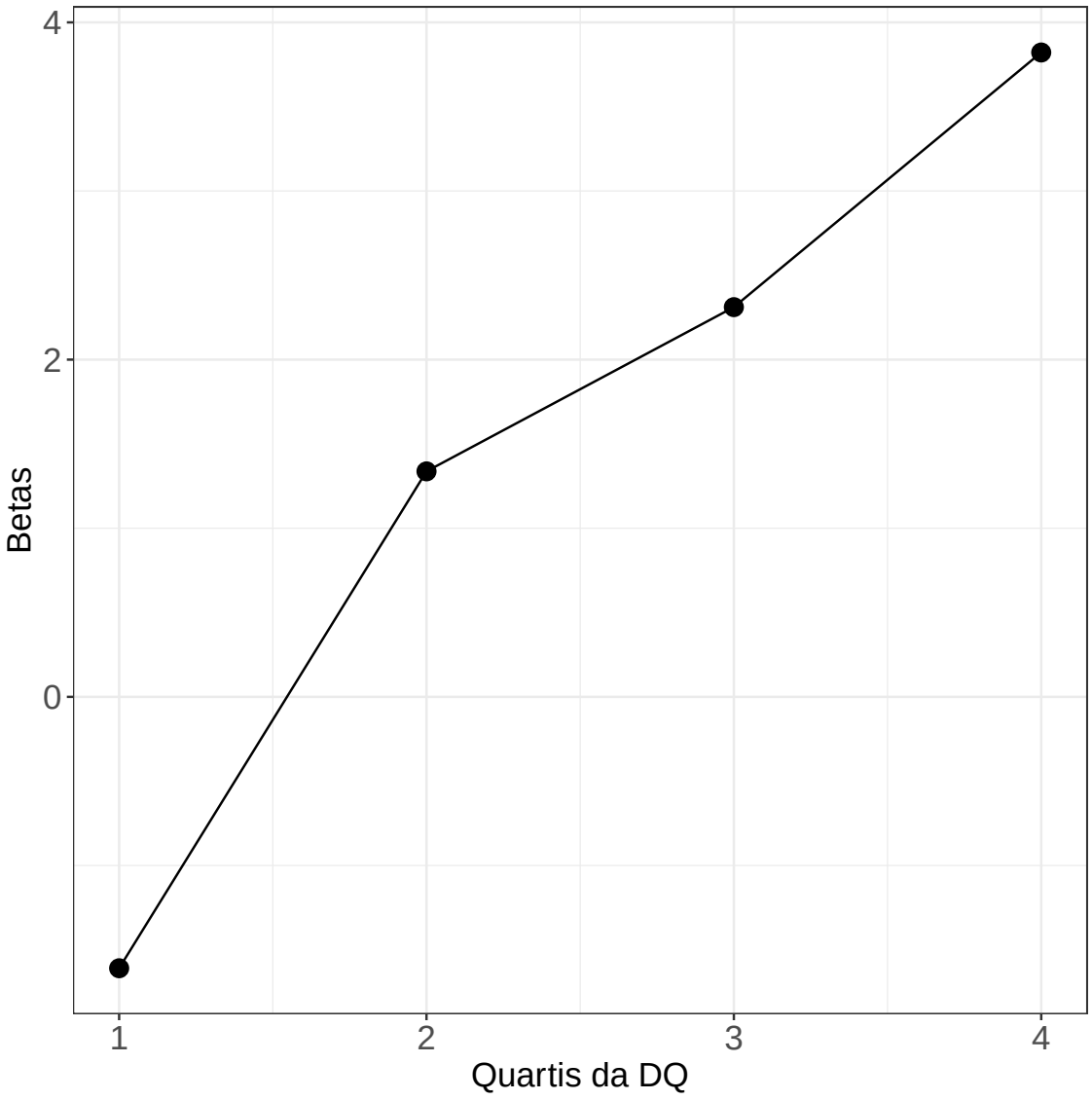


Figura 8.8: Linearidade

Outra forma de testar a linearidade é incluir no modelo termos quadráticos e cúbicos para a diferença de qualidade e então comparar ao modelo original com o teste da razão de verossimilhanças. Pelo teste nem termos quadráticos nem cúbicos trouxeram acréscimos significativos ao modelo (p-valor = 0,8115 e 0,4395).

8.2.2 Análise de Resíduos

Também para avaliar o ajuste do modelo, é feita uma análise dos resíduos, que são obtidos ao calcular a diferença entre os valores observados (que são exatamente 0 ou 1, uma vez que a variável resposta é dicotômica) e os estimados correspondentes (valores entre 0 e 1, já que representam uma probabilidade). Os resíduos aqui utilizados são a versão padronizada do tipo *deviance*, que medem a discordância entre o máximo das funções de verossimilhança observada e estimada (Agranonik, 2005), ou seja a contribuição de cada observação para a *deviance* do modelo (Paula, 2013).

Independência

A Figura (8.9) plota os resíduos contra o índice das observações, isso é, contra a ordem de coleta. Verifica-se, visualmente, que os resíduos são aleatórios, ou seja, não apresentam nenhum padrão de comportamento, confirmando que não há problemas de dependência entre as observações. Alguns confrontos têm resíduos maiores do que 2, principalmente entre os valores negativos, indicando que o valor estimado foi maior do que o verdadeiro. Esses valores podem ser *outliers*, isso é, disputas cujo o resultado foi fora do padrão. No entanto, o futebol é popularmente conhecido por ser "uma caixinha de surpresas", por isso, já era esperado que ocorressem *outliers*, principalmente valores negativos, pois foram times visitantes que surpreenderam, popularmente chamados de "zebras". Além disso, nota-se que existem mais valores positivos do que negativos, isso porque o mandante se classificou em 63% dos confrontos.

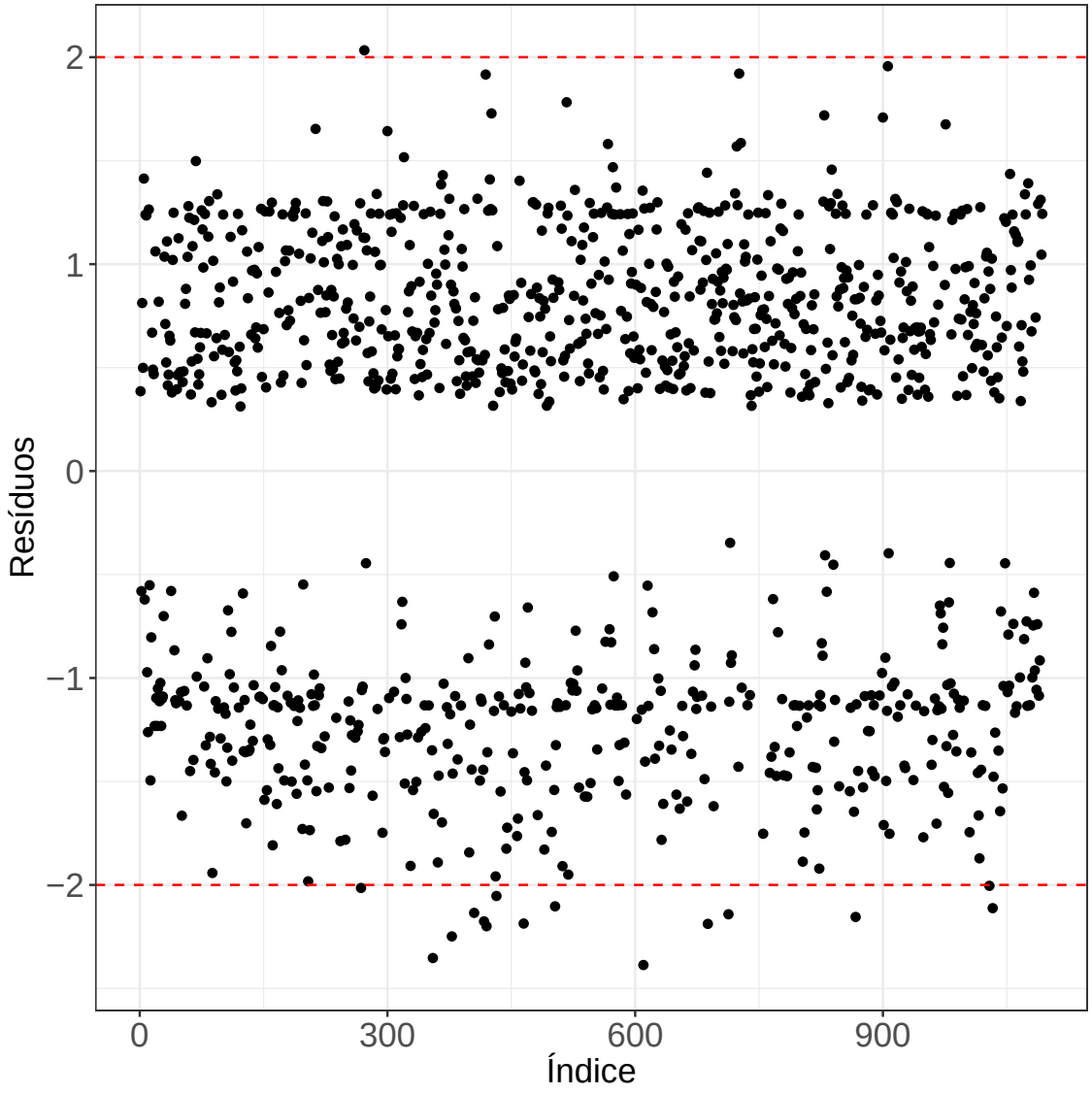


Figura 8.9: Resíduos x índice

Outro indício de independência é dado pelo teste de autocorrelação de Durbin-Watson. Este teste poderia indicar problemas de dependência entre os resíduos (PortalAction, 2017b), no entanto não rejeita a hipótese nula $H_0 : \rho = 0$, evidenciando que não há correlação entre as observações (p-valor = 0,974).

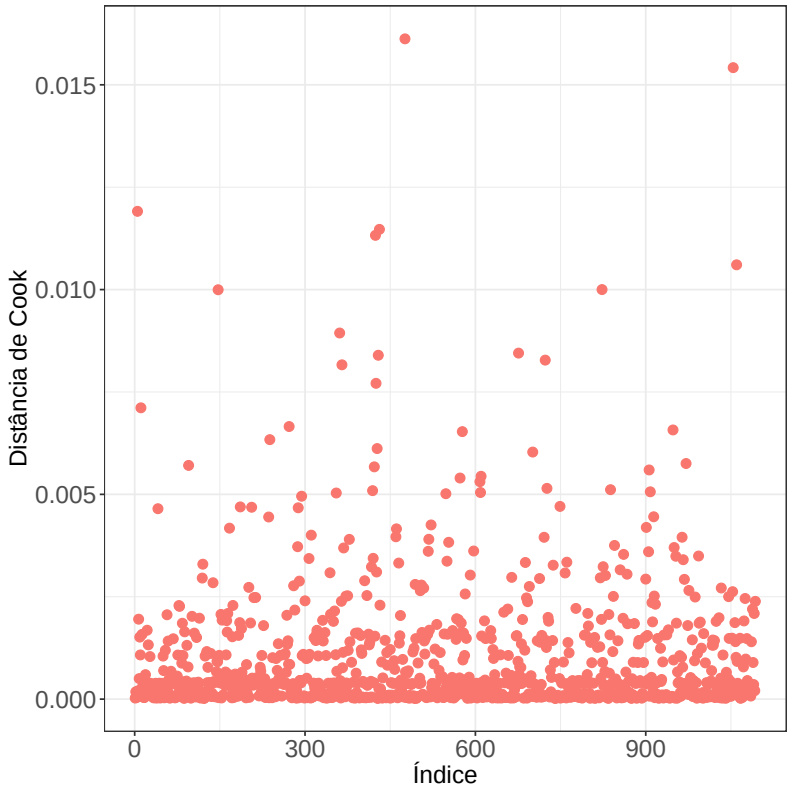
Influência e Alavancagem

A Figura (8.9) identificou a existência de valores atípicos na resposta, aqueles que apresentam resíduos grandes. Por isso é preciso uma análise mais aprofundada que permita verificar as consequências trazidas por essas observações. Estes pontos podem ser influentes, causando mudanças substanciais no modelo, por exemplo, nos valores ajustados ou nas estimativas dos coeficientes. Observações discrepantes também podem ser pontos de alavanca, resultado de uma combinação de valores inesperados para as variáveis explicativas (PortalAction, 2017c). A Figura (8.10) apresenta dois gráficos, que permitem avaliar as observações em cada um dos conceitos descritos.

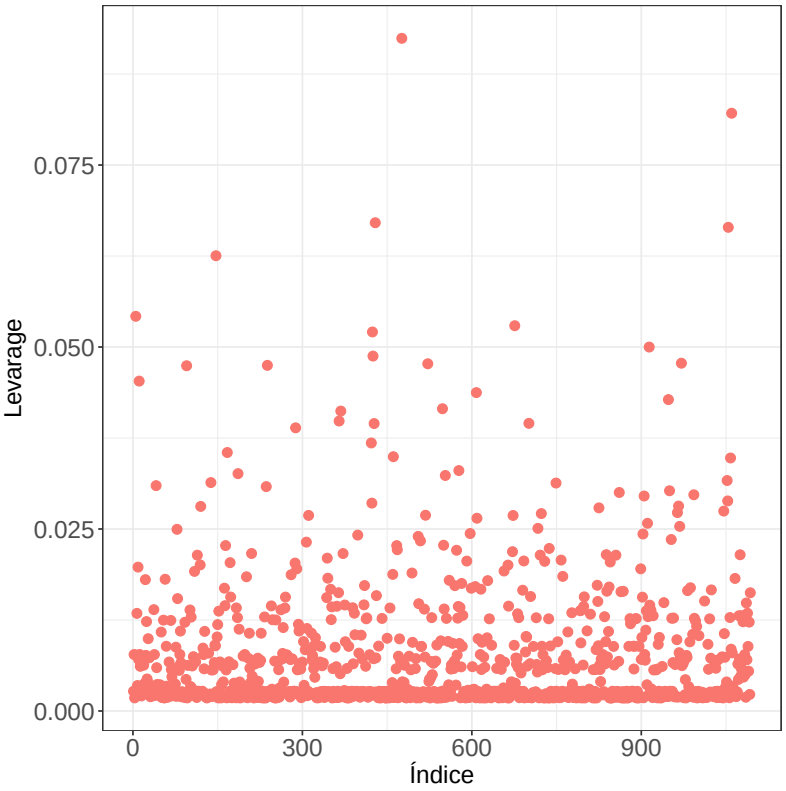
O Gráfico (a) exibe as distâncias de Cook correspondentes à cada observação, enquanto o Gráfico (b) apresenta os *levarage*. Verifica-se que existem valores atípicos para as duas medidas. Como não há valores de referência para indicar se esses pontos representam, ou não, um problema, analisou-se, individualmente, os maiores valores.

A observação de índice 846 é a de maior valor nas duas medidas. Essa, corresponde ao confronto entre Ipatinga e Santos, que aconteceu nas quartas de final, foi decidido nos pênaltis e vencido pelo time mineiro. Entende-se que foi, de fato, um confronto atípico, com um resultado atípico, apesar de o seu resíduo ser 1,30, portanto não considerado *outlier*. O mesmo acontece com os outros pontos assinalados nos dois gráficos: são disputas com características ou resultado atípicos, mas nenhum deles apresenta resíduos *outliers*.

Então, foi considerado que esses confrontos atípicos devem ser mantidos no modelo, pois no futebol espera-se que aconteçam algumas “zebras”; em outras palavras, a ocorrência de eventos atípicos é uma característica dos dados aqui estudados.



(a) Distância de Cook



(b) Levarage

Figura 8.10: Medidas de influência

Resíduos

A regressão logística, diferente da linear, não apresenta resíduos que seguem distribuição Normal. Por isso, um gráfico QQplot, que compara os resíduos com essa distribuição não é adequado. Uma alternativa é o Envelope Simulado, gráfico que faz simulações a partir do modelo estimado e então calcula um intervalo de confiança para os resíduos. A Figura (8.11) apresenta o envelope para o modelo, no qual, conforme descrito pela legenda, a linha cinza indica os quantis da distribuição Normal, os resíduos são os pontos pretos, os limites do intervalo de confiança estão representados pelas linhas vermelhas e entre elas a linha preta pontilhada é a mediana dos resíduos simulados.

Então, observa-se que, apesar de não estar bem adequado aos quantis da distribuição Normal, os resíduos do modelo estão contidos no intervalo simulado, indicando que o modelo está bem ajustado.

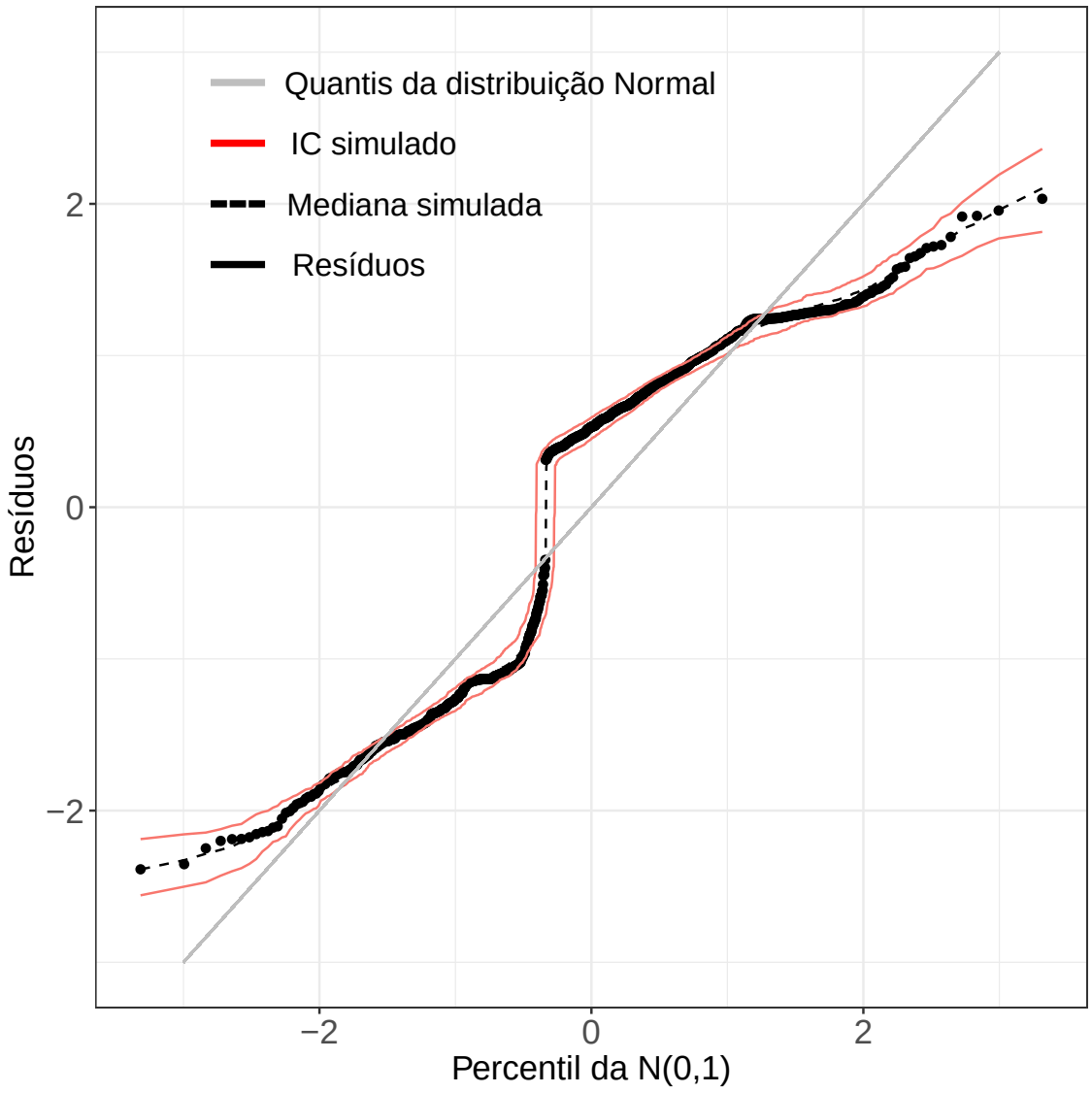
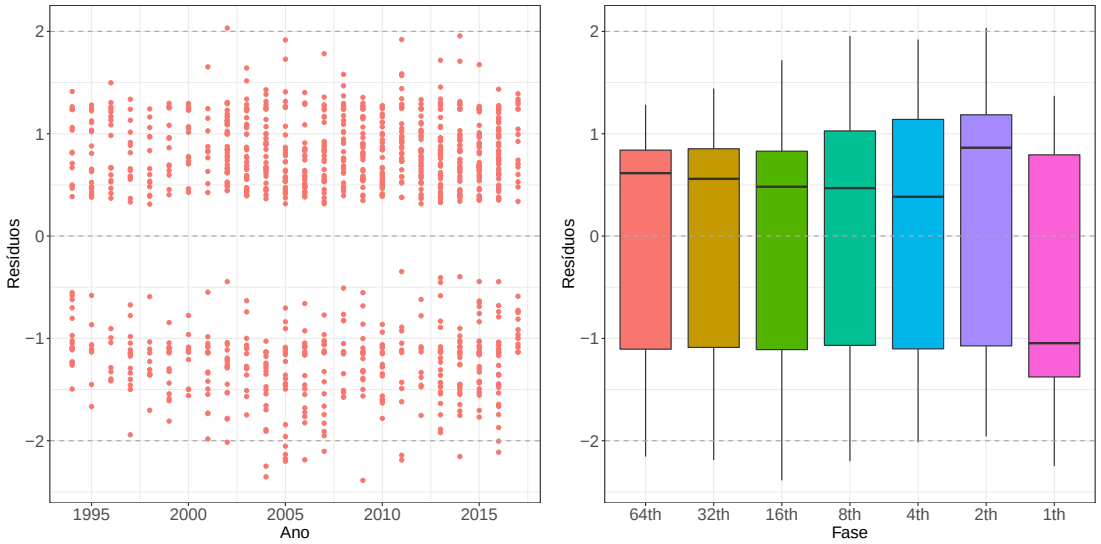


Figura 8.11: Envelope simulado

A partir das diversas análises feitas, foi concluido que o ajuste está adequado; apesar disso, ele poderia ser melhorado com a inclusão de outras variáveis que trouxessem acréscimos relevantes. Então, uma forma de verificar se as variáveis que não foram incluídas no modelo ainda poderiam ter contribuições significativas é plotar os resíduos contra essas variáveis. Conforme discutido anteriormente, é possível que exista uma relação significativa entre o time classificado e as variáveis ano e fase. A fim de investigar essa associação, a Figura (8.12) apresenta a distribuição dos resíduos nos diferentes anos (Gráfico (a)) e fases (Gráfico (b)).

Observa-se que o comportamento padrão dos resíduos se mantém ao longo dos anos, bem como para as fases. Isso indica que nenhuma dessas variáveis têm contribuições significativas para dar ao modelo. Ressalta-se que a mediana da fase final é menor do que as outras, entretanto, essa fase tem apenas 19 observações. Portanto, considera-se que a inclusão da diferença de qualidade foi suficiente para suprir essa informação.



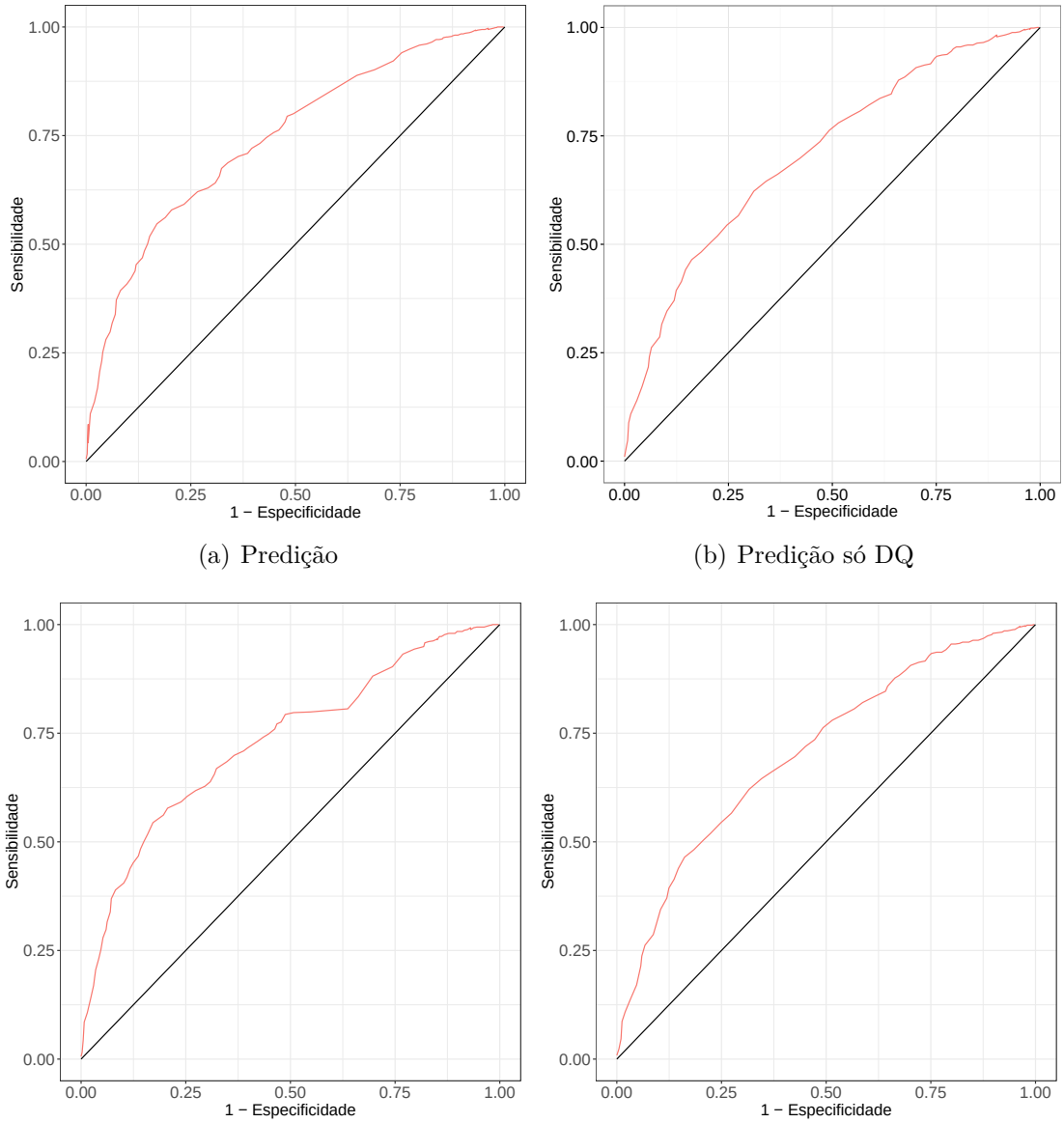
(a) Ano

(b) Fase

Figura 8.12: Resíduos

8.2.3 Predição/Previsão

Uma forma de mensurar a qualidade do ajuste do modelo é analisar a sua capacidade preditiva dentro e fora da amostra, a partir da curva ROC. Essa medida pode ser utilizada com todos os dados da amostra, avaliando a predição, ou também analisando a previsão, para dados fora da amostra. Isso é feito através de um procedimento *leave-one-out*, isto é, retirando-se as observações, uma a uma, estimando o modelo sem esta informação e então prevendo o resultado para ela. Além disso, foram avaliadas a predição e a previsão com o modelo que considera somente a diferença de qualidade entre os times como variável explicativa, pois, para um confronto futuro seria a única informação disponível. A Figura (8.13) apresenta as quatro curvas estimadas. Visualmente, nota-se que a curva do modelo de predição (Gráfico (a)) é mais ampla que a da previsão (Gráfico (c)), que, por sua vez, é maior do que as curvas dos modelos que tem como covariável apenas a diferença de qualidade (Gráficos (b) e (d)).



(a) Predição

(b) Predição só DQ

(c) Previsão

(d) Previsão só DQ

Figura 8.13: Curva ROC

As áreas abaixo da curva (AUC), que estimam a capacidade preditiva do modelo, são expostas na Tabela (8.9) com valores respectivamente 0,7439, 0,7094, 0,7294 e 0,7073. Ou seja, em todos os casos o modelo é melhor do que a moeda para acertar qual é o time vencedor do confronto.

Tabela 8.9: Áreas abaixo da curva ROC (AUC) para os modelos de predição e previsão completos e restritos

Modelo	AUC
Predição (1)	0,7439
Predição só DQ (2)	0,7094
Previsão (3)	0,7294
Previsão só DQ (4)	0,7073

Os resultados obtidos pelo teste de DeLong são apresentados na Tabela (8.10), que tem como hipótese nula a igualdade entre duas AUC's, tem-se que o modelo preditivo é significativamente diferente das capacidades de previsão dos modelos de completo (p-valor $< 0{,}0001$) bem como do restrito (p-valor $= 0{,}0004$) e da predição com o modelo restrito (p-valor $= 0{,}0009$). A diferença entre as previsões dos modelos completo e restrito também é significativa (p-valor $= 0{,}0343$). A capacidade do modelo de predição restrito é significativamente diferente das capacidades dos dois modelos de previsão (p-valores 0,0552 e menor que 0,0001)

Tabela 8.10: Testes de AUC entre predição e previsão nos modelos completos e restritos

Modelos	p-valor
1x2	<0,0001
1x3	0,0004
1x4	0,0009
2x3	0,0343
2x4	0,0552
3x4	<0,0001

Então, após a obtenção dos valores da acurácia, da sensibilidade e da especificidade, para cada α possível, busca-se maximiza-los. Para os três modelos, o valor de α que maximiza simultaneamente a sensibilidade e a especificidade são diferentes dos que maximizam a acurácia. Isso acontece porque as quantidades de *sucesso* e *fracasso* na amostra são diferentes. Os valores máximos desta medidas, bem como os valores de α que levam à isso, são apresentados na Tabela (8.11).

Para a predição, o α que maximiza os acertos gerais do modelo é 0,5, ou seja, será classificado como vitória do mandante quando a probabilidade estimada estiver acima desse valor. Para esse α, a acurácia é de 0,6935, a sensibilidade é 0,7945 e a especificidade é 0,5199. Por outro lado, se o objetivo for maximizar, ao mesmo tempo, a sensibilidade e especificidade, o α ideal é 0,61. Nesse caso, a sensibilidade é 0,6744; a especificidade é 0,6766 e a acurácia é 0,6752.

Para a predição que considera somente a diferença de qualidade, o melhor α para maximizar os acertos gerais do modelo é 0,46, o mesmo da previsão restrita, cujas sensibilidade e especificidade são 0,8973 e 0,2985 respectivamente e a acurácia é 0,6834. Já para ter, conjuntamente, as maiores sensibilidade e especificidade o α deve ser 0,62, com acurácia de 0,6505 e os acertos positivos e negativos são, respectivamente, 0,6454 e 0,6318. Em comparação com as probabilidades preditas, os erros têm grande amplitude, com média 0, porém máximo 0,3108 e mínimo -0,4177.

O erro médio de previsão com o modelo completo em relação à predição é 0,00008, com máximo absoluto de 0,05. Os mesmos α's da predição maximizam os acertos na previsão. Utilizando 0,5, a acurácia é 0,6898 e a sensibilidade e especifidade são, respectivamente, 0,7931 e 0,5124. Ao maximizar conjuntamente as vitórias e derrotas classificadas corretamente a acurácia é 0,6715, a sensibilidade é 0,6686 e a especificidade 0,6766.

Para a previsão que considera somente a diferença de qualidade, o melhor α para maximizar os acertos gerais do modelo é 0,46, cujas sensibilidade e especificidade são 0,8336 e 0,2985 respectivamente e a acurácia é 0,6825. Já para ter, conjuntamente,

as maiores sensibilidade e especificidade o α deve ser 0,63, com acurácia de 0,6488 e os acertos positivos e negativos são, respectivamente, 0,6382 e 0,6542. Neste caso, porém, os erros, em comparação com as probabilidades preditas, têm maior amplitude, com média 0,00001, mas máximo 0,3157 e mínimo -0,4174. Nota-se que todos os valores para a previsão restrita são próximos aos obtidos pela predição restrita.

Tabela 8.11: Medidas de Capacidade

Medida Maximizada	Medida	Predição	Predição só DQ	Previsão	Previsão só DQ
Acurácia	α	0,5	0,46	0,5	0,46
	Acurácia	0,6935	0,6834	0,6898	0,6825
	Sensibilidade	0,7945	0,8973	0,7931	0,8336
	Especificidade	0,5199	0,2985	0,5124	0,2985
Especificidade e Sensibilidade	α	0,61	0,62	0,61	0,63
	Acurácia	0,6752	0,6505	0,6715	0,6488
	Sensibilidade	0,6744	0,6454	0,6686	0,6382
	Especificidade	0,6766	0,6318	0,6766	0,6542

Apesar de os testes terem evidenciado que existem diferenças significativas entre os três modelos, verificou-se que, ao maximizar as medidas de capacidade, a previsão feita pelo modelo completo tem valores próximos aos da predição. Por outro lado, o modelo de previsão que tem apenas o dado da diferença de qualidade apresenta os menores valores para as diferentes medidas, além de utilizar outros valores de α para obter o máximo, bem como cometer erros maiores. Contudo, esse modelo não dispõe de uma informação relevante que é o tipo de decisão e, mesmo assim, os resultados são relativamente próximos dos demais; por essa razão, considera-se que seus resultados são satisfatórios.

9 Considerações Finais e Discussão

Ao longo deste trabalho, foram encontradas evidências de que, incondicionalmente, o fator doméstico de fato constitui uma vantagem nos confrontos de mata-mata, pois o mandante se classifica em aproximadamente 63% das disputas (significativamente maior que 0,5, p-valor $< 0{,}0001$). Confirmando a hipótese de que decidir um confronto da Copa do Brasil como mandante é, de fato, uma vantagem. Entretanto, quando a decisão é por gol qualificado ou pênaltis o percentual de classificação é cerca de 20% menor do que o percentual geral (diferenças significativa, p-valor $< 0{,}0001$ e p-valor $= 0{,}0113$ respectivamente), ou seja, esses critérios beneficiam significativamente o time visitante, se não dando a vantagem, ao menos equiparando as probabilidades das duas equipes, uma vez que os dois critérios levam a proporções que não são diferentes de 0,5 (p-valores respectivamente 0,2229 e 0,5770).

Com isso, pode-se apontar que o uso de gol qualificado ou pênaltis como critério de desempate diminui a probabilidade de classificação do time mandante. Considerando os efeitos isoladamente ou interagindo com a diferença de qualidade entre os times. Em contrapartida, o aumento na diferença de qualidade tem o efeito inverso, pois, quanto mais positiva é a diferença, melhor é o time mandante, e maior é a sua probabilidade de ser o vencedor do confronto, para todos os critérios de decisão.

No entanto, deve-se levar em consideração que, quando as duas equipes participantes de um confronto tem qualidades iguais, a probabilidade de classificação é igual a 0,5 para ambos os times, não importando como foi a decisão. Ou seja, independente do critério utilizado, os dois participantes têm as mesmas probabilidades de classificação quando suas qualidades são equivalentes, na ordem. Os p-valores são 0,1033, 0,2338, 0,0950 e 0,4491, conforme descrito na Seção 8.1.

Nos exemplos que foram dados, para ilustrar as probabilidades estimadas pela regressão logística, pôde-se notar que o modelo completo, isso é, que tem as informações sobre a diferença de qualidade, o tipo de confronto e as interações, teve boas predições em todas situações. Por outro lado, o modelo restrito, que conta somente com a diferença de qualidade como variável explicativa, tem boas predições nos confrontos que foram decididos por pontuação ou saldo de gols. Entretanto, para os confrontos que utilizaram o gol qualificado ou pênaltis o modelo restrito apresenta mais dificuldade. Isso acontece porque na Copa do Brasil a maior parte (82,43%) dos confrontos foram decididos por pontos ou saldo de gols, então essas situação têm mais influência nas estimativas gerais.

O modelo restrito tem a intenção de verificar como seria utilizar o modelo para fazer uma previsão, ou seja, determinar o resultado de um confronto que ainda não

ocorreu. Para tanto, não seria possível utilizar a informação do critério de decisão, uma vez que este só é definido ao término do confronto, a não ser para identificar as probabilidades que o mandante tem em cada decisão, o que não é de grande utilidade sem que se saiba a probabilidade de eles serem utilizados. Acredita-se então que seria interessante utilizar a informação obtida, sobre as probabilidades estimadas de uma disputa terminar empatada, na Seção 7.1 como uma variável explicativa, de forma que o modelo não sabe qual critério será utilizado, porém tem conhecimento das probabilidades de serem necessários. Esse aspecto pode ser investigado em trabalhos futuros.

Um fator que não foi considerado nesse estudo, por falta de informação disponível, é a distância percorrida pelas equipes. Outros trabalhos, como os de Pollard (2006) e Pollard et al. (2008), concluíram que essa é uma variável relacionada à vantagem de jogar um único jogo em seu domínio. Conforme a discussão feita na Seção 1, na Copa do Brasil esse efeito pode ser ainda mais importante, devido à inclusão de times de todas regiões, fazendo com que com tenha grandes distâncias e mudanças climáticas.

Contudo, deve-se ressaltar que, nesse campeonato, os confrontos são disputados em dois jogos, portanto os dois times precisam viajar, cada um para uma partida, mas ambos percorrem a mesma distância. Seria necessário obter informações sobre as distâncias para que se pudesse concluir se há diferença entre fazer uma viagem para o primeiro ou para o segundo jogo. Além disso, é possível que exista uma relação também com a qualidade dos times, uma vez que times de maior expressão geralmente têm mais dinheiro e proporcionam mais conforto para seus atletas, diminuindo o efeito das distâncias percorridas. Ademais, as equipes de alta qualidade estão concentradas nas regiões Sul e Sudeste, de modo que nesses confrontos com qualidades similares, esse efeito pode ser menor. Essa é mais uma possível causa para a hegemonia desses times na Copa do Brasil. De todo modo, é uma variável que seria interessante considerar em estudos futuros, para investigar qual a sua influência na vitória de um time em um confronto de mata-mata, ou se não é relevante, diferentemente dos jogos únicos.

Outro ponto a ser considerado são os critérios utilizados para mensurar a qualidade dos times. Apesar de haver outros possíveis, optou-se por utilizar os critérios da CBF que foram, inicialmente, considerados os mais adequados. No entanto sabe-se que é uma *proxy* imperfeita para a variável latente "qualidade". Atualmente, a Libertadores e a Sul-Americana, são consideradas apenas como uma pontuação bônus, para os times que não puderam participar da Copa do Brasil por conflitos de calendário. Contudo, os times que participam dos campeonatos continentais, geralmente, têm esses como prioridade, usualmente, resultando em um pior desempenho nas competições nacionais, consequentemente tendo uma menor pontuação no ranking do ano subsequente. Isto é, não levar em conta o desempenho nesses campeonatos pode estar subestimando a qualidade de alguns times. Por exemplo, em 2013 o Atlético-MG foi campeão da Libertadores, porém ficou em apenas décimo-quinto lugar no ranking final daquele ano, isso porque ficou em oitavo lugar no Campeonato Brasileiro e nem participou da Copa do Brasil. Entende-se que, com melhorias nos critérios utilizados, poderia-se ter uma estimativa mais fidedigna das qualidades dos times, o que poderia gerar uma análise melhor.

Os critérios utilizados pelo ranking da Conmebol (Confederação Sul-Americana de Futebol) seriam ainda mais inadequados. O ranking é utilizado apenas para

definir o sorteio da Copa Libertadores, por isso só considera o desempenho nessa competição, dá uma pequena pontuação para os times que foram campeões nacionais, porém não leva em conta a Copa Sul-Americana nem as Copas locais, além de desconsiderar os posicionamentos nos campeonatos nacionais (Conmebol, 2017). Outra possibilidade, seria a utilização do ranking mundial de clubes IFFHS, anualmente divulgado desde 1991. A instituição considera os pontos obtidos pelos clubes ao longo de um ano, fazendo uma ponderação de acordo com os campeonatos disputados (Wikipédia, 2017b). Possivelmente, essa seria a melhor forma de medir a qualidade dos times, pois considera todos os jogos que cada equipe disputou, em estudos futuros seria interessante utilizar esse método.

Na Seção 2, foi descrito que o principal objetivo desse trabalho era estudar a vantagem de decidir um confronto como mandante, bem como determinar a influência do gol qualificado sobre isso. Considera-se que esses objetivos foram atendidos, uma vez que todas as probabilidades de interesse foram satisfatoriamente estimadas, além de que foram encontradas evidências que corroboram todas as hipóteses originais desta pesquisa. Posto isso, entende-se que, para decidir sobre a validade do uso da regra do gol qualificado, deve-se pensar nos objetivos do campeonato.

Um argumento popularmente utilizado é de que esse critério não é justo, por exemplo, que 2x1 não é melhor resultado do que 1x0. O futebol de fato está diferente da época em que a norma foi criada. Nos dias atuais, os times têm mais probabilidades de marcar gols jogando como visitantes e até mesmo de vencer o jogo. Apesar disso, os números ainda indicam a existência de vantagem significativa para o mandante.

No entanto, ao longo desse estudo, notou-se que essa regra serve como uma transição entre o saldo de gols e a disputa de pênaltis. Portanto, para tirar o gol qualificado seria necessário encontrar outro critério para substituí-lo, a fim de impedir que aconteça uma queda brusca na vantagem do mandante. Contudo, seria preciso analisar minuciosamente cada possibilidade. Comparando com os campeonatos em que elas são utilizadas ou testando em competições de menor relevância, como das categorias de base, para então, tendo visto na prática, poder considerar a aplicação de uma mudança.

Uma possibilidade, que era válida, porém não foi utilizada, na final da Copa Sul-Americana de 2017, é a prorrogação. Entretanto, entende-se que isso só aumentaria ainda mais a vantagem do mandante, que teria mais tempo jogando em seu estádio, o que já é comprovadamente uma vantagem. Além disso, o tempo extra poderia ser utilizado mantendo a regra do gol qualificado, de forma que um time teria mais tempo como mandante, em contraponto o outro teria mais tempo para marcar um gol qualificado. Nesse caso, não é possível saber se as estimativas e conclusões obtidas neste trabalho permaneceriam válidas.

Outra opção, utilizada em alguns campeonatos, é retirar, além do gol qualificado, o saldo de gols, então a disputa de pênaltis é o único critério de desempate. Viu-se que essa regra também equipara as probabilidades dos dois times. No entanto, traria-se de volta o argumento de ser injusto, vencer por 1x0 é igual à vencer por 6x0? Bem como o gol qualificado, o saldo de gols é um estágio de transição entre a classificação por pontuação e a disputa de pênaltis. Dessa forma, entende-se que os quatro critérios são importantes para que não haja uma mudança brusca nos resultados.

Também foi visto a importância que o primeiro jogo tem para as probabilidades de classificação. Esse resultado é determinante para o resultado final do confronto. Então, se o gol qualificado equilibra as probabilidades de classificação, retirar a possibilidade de uso desse critério tornaria o primeiro jogo ainda mais decisivo. Pois, ao perder, ou até mesmo empatar o primeiro jogo, o time visitante possivelmente perde a esperança de reverter o resultado. De forma que, o segundo jogo acabaria por ser uma mera formalidade.

Então, será que o gol qualificado é de fato injusto? Ou ele está balanceando o favorecimento que o mandante tem simplesmente por jogar a segunda partida em seu estádio? Talvez um time não mereça vencer somente por ter marcado um gol fora de casa, mas ele merece ter menos probabilidades de vencer porque um sorteio definiu que ele seria o time visitante? Pior ainda, nas situações em que não há sorteio, que um time ganha o privilégio de ser mandante porque tem mais qualidade, a equipe visitante já está prejudicada de duas formas, pelo mando de campo e pela qualidade, não merece ter ao menos a possibilidade de usar o gol qualificado para equiparar as probabilidades de vitória? Qual é a emoção em um confronto em que as equipes têm probabilidades tão desiguais?

A Copa do Brasil é popularmente considerada uma competição democrática, porque dá oportunidade para diversos times, de diferentes grandezas e de todas as regiões do país. Apesar disso, os principais times do país são hegemônicos na competição, portanto, todos participam, mas poucos conseguem vencer. Pelo regulamento atual, as primeiras fases têm jogo único, em que o time com pior colocação no ranking da CBF é o mandante, mas só se classifica em caso de vitória. Além disso, o gol qualificado não pode mais ser utilizado, diminuindo cada vez mais a probabilidade de um time ”pequeno´´ vencer um confronto. Dessa forma, as mudanças que a Confederação tem feito estão fazendo com que o torneio perca seu propósito inicial, de incluir à todos, dando uma chance real aos times de menor expressão. A competição se tornou muito extensa, enchendo o calendário dos times ”grandes´´ que participam de outros campeonatos, sendo que as primeiras fases praticamente só servem para constar que os clubes ”pequenos´´ participaram.

Por tudo que foi argumentado, acredita-se que duas coisas deveriam ser feitas. Considerando que o objetivo seja ter a Copa do Brasil como um campeonato justo, minimizando a vantagem do time mandante e assim tornando a disputa mais imprevisível e emocionante para o espectador. Primeiro, o mando de campo deveria sempre ser sorteado, uma vez que decidir em casa é uma vantagem, isso não deve ser dado como um privilégio. Principalmente, não tem sentido que isso seja dado simplesmente por um time ser melhor do que outro de acordo com um ranking, pois está possibilitando uma vantagem ao time que já é privilegiado pela qualidade. Além disso, reitera-se a importância do gol qualificado como um critério que equipara as probabilidades de classificação dos dois times, além de servir de transição, antes da disputa de pênaltis.

Referências Bibliográficas

Agranonik, M. (2005). Técnicas de diagnóstico aplicadas ao modelo de regressão logística. http://hdl.handle.net/10183/128182. [Último acesso em 21-dezembro-2017].

Aitchison, J. e Aitken, C. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, 63(3):413–420.

Allison, P. D. (2014a). Alternatives to the hosmer-lemeshow test. https://statisticalhorizons.com/alternatives-to-the-hosmer-lemeshow-test. [Último acesso em 26-novembro-2017].

Allison, P. D. (2014b). Another goodness-of-fit test for logistic regression. https://statisticalhorizons.com/another-goodness-of-fit-test-for-logistic-regression. [Último acesso em 26-novembro-2017].

Allison, P. D. (2014c). Measures of fit for logistic regression. http://www.statisticalhorizons.com/wp-content/uploads/MeasuresOfFitForLogisticRegression-Slides.pdf. [Último acesso em 26-novembro-2017].

Bolan@Área (2017). Copa do brasil. http://www.bolanaarea.com/gal_copa_do_brasil.htm. [Último acesso em 30-setembro-2017].

CBF (2014). Convenção de pontos do ranking nacional de clubes. https://cdn.cbf.com.br/content/201612/20161212191347_0.pdf. [Último acesso em 03-outubro-2017].

CBF (2017a). Confederação brasileira de futebol. https://www.cbf.com.br/. [Último acesso em 13-julho-2017].

CBF (2017b). Copa do brasil 2018 não terá gol qualificado. https://www.cbf.com.br/noticias/campeonato-copa-brasil-masculino/copa-do-brasil-2018-nao-tera-gol-qualificado?ref=bigfeatured#.WiKZh0qnHIU. [Último acesso em 13-julho-2017].

Collett, D. (2003). *Modelling binary data*. Chapman & Hall.

Conmebol (2017). Ranking conmebol da copa libertadores 2017. http://www.conmebol.com/pt-br/ranking-conmebol-da-copa-libertadores-2017. [Último acesso em 21-dezembro-2017].

da Silva, S. B. (2016). História da copa do brasil. http://www.campeoesdofutebol.com.br/copa_brasil_historia.html. [Último acesso em 22-maio-2017].

de Almeida, L. G., de Oliveira, M. L., e da Silva, C. D. (2011). Uma análise da vantagem de jogar em casa nas duas principais divisões do futebol profissional brasileiro. *Revista Brasileira de Educação Física e Esporte*, 25(1):49–54.

de Andrade, L. (2017). O injusto gol qualificado. http://htesports.com.br/2017/08/o-injusto-gol-qualificado/. [Último acesso em 14-novembro-2017].

DeLon, E., DeLong, D. M., e Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44:837–845.

Hayfield, T. e Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5).

Hosmer, D. W. e Hjort, N. L. (2002). Goodness-of-fit processes for logistic regression: simulation results. *Statistics in Medicine*, 21(18):2723–2738.

Hosmer, D. W. e Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley Series in Probability and Statistics. Wiley, 2 edition.

Loughin, T. M. e Bilder, C. R. (2013). Analysis of categorical data with r. http://www.chrisbilder.com/categorical/Chapter5/AllGOFTests.R. [Último acesso em 20-abril-2017].

Nobre, R. (2016). Futebol, paixão e negócio. http://www.goal.com/br/news/619/especiais/2016/06/10/24444332/futebol-paix~ao-e-negócio. [Último acesso em 16-novembro-2017].

Osius, G. (1994). Evaluating the significance level of goodness-of-fit statistics for large discrete data. http://www.math.uni-bremen.de/~osius/download/papers/Osius1993SignLevelGoF.pdf. [Último acesso em 26-novembro-2017].

Page, L. e Page, K. (2007). The second leg home advantage: Evidence from european football cup competitions. *Journal of Sports Sciences*, 25(14):1547–1556.

Paula, G. A. (2013). Modelos de regressão com apoio computacional. https://www.ime.usp.br/~giapaula/texto_2013.pdf. [Último acesso em 26-dezembro-2017].

Peron, H. L. (2017). Não faz mais sentido o marcado gol fora de casa ser usado para o desempate. http://globoesporte.globo.com/blogs/especial-blog/peron-na-arquibancada/post/nao-faz-mais-sentido-o-marcado-gol-fora-de-casa-ser-usado-para-o-desempate.html. [Último acesso em 14-novembro-2017].

Pollard, R. (1986). Home advantage in soccer: A retrospective analysis. *Journal of Sports Sciences*, 4(3):237–248.

Pollard, R. (2006). Worldwide regional variations in home advantage in association football. *Journal of sports sciences*, 24(3):231–240.

Pollard, R. (2008). Home advantage in football: A current review of an unsolved puzzle. *The open sports sciences journal*, 1(1):12–14.

Pollard, R., Da Silva, C., e Nísio, C. (2008). Home advantage in football in brazil: differences between teams and the effects of distance traveled. *The Brazilian Journal of Soccer Science*, 1(1):3–10.

PortalAction (2017a). Análise de colinearidade e multicolinearidade. http://www.portalaction.com.br/analise-de-regressao/36-analise-de-colinearidade-e-multicolinearidade. [Último acesso em 15-dezembro-2017].

PortalAction (2017b). Diagnóstico de independência. http://www.portalaction.com.br/analise-de-regressao/33-diagnostico-de-independencia. [Último acesso em 21-dezembro-2017].

PortalAction (2017c). Pontos influentes e valores extremos. http://www.portalaction.com.br/analise-de-regressao/34-pontos-influentes-e-valores-extremos. [Último acesso em 17-dezembro-2017].

PortalAction (2017d). Regressão logística. http://www.portalaction.com.br/analise-de-regressao/regressao-logistica. [Último acesso em 27-dezembro-2017].

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., e Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77.

Sánchez, P. A., García-Calvo, T., Leo, F. M., Pollard, R., e Gómez, M. A. (2009). An analysis of home advantage in the top two spanish professional football leagues. *Perceptual and motor skills*, 108(3):789–797.

Seckin, A. e Pollard, R. (2008). Home advantage in turkish professional soccer. *Perceptual and motor skills*, 107(1):51–54.

TabeladoBrasileirão (2017). Tabela do brasileirão. https://www.tabeladobrasileirao.net/. [Último acesso em 04-dezembro-2017].

TradingEconomics (2017). Pib – lista de países. https://pt.tradingeconomics.com/country-list/gdp. [Último acesso em 17-novembro-2017].

Uol (2014). Regra do gol fora de casa ficou ultrapassada, diz blatter. https://esporte.uol.com.br/futebol/ultimas-noticias/2014/10/09/regra-do-gol-fora-de-casa-ficou-ultrapassada-diz-blatter.htm. [Último acesso em 14-novembro-2017].

Wikipédia (2017a). Copa do brasil de futebol. https://pt.wikipedia.org/wiki/Copa_do_Brasil_de_Futebol. [Último acesso em 30-setembro-2017].

Wikipédia (2017b). Ranking mundial de clubes iffhs. https://pt.wikipedia.org/wiki/Ranking_Mundial_de_Clubes_da_IFFHS. [Último acesso em 21-dezembro-2017].