

**INSPER**

**Mestrado em Administração 2011**

**Disciplina de Estatística Multivariada**

# **O futebol e a motivação do torcedor**

**Um estudo aplicado de  
Regressão Logística**

Ramon Barbosa Rosa

---

## Resumo

O futebol é um dos esportes que mais atrai público *in loco*. Nesse estudo, buscamos usar esse esporte como exemplo de aplicação da Regressão Logística para entender as razões que motivam os torcedores à frequentarem os estádios. Utilizando dados de uma edição recente do campeonato brasileiro de futebol, verificamos que os fatores que mais motivam o torcedor à frequentar o estádio são a presença de grandes equipes, a etapa em que o campeonato se encontra e a classificação de seu time de preferência e do time adversário.

## Introdução

No mundo moderno, os esportes se tornam cada vez mais um negócio, com investidores buscando nas diversas modalidades as melhores oportunidades de retorno para seu capital. Em particular, o futebol é um dos esportes mais atraentes para investidores, em razão da atração que exerce sobre os torcedores e das inúmeras possibilidades de negócios que a ele podem ser associados. Exemplos incluem desde a venda de objetos, roupas e uniformes das equipes para torcedores até a negociação de contratos de propaganda envolvidos na transmissão dos jogos ao vivo pela televisão. Em decorrência, a maximização de resultados se torna um objetivo importante, seja na ação dos atletas, seja em todos os outros fatores em torno de uma partida de futebol. Essa busca crescente pelos melhores resultados tem como aliada a ciência, que lança mão dos instrumentos mais modernos para estudar esse esporte. Em particular, diversos estudos na área de Administração têm sido dedicados ao futebol.

No entanto, poucos estudos têm voltado sua atenção para entender os fatores que concorrem para que as partidas de futebol tenham o maior afluxo possível de público aos estádios. Esta questão se torna importante na medida em que o grau de lotação de um estádio durante um partida é um dos melhores indicadores para se aferir o interesse do torcedor e, em decorrência, os resultados de toda a estratégia de negócios envolvida com esse esporte. Nesse sentido, nosso objetivo é ajudar a responder essa questão, trazendo nova luz aos estudos que já se dedicaram sobre o tema.

Este trabalho está organizado como segue. Na próxima seção discutiremos o embasamento para as hipóteses que desejamos validar. Em seguida, descrevemos os dados utilizados e o tratamento dado a alguns deles para ajuste das variáveis ao modelo. Na seção de Metodologia, descrevemos o modelo a ser utilizado para capturar as interações entre as variáveis dependentes e a variável resposta, bem como os procedimentos da Regressão Logística utilizada

para estimação dos parâmetros do modelo. Nas duas seções seguintes discutimos os resultados e a interpretação dos parâmetros do modelo. Por fim concluímos com as implicações teóricas e práticas dos achados e com indicações de direcionamento para futuras pesquisas.

## Hipóteses

Há muito tempo a imprensa especializada discute o antagonismo entre o chamado futebol arte e o futebol de resultados. Futebol arte é uma expressão que tenta definir uma partida caracterizada por belas jogadas com o objetivo de buscar o gol a maior parte do tempo. Em geral, segundo especialistas, está associado à presença de jogadores talentosos em campo e à equipes motivadas e cujos membros se conhecem bem. Nessa linha, seria de se esperar que uma equipe que vem desempenhando bem, marcando muitos gols e acumulando muitos pontos no campeonato teria maior propensão à atrair torcedores ao estádio. Seguindo esse raciocínio o torcedor iria ao estádio motivado pela possibilidade de presenciar um espetáculo entre seu clube e o clube visitante.

*Hipótese 1: Torcedores vão ao estádio motivado pelo histórico recente das equipes participantes.*

Em oposição ao futebol arte, o futebol de resultados é definido pelo esforço mínimo da equipe para obter um objetivo necessário. O futebol de resultados é caracterizado por jogadas de pouca emoção, predominantemente defensivas, resultando em partidas com pouca beleza técnica, com poucas finalizações e muitos empates sem gols. O objetivo imediato é apenas passar para a próxima fase do campeonato ou simplesmente administrar vantagem adquirida nas partidas anteriores ou na própria partida em disputa. Não há consenso se o futebol evoluiu da arte para o pragmatismo ou se as duas formas sempre coexistiram ao longo de sua história. A favor dos que defendem que os dois formatos sempre existiram está a introdução de novas regras (poucas, quando comparadas com outros esportes) para torná-lo mais competitivo. No caso dos que defendem que houve uma evolução do futebol arte para o futebol de resultados, são discutidas diversas razões para esse movimento. Há quem diga que o público se adaptou a este último, de forma que a motivação para ir ao estádio está vinculada não aos últimos resultados, mas ao posicionamento de sua equipe no campeonato em disputa.

*Hipótese 2: Torcedores vão ao estádio porque seu time está bem classificado no campeonato.*

Descreveremos a seguir os dados e a metodologia utilizada para a verificação das hipóteses.

## Dados

Os dados utilizados constituem-se de informações dos 374 jogos do Campeonato Brasileiro de Futebol de 2009, serie A e foram extraídos da base da Confederação Brasileira de Futebol. O campeonato brasileiro começou a ser disputado anualmente a partir de 1971 com 20 clubes. Ao longo dos anos, o número de clubes participantes variou em função da mudança constante de regras, chegando a 94 clubes na edição de 1979. Porém, o campeonato de 2009 teve exatamente o mesmo número de clubes da primeira edição.

Para avaliar a Hipótese 1 de que torcedores vão ao estádio motivados pelo histórico recente de seu time, utilizamos como balizador a quantidade de gols do time da casa (*GLH*) e do visitante (*GLV*) nos últimos três jogos. Além disso, também mensuramos a quantidade de pontos ganhos tanto pelo time da casa (*PGH*) quanto pelo time visitante (*PGV*), igualmente nos últimos três jogos. Tanto no caso dos gols quanto da quantidade de pontos, esperamos uma relação positiva com a variável que mede a lotação do estádio.

No caso da Hipótese 2, estamos interessados em mensurar se torcedores são atraídos ao estádio em razão da classificação de seu time no campeonato em disputa. As variáveis de classificação do time da casa (*CLH*) e classificação do time visitante (*CLV*) visam capturar este efeito. Ressaltamos que as variáveis de classificação são ordinais e estavam originalmente em escala crescente. Para evitar dificuldades na interpretação dos coeficientes, invertemos a ordem de classificação o que nos faz esperar sinal positivo em seu coeficiente.

A base original possui ainda as seguintes variáveis inerentes ao fenômeno de estudo e que podem ajudar a encontrar as respostas desejadas para a questão principal:

- *PT*, que assume os valores 1, 2, 3 e 4 e indica a etapa do campeonato em que a partida ocorre, sendo que 1 é a primeira etapa e 4 a etapa final. Espera-se uma relação positiva com a variável resposta.
- *CLS*, que assume o valor 1 (um) se a partida for considerada um clássico, e 0 (zero) se não for. Partidas clássicas são aquelas disputadas entre grandes e tradicionais equipes. Espera-se que o sinal de seu coeficiente seja positivo.
- *BIG*, que assume o valor 1 (um) se a partida possui alguma grande equipe dos estados de São Paulo ou Rio de Janeiro. Equipes desses estados historicamente tem melhor desempenho no campeonato brasileiro, bem como nos campeonatos estaduais. Além disso, contam os melhores atletas e possuem grandes torcidas.

- *MTO*, que assume o valor 1 (um) se a equipe visitante está entre as dez maiores torcidas do Brasil, ou 0 (zero) caso contrário. Esta variável difere da variável *BIG* na medida em que entre as dez maiores torcidas, quatro são de equipes de MG e RS. Esta é a única informação que não estava na base original da CBF e foi extraída dos resultados da pesquisa Datafolha 2009. Acredita-se que a existência de uma grande torcida compense a falta de mando de campo da equipe visitante e contribua para que a partida tenha um grande público presente no estádio. Por conta disso, espera-se que o sinal do coeficiente dessa variável seja positivo.
- *PRC*, que mensura o preço médio dos ingressos à disposição do público na partida. Assume-se que preços mais baixos incentivem a presença do torcedor no estádio e, portanto, espera-se que o sinal do coeficiente dessa variável seja negativo.

Além das variáveis inerentes ao fenômeno do futebol descritas nos parágrafos anteriores, utilizamos uma série de variáveis de controle que são geralmente associadas à existência de público nos estádios em dias de jogos:

- *DSudeste*, que assume valor 1 (um) se o jogo realizou-se em algum dos estados da Região Sudeste (São Paulo, Minas Gerais, Rio de Janeiro e Espírito Santo) ou 0 (zero) caso tenha ocorrido em outro estado. A inclusão dessa variável justifica-se pela tanto pela tradição dos clubes do Sudeste, quanto pela existência de estádios com melhor infraestrutura, bem como por ser a região com a maior população do Brasil. Todos esses fatores têm sido indicados como elementos que concorrem para o maior afluxo de público ao estádio, sendo esperada uma relação positiva com a variável resposta.
- *RPC*, que mede a renda per capita (em Reais) da cidade onde ocorre o jogo. Admite-se que em municípios com maior renda, a população local possua mais recursos disponíveis para adquirir os ingressos, sendo esperada relação positiva com a variável resposta.
- *POP*, indicando a quantidade de habitantes da cidade em que ocorre o jogo. Jogos realizados em municípios com maior população em tese teriam maior propensão a terem público máximo, sendo esperado sinal positivo em relação à variável resposta.
- *RAN*, mensurando a quantidade de chuva em milímetros no dia jogo. Jogos em dias de chuva seriam um desestímulo para o torcedor, que preferiria ficar em casa assistindo o jogo pela televisão.
- *DInverno*, que assume valor 1 (um) se o jogo se realizou no Inverno ou 0 (zero) se ocorreu em outra estação. Da mesma forma que no caso da chuva, espera-se que o frio afugente o torcedor dos estádios.

- *WND*, que assume valor 1 (um) se o jogo se realizou em um sábado ou domingo, ou 0 (zero) caso contrário. Espera-se que jogos em fins de semana atraiam mais público por conta da disponibilidade de tempo dos torcedores.
- *NGT*, que assume valor 1 (um) se o jogo ocorreu após às 21h00min ou 0 (zero) se antes desse horário. Jogos que ocorrem muito tarde tenderiam a desestimular a presença do torcedor, que tem compromissos com o trabalho no dia seguinte, de maneira que se espera sinal negativo no coeficiente dessa variável.

Por fim, são necessárias duas considerações sobre o tratamento dos dados. Na base de dados original haviam duas observações com valores ausentes para a variável que indica classificação do time visitante (*CLV*). Por ser um número muito pequeno de observações comparativamente ao tamanho da amostra, o risco de viés na ausência dos dados ficaria minimizado. Dessa forma, seguindo Hair et al (1998) <sup>1</sup>, optamos por não o realizar nenhum teste de casualidade e imputar os dados faltantes através de regressão linear simples<sup>2</sup>, considerando *CLV* como variável resposta e *GLV* (Gols do time visitante nos últimos três jogos), sendo esta última a variável independente. O segundo ponto refere-se a detecção de *outliers* univariados. Os resultados dos testes com *Box-Plot* e com a Distância de Hotelling realizados com as variáveis *PRC*, *POP*, *GLH*, *GLV*, *PRC* e *RAN* resultaram em uma média de 48 *outliers* para cada variável no primeiro critério e 43 no segundo. Analisados em profundidade, concluímos por não excluí-los da amostra, pois não são nem dados coletados incorretamente nem erros de preenchimento, mas são característicos dos fenômenos no qual se inserem.

A tabela 1 apresenta as estatísticas descritivas e os resultados da avaliação da correlação entre as variáveis não categóricas da amostra.

**Tabela 1:** MÉDIA, DESVIO PADRÃO E CORRELAÇÃO ENTRE AS VARIÁVEIS NÃO CATEGÓRICAS.

	Média	Desvio Padrão	RPC	POP	CLH	CLV	PGH	PGV	GLH	GLV	PT	PRC	RAN
RPC	24,440	18,758	1,000										
POP	4,781.837	6,272.672	-0,109	1,000									
CLH	10,64706	5,714243	-0,023	-0,256 *	1,000								
CLV	10,21271	5,818227	-0,002 *	0,017	-0,044	1,000							
PGH	3,842246	2,250822	-0,010	0,151 **	-0,411 ***	0,028	1,000						
PGV	4,409091	2,236913	-0,009	-0,072	0,085	-0,455 ***	-0,040	1,000					
GLH	4,173797	2,048803	0,051	-0,042	-0,263 ***	-0,007	0,575 ***	-0,064	1,000				
GLV	4,451872	2,108729	-0,064 *	-0,009	0,044	-0,301 ***	0,021	0,592 ***	0,025	1,000			
PT	2,526738	1,090088	-0,010	-0,019	0,008	0,032	-0,011	0,018	-0,059	-0,030	1,000		
PRC	17,81305	8,765974	-0,045 **	0,527 ***	-0,374 ***	-0,097 **	0,167 ***	0,020	0,032	0,047 *	0,018	1,000	
RAN	4,393048	10,88607	-0,002 **	-0,066	-0,083	0,041	0,029	-0,079	0,084	-0,025	0,078 ***	-0,026	1,000

Spearman's rank correlation coefficients results: \*  $p < 0,10$ ; \*\*  $p < 0,05$ ; \*\*\*  $p < 0,01$

A correlação positiva e significativa entre *PGH* e *GLH* ( $p = 0,575$ ) e entre *PGV* e *GLV* ( $p = 0,592$ ) nos indicou que deveríamos testar a hipótese 1, que mensura o interesse do torcedor pelo

<sup>1</sup> Hair, Black, Babin & Anderson (2009) indicam como “regra de bolso” que dados ausentes que totalizem menos de 10% das observações podem ser ignorados, exceto quando ocorrem de forma não aleatória.

<sup>2</sup> Estatísticas da regressão entre *CLV* e *GLV*:  $R^2 = 0,54$ ;  $F = 19,39$  ( $p = 0,000$ ); Intercepto = 1,116 (erro padrão = 1,115;  $p = 0,000$ ); Coeficiente de *GLV* = -0,937 (erro padrão = 0,213;  $p = 0,000$ ).

futebol espetáculo, com dois modelos distintos, para evitar os impactos da multicolinearidade. Já com relação às variáveis *POP* e *PRC*, que também apresentaram alta correlação ( $p = 0,527$ ), optamos por manter ambas nos modelos testados, pois pertencem a categorias diferentes (*POP* é variável de controle) e medem fenômenos completamente distintos. No total, buscamos ajustar quatro modelos diferentes, sendo que os modelos 2 e 4 também buscam capturar o efeito regional dos diferentes fatores na variável dependente. A Tabela 2 resume as informações das variáveis utilizadas, sua função de cada uma nos modelos e o sinal esperado de seus coeficientes.

**Tabela 2: RESUMO DE INFORMAÇÕES DAS VARIÁVEIS E MODELOS NOS QUAIS SE APLICAM**

Variável	Descrição	Tipo	Função	Sinal esperado	Modelo 1	Modelo 2	Modelo 3	Modelo 4
LOT	Se o estádio está lotado	Categórica	Resposta	Não se aplica	x	x	x	x
DSudeste	Valor 1 se a partida é na região Sudeste	Categórica	Controle	Positivo		x	x	x
RPC	Renda per capita anual do município em que ocorre o jogo (em Reais)	Númerica	Controle	Positivo	x	x		x
POP	População da cidade em que ocorre o jogo	Númerica	Controle	Positivo	x	x	x	x
WND	Valor 1 se a partida ocorreu no final de semana	Categórica	Controle	Positivo	x	x	x	x
NGT	Valor 1 se a partida ocorreu após 21:00	Categórica	Controle	Negativo	x	x	x	x
RAN	Quantidade de chuva (em mm)	Númerica	Controle	Negativo	x	x	x	x
DInverno	Valor 1 se Inverno	Categórica	Controle	Negativo	x	x	x	x
CLH	Classificação do time da casa	Númerica	Causal	Positivo	x	x	x	x
CLV	Classificação do time visitante	Númerica	Causal	Positivo	x	x	x	x
GLH	Gols marcados pelo time da casa nos 3 últimos jogos	Númerica	Causal	Positivo		x		
GLV	Gols marcados pelo time visitante nos 3 últimos jogos	Númerica	Causal	Positivo		x		
PGH	Pontos ganhos pelo time da casa nos 3 últimos jogos	Númerica	Causal	Positivo	x			x
PGV	Pontos ganhos pelo time visitante nos 3 últimos jogos	Númerica	Causal	Positivo	x			x
PT	Etapa do campeonato (categorias 1, 2, 3 e 4)	Númerica	Causal	Positivo	x	x	x	x
CLS	Valor 1 se a partida for considerada um clássico	Categórica	Causal	Positivo	x	x	x	x
BIG	Valor 1 se a partida possui alguma grande equipe de SP ou RJ	Categórica	Causal	Positivo	x	x	x	x
MTO	Valor 1 se visitante está entre as dez maiores torcidas do Brasil	Categórica	Causal	Positivo	x	x	x	x
PRC	Preço médio do ingresso	Númerica	Causal	Negativo	x	x	x	x

## Metodologia

Como desejamos medir as razões pelos quais os torcedores vão ao estádio, utilizamos um modelo que pressupõe uma variável resposta dicotômica. Assim a variável *LOT* assume o valor 0 (zero) se o estádio não estava lotado no dia do jogo ou 1 (um) se estava lotado. A Regressão Logística é o método indicado para utilização com variáveis resposta dicotômicas. Nesse caso, nosso objetivo é determinar a probabilidade  $P$  de que o estádio esteja lotado, sendo que  $P$  expressa a probabilidade condicional de que  $Y$ , a variável resposta seja igual a 1, dadas as variáveis explicativas do modelo, ou seja,

$$P = E(Y = 1 / \mathbf{x}) = \frac{e^Z}{1 + e^Z}$$

em que  $Z$  é o polinômio que representa a variável resposta do modelo geral em função das variáveis explicativas (indicadas pelo vetor  $x$ ) e seus respectivos parâmetros (betas) que queremos estimar, conforme abaixo.

$$Z = \beta_0 + \beta_1 DSudeste + \beta_2 RPC + \beta_3 POP + \beta_4 RAN + \beta_5 Dinverno + \beta_6 WND + \beta_7 NGT + \beta_8 CLH + \beta_9 CLV + \beta_{10} GLH + \beta_{11} GLV + \beta_{12} PGH + \beta_{13} PGV + \beta_{14} PT + \beta_{15} CLS + \beta_{16} BIG + \beta_{17} MTO + \beta_{18} PRC$$

O modelo a ser testado consiste, portanto, de uma equação com um intercepto e dezoito coeficientes e tem caráter linear nas variáveis, sendo que todas as variáveis estão em nível e possuem o mesmo peso. Retornando ao início<sup>3</sup>, se  $P$ , representa a probabilidade de o estádio lotar, então  $(1 - P)$ , a probabilidade de não lotar, pode ser representada por

$$1 - P = \frac{1}{1 + e^Z}$$

E, portanto, podemos indicar

$$\frac{P}{1 - P} = \frac{1 + e^Z}{1 + e^{-Z}} = e^Z$$

como sendo a razão de chances (*Odds Ratio*) a favor de o estádio lotar em função de todas as variáveis do modelo, contidas em  $Z$ . No entanto, dado que  $Z$  expressa uma relação linear, a interpretação do modelo fica facilitada se determinarmos o logaritmo natural dos dois lados da equação, ou seja,

$$L = \ln\left(\frac{P}{1 - P}\right) = \ln\left(\frac{1 + e^Z}{1 + e^{-Z}}\right) = \ln(e^Z) = Z$$

em que  $L$  indica o logaritmo da razão de chances entre o estádio lotar e não lotar, o que facilita também a estimação e interpretação. Nesse caso, em função das propriedades matemáticas do modelo, se  $L$  for positivo, quanto maior o valor atribuído à uma ou mais variáveis dependentes, maior a chance de que a variável resposta de interesse (no caso, a lotação do estádio) se torne igual a um.

## Resultados

Em linha com os objetivos a serem validados, elaboramos quatro modelos, dois para a Hipótese 1 e dois para a Hipótese 2. Os resultados da Tabela 2 nos mostram as estimativas dos

<sup>3</sup> O raciocínio que se segue está baseado em Gujarati (2006).



coeficientes para os quatro modelos. Todos os modelos são significativos, conforme demonstrando pela estatística de Razão de Verossimilhança (LR), ou seja, para todos rejeitamos a hipótese nula de que todos os coeficientes são simultaneamente igual à zero.

Os modelos 1 e 2 tentam buscar validar a hipótese de que torcedores vão os estádios motivados pela performance recente das equipes em disputa e, portanto, fatores relacionados à esta ajudariam a prever a possibilidade de os estádios lotarem em dias de jogos.

Em primeiro lugar vemos que embora tenham todas as variáveis de controle tenham o sinal esperado, nenhuma delas foi considerada significativa no Modelo 1. Esse resultado, embora contra intuitivo, é consistente com os demais resultados que virão a seguir. Além disso, já nos dá algumas pistas sobre o que vem adiante, ou seja, fatores não relacionados diretamente à dinâmica do esporte não tem nenhum impacto significativo na decisão do torcedor em frequentar o estádio.

Observamos também que as variáveis que medem a classificação das equipes no campeonato foram significativas tanto para o time da casa ( $\hat{\beta} = 0,151$ ;  $p < 0,01$ ) como do time visitante ( $\hat{\beta} = 0,089$ ;  $p < 0,1$ ) e apresentaram o sinal esperado em relação à variável dependente, indicando que estes fatores contribuem de forma consistente para a lotação do estádio. O mesmo comportamento foi observado para as variáveis PT ( $\hat{\beta} = 0,515$ ;  $p < 0,05$ ) e BIG ( $\hat{\beta} = 0,939$ ;  $p < 0,1$ ). No primeiro caso, quanto mais avançado o campeonato, maior a chance de o estádio ficar lotado. No entanto, o resultado previsto pela hipótese 1 para as variáveis que mensuram quantidade de gols nos últimos três jogos (GLH e GLV), e que medem o interesse do torcedor pelo futebol espetáculo não foram significativas. O preço médio dos ingressos também parece não interferir na decisão do torcedor em frequentar o estádio, assim como é indiferente para o resultado final o fato de o time visitante possuir uma grande torcida.

Resultados muito similares foram obtidos no modelo 2. Esse modelo visa entender se o fato do jogo ocorrer na Região Sudeste tem consequências para a lotação do estádio e os resultados mostram que sim, já que o coeficiente é positivo e significativo ( $\hat{\beta} = 1,586$ ;  $p < 0,05$ ). O outro objetivo desse modelo era verificar se o histórico recente de pontos obtidos pelas equipes é importante para motivar o torcedor à ir ao estádio. Para tanto, introduzimos nesse modelo duas variáveis que mensuram pontos ganhos pelo time da casa (PGH) e pelo time visitante (PGV) nos últimos três jogos, retirando às variáveis GLH e GLV, que mensuram a quantidade de gols dos dois times. No entanto, ao contrário do previsto na primeira hipótese e na mesma linha do que observamos no Modelo 1, com relação à quantidade de gols, os coeficiente de PGH e PGV não foram significativos. Esses resultados, nos levam a rejeitar a Hipótese 1 de que o futebol espetáculo é que motiva o torcedor à lotar os estádios.

Os resultados dos modelos 3 e 4 foram altamente consistentes com o previsto na Hipótese 2. Particularmente para o Modelo 4, as classificações da equipe da casa ( $\hat{\beta} = 0,154$ ;  $p < 0,01$ ) e do time visitante ( $\hat{\beta} = 0,081$ ;  $p < 0,1$ ) tiveram o sinal esperado na relação com a variável resposta e foram bastante significativas. Da mesma forma como nos modelos 1 e 2, a etapa do campeonato ( $\hat{\beta} = 0,521$ ;  $p < 0,05$ ) e a presença de uma grande equipe de SP ou RJ ( $\hat{\beta} = 0,865$ ;  $p < 0,1$ ) também foram significativas, embora em menor grau. O Modelo 4 também mostrou que o efeito Regional influi na lotação dos estádios ( $\hat{\beta} = 1,471$ ;  $p < 0,05$ ). Dessa forma temos evidências suficientes para não rejeitar a Hipótese 2 de que o torcedor é pragmático e que os estádios tendem a lotar quando o time da casa está bem classificado no campeonato, mesmo que o tamanho dos coeficientes indique o impacto das variáveis *DSudeste*, *BIG* e *PT* na variável resposta sejam mais fortes do que o das variáveis *CLH* e *CLV*.

**Tabela 3** – FATORES QUE DETERMINAM A LOTAÇÃO NOS ESTÁDIOS; VARIÁVEL DEPENDENTE IGUAL A 1, SE ESTÁDIO LOTADO, OU 0 SE ESTÁDIO NÃO LOTADO NO DIA DO JOGO; COEFICIENTES CALCULADOS VIA REGRESSÃO LOGÍSTICA (MÉTODO DE OTIMIZAÇÃO: STATA'S MODIFIED NEWTON-RAPHSON); ERROS PADRÃO EM ITÁLICO.

Variáveis		Modelo 1 Em busca do espetáculo nacional	Modelo 2 Em busca do espetáculo regional	Modelo 3 Pragmatismo Nacional	Modelo 4 Pragmatismo Regional
Variáveis de controle	Constante	-2,445 1,556	-2,520 1,575	-2,281 * 1,256	-3,003 1,613
	DSudeste - Valor 1 se a partida é na região Sudeste		1,586 ** 0,639		1,471 ** 0,644
	RPC - Renda per capita anual do município em que ocorre o jogo (em Reais)	0,000 0,000	0,000 0,000	0,000 0,000	0,000 0,000
	POP - População da cidade em que ocorre o jogo	0,000 0,000	0,000 0,000	0,000 0,000	0,000 0,000
	RAN - Quantidade de chuva (em mm)	-0,009 0,020	-0,009 0,019	-0,008 0,020	-0,009 0,019
	DIverno - Valor 1 se for inverno no dia do jogo	-0,318 0,490	-0,262 0,493	-0,301 0,480	-0,224 0,498
	WND - Valor 1 se a partida ocorreu no final de semana	0,079 0,649	0,049 0,660	0,021 0,647	0,189 0,650
	NGT - Valor 1 se a partida ocorreu após 21:00	-0,948 1,015	-0,736 1,032	-0,997 1,010	-0,610 1,031
Variáveis do fenômeno	CLH - Classificação do time da casa	0,151 *** 0,050	0,170 *** 0,049	0,178 *** 0,048	0,154 *** 0,046
	CLV - Classificação do time visitante	0,089 * 0,045	0,066 0,042	0,064 0,041	0,081 * 0,046
	GLH - Gols marcados pelo time da casa nos 3 últimos jogos		0,024 0,112		
	GLV - Gols marcados pelo time da casa nos 3 últimos jogos		-0,109 0,042		
	PGH - Pontos ganhos pelo time da casa nos 3 últimos jogos	0,156 0,099			0,119 0,103
	PGV - Pontos ganhos pelo time visitante nos 3 últimos jogos	-0,126 0,099			-0,115 0,114
	PT - Etapa do campeonato (categorias 1, 2, 3 e 4)	0,515 ** 0,219	0,503 ** 0,225	0,500 ** 0,218	0,521 ** 0,223
	CLS - Valor 1 se a partida for considerada um clássico	0,757 0,686	0,675 0,689	0,898 0,663	0,603 0,698
	BIG - Valor 1 se a partida possui alguma grande equipe de SP ou RJ	0,939 * 0,485	0,904 * 0,498	0,964 * 0,482	0,865 * 0,498
	MTO - Valor 1 se visitante está entre as dez maiores torcidas do Brasil	-0,373 0,478	-0,362 0,485	-0,391 0,471	-0,342 0,490
	PRC - Preço médio do ingresso	0,016 0,025	0,015 0,025	0,017 0,025	0,014 0,025
Avaliação do modelo	N	374	374	374	374
	LR chi2(g.l. <sup>a</sup> )	57,710 ***	62,410 ***	54,370 ***	63,520 ***
	Pseudo R <sup>2</sup>	0,264	0,286	0,249	0,291
	Log likelihood	-80,406	-78,057	-82,076	-77,504

\*  $p < 0,10$ ; \*\*  $p < 0,05$ ; \*\*\*  $p < 0,01$

<sup>a</sup>Graus de liberdade do teste Likelihood Rate: 15 para o Modelo 1 e 16 para os modelos 2, 3 e 4.

## Discussão

Na comparação entre os modelos 3 e 4 este último possui maior poder explicativo em razão da maior estatística Pseudo R<sup>2</sup><sup>4</sup>, motivo pelo qual vamos utilizá-lo como base para discutir os resultados à luz dos objetivos do trabalho.

Como vimos na seção dedicada à metodologia, o coeficiente de um modelo logístico é o logaritmo natural da razão de chances. Dessa forma, fazendo a operação inversa, ou seja, calculando-se o valor da função exponencial no coeficiente obtido<sup>5</sup> chegamos à razão de chances entre o estádio lotar e não lotar para cada fator, considerando-se todos os outros fixos. A Tabela 4 traz os resultados da razão de chances calculada para os coeficientes significativos do Modelo 4:

**Tabela 4:** RAZÃO DE CHANCES A FAVOR DE O ESTÁDIO LOTAR, POR VARIÁVEL.

Variável	Razão de chances
Dsudeste - Valor 1 se a partida é na região Sudeste	335,4%
CLH - Classificação do time da casa	16,7%
CLV - Classificação do time visitante	8,4%
PT - Etapa do campeonato (categorias 1, 2, 3 e 4)	68,3%
BIG - Valor 1 se a partida possui alguma grande equipe de SP ou RJ	137,5%

A interpretação desses novos indicadores é direta. Jogos realizados na região Sudeste têm 335,4% mais de chances de lotar o estádio quando comparados a jogos realizados em outras regiões. Interpretação semelhante ocorre com a variável BIG, ou seja, a presença de uma grande equipe de SP ou RJ aumenta a chance de lotação em 137,5% quando comparados aos jogos sem alguma dessas equipes. Quanto ao momento do campeonato, vemos que a cada nova etapa a chance de lotação aumenta em 68,3%, ou seja, o torcedor interessa-se pelos jogos quanto mais próximos da rodada final o campeonato estiver. Por outro lado, para cada nova posição que equipe da casa sobe na classificação do campeonato antes do jogo, a chance do estádio lotar é 16,7%. No caso do time visitante, a possibilidade do estádio lotar aumenta 8,4% para cada nova posição cima na classificação geral.

<sup>4</sup> Infelizmente a documentação do pacote estatístico STATA, que utilizamos para executar a Regressão Logística, limita-se a descrever como interpretar a saída da estatística Pseudo R<sup>2</sup>, não informando qual das possíveis modalidades dessa estatística que o programa adota.

<sup>5</sup> Exemplo: a razão de chances  $R$  para o coeficiente da variável PT é calculada como  $R = [(e^{1,4171}) - 1] \times 100 = 68,3\%$

Na seção anterior verificamos que não temos elementos para aceitar a hipótese de que os torcedores buscam os estádios em função da possibilidade de um grande espetáculo, mas sim porque seu time está bem classificado no campeonato. Uma das possíveis explicações é que a perspectiva de ser o campeão do torneio traz ao torcedor maior prazer do que presenciar um bom futebol. Ou seja, o torcedor tende a ir ao estádio não por se constituir em forma de lazer, como seria uma ida ao cinema ou ao teatro, mas pela expectativa de, em uma data futura, poder comemorar junto com amigos e colegas o título do campeonato, independente da forma como foi obtido. Esta conclusão traz algumas consequências, uma delas é que a Confederação Brasileira de Futebol tem opções de estratégia muito limitadas, caso esteja interessada em aumentar o comparecimento do torcedor aos estádios. Uma dessas opções, por exemplo, é aumentar os investimentos em Marketing para posicionar o futebol como um esporte cada vez mais competitivo, buscando realçar na mente do torcedor em potencial os aspectos mais primitivos de uma disputa e os benefícios relacionados.

No geral, geral, apesar de confirmar a hipótese da visão pragmática do torcedor, vemos que a disputa entre grandes equipes e a proximidade das finais atrai os torcedores mais até do que a própria classificação da equipe.

## Conclusões

Verificamos que os fatores mais importantes que contribuem para lotar um estádio são a classificação do time da casa no campeonato, a etapa do campeonato e se a partida for considerada um clássico. A propensão para lotação do estádio também aumenta se a partida for realizada em algum dos estados da Região Sudeste, exceto o Espírito Santo, que não fazia parte da amostra.

Por outro lado, o desempenho recente das equipes envolvidas, expresso na quantidade de gols marcados ou pontos adquiridos pelas equipes em disputa parecem não ser importantes para incentivar a ida de torcedores aos estádios.

Por fim, fatores naturais como temperatura, chuva ou horário do jogo também não afetam a disposição do torcedor em frequentar o estádio.

As descobertas acima evidenciam a preferência do torcedor pelo futebol de resultados em contraposição ao futebol espetáculo. A implicação em termos de possíveis estratégias a serem utilizadas pelos organizadores de campeonatos de futebol, visando aumentar a lotação dos estádios, são bem limitadas. Para os clubes, o investimento na formação de equipes cada mais competitivas é a estratégia mais indicada.

## Limitações e indicações para pesquisas futuras

Possíveis generalizações dos resultados obtidos neste trabalho deverão considerar que utilizamos como base os dados de uma edição do campeonato brasileiro de futebol. Este campeonato tem algumas características que o tornam singular quando comparados aos campeonatos nacionais de outros países da América do Sul ou Europa. Por exemplo, o tamanho do território brasileiro implica em longos deslocamentos das torcidas dos times visitantes para estar presente nos jogos de seus times, quando realizados em outros estados. Pesquisas futuras poderiam levar esse fator em consideração. Outra possível estratégia para reforçar os resultados encontrados nesse trabalho seria fazer um estudo econométrico de painel com campeonatos brasileiros de vários anos consecutivos e repetir o mesmo estudo com os campeonatos estaduais. Por último, pode ser interessante incluir nos estudos futuros variáveis que mensurem a presença de craques nos times em disputa, bem como o nível de qualidade, segurança e existência de transporte público para os estádios.

## BIBLIOGRAFIA CONSULTADA

GUJARATI, D. – Basic Econometrics, 4<sup>th</sup> ed, *The McGraw Hill Company* (2004)

HAIR, J. F., BLACK, W. C.; BABIN, B. J., & ANDERSON, R. E. – Multivariate Data Analysis, 7<sup>th</sup> ed, *Prentice Hall* (2009)

HOSMER, D. W., LEMESHOW, S. – Applied Logistic Regression, 2<sup>th</sup> ed, *Wiley Interscience Publication* (2000)

PAMPEL, F. C. – Logistic Regression: a Primer – *SAGE Publications* (2000)