UNIVERSIDADE FEDERAL DE JUIZ DE FORA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

Vanessa Rebello Morani

QUE FATORES ESTÃO ASSOCIADOS AO RESULTADO DE UM JOGO DE VOLEIBOL? – Análise Estatística da Superliga Masculina 2011/12

Juiz de Fora

2013

Vanessa Rebello Morani

QUE FATORES ESTÃO ASSOCIADOS AO RESULTADO DE UM JOGO DE VOLEIBOL? – Análise Estatística da Superliga Masculina 2011/12

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito parcial para a obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. Luiz Cláudio Ribeiro

Juiz de Fora

2013

Ficha catalográfica elaborada através do Programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Rebello Morani, Vanessa.
QUE FATORES ESTÃO ASSOCIADOS AO RESULTADO DE UM JOGO DE VOLEIBOL? ? Análise Estatística da Superliga Masculina 2011/12 / Vanessa Rebello Morani. -- 2013.
32 p.

Orientador: Luiz Cláudio Ribeiro
Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Juiz de Fora, ICE/Engenharia, 2013.

1. Voleibol. 2. Análise de Jogo. 3. Regressão Logística. I. Ribeiro, Luiz Cláudio, orient. II. Título.

Vanessa Rebello Morani

QUE FATORES ESTÃO ASSOCIADOS AO RESULTADO DE UM JOGO DE VOLEIBOL? – Análise Estatística da Superliga Masculina 2011/12

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito parcial para a obtenção do grau de Bacharel em Estatística.

Aprovada em 27 de Agosto de 2013

BANCA EXAMINADORA

_____

Luiz Cláudio Ribeiro - Orientador
Doutor em Demografia – UFMG

_____

Camila Borelli Zeller
Doutora em Estatística – UNICAMP

_____

Francisco Zacaron Werneck
Mestre em Educação Física - UGF

AGRADECIMENTOS

A Deus, por me dar força e coragem para enfrentar todas as dificuldades.

À minha família, pelo carinho, apoio e compreensão.

Aos meus colegas de curso, especialmente à Leiliane, Manoel, Jéssica e Franciele, pela ajuda em momentos difícies e fundamentais.

À minha amiga Sarah, por estar ao meu lado durante todos esses anos, nos momentos bons e ruins e por, mesmo longe, sempre acreditar em mim.

Ao meu querido amigo Cadu, pela disposição em ajudar, por estar sempre disponível para me fazer companhia, por me divertir e fazer tudo parecer mais fácil.

À minha irmã e alma gêmea Raquel, pois sem ela eu não chegaria a lugar nenhum.

Aos professores do Departamento pela paciência e disposição em ajudar.

Ao Maurício Bara e Zacaron, pela confiança, apoio e contribuição com o trabalho.

À professora Camila, por ser um exemplo de profissional, dedicada e atenciosa.

Ao professor e orientador Luiz Cláudio, pelos ensinamentos, pela paciência e por acreditar em mim.

RESUMO

Objetivo: Identificar quais fundamentos associados ao jogo permitem discriminar as equipes que alcançaram ou não as quartas de final da Superliga e as equipes vencedoras e perdedoras das partidas.

Método: Esse estudo foi realizado tendo como base amostral os jogos da Superliga Masculina de Voleibol 2011/12. Para comparar as médias dos aproveitamentos dos fundamentos técnicos (ataque, saque, bloqueio, levantamento, recepção e defesa) em função das quartas de final (classificados vs. não classificados) e do resultado do jogo (vitória vs.derrota), utilizou-se o teste t de Student. Para testar associação entre o local e o resultado do jogo, utilizou-se o teste qui-quadrado. Foi realizada uma regressão logística a fim de avaliar a significância de um fundamento estudado, quando controlado pelas demais variáveis, sobre a probabilidade de a equipe se classificar para as quartas de final ou sair vencedora de uma partida. Foram adotados nível de significância de 5% e intervalo de confiança de 95%.

Resultados: Segundo o teste t de Student, os fundamentos ataque, saque, levantamento e defesa estão significativamente associados às Quartas de Final. Após analisada a regressão logística, temos que, para cada incremento de uma unidade na variável explicativa ataque, a chance de classificação aumenta 13% e para cada aumento na unidade da variável defesa, as chances de classificação crescem 8,5%. Todos os fundamentos estudados, além do local do jogo, estão significativamente associados ao resultado, de acordo com os testes t de Student e qui-quadrado. Interpretando o modelo de regressão logística, temos que, para cada incremento na unidade da variável ataque, as chances de vitória aumentam 30,3%. Em relação à defesa, para cada aumento de uma unidade, a chance de vencer fica 20,9% maior. Para cada aumento de uma unidade na variável saque, aumenta-se 18,2% a chance de vitória. Para cada incremento de uma unidade da variável bloqueio, as chances de vitória crescem 15,6%.

Conclusão: os aproveitamentos de ataque e defesa apareceram como preditores da classificação para as quartas de final e a combinação das variáveis ataque, defesa, saque e bloqueio foi preditora do resultado das partidas. Os modelos de regressão lógistica para as quartas de final e resultado do jogo tiveram uma porcentagem de acerto de 74,2% e 81,6%, respectivamente.

Palavras-chave: voleibol, análise de jogo, regressão logística.

ABSTRACT

Objective: To identify which fundamentals associated to the match are able to discriminate the teams that have reached or not the Super-league quarter finals and the winners and losers of the matches.

Method: This paper was based on a sampling of Men's Volleyball Super-league 2011/12 matches. To compare the means of the fundamentals performances (attack, serving, blocking, setting, reception and defense) according to the quarter finals (classified vs. unclassified) and the match result (win vs. loss), we used the t Student test. To test the association between the match location and match result, we used the chi-square test. It was applied a logistic regression to assess the significance of a studied fundamental, when controlled by the other variables, over the likelihood of the team to qualify for quarter finals or to win a match. We adopted a significance level of 5% and a confidence interval of 95%.

Results: According to the t Student test, the fundamentals attack, serving, setting and defense are significantly associated with the quarter finals. After analyzing the logistic regression, we have that for each increase of one unit in the explanatory variable attack, the chance of ranking increases 13% and for each unit increment in the variable defense, the chances of ranking grow 8.5%. All studied fundamentals, besides match location, are significantly associated with the match result, according to the t Student and chi-square tests. Interpreting the logistic regression model, we have that for each unit increase in the variable attack, the chances of winning increase 30.3%. Regarding the defense, for each one-unit increment, the chance of winning is 20.9% higher. For each extra one-unit in the variable serve, the chance of winning increases 18.2%. For each increment of one unit of the variable block, the chances of victory increase 15.6%.

Conclusion: the attack and defense performances showed as predictors of the quarterfinals classification and the combination of the variables attack, defense, serving and blocking was predictor of the match result. The logistic regression models for the quarter finals and match result had a correct overall percentage of 74.2% and 81.6%, respectively.

Keywords: volleyball, match analysis, logistic regression.

SUMÁRIO

1 INTRODUÇÃO

Os jogos esportivos coletivos são caracterizados por haver situações frequentes de grande imprevisibilidade e isso faz com que os seus praticantes e treinadores necessitem adotar permanentemente atitudes táticas e estratégicas (GRECO & CHAGAS, 1992).

De acordo com a Federação Internacional de Voleibol (FIVB), o esporte apresenta seis fundamentos: saque, recepção, levantamento, ataque, bloqueio e defesa. O percentual de aproveitamento destes fundamentos durante o jogo determina o resultado final da partida.

Dentre os fundamentos do vôlei, o ataque é o que apresenta a maior correlação com o sucesso (BELLENDIER, 2002; MARCELINO & MESQUITA, 2006; MARCELINO, et al., 2008). Compreender a combinação ideal dos fundamentos auxilia no alcance do sucesso nessa modalidade esportiva (BAACKE, 1972).

A análise de jogo tem por objetivo quantificar e analisar os eventos que ocorrem durante a competição e permite identificar pontos fortes e fracos da própria equipe e do adversário, gerando implicações para a elaboração de treinamentos e táticas de jogo (CARLING et al., 2009). A análise das estatísticas de jogo é uma das ferramentas que possibilitam o entendimento do comportamento individual e coletivo durante a competição (HUGHES & BARTLETT, 2002).

Para o planejamento e o controle do treinamento, para o teste de sistemas de ataque, de defesa e de cobertura e para determinar o coeficiente de êxito dos jogadores, muitas competições deveriam ser anotadas e avaliadas estatisticamente. Os resultados podem proporcionar dados importantes ao técnico, ainda durante o jogo para que possam ser tomadas medidas táticas adequadas (DURRWACHTER, 1984).

Atualmente, treinadores de voleibol podem ter acesso e fazer uso das estatísticas do jogo durante e após cada set, podendo alterar seus planos de acordo com os resultados obtidos (GARCIA-HERMOSO et al., 2013). Por isso a análise estatística é tão importante no voleibol.

Existem diversos estudos sobre análise de jogo no futebol (LAGO-PEÑAS, et al., 2010), futebol americano (COHEA & PAYTON, 2011), rugby (ORTEGA et al., 2009), dentre outros. No voleibol, existem variados estudos sobre a análise de sets (GARCIA-HERMOSO et al.; 2013; MARCELINO et al., 2010; MARELIC et al., 2004;

ROMERO et al., 2012), análise de levantadores e do levantamento (AFONSO, et at., 2012; MATIAS & GRECO, 2011).

A análise de jogo é de extrema importância no voleibol e, por isso, o objetivo do presente estudo foi analisar os jogos da temporada da Superliga Masculina de Voleibol (Série A) – Temporada 2011/2012 e, através de métodos estatísticos, identificar quais fundamentos associados ao jogo permitem discriminar as equipes que alcançaram ou não as quartas de final da competição e, principalmente, as equipes vencedoras e perdedoras das partidas.

2 MÉTODO

2.1 Amostra

Os dados coletados foram fornecidos pela SCConsultoria, empresa privada dedicada à medida de performance das equipes de voleibol da Superliga, promovida pela Confederação Brasileira de Voleibol – CBV. Os dados, de domínio público, estão disponíveis no site da CBV (www.cbv.com.br). No total, 12 equipes disputaram 567 sets em 148 partidas, gerando, por jogo, aproximadamente 8 pontos de saque, 95 de ataque, 18 de bloqueio, 52 ações excelentes de levantamento, 64 de defesa e 71 de recepção.

2.2 Variáveis da análise

As variáveis estudadas são divididas entre variáveis dependentes, e independentes (Tabela 1).

As variáveis dependentes analisadas são: Resultado do Jogo (vitória vs. derrota) e Quartas de Final (classificados vs. não classificados).

As variáveis independentes são: Local do Jogo (em casa vs. fora de casa), Aproveitamento do Ataque (aqui, incluem-se ações de ataque após a recepção do saque adversário e ações de ataque após defesa, ou contra-ataque), Aproveitamento do Bloqueio, Aproveitamento do Saque, Aproveitamento do Levantamento, Aproveitamento da Defesa e Aproveitamento da Recepção.

Algumas variáveis não existiam inicialmente e foram criadas a partir de outras já existentes no banco de dados. Uma delas foi a variável “quartas de final”, onde, a partir da colocação de cada equipe no campeonato, separou-se as classificadas das não classificadas. Assim, criou-se essa variável no banco de dados rotulando como “0” as quatro equipes não classificadas e como “1”, as oito classificadas para as quartas de

final. As demais variáveis criadas foram os "aproveitamentos" dos fundamentos *ataque, bloqueio, saque, levantamento, defesa e recepção*. Elas foram criadas a partir de uma simples proporção entre duas variáveis originais do banco de dados, ou seja, dividindo o número total de pontos feitos (ou execução excelente nos casos de recepção, defesa e levantamento) pelo número total de ações de cada fundamento estudado.

Tabela 1 - Variáveis estudadas

Variáveis	**Descrição**
Variáveis Dependentes	
Resultado do Jogo	vitória vs. derrota
Quartas de Final	classificados vs. não classificados
Variáveis Independentes	
Aproveitamento de Ataque	Total de pontos de ataque/total de ações de ataque
Aproveitamento de Saque	Total de pontos de saque/total de ações de saque
Aproveitamento de Bloqueio	Total de pontos de bloqueio/total de ações de bloqueio
Aproveitamento de Defesa	Total de ações excelentes de defesa/total de ações de defesa
Aproveitamento de Levantamento	Total de ações excelentes de levantamento/total de ações de levantamento
Aproveitamento de Recepção	Total de ações excelentes de recepção/total de ações de recepção

2.3 Procedimentos Estatísticos

Neste trabalho foi utilizada a análise de estatísticas descritivas tais como média, mínimo, máximo e desvio padrão a fim de descrever e resumir os dados estudados pelo resultado do jogo. Analisou-se, também, variações relativas (ou taxas de variação), que nos permitiu verificar qual foi a variação, em porcentagem, das médias dos aproveitamentos dos fundamentos, comparando vitória em relação à derrota da partida e classificação em relação à não classificação para as quartas de final da competição.

Para tratamento e análise dos dados foi utilizada a média das variáveis de aproveitamento dos fundamentos. Com o objetivo de testar a existência de diferenças significativas entre essas médias pelo resultado do jogo ou pela classificação para as quartas de final, utilizou-se o teste t de Student (LARSON & FARBER, 2004). Para testar a significância do local do jogo para o resultado da partida, usou-se o teste qui-quadrado (AGRESTI, 2007). É importante ressaltar que os pressupostos dos testes utilizados não foram violados.

Por fim, foi conduzida uma regressão logística (HOSMER & LEMESHOW, 1989) a fim de avaliar a significância dos fundamentos estudados sobre a probabilidade de a equipe se classificar para as quartas de final e de sair vencedora de uma partida, quando as demais variáveis do modelo eram mantidas constantes. A regressão logística binária é um método de predição multivariado utilizado na explicação de eventos categóricos binários (neste caso, classificado vs. não classificado e vitória vs. derrota). Ela avalia a probabilidade de obtenção de um dos resultados dado um conjunto de variáveis explicativas.

Em todas as análises utilizou-se o nível de significância de 5%.

3 RESULTADOS E DISCUSSÃO

3.1 Estatísticas Descritivas

Ao analisar algumas estatísticas descritivas das variáveis de aproveitamento pelo resultado do jogo (Quadro 1), tem-se que a média de todos fundamentos estudados foram mais elevadas quando as equipes venceram as partidas, o que é esperado. Em relação ao desvio padrão absoluto, nota-se que o fundamento recepção foi o que apresentou maior variação, tanto nas situações de vitória, quanto nas de derrota e o saque obteve o menor desvio padrão. Porém, quando se trata da variabilidade dos dados em relação à média, devemos calcular o coeficiente de variação. Ele é resultado da divisão do desvio padrão pela média. Com isso, temos que o fundamento que apresentou a menor variação foi o ataque e o que mais variou foi o saque.

Quadro 1

Estatísticas Descritivas						
	Aproveitamento	Média	Mínimo	Máximo	Desvio Padrão	CV
Derrota	ataque	45,9%	31,3%	62,9%	5,7%	12,4%
	bloqueio	17,3%	3,8%	35,4%	5,9%	34,1%
	saque	4,1%	0%	11,0%	2,5%	61,0%
	levantamento	26,0%	1,2%	46,1%	8,5%	32,7%
	defesa	48,4%	22,2%	70,2%	7,8%	16,1%
	recepção	48,6%	15,9%	72,0%	11,5%	23,6%
Vitória	ataque	52,9%	37,3%	69,2%	6,1%	11,5%
	bloqueio	21,0%	7,4%	40,5%	6,3%	30,0%
	saque	5,1%	0%	15,0%	2,6%	51,0%
	levantamento	29,9%	7,1%	53,5%	9,9%	33,1%
	defesa	54,5%	34,4%	72,7%	7,2%	13,21%
	recepção	52,9%	22,5%	81,3%	12,2%	23,1%

Fonte dos dados brutos: Confederação Brasileira de Voleibol - CBV

3.2 Análises bivariadas

3.2.1 Classificação para as quartas de final

A Tabela 2 apresenta os resultados da análise da classificação para as quartas de final. Quando testamos os fundamentos do vôlei, concluímos que quatro dos seis fatores analisados aparecem significantemente associados à classificação. São eles o ataque (p = 0,000), o saque (p = 0,009), o levantamento (p = 0,012) e a defesa (p = 0,000). O fundamento recepção apresentou um p-valor próximo do nível de significância adotado, ou seja, dependendo do estudo , ele poderia ser considerado um fator significativo para a classificação.

Tabela 2: Significância dos fatores estudados em relação às Quartas de Final

Quartas de Final	Fatores Estudados					
	ataque	bloqueio	saque	levantamento	defesa	recepção
p-valor	0,000	0,178	0,009	0,012	0,000	0,056

Fonte dos dados brutos: Confederação Brasileira de Voleibol - CBV

3.2.1.1 Variações Relativas

A Tabela 3 apresenta as variações relativas das médias dos aproveitamentos dos fundamentos estudados, comparando se as equipes foram ou não classificadas para as quartas de final. Ao analisá-la, primeiramente, nota-se que o ataque, a defesa e a recepção foram os fundamentos que obtiveram os desempenhos mais elevados, no que diz respeito aos valores das médias. Porém, vê-se que, em relação às taxas de variação, o saque apresentou o maior destaque dentre os fundamentos. As equipes classificadas tiveram as médias do aproveitamento desse quesito 21,55% maior do que as equipes não classificadas. O levantamento apareceu com a segunda maior diferença. As equipes que passaram para a fase das quartas de final tiveram um aproveitamento dessa função 11,79% maior do que aquelas que não passaram. A média de ataque daqueles classificados foi 9,94% superior, quando comparados à dos não classificados. A defesa das equipes também apresentaram uma diferença elevada, nesse caso. A média do aproveitamento desse fundamento para as equipes que passaram da primeira fase foi 9,12% maior do que o daquelas que não passaram. Em relação ao bloqueio, a taxa de variação foi de somente 5,87%. Esse valor pode se dever ao fato de a diferença das médias, para esse caso, não ter sido significativa (p = 0,178).

Tabela 3: Variações Relativas do fundamentos para as Quartas de Final

Quartas de Final	Aproveitamentos					
	ataque	bloqueio	saque	levantamento	defesa	recepção
Não Classificado	0,462	0,184	0,040	0,258	0,484	0,487
Classificado	0,508	0,195	0,049	0,289	0,528	0,516
p-valor	0,000	0,178	0,009	0,012	0,000	0,056
VR (%)	9,94%	5,87%	21,55%	11,79%	9,12%	5,93%

Fonte dos dados brutos: Confederação Brasileira de Voleibol – CBV

3.2.1.2 Regressão Logística

O objetivo da regressão logística é criar um modelo que permita a predição de valores tomados por uma variável categórica (classificação para as quartas de final) a partir de uma série de variáveis explicativas contínuas que, nesse caso, seriam os aproveitamentos dos fundamentos analisados.

No modelo de regressão logística para as quartas de final, primeiramente, foram colocados os efeitos principais e a variável "local do jogo". Somente os fatores ataque, saque e defesa foram significativos.

Foram testadas algumas interações no modelo, tais como entre bloqueio e defesa, levantamento e ataque, saque e bloqueio, saque e defesa e, finalmente, recepção e levantamento, porém nenhuma delas obteve resultado significativo.

Com isso, o modelo final contém os fundamentos ataque e defesa (Tabela 4). Interpretando esse modelo, temos que, mantendo a outra variável constante, para o aumento de uma unidade na variável explicativa ataque, a chance de classificação para as quartas de final (evento de interesse) aumenta 13% (IC: 8%; 18,3%). Para cada incremento na unidade da variável defesa, as chances de classificação crescem 8,5% (IC: 4,7%; 12,4%).

É interessante destacar que o coeficiente da variável ataque foi maior do que o da defesa. De fato, na prática, o ataque é um fundamento mais decisivo.

O modelo proposto foi composto pela equação $Y = -9{,}175 + 0{,}122(ataque) + 0{,}081(defesa)$. Esse modelo obteve uma porcentagem de acerto de 33% nos casos de não classificação e 91,8% nos casos de classificação para as quartas de final, gerando uma porcentagem geral de acerto de 74,2% (Tabela 5).

Tabela 4: Regressão Logística para as Quartas de Final

Variável	β	valor de p	exp(β)	IC 95% para exp(β)	
				lim.inferior	lim. superior
Modelo completo					
local do jogo	-0,395	0.178	0.674	0.379	1,197
ataque	0,106	0.000	1,112	1,058	1,168
bloqueio	0,015	0.520	1,015	0.969	1,063
saque	0,119	0.048	1,127	1,001	1,268
levantamento	0,013	0.454	1,014	0.978	1,050
defesa	0,089	0.000	1,093	1,053	1,134
recepção	0,008	0.512	1,008	0.984	1,034
Modelo parcimonioso					
ataque	0,122	0.000	1,130	1,080	1,183
defesa	0,081	0.000	1,085	1,047	1,124

Fonte dos dados brutos: Confederação Brasileira de Voleibol – CBV

Tabela 5: Poder de predição do modelo de regressão

Quartas de Final	Preditos		Total (%)
Observados	Não Classificado	Classificado	
Não Classificado	29	59	33,0
Classificado	17	190	91,8
Total (%)			74,2

Fonte dos dados brutos: Confederação Brasileira de Voleibol – CBV

3.2.2 Resultado do jogo

As Tabelas 6 e 7 apresentam as estatísticas relacionadas ao resultado da partida. Observa-se que todos os fatores estudados estão significantemente associados ao resultado. Dentre esses fatores, apesar de significativo, o menos expressivo foi o local do jogo (p = 0,048). Ou seja, esse fator parece ser importante, porém não se pode afirmar que seja tão fundamental para uma equipe vencer quanto os demais.

Somente a equipe campeã, *Cruzeiro,* apresentou diferença significativa nos resultados (p = 0,016). Ou seja, há indícios de que, para essa equipe em especial, jogar dentro de casa foi um fator decisivo.

Tabela 6: Resultado do jogo em função do local do jogo

Local do Jogo	Resultado do Jogo				Total
	Derrota		Vitória		
	n	%	n	%	
Fora de Casa	82	55.8%	65	44.2%	147
Em casa	66	44.3%	83	55.7%	149
Total	148	50.0%	148	50.0%	296

Fonte dos dados brutos: Confederação Brasileira de Voleibol – CBV

Tabela 7: Significância dos fatores estudados em relação ao Resultado do Jogo

Resultado do Jogo	Fatores Estudados						
	ataque	bloqueio	saque	levantamento	defesa	recepção	local do jogo
p-valor	0,000	0,000	0,001	0,000	0,000	0,002	0,048

Fonte dos dados brutos: Confederação Brasileira de Voleibol – CBV

3.2.2.1 Variações Relativas

A Tabela 8 apresenta as variações relativas das médias dos aproveitamentos de todos os fundamentos pelo resultado do jogo. Mais uma vez, as médias de ataque, defesa e recepção obtiveram os valores mais elevados. No entanto, o saque, novamente, obteve a maior taxa de variação. O aproveitamento das equipes que sairam vitoriosas das partidas foi, em média, 24,15% maior quando comparado ao aproveitamento desse fundamento quando as equipe não venceram. Diferentemente do caso anterior, em relação ao resultado da partida, a variação relativa do fundamento bloqueio foi alta. A média das equipes vencedoras foi 21,83% maior do que a média das equipes perdedoras. Isso pode nos levar a crer que se a equipe saca e bloqueia bem durante um jogo, as chances de ela vencer o jogo são altas. Os fundamentos ataque e levantamento também chamam atenção. Ambos tiveram aproveitamentos aproximadamente 15% mais elevados quando as equipes venceram as partidas. Novamente, parece que quanto mais bem executados forem esses fundamentos por uma equipe, maiores são as chances de ela sair-se vencedora de um jogo. De acordo com essa análise, é razoável afirmar que esses fundamentos são importantes para uma equipe vencer uma partida de voleibol.

Tabela 8: Variações Relativas dos fundamentos para o Resultado do Jogo

Resultado do Jogo	Aproveitamentos					
	ataque	bloqueio	saque	levantamento	defesa	recepção
Derrota	0,459	0,172	0,041	0,260	0,484	0,486
Vitória	0,529	0,210	0,051	0,299	0,545	0,529
p-valor	0,000	0,000	0,001	0,000	0,000	0,002
VR (%)	15,18%	21,83%	24,15%	15,15%	12,65%	8,78%

Fonte dos dados brutos: Confederação Brasileira de Voleibol – CBV

3.2.2.2 Regressão Logística

A princípio, foram colocados no modelo de regressão a variável "local do jogo" e os seis fundamentos estudados. O local da partida, que pelo teste qui-quadrado realizado anteriormente, foi significativamente associado ao resultado do jogo, não obteve resultado significativo após a regressão logística. Os fatores "levantamento" e "recepção" também foram não significativos. De fato, esses dois fundamentos apresentaram forte correlação com as outras variáveis independentes estudadas, quando analisadas as correlações entre elas. É importante verificar a existência de alta correlação entre as variáveis, pois o modelo de regressão é sensível à colinearidade entre as mesmas (HAIR et al., 1998).

Foram testadas, também, algumas interações que poderiam ser interessantes no modelo proposto, tais como entre bloqueio e defesa, levantamento e ataque, saque e bloqueio e saque e defesa, além de recepção e levantamento. Nenhuma delas, porém, apresentou resultado significativo.

Na regressão logística, o coeficiente de uma variável explicativa é interpretado pela inclinação ou pela taxa de mudança da função logito para cada incremento de uma unidade no valor dessa variável com as demais variáveis explicativas fixas. No contexto do presente trabalho, temos que para o aumento de uma unidade na variável explicativa ataque, a chance de vitória (evento de interesse) aumenta em 30,3% (IC: 21,5%; 39,8%). Para cada incremento na unidade da variável bloqueio, as chances de vitória crescem 15,6% (IC: 8,6%; 23%). Em relação ao saque, para cada incremento de uma unidade, a chance de vencer fica 18,2% maior (IC: 2,5%; 36,2%). Para cada aumento de uma unidade na variável defesa, aumenta-se 20,9% a chance de vitória (IC: 14,4%; 27,7%). Em cada um dos quatro casos, teríamos as demais variáveis fixadas. Todos os fundamentos foram altamente significativos (Tabela 9).

Nesse caso, o ataque também apresentou o maior coeficiente, reforçando a ideia de que esse fundamento é o mais importante dentre os demais.

O modelo final foi composto pela equação $Y = -26{,}323 + 0{,}265(ataque) + 0{,}145(bloqueio) + 0{,}167(saque) + 0{,}189(defesa)$. Esse modelo obteve uma porcentagem de acerto de 81% nos casos de derrota e 82,2% nos casos de vitória, o que gerou uma porcentagem geral de acerto de 81,6% (Tabela 10).

Tabela 9: Regressão Logística para o Resultado do Jogo

Variável	β	valor de p	exp(β)	IC 95% para exp(β)	
				lim. inferior	lim. superior
Modelo completo					
local do jogo	-0,256	0.467	0,774	0,388	1,543
ataque	0,253	0.000	1,288	1,197	1,386
bloqueio	0,150	0.000	1,162	1,089	1,240
saque	0,177	0.017	1,194	1,032	1,380
levantamento	0,014	0.526	1,014	0,971	1,059
defesa	0,197	0.000	1,218	1,149	1,290
recepção	0,017	0.251	1,018	0,988	1,048
Modelo parcimonioso					
ataque	0,265	0.000	1,303	1,215	1,398
bloqueio	0,145	0.000	1,156	1,086	1,230
saque	0,167	0.021	1,182	1,025	1,362
defesa	0,189	0.000	1,209	1,144	1,277

Fonte dos dados brutos: Confederação Brasileira de Voleibol – CBV

Tabela 10: Poder de predição do modelo de regressão

Resultado do Jogo	Preditos		Total (%)
Observados	Derrota	Vitória	
Derrota	119	28	81,0
Vitória	26	120	82,2
Total (%)			81,6

Fonte dos dados brutos: Confederação Brasileira de Voleibol - CBV

3.3 Ações de Jogo

As equipes que venceram as partidas conquistaram, em média, mais pontos de ataque (49,59 ± 20,23), de bloqueio (9,93 ± 7,09) e saque (4,53 ± 4,44) do que as equipes derrotadas (45,4 ± 22,88; 7,83 ± 7,17; 3,41 ± 4,44, respectivamente). Foi também nas equipes que venceram as partidas que se observou maiores médias de defesas e levantamentos excelentes (33,99 ± 18,52 vs. 30,67 ± 19,47) e (26,88 ± 18,62 vs. 25,09 ± 19,14), respectivamente. Somente a média do fundamento recepção

excelente foi superior para as equipes que perderam as partidas (36,61 ± 22,39 vs. 34,39 ± 21,67) (MARCELINO et al., 2010).

3.4 Análise por equipe

Foi feita uma análise da ordenação dos percentuais de vitória - dentro e fora de casa - e dos aproveitamentos dos seis fundamentos pelas equipes estudadas (Quadro 2).

É interessante notar que a primeira colocada da Superliga, Cruzeiro, teve o maior percentual de vitórias dentro de casa (93,3%), enquanto que a segunda colocada, a equipe Vôlei Futuro, teve o maior percentual de vitórias fora de casa (73,3%). Esses números parecem indicar que o local do jogo não está associado ao resultado.

Outro fato que chamou atenção foi que três equipes classificadas para a fase de quartas de final, dentre as doze que disputaram a Superliga de vôlei, obtiveram as piores médias de aproveitamentos em três diferentes fundamentos. A equipe de São Bernardo do Campo, BMG, que ocupou oitava posição, obteve a pior média de ataque (0,440). A equipe VIVO, de Minas Gerais, terceira colocada, apareceu com a pior média de bloqueio (0,178) e a quarta colocada, RJX, com a pior média de recepção (0,453).

Fato semelhante ao citado anteriormente aconteceu com equipes não classificadas para as quartas de final. A equipe de Volta Redonda, décima colocada, obteve a terceira melhor média de levantamento (0,321) e a equipe de Montes Claros, décima primeira, obteve as terceiras e quartas maiores médias dos fundamentos recepção e saque (0,544 e 0,051), respectivamente.

Os resultados dessa análise evidenciam que não somente um fator e sim, um conjunto deles, podem levar uma equipe a obter sucesso no voleibol.

Quadro 2

Classificação	Percentual de vitórias			Aproveitamentos				
	Em casa	Fora	Ataque	Bloqueio	Saque	Levantamento	Defesa	Recepção
SADA CRUZEIRO	**93,3 (1)**	53,8 (5)	**0,547 (1)**	0,190 (6)	0,049 (6)	0,336 (2)	**0,562 (1)**	0,557 (2)
VOL. FUTURO	61,5 (6)	**73,3 (1)**	0,528 (2)	**0,215 (1)**	0,054 (2)	0,223 (11)	0,5312(3)	0,538 (4)
VIVO/MINAS	76,9 (2)	57,1 (4)	0,518 (4)	**0,178(12)**	0,050 (5)	0,319 (4)	0,496 (9)	0,512 (6)
RJX	69,2 (5)	42,9 (6)	0,501 (6)	0,195 (4)	0,052 (3)	0,285 (6)	0,537 (2)	**0,453(12)**
SESI-SP	75 (3)	58,3 (3)	0,511 (5)	0,202 (2)	0,040 (9)	0,266 (8)	0,528 (5)	**0,581 (1)**
CIMED	71,4 (4)	63,6 (2)	0,522 (3)	0,189 (7)	**0,062 (1)**	**0,345 (1)**	0,510 (7)	0,532 (5)
MEDLEY CAMPINAS	50 (8)	41,7 (7)	0,488 (7)	0,196 (3)	0,040(8)	0,306 (5)	0,524 (6)	0,507 (7)
BMG - SBC	53,8 (7)	33,3 (9)	**0,440(12)**	0,193 (5)	0,040(7)	0,234 (10)	0,531 (4)	0,473 (9)
VOLTA REDONDA	18,2 (11)	36,4 (8)	0,469 (9)	0,180(11)	**0,033(12)**	0,321 (3)	0,490(10)	0,466(10)
MONTES CLAROS	36,4 (9)	27,3(10)	0,472 (8)	0,184(10)	0,051 (4)	0,280 (7)	0,499 (8)	0,544 (3)
UFJF	**16,7(12)**	20 (11)	0,444(11)	0,186 (9)	0,037(11)	0,237 (9)	0,479(11)	0,461(11)
LONDRINA	20 (10)	**8,3 (12)**	0,463(10)	0,187 (8)	0,038(10)	**0,197 (12)**	**0,466(12)**	0,478 (8)

Fonte dos dados brutos: Confederação Brasileira de Voleibol - CBV

4 CONCLUSÃO

Os fatores associados à classificação para as quartas de final foram o ataque e a defesa. Tem-se que, para cada incremento de uma unidade na variável explicativa ataque, a chance de classificação para as quartas de final aumenta 13%, mantendo a outra variável constante. Para cada aumento na unidade da variável defesa, as chances de classificação crescem 8,5%, com a outra variável mantida fixa. A combinação desses fundamentos apresentou um poder preditivo total de 74,2%.

Os fundamentos ataque, defesa, saque e bloqueio apareceram associados ao resultado do jogo, nessa ordem de importância. Com as demais variáveis constantes, tem-se que, para cada incremento na unidade da variável ataque, as chances de vitória aumentam 30,3%. Em relação à defesa, para cada aumento de uma unidade, a chance de vencer fica 20,9% maior. Para cada aumento de uma unidade na variável saque, aumenta-se 18,2% a chance de vitória. Finalmente, para cada incremento de uma unidade da variável bloqueio, as chances de vitória crescem 15,6%. O poder total de predição desse conjunto de variáveis foi de 81,6%.

REFERÊNCIAS

1. AFONSO, J.; ESTEVES, F., ARAÚJO, R., THOMAS, L., & MESQUITA, I. Tactical determinants of setting zone in elite men's volleyball. Journal of Sports Science and Medicine, v. 11, p. 64-70, 2012.
2. AGRESTI, A. An introduction to categorical data analysis. 2nd edition. Wiley, 2007.
3. BAACKE, H. Mini volleyball. In: CONFEDERAÇÃO Brasileira de Voleibol. Manual do Treinador. Brasília: Secretaria de Educação Física e Desportos; Subsecretária de Desportos. 1972.
4. BELLENDIER, J. Ataque de rotación en el voleibol, un enfoque actualizado. *Efdeportes – Revista Digital*, ano 8, 51. Disponivel em http://www.efedeportes.com/efd60. 2002.
5. CARLING, C.;_REILLY, T.; WILLIAMS, A. M. Performance Assessment for Field Sports. Routledge, 2009.
6. COHEA, C; PAYTON, M. E. Relationships Between Player Actions and Game Outcomes in American Football Sportscience 15, 19-24, 2011.
7. DURRWACHTER, G. Voleibol: treinar jogando. Rio de janeiro: Ao Livro Técnico, 1984.
8. GARCIA-HERMOSO, A.; ROMERO, C. D.; SAAVEDRA, J. M. Discriminatory Power of Game Related Statistics in 14-15 Year Age Group Male Volleyball, According to Set. Perceptual and Motor Skills: Volume 116, Issue, pp. 132-143, 2013.
9. GRECO, P. J.; CHAGAS, M. H. Considerações teóricas da tática nos jogos esportivos coletivos. Revista Paulista de Educação Física, São Paulo, v. 6, n. 2, p. 47-58, jul./dez. 1992.
10. HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. Multivariate data analysis, 5 ed. Upper Saddle River, NJ: Prentice-Hall. 1998.
11. HOSMER, D. W; LEMESHOW, S. Applied logistic regression. New York: JohnWiley & Sons, Inc., 1989.
12. HUGHES, M.; BARTLETT, R. The use of performance indicators in performance analysis. Journal of Sports Science 20, 739 – 754, 2002.

13. LAGO-PEÑAS, C; LAGO-BALLESTEROS, J; DELLAL, A. Game-related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league. Journal of Sport Science and Medicine, 10 (2), p.254-261, 2010.
14. LARSON, R; FARBER, B. Estatística Aplicada. 2ª. ed. São Paulo, Pearson Prentice Hall, 2004.
15. MARCELINO, R.; MESQUITA, I. Characterizing the efficacy of skills in high performance competitive volleyball. *World Congress of Performance Analysis of Performance,* 7, 491-496, 2006.
16. MARCELINO, R.; MESQUITA, I.; AFONSO, J. The weight of terminal actions in volleyball: contributions of the spike, serve, and block for the teams' rankings in the World League 2005. *International Journal of Performance Analysis in Sport,* 8, 1-7, 2008.
17. MARCELINO, R.; MESQUITA, I.; SAMPAIO, J.; MORAES, J. C. Estudo dos indicadores de rendimento em voleibol em função do resultado do set. Revista brasileira de educação física e esporte, São Paulo, v. 24 n. 1, p. 69-78, Jan/Mar, 2010.
18. MARELIC, N.; RESETAR, T.; JANKOVIC, V. Discriminant analysis of the sets won and the sets lost by one team in A1 Italian volleyball league-a case study. Kinesiolog**y,** Zagreb, v.36, n.1, p.75-82, 2004.
19. MATIAS, C. J. A. S.; GRECO, P. J. Conhecimento tático-estratégico dos levantadores brasileiros campeões de voleibol: da formação ao alto nível. Rev. bras. educ. fís. esporte (Impr.), São Paulo , v. 25, n. 3, Sept. 2011 .
20. ORTEGA, E.; VILLAREJO, D; PALAO, J.M. Differences in game statistics between winning and losing rugby teams in the Six Nations Tournament. Journal of Sports Science and Medicine 8, 523-527, 2009.
21. ROMERO, C. D.; GARCÍA-HERMOSO, A. El set cerrado en voleibol. Diferencias y poder discriminatorio de las acciones finales en etapas de formación. Retos. Nuevas tendencias en Educación Física, Deporte y Recreación, n. 21, p. 67-70, 2012.

APÊNDICE A

Teste t de duas amostras para verificar diferença entre médias

Um **teste t de duas amostras** é usado para testar a diferença entre duas médias populacionais μ_1 e μ_2 quando uma amostra é selecionada aleatoriamente de cada população. Para realizar esse teste, cada população deve ser normalmente distribuída, as amostras devem ser independentes e o tamanho de pelo menos uma delas deve ser inferior a 30. A estatística teste padronizada é

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

Se as variâncias populacionais forem iguais, então

$$g.l. = n_1 + n_2 - 2$$

e

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Se as variâncias populacionais não forem iguais, então g.l. é o menor dentre $n_1 - 1$ ou $n_2 - 1$ e

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Deve-se estabelecer as hipóteses nula e alternativa (H_0 e H_1) e especificar o nível de significância (α) para interpretar a decisão no contexto estudado.

Para exemplificar a aplicação do teste t, usaremos os dados das Tabelas A.1 e A.2 abaixo. Suponhamos que as distribuições do aproveitamento de saque sejam normalmente distribuídas nas duas populações (vitória e derrota) e que essas distribuições sejam independentes. Assumindo que as variâncias populacionais sejam iguais, queremos testar se a variável *aproveitamento de saque* é significativa para o resultado do jogo. Ou seja, testaremos se as médias dessa variável são diferentes nos casos de derrota e vitória.

Assim, as hipóteses nula e alternativa são:

$$H_0 = \mu_1 = \mu_2 \text{ e } H_1 = \mu_1 \neq \mu_2$$

Tendo em vista que assumimos variâncias iguais, usamos g.l. = 148 + 146 - 2 = 292. Uma vez que o teste é bicaudal com g.l. = 292 e α = 0,05, temos que os valores críticos são -1,96 e 1,96. Portanto, as regiões de rejeição são t < -1,96 e t > 1,96. Temos ainda que as médias de aproveitamento de saque nos casos de derrota e vitória são $\bar{x}_1 =$ 0,041 e $\bar{x}_2 =$ 0,051, respectivamente. Os desvios padrão são σ_1= 0,024 para o jogos terminados em derrota e σ_2 = 0,026 para jogos terminados em vitória (Tabela A.1). O erro padrão é $\sigma_{\bar{x}_1 - \bar{x}_2}$ = 0,003.

Usando o teste t, calculamos a estatística padronizada e temos t = -3.320, que se encontra na região de rejeição. Portanto, rejeitamos a hipótese nula de que as médias de aproveitamento de saque sejam iguais nos casos de vitória e de derrota. A um nível de 5% de significância, há evidências de que o fundamento saque esteja associado ao resultado do jogo.

Conclui-se, com o intervalo de confiança de 95%, que a média do aproveitamento de saque dos times que perderam os jogos foi de 0,4% a 1,6% menor do que a média desse fundamento dos times que saíram vencedores das partidas (Tabela A.2).

Tabela A.1: Estatísticas do Resultado do Jogo

Variável	Resultado do Jogo	n	média	desvio padrão
saque	derrota	148	0,041	0,024
	vitória	146	0,051	0,026

Fonte dos dados brutos: Confederação Brasileira de Voleibol - CBV

Tabela A.2: Teste t de Student

Variável	t	g.l.	valor de p	diferença média	erro padrão	IC (95%)	
						lim. inferior	lim. superior
saque	-3,320	292	0,001	-0,010	0,003	-0,016	-0,004

Fonte dos dados brutos: Confederação Brasileira de Voleibol - CBV

APÊNDICE B

Teste Qui-Quadrado de Independência

Para verificar a significância da associação entre duas variáveis categóricas (ou qualitativas), deve-se utilizar o teste Qui-Quadrado.

Para isso, deve-se testar a hipótese nula H_0 de que, na população, as variáveis não estão associadas. Deve-se também estipular o nível de significância (usualmente igual a 5%), calcular o valor da estatística X^2 comparando os valores observados com os esperados considerando-se os respectivos graus de liberdade. Por fim, em função do Qui-Quadrado calculado, obtém-se o valor de p. Se o valor de p for menor do que o nível de significância estipulado, rejeita-se a hipótese nula e assume-se que há associação entre as variáveis na população.

Como forma de exemplificar a aplicação do teste Qui-Quadrado, usaremos os dados da Tabela B.1 abaixo. Observa-se que, na prática (na nossa amostra), a proporção de vitórias, independentemente de o time ter jogado dentro ou fora de casa é de 50%. Assim, esperar-se-ia que 50% dos 147 jogos fora de casa e 50% dos 149 jogos dentro de casa terminassem em vitória, se estas variáveis não estivessem associadas.

Tabela B.1: Resultado do jogo em função do local do jogo

Local do Jogo	Resultado do Jogo				Total
	Derrota		Vitória		
	n	%	n	%	
Fora de Casa	82	55.8%	65	44.2%	147
Em casa	66	44.3%	83	55.7%	149
Total	148	50.0%	148	50.0%	296

Fonte dos dados brutos: Confederação Brasileira de Voleibol - CBV

Fazendo os devidos cálculos, obtêm-se as frequências esperadas na Tabela B.2 abaixo.

Tabela B.2: Frequências observadas x Frequências esperadas

Local do Jogo	Resultado do Jogo				Total
	Derrota		Vitória		
	fo	fe	fo	fe	
Fora de Casa	82	73,5	65	73,5	147
Em casa	66	74,5	83	74,5	149
Total	148	148	148	148	296

Fonte dos dados brutos: Confederação Brasileira de Voleibol – CBV

Comparando-se os valores obsevados com os valores esperados, calcula-se, então, o valor da estatística X^2, através da seguinte fórmula:

$$X^2 = \sum \frac{(fo - fe)^2}{fe}$$

Onde:

fo = frequência observada

fe = frequência esperada

Assim, encontra-se o valor de X^2 igual a 3,906. Sob a hipótese nula de que não há associação entre as variáveis, espera-se que o valor de X^2 seja igual a zero. Todavia, as diferenças entre os valores observados e esperados levaram a um valor igual 3,906. A probabilidade de encontrarmos um valor superior a este, dado que esperávamos zero, é muito pequena (valor de p < 0,000), portanto, deve-se rejeitar H_0 e assumir que, na população, as variáveis estão associadas.

Considerando a hipótese nula (H_0) de que as probabilidades de célula sejam iguais a certos valores fixos x, para uma amostra de tamanho n, com contagens de célula fo, os valores $fe = nx$ são as frequências esperadas. Eles representam os valores esperados $E(fo)$ quando H_0 é verdadeira.

Esta notação refere-se a tabelas 2x2, mas concepções semelhantes se aplicam a um conjunto de contagens para uma única variável categórica ou para tabelas nxn.

Para julgar se os dados contradizem H_0, comparamos fo com fe. Se H_0 é verdadeira, fo deve estar próximo de fe em cada célula. Quanto maior a diferença {fo

- $fe\}$, mais forte será a evidência contra H_0. O teste estatístico utilizado para fazer estas comparações tem distribuição Qui-Quadrado para grandes amostras.

Estatística de Pearson e Distribuição Qui-Quadrado

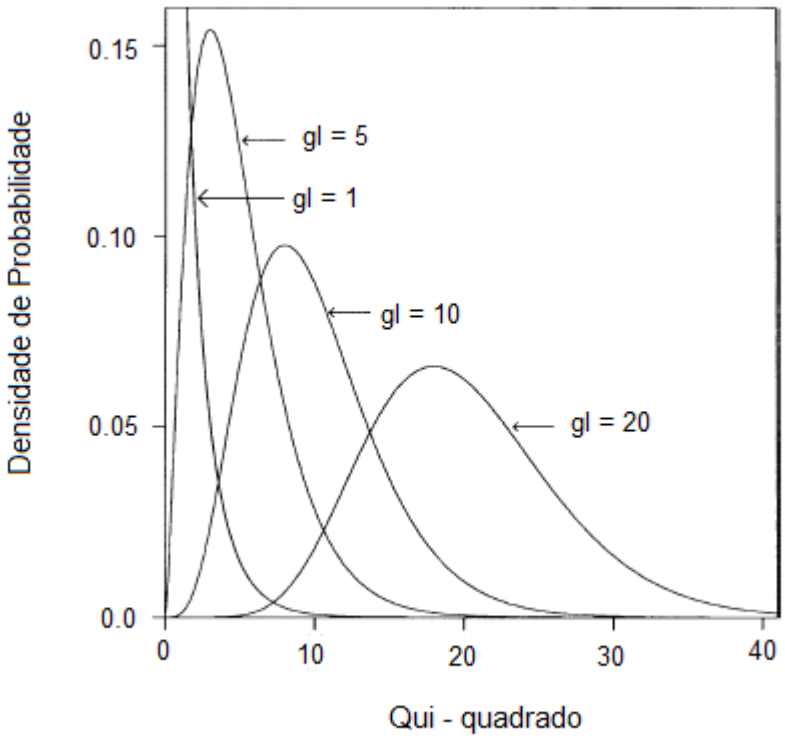
A estatística do Qui-Quadrado de Pearson para testar H_0 é:

$$X^2 = \sum \frac{(fo - fe)^2}{fe}$$

Ela foi proposta em 1900 por Karl Pearson, o estatístico britânico conhecido também pela estimativa da correlação de Pearson, entre muitas outras contribuições. Esta estatística tem o seu valor mínimo igual a zero quando $fo = fe$. Para uma amostra de tamanho fixo, maiores diferenças $\{fo - fe\}$ produzem maiores valores de X^2 e evidências mais fortes contra H_0.

Uma vez que maiores valores de X^2 contradizem a H_0, o valor de p é a probabilidade nula de que X^2 seja pelo menos tão grande quanto o valor observado. A estatística X^2 tem aproximadamente uma distribuição Qui-Quadrado, para grandes valores de amostra n. O valor de p é a probabilidade Qui-Quadrado acima da cauda direita do valor de X^2 observado. A aproximação do Qui-Quadrado melhora conforme fe aumenta, e $\{fe \geq 5\}$ é geralmente suficiente para uma boa aproximação.

A distribuição Qui-Quadrado é concentrada em valores não negativos. Sua média é igual aos seus graus de liberdade (gl), e o seu desvio padrão é igual a $\sqrt{(2gl)}$. Conforme os gl aumentam, a distribuição tende a se concentrar em torno de valores maiores e ficar mais espalhada. A distribuição se inclina para a direita e tende-se a uma distribuição Normal de acordo com o aumento dos gl. A Figura abaixo mostra densidades de Qui-Quadrado com gl = 1, 5, 10 e 20.



O valor dos gl é igual à diferença entre o número de parâmetros das hipóteses alternativa e nula.

APÊNDICE C

Modelo de Regressão Logística

Para estudar a relação entre uma variável resposta e duas ou mais variáveis explicativas, quando a variável resposta apenas assume valores binários, ou seja, Y=0 ou Y=1, utiliza-se o Modelo de Regressão Logística, que pode ser definido como:

$$Y_i = \pi(x_i) + \epsilon_i \quad (1)$$

onde ϵ_i é o erro aleatório e assume-se que $Y_i \sim Ber(\pi(x_i))$, ou seja, a variável resposta assume o valor 1 para o evento de interesse (sucesso) e o valor 0 para o evento complementar (fracasso), com probabilidades $\pi(x_i) = P(Y = 1|x_i)$ e 1 - $\pi(x_i) = P(Y = 0|x_i)$, respectivamente. A probabilidade de sucesso deste modelo é:

$$\pi(x_i) = \pi_i = \frac{e^{\beta_0+\beta_1 x_{i_1}+\beta_2 x_{i_2}+\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i_1}+\beta_2 x_{i_2}+\cdots+\beta_p x_{ip}}}, \quad (2)$$

e de fracasso:

$$1 - \pi(x_i) = 1 - \pi_i = \frac{1}{1+ e^{\beta_0+\beta_1 x_{i_1}+\beta_2 x_{i_2}+\cdots+\beta_p x_{ip}}}. \quad (3)$$

A quantidade a ser modelada em todo problema de regressão, é o valor esperado da variável resposta dados os valores das variáveis explicativas, ou seja, $E(Y| x_i)$. Dado que no modelo de Regressão Logística, a variável resposta é binária, temos que:

$$0 \leq E(Y|x_i) = 1P(Y_i = 1|x_i) + 0P(Y_i = 0|x_i) = \pi_i \leq 1$$

Além disso, devido à natureza da variável resposta, tem-se que o erro ϵ_i pode assumir somente um dos dois possíveis valores, isto é, ϵ_i = 1 - π_i, para y_i = 1 ou ϵ_i = - π_i para y_i = 0. Assim, segue-se que ϵ_i tem distribuição com média zero e variância π_i(1 - π_i).

A transformação de π_i é de extrema importância para o estudo de regressão logística, e é denominada Transformação Logito. Ela é definida por:

$$g(x_i) = ln\frac{(\pi(x_i))}{(1-\pi(x_i))} = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2} + \ldots + \beta_p x_{i_p}. \qquad (4)$$

Esta transformação é relevante uma vez que o logito é linear em β_0 e β_j, $j = 1, \ldots, p$ pode ser contínuo e variar de -∞ a +∞ dependendo dos valores assumidos pelas variáveis explicativas.

Estimação por Máxima Verossimilhança

Supondo que (x_i, y_i) seja uma amostra independente com n pares de observações, y_i representa o valor observado da variável resposta dicotômica e x_i é o valor observado da variável explicativa da i-ésima observação em que $i = 1, \ldots, n$. Para o ajuste do modelo de regressão logística, segundo a equação (1 a 3), é necessário estimar os parâmetros β_0 e $\beta_j, j = 1, \ldots, p$.

O método de Máxima Verossimilhança é utilizado para estimar os parâmetros do modelo de regressão.

A função de distribuição da probabilidade de Y_i para o modelo de regressão logística com $Y_i \sim Ber(\pi(x_i))$, é dada por:

$$f(y_i, \pi_i) = \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}, \; y_i = 0 \text{ ou } y_i = 1. \qquad (5)$$

Tendo em vista que $Y_1, Y_2, \ldots, Y_n$ são independentes, a função de verossimilhança é obtida pelo produto dos termos dados na expressão acima e é definida por:

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}, \qquad (6)$$

onde denota-se por $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ o vetor de parâmetros desconhecidos. Assim, a função de log verossimilhança é dada por:

$$l(\beta) = \ln[L(\beta)] = \sum_{i=1}^{n} y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]. \qquad (7)$$

O princípio da Máxima Verossimilhança é obter o valor de β que maximize $L(\beta)$ ou equivalentemente $l(\beta)$. Dessa forma, deriva-se $l(\beta)$ em relação a cada parâmetro, obtendo o sistema de equações abaixo:

$$\frac{\partial l(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} [y_i - \pi(x_i)] = 0, \qquad (8)$$

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}[y_i - \pi(x_i)] = 0, \qquad (9)$$

Onde $\pi(x_i)$ é dada pela expressão (2).

Nota-se que as equações em (8) e (9) são não lineares em β_0 e β_j $j = 1, \ldots, j$, assim são necessários métodos iterativos para resolução do sistema de equações.

Intervalos de confiança

O intervalo de $100(1 - \alpha)\%$ de confiança para β_0 e β_j, $j = 1, \ldots, p$, é dado por:

$$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \widehat{SE}(\hat{\beta}_j), \quad j = 1, \ldots, p,$$

e para o intercepto,

$$\hat{\beta}_0 \pm z_{1-\frac{\alpha}{2}} \widehat{SE}(\hat{\beta}_0),$$

onde $\widehat{SE}(.)$ é o desvio padrão estimado.

Interpretação dos coeficientes

O coeficiente de uma variável explicativa na Regressão Logística representa a inclinação ou a taxa de mudança da função logito para cada incremento de uma unidade no valor da mesma, dado que as demais variáveis estão fixas, podendo ser representado dessa forma:

$$\beta_j = g(x_i + 1) - g(x_i)\,,$$

onde g(.) representa a função descrita em (4). O coeficiente β_j pode ser interpretado como o logaritmo da razão de chances e $exp(\beta_j)$ como sendo a própria razão de chances.

A razão de chances é uma medida que representa o quanto é mais provável se observar o evento de interesse para um indivíduo do que para outro, assumido como referência. Pode-se dizer, então, que para os indivíduos com $X_i = x_i + 1$ o evento de interesse tem $exp(\beta_1)$ vezes a chance daqueles que assumem $X_i = x_i$. A razão de chances e o logaritmo da razão de chances são mostrados a seguir.

Quando a variável explicativa é dicotômica, tem-se que a chance de resposta, dado que $x = 1$, é $\pi(1)/[1 - \pi(1)]$ da mesma forma, quando $x = 0$, é $\pi(0)/[1 - \pi(0)]$. O logaritmo da razão é dado por:

$$g(1) = ln\,\pi(1)/[1 - \pi(1)] \;\text{ e }\; g(0) = ln\,\pi(0)/[1 - \pi(0)]\,.$$

A razão de chances é denotada por RC e definida como:

$$RC = \frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]}.$$

O logaritmo da razão de chances é:

$$\ln(RC) = ln\left[\frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}\right] = g(1) - g(0)$$

Usando a expressão para o modelo de Regressão Logística, definido em (2) e (3), a razão de chances será dada por:

$$RC = \frac{\left(\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}\right)/\left(\frac{1}{1+\exp(\beta_0+\beta_1)}\right)}{\left(\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right)/\left(\frac{1}{1+\exp(\beta_0)}\right)} = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1),$$

E o logaritmo da razão de chances será dado por:

$$\ln(RC) = \ln[\exp(\beta_1)] = \beta_1$$

O intervalo de confiança, com nível de confiança $100(1-\alpha)\%$ para razão de chances é obtido calculando inicialmente o intervalo para β_1 e aplicando a exponencial:

$$\exp[\hat{\beta}_1 \pm z_{1-\frac{\alpha}{2}}\widehat{SE}(\hat{\beta}_1)],$$

onde $\widehat{SE}(\hat{\beta}_1)$ é o desvio padrão estimado de $\hat{\beta}_1$.