MODELO DE PREVISÃO DE DESEMPENHO DE TRIATLETAS COM A UTILIZAÇÃO DA ANÁLISE DISCRIMINANTE

TRIATHLETES PERFORMANCE PREDICTION MODEL USING DISCRIMINANT ANALYSIS

Domingos Rodrigues Pandeló Júnior*

Resumo: A previsão do desempenho esportivo é relevante para a identificação de talentos e no estabelecimento de estratégias de treinamento. O objetivo do presente artigo é estabelecer um modelo que tenha a capacidade de predizer o desempenho de triatletas. Para isso foi utilizada a análise discriminante, que é uma técnica de análise multivariada. Trabalhou-se com 21 voluntários, sendo 7 profissionais e 14 amadores, todos do sexo masculino. Foram selecionadas variáveis antropométricas, fisiológicas e de treinamento fáceis de serem mensuradas, sem a necessidade de utilização de laboratórios específicos. O presente estudo mostrou que com a utilização de algumas variáveis pode-se buscar inferir a *performance* de triatletas. A previsão de *performance* é de vital importância, quer seja para a detecção de talentos, quer seja para a estruturação do treinamento, o que mostra a importância do desenvolvimento de modelos desse tipo.

Palavras-chave: triatlo, desempenho, treinamento, detecção de talentos.

Abstract: The prediction of sport performance is relevant for identifying talent and establishing training strategies. The aim of this paper is to establish a model that has the ability to predict the performance of triathletes. It was used discriminant analysis, which is a multivariate analysis technique. 21 volunteers, 7 professionals and 14 amateurs, all male, were selected. Anthropometric, physiological and training variables, easy to be measured, without the need to use specific laboratories were selected. This study showed that the use of some variables may help to infer the performance of triathletes. The expected performance is vital, whether for the detection of talent, whether for the structuring of training, which shows the importance of developing models of this kind.

Keywords: triathlon , performance, training, talent detection..

1 Introdução

O objetivo do presente artigo é estabelecer um modelo que tenha a capacidade de predizer o desempenho de triatletas. Para isso foi utilizada a análise discriminante, que é uma técnica de análise multivariada. Trata-se de um modelo de regressão do tipo linear, que busca encontrar a melhor equação capaz de predizer o que se quer. No caso, a busca é por um modelo capaz de mensurar a *performance* potencial com base em poucas variáveis preditivas. Foram escolhidas variáveis antropométricas, fisiológicas, e de treinamento, de fácil acesso. A escolha é

* Cursou a Academia da Força Aérea (AFA), Cursou Engenharia de Produção, com especialização em PO (ITA), graduado em administração pública (EAESP/FGV), mestre em planejamento e finanças públicas EAESP/FGV. Graduado em educação física (FEFIS/UNIMES), especialista em ciências do esporte (UNIFESP/SP), doutorando em ciências do movimento e reabilitação (UNIFESP/SP). Fundador do Centro de Alta Performance. (pandelo@hotmail.com).

justificada pela intenção de permitir que o modelo seja replicado e utilizado pelo maior número de pessoas possível.

A previsão do desempenho esportivo é relevante para a identificação de talentos e no estabelecimento de estratégias de treinamento. Diversos estudos já utilizaram a análise discriminante para classificar e prever desempenho de atletas. Pode-se citar, por exemplo, o trabalho de Le Meur e colaboradores (2013), que usou a técnica para prever o *overreaching* em atletas de *endurance*; o de Saavedra e colaboradores (2010), que a usou para a previsão da *performance* de jovens nadadores; e ainda o de Opstoel e colaboradores (2015), que utilizou a análise discriminante para, através de variáveis antropométricas e de *performance*, classificar jovens atletas de elite em nove modalidades esportivas (LE MEUR *et al*, 2013; SAAVEDRA, ESCALATE, RODRIGUES, 2010; OPSTOEL *et al*, 2015).

2 Metodologia e Procedimentos

2.1 Amostra

Para se chegar ao número ideal da amostra, trabalhou-se com o *software* G* Power. Para um α de 0,05 e um β de 0,80, o número mínimo de participantes teria que ser 16. No presente estudo, trabalhou-se com 21 voluntários, sendo 7 profissionais e 14 amadores, todos do sexo masculino. Optou-se por trabalhar com voluntários do sexo masculino, pois desta forma seria mais fácil a discriminação, pelo modelo, do desempenho. Caso atletas do sexo feminino fossem incluídas, aumentaria a variabilidade dos resultados em função das diferenças de desempenho entre os sexos, o que tornaria a classificação mais complexa, bem como sujeita a uma maior taxa de erro. Os 21 participantes foram divididos em dois grupos: grupo 1 (profissionais) e grupo 2 (amadores).

2.2 Variáveis selecionadas

Foram selecionadas variáveis antropométricas, fisiológicas e de treinamento fáceis de serem mensuradas, sem a necessidade de utilização de laboratórios específicos. As variáveis foram selecionadas com base em estudos anteriores (GILINSKY *et al*, 2014; HUE, 2003; LAURSEN, RHODES, 2001; KNECHTLE *et al*, 2012; MILLET *et al*, 2002; KNECHTLE *et al*, 2010). As variáveis selecionas foram o índice de massa corporal (IMC); idade (I); frequência cardíaca de repouso (BPM); o número de anos de prática do triatlo (APT); o consumo máximo de oxigênio (VO_{2Max}) – neste estudo calculado de forma indireta a partir de corrida de 3.000m, em máximo esforço; a distância semanal de treino em natação (DNS); a distância semanal de treino em ciclismo (DCiS); a distância semanal de treino em corrida (DCoS); e a idade dos atletas.

Os dados relativos aos atletas foram colhidos por meio de questionário, após a explicação sobre os objetivos do experimento, bem como a assinatura do Termo Livre de

Consentimento Esclarecido. A pesquisa foi cadastrada na Plataforma Brasil e submetida ao comitê de ética da Universidade Metropolitana de Santos (UNIMES), tendo sido aprovada sob o número 48748015.8.0000.5509.

A tabela 1 mostra a média, o desvio padrão e o coeficiente de variação das variáveis utilizadas na construção do modelo, bem como os estudos em que se utilizaram os mesmos indicadores. Pode-se observar uma maior variabilidade, medida pelo coeficiente de variação, nos anos de prática de triatlo e na distância semanal de natação. A menor variabilidade foi observada no IMC. Os valores apresentados referem-se às médias dos dois grupos (profissionais e amadores) em conjunto.

Tabela 1. Média e desvio padrão das variáveis selecionadas

Variável	Média	Desvio Padrão	Coeficiente de Variação	Referência
Índice de massa corporal (IMC)	23,51	1,72	0,07	Gilinky et al. (2014) Knechtle et al. (2012) Knechtle et al. (2010)
Idade (I)	38,71	8,7	0,22	Gilinky et al. (2014) Millet et al. (2002) Knechtle et al. (2012) Knechtle et al. (2010)
Frequência cardíaca de repouso (BPM)	50,14	9,92	0,20	Laursen & Rhodes (2001)
Consumo máximo de oxigênio (VO_{2Max})	52,73	8,00	0,15	Hue(2003); Laursen & Rhodes (2001); Millet et al. (2002)
Anos de prática de triatlo (APT)	11,29	7,18	0,64	Gilinky et al. (2014)
Distância natação semanal km (DNS)	11,00	7,14	0,65	Gilinky et al. (2014) Hue(2003) Millet et al. (2002)
Distância ciclismo semanal km (DCiS)	221,90	93,95	0,42	Gilinky et al. (2014) Hue(2003) Millet et al. (2002)
Distância corrida semanal km (DCoS)	49,19	15,64	0,32	Gilinky et al. (2014) Hue(2003) Millet et al. (2002) Knechtle et al. (2010)

Fonte: Elaboração própria

2.3. A técnica utilizada

A análise discriminante é indicada para a classificação de um determinado elemento em um dos grupos previamente trabalhados. No caso da previsão de desempenho de atletas, pode-se, com base em variáveis fisiológicas e antropométricas, por exemplo, buscar a classificação de atletas em grupos de *performance*.

O objetivo da análise discriminante é criar uma função que busque maximizar a variância entre os grupos e, simultaneamente, minimizar a variância dentro dos grupos (HAIR JF *et al*, 2006). Pelo objetivo da análise discriminante, já se pode perceber a importância da escolha da amostra e dos regressores.

A amostra deve ser a mais homogênea possível, com exceção do item que se pretende discriminar, no caso deste artigo, o nível de *performance*.

A função discriminante representa o escore discriminante Z, que é a soma dos regressores selecionados estatisticamente pelo modelo, ponderado pelos seus respectivos pesos.

$$Z = c + p_1 v_1 + p_2 v_2 + ... p_n v_n$$

Onde *Z* é o escore discriminante, *c* a constante do modelo, *p* o peso de cada variável, *v*, selecionada.

Como salientado por Hair e colaboradores (2006), a análise discriminante é relativamente robusta a algumas violações exigidas pelos modelos de análise multivariada, tais como normalidade, linearidade e homocedasticidade. Todavia, em função do tamanho amostral ser relativamente pequeno (menor que 30, o que caracteriza pequenas amostras) e após efetuar alguns testes de normalidade e linearidade, optou-se por trabalhar com os valores transformados, com o auxílio do logaritmo natural (ln). Dessa forma, os dados foram ajustados com a finalidade de minimizar eventuais problemas advindos da violação de algumas premissas clássicas do modelo.

Todas as análises estatísticas foram efetuadas no SPSS 21. Os testes de normalidade efetuados foram o Kolmogorov-Smirnov e o Shapiro-Wilk. Já para se avaliar a linearidade, utilizou-se a função de gráficos de dispersão (scatterplot). A análise da homocedasticidade foi efetuada pelo teste de Levene.

3. Resultados

Após a inserção dos dados e realização dos testes básicos para verificar a adequação da amostra às premissas clássicas da análise multivariada, iniciaram-se as análises para a construção do modelo. A Tabela 2 mostra as respostas de cada variável isoladamente.

Tabela 2. Nível de significância e tamanho do efeito

	Lambda de Wilks	F	Sig.	Hedges 'g
ln (IMC)	,741	6,647	,018	1,19
ln (I)	,838	3,674	,070	0,89
lnbpm	,472	21,294	,000	2,33
lnVO_{2Max}	,572	14,243	,001	1,75
lnanostriatlo	,980	,392	,539	0,29
lndistnatcsem	,324	39,623	,000	2,91
lndistcicsem	,694	8,366	,009	1,34
lndistcorsem	,650	10,216	,005	1,48

Fonte: Elaboração própria

As respostas da Tabela 3 nos mostram os coeficientes de classificação com base nas variáveis selecionadas pelo modelo de análise discriminante para os grupos profissionais e amadores.

Tabela 3. Coeficientes de função de classificação (funções discriminantes lineares de Fisher)

	Grupo	
	Profissionais	**Amadores**
Lndistnatcsem	29,253	20,573
Lndistcicsem	46,751	41,400
(Constante)	-176,098	-125,819

Fonte: Elaboração própria

A Tabela 4 nos mostra o centroide de cada grupo. Pode-se observar que o centroide (ponto médio) do Grupo 1 (Profissionais) ficou em 2,338, ao passo que o centroide do Grupo 2 (Amadores) ficou em -1,169.

Tabela 4. Funções em centroides de grupo

Grupo	**Função**
1,00	2,338
2,00	-1,169

Fonte: Elaboração própria

A Tabela 5 mostra os resultados referentes aos modelos Original e Com validação cruzada, para ambos os grupos.

Tabela 5. Resultados da classificação [a]

		Grupo	**Associação ao grupo prevista**		**Total**
			Profissionais	**Amadores**	
Original	Contagem	1,00	7	0	7
		2,00	0	14	14
	%	1,00	100,0	,0	100,0
		2,00	,0	100,0	100,0
Com validação cruzada[b]	Contagem	1,00	7	0	7
		2,00	1	13	14
	%	1,00	100,0	,0	100,0
		2,00	7,1	92,9	100,0

a. 100,0% de casos originais agrupados corretamente classificados.
b. A validação cruzada é feita apenas para os casos da análise. Na validação cruzada, cada caso é classificado pelas funções derivadas de todos os casos diferentes desse caso.

Fonte: Elaboração própria

4. Discussão

Observando a relevância de cada variável para explicar o fenômeno estudado (Tabela 2), pode-se verificar que, tomadas de forma isolada, as variáveis mais relevantes para a explicação e classificação do desempenho dos triatletas, dentre as selecionadas no presente

estudo com base no nível de significância, foram, respectivamente: distância de natação semanal, frequência cardíaca de repouso e consumo máximo de oxigênio.

Quando complementamos a análise com a mensuração do tamanho do efeito, com a utilização do Hedges 'g, algumas informações interessantes aparecem. Pode-se ver que o índice de natação semanal parece ter uma grande capacidade de discriminação, assim como a frequência cardíaca de repouso e o consumo máximo de oxigênio. Até aqui o resultado ficou similar à análise pelo nível de significância medido pelo valor de *P*. Porém, quando se analisa o tamanho do efeito, pode-se verificar que a distância semanal de corrida e de ciclismo também são variáveis relevantes.

Do ponto de vista do rigor da análise, a inclusão de algum tipo de análise do tamanho do efeito é fundamental para uma melhor avaliação prática do fenômeno que se pretende estudar (CUMMING, 2013). No presente artigo, optou-se por trabalhar com o Hedges 'g e não com o Cohen's d em função do pequeno número da amostra, bem como pelo tamanho diferente de cada grupo (ELLIS, 2010).

Na técnica de análise discriminante (Tabela 3), pode-se optar pelo método *enter*, no qual todas as variáveis que atenderem os pré-requisitos mínimos do modelo são forçadas a entrar na função discriminante final, ou o método *stepwise*, que foi o adotado neste artigo e no qual as variáveis só permanecem no modelo se contribuírem, de forma efetiva para a melhoria de sua capacidade preditiva.

Assim, pode-se observar que as variáveis mais relevantes para a construção do modelo discriminante foram a distância semanal de natação e a distância semanal de ciclismo. Tal informação parece ser bem relevante, pois tais variáveis não são antropométricas nem fisiológicas (embora afetem essas variáveis), o que parece indicar que um bom desempenho em provas de triatlo parece depender, de forma significativa, da estratégia de treinamento. É óbvio que quando se pretende distinguir, discriminar desempenho de atletas para classificá-los em dois grupos (profissionais e amadores), diversas variáveis fisiológicas e antropométricas, além das de treinamento, parecem ser importantes. O que este modelo nos diz é que, quando analisado de forma conjunta, levando-se em consideração as correlações entre as variáveis, a interação entre as mesmas, pode-se, com a análise do treino semanal de natação e ciclismo, inferir de forma bastante precisa o desempenho de um triatleta.

Pode-se observar, ainda, a existência de duas funções discriminantes. Uma para o grupo de profissionais e outra para amadores. Ambas são funções lineares com uma constante e duas variáveis explicativas (o logaritmo natural da distância de natação semanal e da distância de ciclismo semanal). Com base nas duas equações e levando-se em consideração o valor dos centroides (encontrados na Tabela 4), pode-se efetuar a classificação de um atleta num dos dois grupos.

Com base nas informações do centroide de cada grupo (Tabela 4), pode-se, com facilidade, calcular o *z* escore de corte, em termos de desempenho. Para tanto, basta fazer uma média ponderada dos centroides de cada grupo pelo número da amostra em cada grupo. No caso do presente artigo, o *Z* escore de corte seria zero. Valores acima de zero estariam mais próximos do desempenho de atletas amadores, e valores abaixo de zero estariam mais próximos do desempenho de atletas profissionais.

Como a amostra era reduzida, optou-se por trabalhar com os modelos original e validação cruzada. A validação cruzada é importante para se testar a capacidade preditiva do modelo (JOHNSON; WICHEM, 1992). Com a validação cruzada, são efetuados *n* modelos, sendo que *n* é o número da amostra utilizada. Assim, cada atleta da amostra é testado num modelo construído sem a sua presença na base de dados. Tal técnica é importante quando não se tem uma amostra grande o suficiente para se separar um grupo para teste. Em geral, o grau de acerto no modelo original é maior do que no com validação cruzada, pois no modelo original os mesmos atletas que fizeram parte da amostra para a construção do modelo foram testados e classificados com o mesmo modelo. Já com a validação cruzada, o percentual de acerto tende a ser menor, mas é uma situação mais próxima da real e é de extrema importância, especialmente quando se tem uma pequena base de dados para se trabalhar.

A capacidade preditiva do modelo, com base nos números apresentados na Tabela 5, pode ser considerada muito boa. Ocorreu apenas um pequeno percentual de má classificação e, no caso específico, foi de um atleta que, embora seja amador, tem resultados muito bons, o que o aproxima de profissionais.

O presente estudo mostrou que, com a utilização de algumas variáveis antropométricas, fisiológicas e de treinamento, pode-se buscar inferir a *performance* de triatletas. A previsão de *performance* é de vital importância, quer seja para a detecção de talentos, quer seja para a estruturação do treinamento. Evidentemente, como todo e qualquer modelo, este aqui apresentado tem as suas limitações, mas isso faz parte da essência da modelagem de dados. O que o pesquisador, ou usuário, não pode perder de vista é que um modelo é uma simplificação da realidade e é construído com o intuito de facilitar o processo de análise e tomada de decisão.

A contribuição do presente artigo é mostrar uma possibilidade a ser explorada com a construção de modelos para a previsão de desempenho e classificação de atletas, com base numa técnica de análise multivariada. Outras variáveis mais complexas podem ser utilizadas, e, eventualmente, melhores resultados podem ser obtidos em função disso. Neste estudo, optou-se por trabalhar com variáveis mais simples, de fácil mensuração, sem a necessidade de testes específicos em laboratórios.

Referências

CUMMING G. **Understanding the new statistics:** effect sizes, confidence intervals, and meta-analysis**:** Routledge; 2013.

ELLIS PD. **The essential guide to effect sizes:** statistical power, meta-analysis, and the interpretation of research results**:** Cambridge University Press; 2010.

GILINSKY N, HAWKINS KR, TOKAR TN, COOPER JA. **Predictive variables for half-Ironman triathlon performance.** J Sci Med Sport. 2014 May;17(3):300-5. PubMed PMID: 23707141. Epub 2013/05/28. eng.

HUE O. **Prediction of drafted-triathlon race time from submaximal laboratory testing in elite triathletes**. Can J Appl Physiol. 2003 Aug;28(4):547-60. PubMed PMID: 12904633. Epub 2003/08/09. eng.

HAIR JF, BLACK WC, BABIN BJ, ANDERSON RE, TATHAM RL. **Multivariate data analysis:** Pearson Prentice Hall Upper Saddle River**, NJ**; 2006.

JOHNSON RA, WICHERN DW. **Applied multivariate statistical analysis:** Prentice hall Englewood Cliffs**, NJ**; 1992.

KNECHTLE B, KNECHTLE P, WIRTH A, ALEXANDER RUST C, ROSEMANN T. **A faster running speed is associated with a greater body weight loss in 100-km ultra-marathoners**. J Sports Sci. 2012;30(11):1131-40. PubMed PMID: 22668199. Epub 2012/06/07. eng.

KNECHTLE B, WIRTH A, BAUMANN B, KNECHTLE P, ROSEMANN T, OLIVER S. **Differential correlations between anthropometry, training volume, and performance in male and female Ironman triathletes**. J Strength Cond Res. 2010 Oct;24(10):2785-93. PubMed PMID: 20571444. Epub 2010/06/24. eng.

LE MEUR Y, HAUSSWIRTH C, NATTA F, COUTURIER A, BIGNET F, VIDAL PP. **A multidisciplinary approach to overreaching detection in endurance trained athletes**. J Appl Physiol (1985). 2013 Feb;114(3):411-20. PubMed PMID: 23195630. Epub 2012/12/01. eng.

LAURSEN PB, RHODES EC. **Factors affecting performance in an ultraendurance triathlon**. Sports Med. 2001;31(3):195-209. PubMed PMID: 11286356. Epub 2001/04/05. eng.

MILLET GP, CANDAU RB, BARBIER B, BUSSO T, ROUILLON JD, CHATARD JC. **Modelling the transfers of training effects on performance in elite triathletes**. Int J Sports Med. 2002 Jan;23(1):55-63. PubMed PMID: 11774068. Epub 2002/01/05. eng.

OPSTOEL K, PION J, ELFERINK-GEMSER M, HARTMAN E, WILLEMSE B, PHILIPPAERTS R, ET AL. **Anthropometric characteristics, physical fitness and motor coordination of 9 to 11 year old children participating in a wide range of sports**. PLoS One. 2015;10(5):e0126282. PubMed PMID: 25978313. Pubmed Central PMCID: PMC4433213. Epub 2015/05/16. eng.

SAAVEDRA JM, ESCALANTE Y, RODRIGUEZ FA. **A multivariate analysis of performance in young swimmers**. Pediatr Exerc Sci. 2010 Feb;22(1):135-51. PubMed PMID: 20332546. Epub 2010/03/25. eng.

Artigo recebido em: 08/11/2017
Artigo aceito para publicação em: 05/12/2017