REVISTA DE ESTATÍSTICA
3º QUADRIMESTRE DE 2001

APLICAÇÃO DE MÉTODOS ESTATÍSTICOS NO DESPORTO: ANÁLISE DO CAMPEONATO DE FUTEBOL, EM PORTUGAL

STATISTICAL METHODS APPLIED TO SPORTS: ANALISYS OF THE PORTUGUESE SOCCER CHAMPIONSHIP

Autor: Paulo Almeida Pereira
- Prof. Auxiliar da Universidade Católica Portuguesa, Pólo de Viseu
Instituto Universitário de Desenvolvimento e Promoção Social

***RESUMO*:**

- Métodos estatísticos, como os modelos de regressão e a análise factorial de componentes principais são aplicados ao estudo dos resultados do campeonato nacional de futebol. São desenvolvidos dois modelos, utilizando, como variável dependente, os resultados dos jogos e como variáveis independentes, dados estatísticos disponíveis para o comportamento das equipas intervenientes. Após a eliminação de *outliers* e a validação dos pressupostos dos modelos de regressão, calculam-se as estimativas dos resultados dos jogos e respectivos intervalos de confiança, a 95%, que originam classificações previstas pelos modelos, que são comparadas com os valores observados.

***PALAVRAS-CHAVE*:**

- *Modelos de regressão, análise factorial, desporto, campeonato nacional de futebol, estimativa das classificações.*

ABSTRACT:

- Statistical methods, such as regression models and factorial analysis of principal components are applied to the study of portuguese soccer championship results. Two models are developed, using, as dependent variable, the results of soccer games and as independent variables, available statistical data for the behavior of the intervening teams. After outliers elimination and validation of the regression models assumptions, the estimation of results for the soccer games and 95% confidence intervals are calculated, which originate classification predictions by the models that are compared with the observed values.

KEY-WORDS:

- *Regression models, factorial analysis, sports, portuguese soccer championship, classification estimation.*

1. INTRODUÇÃO

O desporto gira em torno de estatísticas, tendo sempre por base um sistema numérico de resultados, por exemplo, os golos num jogo de futebol. Os acontecimentos desportivos podem, assim, constituir um tópico de investigação estatística, o que vem a acontecer há, pelo menos, três décadas (Machol *et al*., 1976; Ladany e Machol, 1977). Devido à sua relevância, a *American Statistical Association* criou, em 1992, uma secção denominada *Statistics in Sports*.

Existem publicações, nesta área, dispersas por vários jornais de estatística, tendo surgido, recentemente, uma importante compilação de aplicações, na área da estatística, a acontecimentos desportivos (Bennett, 1998), em que são apresentados trabalhos sobre várias modalidades desportivas: Futebol Americano, Basebol, Basquetebol, Futebol, Golfe, Hóquei no Gelo, Ténis e Atletismo.

As aplicações estatísticas a eventos desportivos são também uma poderosa ferramenta para cativar a atenção dos estudantes: pela familiaridade com os desportos, o contexto da aplicação é mais facilmente compreendido, podendo ser dado o devido ênfase à análise estatística e à sua percepção crítica.

Com o desenvolvimento das tecnologias e sistemas de informação, temos presentemente uma grande disponibilidade de dados, nomeadamente estatísticos. A *Infordesporto* (Infordesporto, *site* na *Internet*) é uma dessas bases de dados, que disponibiliza uma informação completa sobre várias estatísticas relativas aos jogos de futebol do campeonato nacional, em Portugal, desde a época de 1998/99.

A análise de regressão, sendo um método estatístico que utiliza a relação entre duas ou mais variáveis quantitativas, permite estimar uma variável a partir de outras. O Modelo de Regressão Linear, discutido por alguns autores – indicam-se dois, a título de exemplo (Neter *et al*., 1996; Draper e Smith, 1981) –, é um meio formal de exprimir essa relação estatística entre variáveis independentes, de previsão ou explicativas e uma variável dependente, de resposta ou explicada. Encontram-se algumas aplicações destes modelos a competições desportivas em publicações recentes (Smith, 1999; Szymanski e Smith, 1997; Gray, 1997; Kahane, 1997; Hamilton, 1997) que, usualmente, se cingem à análise da influência de algumas, poucas, variáveis nos resultados desportivos.

Neste trabalho, pretende aplicar-se o modelo de regressão aos campeonatos de futebol, em Portugal, de 1998/1999 e 1999/2000, de modo a estudar a relação entre os dados estatísticos para os jogos e os resultados obtidos pelas equipas intervenientes. Esta análise passa pela descrição do desenvolvimento prático de um modelo de regressão, permitindo exemplificar e detalhar os passos conducentes à validação e aplicação do modelo. O modelo permite desenvolver previsões sobre qual deveria ter sido o resultado dos jogos, de acordo com as estatísticas disponíveis, e compará-lo com o resultado efectivamente observado, originando classificações previstas pelo modelo.

A manipulação de grandes quantidades de dados estatísticos só é possível com adequados meios informáticos. O *software* utilizado neste estudo é o *SPSS - Statistical Package for Social Sciences*, para o qual se refere um texto de apoio (Pestana e Gageiro, 2000), entre vários existentes.

2. DADOS ESTATÍSTICOS

A *Infordesporto* disponibiliza uma informação estatística bastante completa sobre os jogos do campeonato nacional de futebol da I divisão, desde a época de 1998/99. Para aplicação e desenvolvimento de um modelo de regressão é necessário, em primeiro lugar, sistematizar os dados disponíveis, tendo por objectivo relacionar o resultado de um jogo de futebol (variável dependente) com o comportamento das duas equipas antagonistas, descrito pelas denominadas variáveis independentes: a informação estatística disponível para os intervenientes no jogo. No Quadro n.º 1, introduz-se a lista das estatísticas disponíveis para cada jogo, agrupadas em categorias e representadas por abreviaturas, siglas essas que serão utilizadas ao longo do texto.

Quadro n.º 1. Estatísticas disponíveis para cada jogo

Gerais:		Guarda-redes:	
RELV	estado do relvado [1]	DC	defesas completas
ESPEC	n.º de espectadores	DI	defesas incompletas
TJOGO	tempo jogado (% do tempo total)	SC	saídas completas
Tempos de posse de bola (min.):		SI	saídas incompletas
DEF	na defesa	Jogadores:	
MCD	no meio campo defensivo	FC	faltas cometidas
MCO	no meio campo ofensivo	FS	faltas sofridas
ATA	no ataque	P	perdas de bola
TOT	total	R	recuperações de bola
POSSE	posse de bola (% do total)	AT	Ataques
Disciplinares:		CR	Cruzamentos
CA	cartões amarelos	RE	Remates
CV	cartões vermelhos	AS	Assistências

[1] Utilizou-se a seguinte escala ordinal: 1-mau, 2-razoável, 3-bom, 4-muito bom, 5-excelente.

Destas várias estatísticas analisadas num jogo de futebol, as estatísticas gerais e a percentagem de posse de bola apresentam um valor para cada jogo; todas as restantes apresentam dois valores, o primeiro associado a uma das equipas e o segundo associado à outra: sendo precedidas pela letra "C", quando dizem respeito à equipa que joga em casa e pelo prefixo "F", quando associadas à equipa que actua fora do seu

terreno. Deste modo, o resultado de cada jogo está relacionado com estas 42 estatísticas, denominadas variáveis independentes, para efeito do modelo de regressão.

A análise reporta-se a duas épocas (1998/1999 e 1999/2000) do campeonato nacional de futebol, constituindo cada jogo uma observação que irá integrar o modelo de regressão, pelo que existem, no total, 612 observações destas 42 variáveis. O Modelo de Regressão relaciona, estatisticamente, uma variável dependente com variáveis independentes, exprimindo a tendência da primeira para variar com as segundas, de um modo sistemático.

3. MODELO DE REGRESSÃO

A hipótese formulada neste estudo consiste em estabelecer uma relação estatística, utilizando a análise de regressão, entre o resultado de um jogo de futebol e as estatísticas disponíveis para as variáveis que estão relacionadas com o comportamento das equipas intervenientes, de modo a servir dois objectivos:

- descrição, através do desenvolvimento de um modelo válido e utilizável para caracterizar a relação entre as variáveis;
- controlo, de modo a verificar se os resultados efectivamente obtidos são, ou não, diferentes dos previstos pelo modelo.

A fórmula geral da curva de regressão para o modelo é:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad i = 1, 2, ..., n \quad (1)$$

- n é o número de observações: corresponde aos 612 jogos, das épocas de 1998/1999 e 1999/2000, do campeonato nacional de futebol;
- Y_i é a variável dependente (i representa a observação i em n observações): serão testadas duas medidas do resultado de cada jogo de futebol: a primeira, denominada VED, consiste numa escala ordinal, em que 2 = vitória, 1 = empate e 0 = derrota, definida desta forma, para garantir a simetria da variável; a segunda, representada pela sigla GMGS, é dada pela diferença, entre Golos Marcados e Golos Sofridos, para a equipa que actua em casa;
- X_{p-1} são as variáveis independentes ou explicativas: correspondem às observações das 42 estatísticas apresentadas no Quadro n.º 1, para cada jogo;
- β_k são os parâmetros do modelo, medindo a variação média do valor esperado da variável dependente Y, com o aumento de uma unidade na variável associada, quando todas as outras variáveis explicativas no modelo permanecem constantes;
- ε_i é o termo de erro aleatório, representando as variáveis com poder explicativo sobre a variável dependente omitidas pelo modelo.

No modelo, uma componente da variação na variável dependente é explicada pelas variáveis independentes que o constituem, sendo dada pelo coeficiente de determinação (r^2), mas existe uma outra componente que reflecte a nossa ignorância referente a factores explicativos não contemplados.

4. DESENVOLVIMENTO DOS MODELOS DE REGRESSÃO

4.1. DADOS ESTATÍSTICOS: VARIÁVEIS DO MODELO

Os dados em análise constituem um estudo observacional explicativo. Uma regra empírica para situações deste tipo é que devem existir 6 a 10 observações por cada variável independente, o que sucede neste estudo.

4.1.1. SELECÇÃO DAS VARIÁVEIS INDEPENDENTES A INCLUIR NOS MODELOS

É também importante reduzir o número de variáveis a incluir no modelo final (Miller, 1990), devido à dificuldade em compreender um modelo com muitas variáveis e ao aumento na variabilidade associada aos coeficientes do modelo, provocado pela correlação entre as variáveis – o que sucede, normalmente, num estudo observacional explicativo –, que diminui as capacidades descritivas e de controlo do modelo (os valores estimados para a variável dependente apresentam grande variância).

Utilizam-se procedimentos para seleccionar e testar sub-grupos de variáveis independentes importantes, com o objectivo de desenvolver um modelo com um menor número de variáveis, escolhendo as que explicam "melhor" a variável dependente.

4.1.2. ANÁLISE FACTORIAL DE COMPONENTES PRINCIPAIS

No decorrer do estudo que permite a selecção das variáveis independentes relevantes para o modelo, verifica-se a existência de uma forte correlação entre algumas variáveis, o que impede a sua utilização simultânea nos modelos de regressão. Este facto, que ocorre para qualquer uma das variáveis dependentes (VED ou GMGS), obriga à selecção de apenas uma variável, entre as várias que estão correlacionadas entre si, para integrar o modelo de regressão, situação que originaria uma perda da informação, presente nas restantes variáveis eliminadas do modelo, que se podem considerar, à partida, importantes. Para ultrapassar esta contrariedade, recorreu-se à aplicação das técnicas de análise factorial de componentes principais, de modo a reduzir o número de variáveis, minimizando a perda de informação.

As variáveis com forte correlação, entre si, são as que representam as estatísticas de ataque de qualquer das equipa (tempos de posse de bola e recuperações,

ataques, remates e cruzamentos) e as disciplinares para as equipas que actuam fora do seu terreno (cartões amarelos e vermelhos).

De acordo com as correlações existentes entre variáveis, existem cinco grupos de variáveis, apresentados no Quadro n.º 2, aos quais será aplicada a análise factorial de componentes principais, de modo a transformar as variáveis iniciais (X_1, X_2, ..., X_p), correlacionadas entre si, num menor número de variáveis não correlacionadas (F_1, F_2, ..., F_p), designadas por factores principais ou, simplesmente, por factores. Cada variável é expressa como combinação linear dos factores que lhe são comuns:

$$\begin{aligned} X_1 &= b_{11}F_1 + b_{12}F_2 + ... + b_{1k}F_k + U_1 \\ X_2 &= b_{21}F_1 + b_{22}F_2 + ... + b_{2k}F_k + U_2 \\ &... \\ X_p &= b_{p1}F_1 + b_{p2}F_2 + ... + b_{pk}F_k + U_p \end{aligned} \qquad (2)$$

Onde b_{ij} são os coeficientes, factores ou *loadings*, que correlacionam as variáveis originais com os factores comuns e U_i representam os factores únicos, ou seja, a parte de uma variável que não é explicada pelos factores comuns.

Os primeiro e segundo grupos de variáveis são constituídos pelas estatísticas de ataque (ataques, remates, cruzamentos e recuperações), para as equipas que jogam em casa e fora, respectivamente CATA e FATA; o terceiro e quarto grupos integram os tempos de posse de bola (no meio campo defensivo, no ataque e total), também para as equipas que actuam em casa – CTAT – e fora – FTAT –, respectivamente; o quinto grupo apresenta as estatísticas disciplinares (cartões amarelos e vermelhos), da equipa que actua fora de casa – FCAV.

Da análise do quadro, verifica-se que todas as variáveis de cada grupo apresentam correlações positivas entre si e, como dado complementar, deve ser referido que todas as correlações apresentadas são significativas, para um nível de significância de 1%.

Quadro n.º 2. Grupos de variáveis sujeitas a análise factorial: correlações entre as variáveis integrantes de cada grupo

Grupo 1 – **CATA**

	CAT	CRE	CCR	CR
CAT	1,00	0,59	0,74	0,35
CRE	0,59	1,00	0,53	0,18
CCR	0,74	0,53	1,00	0,28
CR	0,35	0,18	0,28	1,00

Grupo 2 – **FATA**

	FAT	FRE	FCR	FR
FAT	1,00	0,59	0,73	0,39
FRE	0,59	1,00	0,50	0,21
FCR	0,73	0,50	1,00	0,25
FR	0,39	0,21	0,25	1,00

Grupo 3 – **CTAT**

	CMCO	CATA	CTOT
CMCO	1,00	0,49	0,77
CATA	0,49	1,00	0,70
CTOT	0,77	0,70	1,00

Grupo 4 – **FTAT**

	FMCO	FATA	FTOT
FMCO	1,00	0,51	0,77
FATA	0,51	1,00	0,64
FTOT	0,77	0,64	1,00

Grupo 5 – **FCAV**

	FCA	FCV
FCA	1,00	0,52
FCV	0,52	1,00

De modo a prosseguir com a análise factorial, deve testar-se e rejeitar a hipótese da matriz de correlações ser a matriz identidade: utilizando o teste de esfericidade de Bartlett que, para todos os grupos de variáveis, apresenta um nível de significância de 0,000, pode rejeitar-se a hipótese testada, mostrando a existência de correlação entre as variáveis.

Uma medida da adequação dos dados (MAD) para a análise factorial de componentes principais consiste na estatística Kaiser-Meyer-Olkin (KMO), que compara as correlações simples com as parciais entre as variáveis, devendo os coeficientes de correlação parciais ser pequenos. Kaiser adjectiva os valores de KMO, de acordo com a adequação dos dados para a realização da análise factorial, como:

KMO	<0,5	0,5-0,6	0,6-0,7	0,7-0,8	0,8-0,9	1,0
MAD	Inaceitável	Má	Razoável	Média	Boa	Muito boa

No Quadro n.º 3 são apresentados os valores da KMO para os cinco grupos de variáveis, que sugerem dados adequados, de forma razoável ou média, para a aplicação da análise factorial, com excepção do grupo FCAV, em que essa medida é adjectivada como má, embora sendo ainda possível realizar a referida análise.

Quadro n.º 3. Adequação dos dados para a análise factorial

Grupos de variáveis	KMO	MAD
CATA	0,715	Média
FATA	0,698	Razoável
CTAT	0,614	Razoável
FTAT	0,660	Razoável
FCAV	0,500	Má

Em cada grupo de variáveis pode, ainda, ser analisada a adequação de cada variável para utilização da análise factorial, através dos valores das correlações da diagonal principal da matriz anti-imagem, que consistem numa medida da adequação dos dados (MAD), para cada variável, introduzida no Quadro n.º 4. Valores da MAD elevados, tal como os obtidos neste estudo, permitem concluir que é possível a realização da análise factorial.

Quadro n.º 4. Adequação das variáveis para a análise factorial

CATA		FATA		CTAT		FTAT		FCAV	
Variável	MAD	Variável	MAD	Variável	MAD	Variável	MAD	Variável	MAD
CAT	0,66	FAT	0,64	CMCD	0,63	FMCD	0,66	FCA	0,50
CRE	0,81	FRE	0,80	CATA	0,67	FATA	0,77	FCV	0,50
CCR	0,70	FCR	0,69	CTOT	0,57	FTOT	0,61		
CR	0,82	FR	0,74						

Depois de verificar a possibilidade de executar adequadamente a análise factorial, prossegue-se com a extracção dos factores a partir das variáveis para os cinco grupos definidos. O primeiro passo consiste, precisamente, na definição do número de componentes (ou factores) retidos, necessários para descrever os dados, através da aplicação de três regras:

- a proporção da variância explicada pelas componentes deve ser superior a 60%;
- a variância de cada componente (valores próprios ou *eigenvalues*) retida deve ser superior à unidade, ou seja, à variância média;
- no *screeplot* (gráfico da variância explicada em função das componentes), os pontos de maior declive correspondem às componentes retidas.

De acordo com os procedimentos referidos, como pode ser observado no Quadro n.º 5, foram retidos dois factores nos dois primeiros grupos de variáveis e apenas um factor nos restantes, assinalados a *negrito*.

Quadro n.º 5. Factores (componentes) retidos a partir das variáveis originais

CATA	Valores próprios		
Factor	Total	% de σ^2	% cum.
1	**2,40**	**60,0**	**60,0**
2	**0,86**	**21,6**	**81,6**
3	0,49	12,1	
4	0,25	6,3	

FATA	Valores próprios		
Factor	Total	% de σ^2	% cum.
1	**2,40**	**59,9**	**59,9**
2	**0,85**	**21,1**	**81,0**
3	0,51	12,8	
4	0,25	6,2	

CTAT	Valores próprios		
Factor	Total	% de σ^2	% cum.
1	**2,31**	**77,1**	**77,1**
2	0,52	17,2	
3	0,17	5,7	

FTAT	Valores próprios		
Factor	Total	% de σ^2	% cum.
1	**2,28**	**76,0**	**76,0**
2	0,51	16,9	
3	0,21	7,1	

FCAV	Valores próprios		
Factor	Total	% de σ^2	% cum.
1	**1,52**	**76,0**	**76,0**
2	0,48	24,0	

No quadro apresentam-se os valores próprios, a percentagem de variância associada a cada factor e a percentagem cumulativa de variância explicada pelas componentes retidas. A variância explicada pelos factores retidos, nos vários grupos de variáveis, varia entre 76% e 82%, valores que são bastante razoáveis para a aplicação de análise factorial.

No Quadro n.º 6 apresentam-se os valores das comunalidades, que correspondem à proporção da variância de cada variável explicada pelas componentes principais retidas. Todas as variáveis apresentam uma forte relação com os factores retidos, uma vez que os valores observados para as comunalidades são elevados, o menor valor é de 65%, sendo a maioria dos valores superiores a 75%.

Quadro n.º 6. Comunalidades (Com.) associadas a cada variável

CATA		**FATA**		**CTAT**		**FTAT**		**FCAV**	
Variável	Com.	Variável	Com.	Variável	Com.	Variável	Com.	Variável	Com.
CAT	0,82	FAT	0,83	CMCD	0,74	FMCD	0,77	FCA	0,76
CRE	0,70	FRE	0,68	CATA	0,68	FATA	0,65	FCV	0,76
CCR	0,77	FCR	0,75	CTOT	0,89	FTOT	0,86		
CR	0,98	FR	0,99						

Após a aplicação e validação de todos os passos da análise factorial de componentes principais, finalmente, procede-se à determinação da matriz dos componentes, que apresenta os coeficientes, factores ou *loadings*, que relacionam as variáveis com os factores retidos.

Quadro n.º 7. Matriz de componentes

	Factor CATA			Factor FATA			Factor		Factor		Factor	
Var.	1	2	Var.	1	2	Var.	CTAT	Var.	FTAT	Var.	FCAV
CAT	0,903	-0,067	FAT	0,909	-0,048	CMCD	0,860	FMCD	0,876	FCA	0,872
CRE	0,770	-0,328	FRE	0,767	-0,297	CATA	0,825	FATA	0,809	FCV	0,872
CCR	0,866	-0,127	FCR	0,843	-0,200	CTOT	0,945	FTOT	0,926		
CR	0,493	0,858	FR	0,520	0,846						

Por exemplo, para o primeiro grupo de variáveis:

$$CAT = 0{,}903CATA_1 - 0{,}067CATA_2 \qquad CRE = 0{,}770CATA_1 - 0{,}328CATA_2 \quad (3)$$
$$CCR = 0{,}866CATA_1 - 0{,}127CATA_2 \qquad CR = 0{,}493CATA_1 + 0{,}858CATA_2$$

Os valores próprios das componentes retidas correspondem à soma dos quadrados dos coeficientes das variáveis, para cada factor (soma em coluna), enquanto que as comunalidades podem ser obtidas pela soma dos quadrados dos coeficientes dos factores, para cada variável (soma em linha).

Para os dois primeiros grupos de variáveis, em que foram retidos dois factores, procede-se ainda à rotação da matriz dos componentes, de modo a extremar os valores dos coeficientes, para associar, indubitavelmente, cada variável a apenas um factor. Nos grupos de variáveis CATA e FATA, embora já exista uma separação clara entre as variáveis associadas a cada factor, realizou-se a rotação da matriz de componentes, utilizando o método *Varimax*, com a normalização de Kaiser, obtendo-se, no Quadro n.º 8, a matriz de componentes, após rotação.

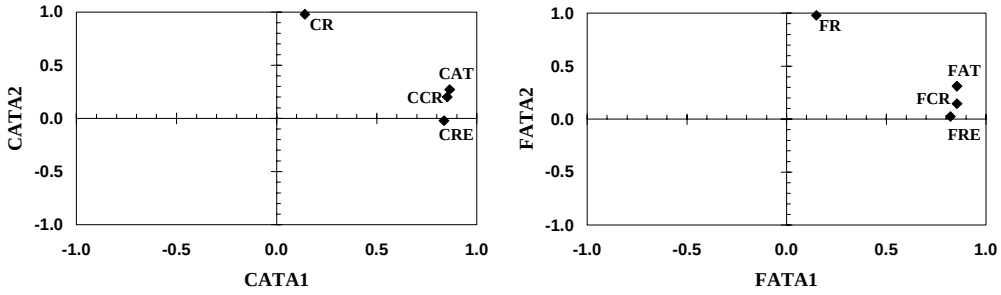
O Gráfico n.º 1 ilustra a forma como as variáveis estão associadas a cada um dos factores: para ambos os grupos de variáveis, representativos das estatísticas de ataque para as equipas que actuam em casa e fora, os ataques, remates e cruzamentos são representados pelo factor 1 e as recuperações estão ligadas ao factor 2.

No desenvolvimento dos modelos de regressão, as estatísticas de ataque são substituídas pelos factores retidos que lhes estão associados, de modo a eliminar a correlação existente entre as variáveis, não perdendo a informação por elas representada.

Quadro n.º 8. Matriz de componentes após rotação

	Factor			Factor	
Var.	CATA1	CATA2	Var.	FATA1	FATA2
CAT	0,864	0,271	FAT	0,855	0,311
CRE	0,836	-0,021	FRE	0,822	0,026
CCR	0,852	0,202	FCR	0,855	0,145
CR	0,141	0,979	FR	0,149	0,981

Gráfico n.º 1. Associação entre componentes retidos e variáveis



4.2. MODELOS DE REGRESSÃO

4.2.1. VARIÁVEIS UTILIZADAS

O objectivo deste estudo passa pelo desenvolvimento de modelos que relacionem a variável dependente, resultado de um jogo de futebol, com as variáveis independentes em análise. Foram construídos modelos, utilizando duas variáveis dependentes que exprimem, de modo diferente, o resultado de um jogo. Tal como referido anteriormente, o Modelo 1 tem como variável dependente "VED", que apresenta três resultados possíveis, para cada jogo: 2 = vitória, 1 = empate e 0 = derrota, reportados à equipa que actua no seu terreno, sendo o resultado da equipa que joga fora complementar a este; o Modelo 2 utiliza a variável dependente "GMGS", que resulta da diferença entre Golos Marcados e Golos Sofridos pela equipa que joga perante o seu público.

As 42 estatísticas disponíveis sobre cada jogo de futebol foram reduzidas; através da aplicação da análise factorial de componentes principais, previamente descrita, a algumas variáveis correlacionadas entre si. No Quadro n.º 9 estão listadas as variáveis resultantes, que serão utilizadas na construção dos modelos de regressão.

Na análise subsequente serão utilizadas estas 33 variáveis independentes, cuja nomenclatura resulta da anteriormente apresentada, no Quadro n.º 1, com a introdução do prefixo "C", quando a estatística diz respeito à equipa que actua em casa e do prefixo "F", quando se reporta à equipa que joga fora. As estatística gerais e a percentagem de posse de bola da equipa que actua no seu terreno ("POSSE") dizem respeito a cada jogo. Assinalam-se a negrito as variáveis que resultam de factores determinados na análise factorial de componentes principais.

Quadro n.º 9. Variáveis independentes para cada jogo

Gerais:		Tempos de posse:		Guarda-redes:		Jogadores:	
RELV		Casa	Fora	Casa	Fora	Casa	Fora
ESPEC		CDEF	FDEF	CDC	FDC	CFC	FFC
TJOGO		CMCD	FMCD	CDI	FDI	CFS	FFS
Disciplinares:		**CTAT**	**FTAT**	CSC	FSC	CP	FP
Casa	Fora	POSSE		CSI	FSI	***CATA1***	***FATA1***
CCA	**FCAV**					***CATA2***	***FATA2***
CCV						CAS	FAS

Para um dos jogos não estão disponíveis dados estatísticos, pelo que existem 611 observações, relativas aos jogos analisados, das variáveis em análise, com a excepção de quatro encontros, em que não estão disponíveis dados para algumas variáveis (*missing values*). Nesta situação, optou-se pela utilização dos valores médios da equipa em análise, em casa ou fora, conforme o caso, de modo a não se perder a informação estatística para outras variáveis da observação em causa.

4.2.2. CONSTRUÇÃO DOS MODELOS DE REGRESSÃO

Este constitui o passo decisivo do estudo da relação estatística entre as variáveis, embora seja descrito de forma bastante resumida.

Ambos os modelos de regressão, inicialmente, integram todas as variáveis independentes, bem como as formas funcionais não lineares da relação destas com a variável dependente, que introduzem melhorias na descrição dos dados, o que ocorre esporadicamente, para algumas variáveis, com a função quadrática, que provoca aumentos significativos no coeficiente de determinação, quando comparado com o valor do mesmo para a relação linear.

A utilização desta forma pressupõe a realização de uma transformação das variáveis, de modo a reduzir a correlação entre os termos da função quadrática, que se consegue do seguinte modo: sendo $X_{i,p}$ a observação *i* de uma variável deste tipo, a sua relação com a variável dependente será:

$$Y_i = \beta_0 + ... + \beta_{p1} x_{i,p} + \beta_{p2} x_{i,p}{}^2 + \cdots + \varepsilon_i \quad \text{com} \quad x_{i,p} = X_{i,p} - \overline{X} \quad (4)$$

Através de um processo de análise sistemática da importância de cada variável nos modelos desenvolvidos, vão sendo eliminadas, passo a passo, variáveis que não apresentam relevância, de acordo com os critérios de análise da significância das variáveis independentes, de maximização do coeficiente de determinação ajustado e utilizando o procedimento *Forward Stepwise* que, essencialmente, desenvolve uma sequência de modelos de regressão, adicionando ou retirando em cada passo uma

variável independente. A aplicação deste procedimento é descrita, pormenorizadamente, por vários autores (Freedman, 1983; Pope e Webster, 1972).

No Quadro n.º 10 são apresentados os resultados mais significativos para os modelos de regressão inicialmente construídos, bem como as variáveis independentes seleccionadas para integrar estes modelos e respectivos níveis de significância.

Quadro n.º 10. Modelos de Regressão

Modelo 1: Variável dependente - VED

		g.l.	*SS*	*MS*
Coeficiente de Determinação: r^2 = 0,334				
Estimativa do desvio padrão: $\sqrt{MSE}$ = 0,663	Regressão	20	130,00	6,500
F = 14,8 ⇒ Significância F =0,00	Resíduos	590	259,14	0,439

	Var. i	b_i	$s(b_i)$	*Sig. t*
Constante		4,020	0,465	0,000
Gerais	RELV	0,061	0,028	0,027
	ESPEC	$3{,}12e^{-6}$	$2{,}98e^{-6}$	0,295
Disci-	CCA	-0,025	0,019	0,181
plinares	CCV	-0,250	0,078	0,001
	FCAV	0,067	0,030	0,025
Tempos	CDEF	0,069	0,033	0,036
de	FDEF	-0,060	0,030	0,048
posse	POSSE	-4,483	0,775	0,000
Guarda-	CSC	0,022	0,010	0,027
-redes	FSC	-0,038	0,010	0,000

	Var. i	b_i	$s(b_i)$	*Sig. t*
Jogador	CFC	0,013	0,006	0,021
	CFC^2	-0,001	0,001	0,047
	CP	-0,006	0,004	0,121
	CATA1	0,108	0,042	0,009
	CATA2	0,199	0,040	0,000
	CAS	0,118	0,019	0,000
	FFC	-0,011	0,005	0,034
	FATA1	-0,096	0,040	0,018
	FATA2	-0,195	0,039	0,000
	FAS	-0,142	0,023	0,000

Modelo 2: Variável dependente - GMGS

Coeficiente de Determinação: r^2 = 0,422		g.l.	*SS*	*MS*
Estimativa do desvio padrão: $\sqrt{MSE}$ = 1,290	Regressão	18	718,93	39,941
F = 24,0 ⇒ Significância F =0,00	Resíduos	592	984,49	1,663

	Var. i	b_i	$s(b_i)$	*Sig. t*
Constante		-0,732	0,968	0,450
Gerais	RELV	0,096	0,054	0,077
	ESPEC	$7{,}80e^{-6}$	$5{,}77e^{-6}$	0,177
	TJOGO	3,658	1,492	0,014
Disci-	CCA	-0,056	0,036	0,126
plinares	CCV	-0,523	0,151	0,001
	FCAV	0,159	0,059	0,007
Guarda-	CSC	0,060	0,019	0,002
-redes	FSC	-0,099	0,018	0,000

	Var. i	b_i	$s(b_i)$	*Sig. t*
Jogador	CFC	0,035	0,010	0,001
	CP	-0,012	0,008	0,152
	CATA1	-0,014	0,069	0,835
	CATA2	0,377	0,078	0,000
	CAS	0,305	0,043	0,000
	CAS^2	0,031	0,011	0,004
	FFC	-0,037	0,010	0,000
	FATA1	-0,014	0,069	0,840
	FATA2	-0,384	0,077	0,000
	FAS	-0,340	0,044	0,000

g.l. – graus de liberdade *SS* e *MS* – somatório e média do somatório dos quadrados.

b_i e $s(b_i)$ – estimativas do coeficiente e do seu desvio padrão para a variável i.

Sig. *t* – nível de significância do teste *t* de *Student*.

Verifica-se a exclusão, do modelo 1 – VED –, das variáveis percentagem de tempo de jogo, tempos de posse de bola no meio campo defensivo e no ataque, defesas e saídas incompletas dos guarda-redes, faltas sofridas e perdas de bola da equipa que joga fora. A variável faltas cometidas pela equipa que actua em casa surge na forma quadrática.

O coeficiente de determinação indica que apenas 33,4% da variação que ocorre na variável dependente VED é explicada pelas variáveis incluídas no modelo, o teste *F*, à significância global do modelo, é validado por apresentar significância nula.

A estatística, ou variável, cuja estimativa do coeficiente apresenta valor positivo contribui positivamente para a vitória da equipa que actua em casa, tendo as estimativas negativas o efeito contrário: uma variação de uma unidade na variável independente provoca uma variação média esperada na variável dependente igual ao valor da estimativa do coeficiente. A significância do teste *t* de *Student* para cada variável indica-nos a probabilidade dessa variável tomar um valor nulo no modelo, não sendo significante. Existem três variáveis (ESPEC, CCA e CP), cujos valores da significância são superiores ao estabelecido como desejável, que é de 5%.

O modelo 2 – GMGS –, comparativamente com o anterior, passa a integrar a variável percentagem de tempo de jogo, sendo excluídos os tempos de posse de bola. A variável assistências da equipa que joga em casa aparece na forma quadrática.

O coeficiente de determinação é superior: as variáveis do modelo explicam 42,2% da variação da variável dependente GMGS e o teste *F* valida o modelo.

As variáveis apresentam estimativas para os seus coeficientes com sinal idêntico às do modelo 1, com excepção do factor CATA1, para as estatísticas de ataque da equipa que actua em casa. Existe um número apreciável de variáveis, cuja significância do teste *t* de *Student* apresenta valores da significância superiores ao estabelecido como desejável, que é de 5%, que serão analisadas nos pontos seguintes.

4.2.3. ANÁLISE DE *OUTLIERS*

Os *outliers* são casos extremos influentes em determinada análise estatística. No desenvolvimento dos modelos de regressão importa determinar o conjunto de observações que podem ser consideradas como *outliers*, de modo a equacionar a sua eliminação na construção de modelos subsequentes, com o objectivo de obter refinamentos. A análise de *outliers* (Barnett e Lewis, 1994; Rousseeuw e Leroy, 1987), com o objectivo de identificação das observações que os constituem será efectuada com a ajuda das seguintes estatísticas: resíduos estandardizados, resíduos estudantizados, *leverage*, distância de *Cook*, *dfBetas* estandardizados e *dfFit* estandardizado.

De modo a proceder a esta análise, torna-se premente introduzir a matriz *H*, cuja compreensão é facilitada pela formulação matricial do modelo.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1,p-1} \\ 1 & X_{21} & \dots & X_{2,p-1} \\ \dots & \dots & \dots & \dots \\ 1 & X_{n1} & \dots & X_{n,p-1} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix} \text{ ou } \underset{(n\times 1)}{Y} = \underset{(n\times p)}{X} \cdot \underset{(p\times 1)}{\beta} + \underset{(n\times 1)}{\varepsilon} \qquad (5)$$

A matriz H resulta de:

$$\underset{(n\times n)}{H} = X \cdot \left(X'X\right)^{-1} \cdot X' \text{, sendo } X' \text{ a matriz transposta de } X. \qquad (6)$$

Os resíduos, para cada observação, calculam-se a partir da diferença entre os valores observados e os valores estimados pelo modelo para a variável dependente:

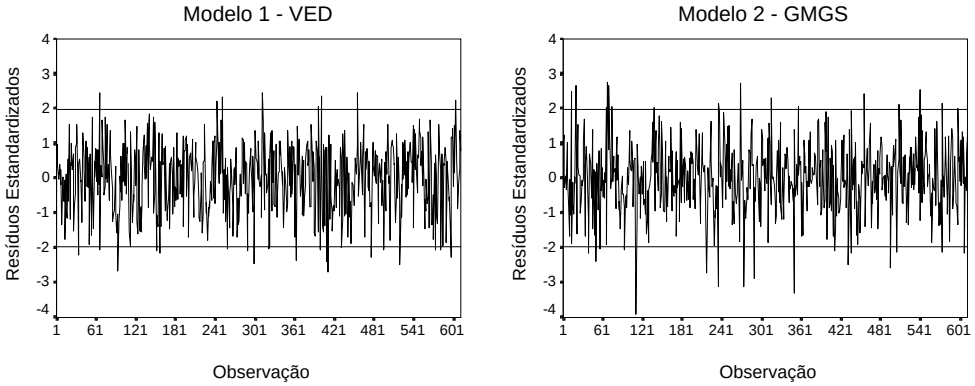
$$e_i = Y_i - \hat{Y}_i \qquad (7)$$

A partir destes valores podem calcular-se os resíduos estandardizados:

$$e_i^{\,*} = \frac{e_i}{\sqrt{MSE}} \qquad (8)$$

Considera-se como *outlier* uma observação em que o resíduo estandardizado tenha valor absoluto superior a 1,96, para um nível de significânca de 5%. Nos dois modelos desenvolvidos, identificam-se algumas observações como *outliers*, representadas pelos pontos que ultrapassam os limites, no Gráfico n.º 2.

Gráfico n.º 2. Resíduos estandardizados



Um primeiro refinamento, que torna os resíduos mais eficazes no reconhecimento de *outliers*, consiste no reconhecimento do facto de que as observações podem apresentar diferentes desvios padrão entre elas. O desvio padrão de uma observação é estimado através da expressão (9), em que h_{ii} representa o elemento da diagonal principal da matriz *H* referente à observação *i*.

$$s(e_i) = \sqrt{MSE(1 - h_{ii})} \tag{9}$$

A razão entre cada resíduo e o seu desvio padrão estimado é denominada resíduo estudantizado:

$$r_i = \frac{e_i}{s(e_i)} \tag{10}$$

Uma segunda melhoria resulta do cálculo dos resíduos para a observação *i*, quando o modelo de regressão se baseia em todos os dados, com a excepção da observação *i*, obtendo-se os resíduos *deleted*:

$$d_i = Y_i - \hat{Y}_{i(i)} \tag{11}$$

Combinando as duas formas introduzidas, podem calcular-se os resíduos estudantizados *deleted*:

$$t_i = \frac{d_i}{s(d_i)} \tag{11}$$

Estes, possibilitam um melhor diagnóstico de *outliers*, que resultam das observações, cujos resíduos estudantizados *deleted* são elevados, em valor absoluto. Pode demonstrar-se que este tipo de resíduos seguem uma distribuição t de Student, pelo que é possível definir um valor crítico, a partir do qual se considera uma observação como *outlier*: para um nível de significância de 5%, esse valor crítico é também de 1,96. Assim, no Gráfico n.º 3, ilustram-se os *outliers* obtidos.

A partir da análise dos resíduos, ilustrada pelos Gráficos n.º 2 e n.º 3, a detecção de *outliers* é semelhante para os dois tipos de resíduos, sendo o seu número superior no Modelo 2 – GMGS –, bem como o valor absoluto dos resíduos.

O *leverage*, dado pelo elemento da diagonal principal (h_{ii}) da matriz *H*, previamente definida, representa a influência da observação *i* na qualidade do ajustamento feito. Quando é superior a duas vezes o seu valor médio, ou seja, $2(p+1)/n$ (utilizando a nomenclatura introduzida na formulação do modelo), a observação é considerada influente. O gráfico n.º 4 ilustra os *outliers* identificados por esta regra.

Gráfico n.º 3. Resíduos estudantizados *deleted*

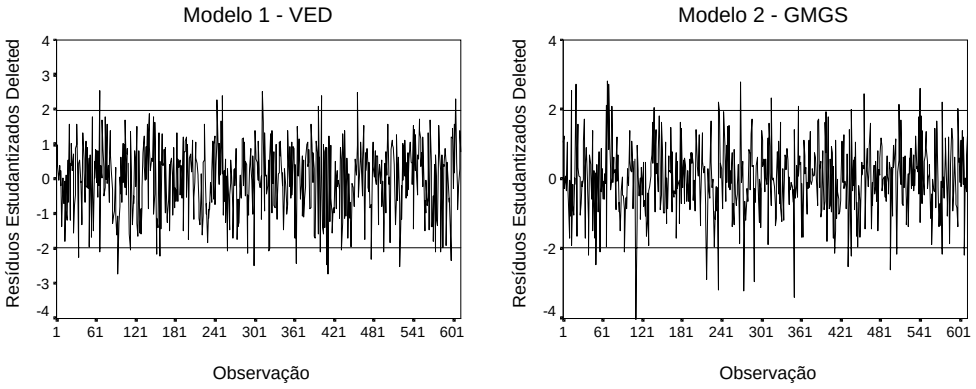
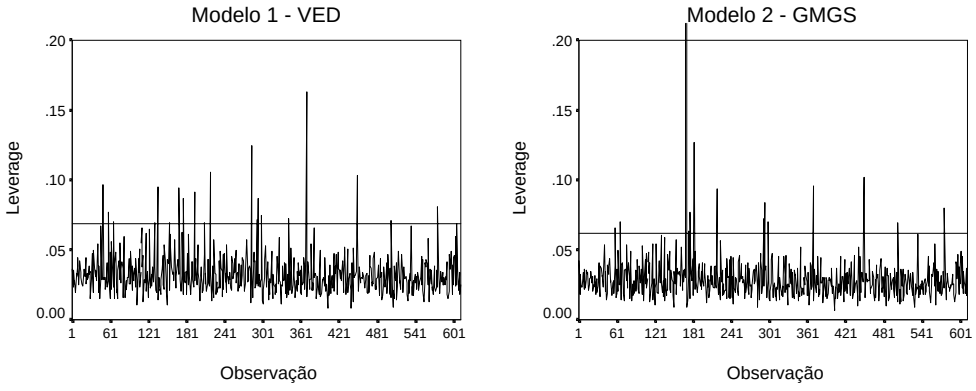


Gráfico n.º 4. *Leverage*



Após a identificação, como *outliers*, de observações, no que diz respeito aos valores das variáveis dependente e independentes, importa verificar a sua influência

no comportamento do modelo, que pode ser quantificada pela distância de *Cook*, *dfBetas* estandardizados e *dfFit* estandardizado: uma observação considera-se influente, se a sua exclusão causar alterações substanciais na função de regressão estimada.

A distância de *Cook* considera a variação provocada nos resíduos de todas as observações, quando a observação *i* é excluída do cálculo dos coeficientes de regressão, podendo ser calculada sem recorrer à estimação de uma nova função de regressão, cada vez que uma observação é excluída, através de uma expressão equivalente:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{(p+1)\cdot MSE} \Leftrightarrow D_i = \frac{{e_i}^2}{(p+1)\cdot MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \qquad (12)$$

Uma observação é considerada influente quando a distância de *Cook* é superior a 4/(*n*-*p*-1). Os respectivos *outliers* são ilustradas pelo Gráfico n.º 5.

O *dfFit* estandardizado representa a diferença entre o valor estimado pelo modelo, para a observação *i*, quando todas as observações são utilizadas e o valor estimado, para a mesma observação, quando o caso *i* é excluído do cálculo da função de regressão que, tal como na equação anterior, pode ser calculado através de uma expressão equivalente, que não obriga ao cálculo da função de regressão, cada vez que uma observação é excluída do modelo.

$$dfFits_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}\, h_{ii}}} \Leftrightarrow dfFits_i = t_i \sqrt{\frac{h_{ii}}{1-h_{ii}}} \qquad (13)$$

Uma observação é considerada *outlier*, quando o valor absoluto do *dfFit* estandardizado é superior a $2\sqrt{(p+1)/n}$. Os resultados, para esta medida da influência das observações, apresentam-se no Gráfico n.º 6.

Gráfico n.º 5. Distância de *Cook*

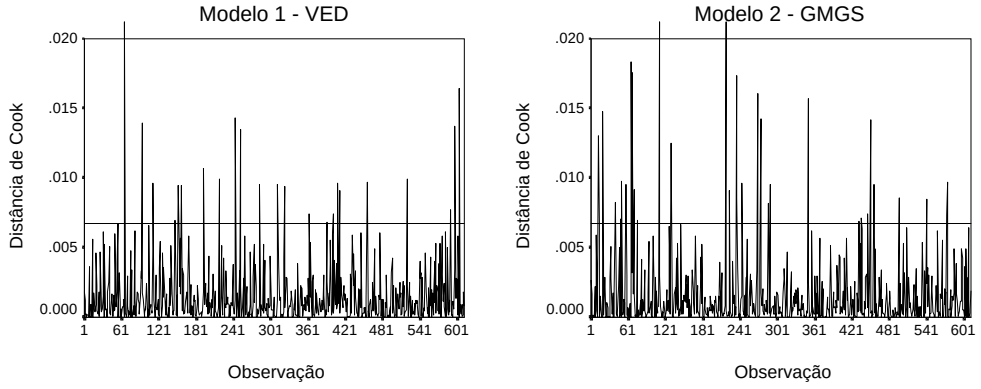
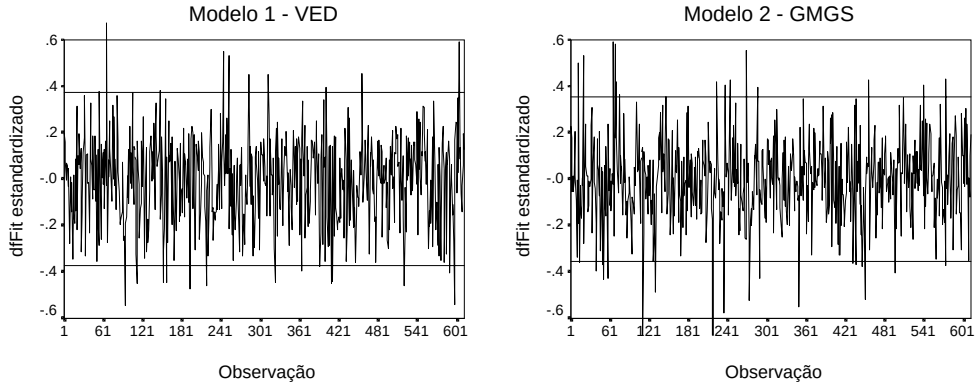


Gráfico n.º 6. dfFit estandardizado



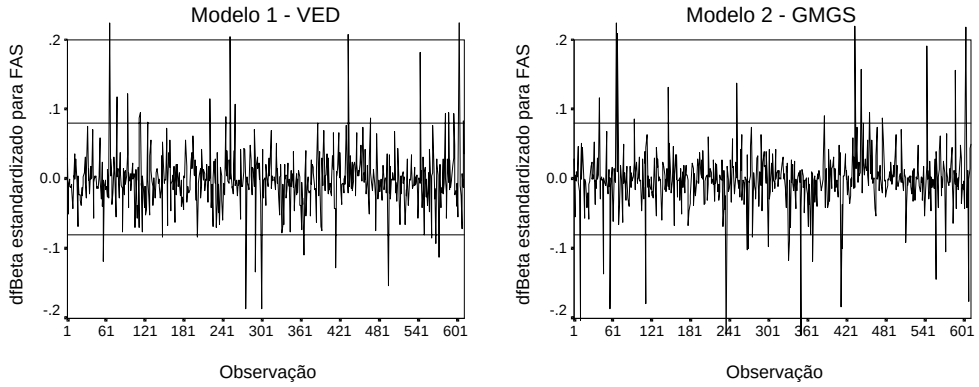
A medida da influência de uma observação *i*, em cada coeficiente da regressão β_k, resulta da diferença entre o valor estimado para o coeficiente de regressão baseado em todas as observações e o mesmo valor omitindo o caso i. O *DfBeta* estandardizado obtém-se, pelo quociente entre essa diferença e a estimativa do desvio padrão do coeficiente de regressão em análise:

$$dfBeta_i = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} c_{kk}}} \qquad k = 0, 1, \ldots, p\text{-}1 \qquad (14)$$

Em que c_{kk} é o *k* elemento da diagonal principal da matriz $(X'X)^{-1}$.

O valor de *DfBeta* é calculado, para todas as observações, para todos os parâmetros e para a constante do modelo. As observações são consideradas *outliers* quando o valor absoluto de *DfBeta* é superior a $2/\sqrt{n}$. O Gráfico n.º 7 permite observar os pontos assim identificados, segundo o critério do *DfBeta*, para os coeficientes associados à variável CAS, a título de exemplo.

Gráfico n.º 7. *dfBeta* estandardizado para a variável CAS



A análise de *outliers* apresentada permite identificar os casos extremos considerados influentes para os modelos, que serão excluídos na construção de novas

funções de regressão. Foram considerados casos extremos influentes as observações que desrespeitam as condições impostas aos resíduos ou, então, que não estejam dentro dos limites impostos para, pelo menos, três dos restantes critérios considerados. A análise de *outliers* processou-se, deste modo, em dois passos sucessivos, de maneira a construir um modelo intermédio, cuja análise, nos mesmos termos do modelo inicial, permitiu a detecção de mais casos extremos influentes, que levaram à construção dos modelos de regressão definitivos. Os critérios estabelecidos permitem a detecção de 120 *outliers* no modelo 1 – VED – e, no modelo 2 – GMGS –, foram considerados 109 *outliers*.

É relevante observar que as observações detectadas como *outliers*, em ambos os modelos, se distribuem de modo uniforme por todas as equipas, quer actuando em casa, quer em jogos fora. Quanto às variáveis que representam os resultados dos jogos, tanto no caso da variável VED, como para a variável GMGS, há poucos empates considerados como outliers, quando comparados com a sua frequência relativa no total de jogos.

4.2.4. REFINAMENTO DOS MODELOS DE REGRESSÃO, ELIMINANDO OS *OUTLIERS*

Foram desenvolvidos novos modelos, excluindo os *outliers* detectados anteriormente, que serão utilizados para explicar as relações estatísticas entre as variáveis em análise. No Quadro n.º 11 apresentam-se os resultados significativos para os modelos de regressão definitivos, o primeiro com 491 observações e o segundo com 502, após a eliminação dos casos extremos influentes, bem como para as variáveis independentes.

As alterações mais importantes no modelo 1 – VED –, relativamente ao inicialmente construído, residem no aumento do coeficiente de determinação: a variação que ocorre na variável dependente VED, explicada pelas variáveis do modelo, aumentou praticamente para o dobro – 62,1% –; na diminuição do desvio padrão e nos níveis de significância associados às variáveis, sendo apenas dois moderadamente superiores a 5%.

As variáveis que contribuem positivamente para a vitória da equipa que actua em casa, como explicado anteriormente, pelo sinal da estimativa do respectivo coeficiente, são o estado do relvado, o número de espectadores, o tempo de posse na defesa, as saídas completas do guarda-redes, as faltas cometidas, as estatísticas de ataque e as assistências, para a equipa que actua em casa e as acções disciplinares para a equipa que joga fora. As variáveis cuja contribuição para a variável dependente é negativa são o tempo de posse na defesa, as saídas completas do guarda-redes, as faltas cometidas, as estatísticas de ataque e as assistências, para a equipa que joga fora e as acções disciplinares, a percentagem de posse de bola e as perdas de bola para a equipa que actua em casa. Analisando criticamente os resultados, não seria talvez de esperar que as faltas cometidas por uma equipa tivessem um contributo positivo para o resultado e que a percentagem de posse de bola influenciasse negativamente o resultado.

Quadro n.º 11 Modelos de Regressão

Modelo 1: Variável dependente - VED

Coeficiente de Determinação: r^2 = 0,621		g.l.	*SS*	*MS*
Estimativa do desvio padrão: $\sqrt{MSE}$ = 0,457	Regressão	20	160,91	8,045
F = 38,5 ⇒ Significância F =0,00	Resíduos	470	98,14	0,209

	Var. i	b_i	$s(b_i)$	*Sig. t*
Constante		5,757	0,366	0,000
Gerais	RELV	0,038	0,022	0,086
	ESPEC	4,52e^{-6}	2,30e^{-6}	0,050
Disci-	CCA	-0,041	0,014	0,004
plinares	CCV	-0,407	0,066	0,000
	FCAV	0,087	0,023	0,000
Tempos	CDEF	0,044	0,027	0,099
de	FDEF	-0,059	0,024	0,015
posse	POSSE	-6,387	0,611	0,000
Guarda-	CSC	0,015	0,008	0,053
-redes	FSC	-0,043	0,008	0,000

	Var. i	b_i	$s(b_i)$	*Sig. t*
Jogador	CFC	0,011	0,004	0,010
	CFC^2	-0,002	0,000	0,000
	CP	-0,006	0,003	0,053
	CATA1	0,067	0,033	0,043
	CATA2	0,202	0,032	0,000
	CAS	0,091	0,015	0,000
	FFC	-0,022	0,004	0,000
	FATA1	-0,211	0,032	0,000
	FATA2	-0,284	0,031	0,000
	FAS	-0,197	0,018	0,000

Modelo 2: Variável dependente - GMGS

Coeficiente de Determinação: r^2 = 0,620		g.l.	*SS*	*MS*
Estimativa do desvio padrão: $\sqrt{MSE}$ = 0,834	Regressão	18	546,54	30,363
F = 43,7 ⇒ Significância F =0,00	Resíduos	483	335,61	0,695

	Var. i	b_i	$s(b_i)$	*Sig. t*
Constante		-0,864	0,711	0,224
Gerais	RELV	0,083	0,039	0,034
	ESPEC	7,51e^{-6}	4,30e^{-6}	0,081
	TJOGO	4,368	1,107	0,000
Disci-	CCA	-0,083	0,026	0,001
Plinares	CCV	-0,496	0,110	0,000
	FCAV	0,120	0,042	0,004
Guarda-	CSC	0,059	0,014	0,000
-redes	FSC	-0,100	0,013	0,000

	Var. i	b_i	$s(b_i)$	*Sig. t*
Jogador	CFC	0,034	0,008	0,000
	CP	-0,012	0,006	0,039
	CATA1	-0,126	0,050	0,011
	CATA2	0,342	0,056	0,000
	CAS	0,270	0,030	0,000
	CAS^2	0,030	0,007	0,000
	FFC	-0,042	0,007	0,000
	FATA1	-0,117	0,049	0,017
	FATA2	-0,501	0,056	0,000
	FAS	-0,350	0,034	0,000

g.l. – graus de liberdade *SS* e *MS* – somatório e média do somatório dos quadrados.

b_i e $s(b_i)$ – estimativas do coeficiente e do seu desvio padrão para a variável i.

Sig. *t* – nível de significância do teste *t* de *Student*.

O novo modelo 2 – GMGS – introduz também um aumento do coeficiente de determinação, relativamente ao modelo inicial, para 62%, diminuição do desvio padrão e níveis de significância inferiores a 5% para as variáveis, com excepção apenas de uma delas.

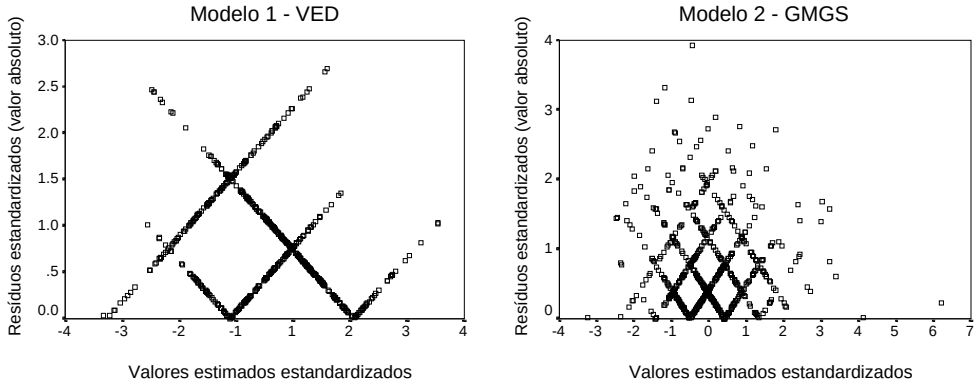
As variáveis partilhadas com o modelo 1 contribuem todas da mesma forma para o resultado do jogo, com excepção do factor CATA1, para as estatísticas de ataque (ataques, cruzamentos e remates) da equipa que actua em casa, que influencia negativamente o resultado da equipa que joga em casa, o que não seria de esperar à partida. A variável tempo jogado (% do tempo total) favorece o resultado para a equipa que actua perante o seu público.

4.2.5. VALIDAÇÃO DOS MODELOS DE REGRESSÃO

Os modelos de regressão devem cumprir determinados pressupostos, cuja verificação valida os modelos desenvolvidos. Deste modo, torna-se necessária a concretização de testes estatísticos, que incluem análise gráfica de resíduos, estudo da multicolinearidade (correlação entre variáveis independentes), análise da homocedasticidade (variância constante dos termos de erro) e medida da auto-correlação, com o objectivo de validar os modelos.

Em primeiro lugar será verificada a homocedasticidade que, etimologicamente significa variância constante. Resultando um resíduo da diferença entre os valores previstos pelo modelo e os valores observados, um dos processos alternativos para analisar a homocedasticidade consiste em observar a relação entre os resíduos estandardizados e os valores estimados estandardizados da variável dependente. No gráfico n.º 8 ilustra-se esta relação, para o valor absoluto dos resíduos estandardizados, que torna mais fácil a análise gráfica.

Gráfico n.º 8. Relação entre resíduos e valores estimados estandardizados



No modelo 1 verifica-se uma dispersão ligeiramente superior para valores estimados inferiores, em valor absoluto, derivada também do tipo de variável dependente. No modelo 2 a amplitude mantém-se aproximadamente constante em relação ao eixo horizontal, pelo que não parece existir variação da dispersão. Os

modelos parecem passíveis de ser considerados, quanto a este primeiro pressuposto, com alguns cuidados para o primeiro modelo.

Um segundo pressuposto a analisar é a inexistência de auto-correlação entre as variáveis independentes, através do teste de Durbin-Watson, que permite verificar se os termos de erro são independentes, ou seja, se o parâmetro de auto-correlação é nulo. A estatística de teste apresenta o valor de 1,85 e de 1,86 para os modelos 1 e 2, respectivamente. Para um nível de significância de 5%, o valor crítico considerado para este teste é de 1,78, pelo que não podemos rejeitar a hipótese, para nenhum dos modelos construídos, de que a auto-correlação seja nula.

Um terceiro pressuposto define que os resíduos devem seguir uma distribuição normal, podendo ser verificado pelo teste Kolmogorov-Smirnov (*K-S*), com a correcção de Lilliefors, apresentado no Quadro n.º 12.

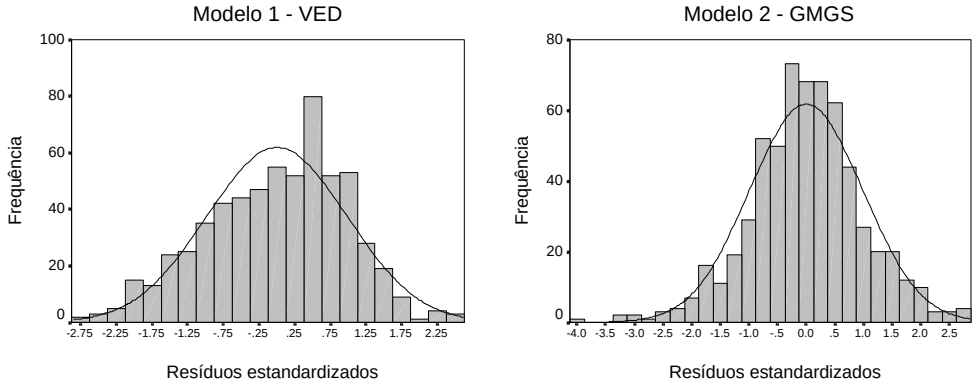
Quadro n.º 12. Teste à *normalidade* dos resíduos estandardizados

	Estatística *K-S* (Lilliefors)	Graus de liberdade	Significância
Modelo 1	0,059	611	0,000
Modelo 2	0,036	611	0,052

Exige-se, normalmente, um nível de significância de 5% para não rejeitar a hipótese dos resíduos seguirem uma distribuição normal, o que sucede apenas para o Modelo 2. Também para este pressuposto, o Modelo 1 exige uma atenção especial.

De modo a complementar o estudo da *normalidade* dos resíduos, apresenta-se o Gráfico n.º 9, que regista a diferença entre o histograma da distribuição das variáveis aleatórias residuais e a distribuição normal, observando-se alguma correspondência entre a distribuição das frequências relativas das várias classes definidas para os resíduos e a curva de distribuição normal, principalmente para o Modelo 2.

Gráfico n.º 9. Histograma dos resíduos estandardizados e distribuição normal

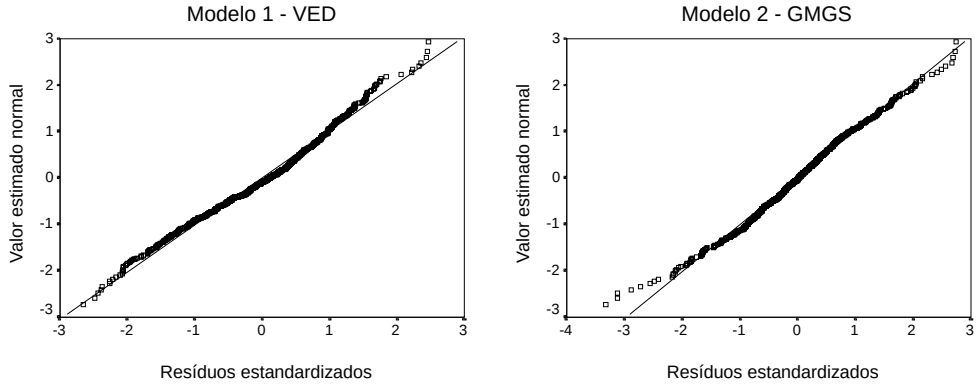


Os desvios à normalidade podem ser ainda observados no Gráfico n.º 10, em que se apresentam os gráficos *Q-Q*, de modo a ilustrar, pelos desvios à linha oblíqua, as diferenças em relação à distribuição normal. Verifica-se que estes desvios não

apresentam magnitude mais elevada para o modelo 1, pelo que podemos prosseguir com a validação dos modelos.

Finalmente, importa verificar o pressuposto da ausência de multicolinearidade, cuja intensidade pode ser estudada, sumariamente, através da análise da correlação entre as variáveis independentes. Os maiores valores absolutos observados para a correlação são entre as variáveis CATA2 e FATA2, cujas maiores componentes são dadas pelas recuperações de bola, com valores de (-0,65) e (-0,62) para os modelos 1 e 2, respectivamente, o que não indicia multicolinearidade.

Gráfico n.º 10. Gráficos *Q-Q* dos resíduos estandardizados



O factor de inflação da variância (FIV) é também uma medida da multicolineariedade, que contabiliza a inflação sofrida pela variância dos coeficientes de regressão estimados, provocada pela correlação entre variáveis. Pode ser demonstrado que este factor, para uma variável *k*, é:

$$(FIV)_k = \left(1 - r_k^{\,2}\right)^{-1} \qquad k = 1, 2, \ldots, p\text{-}1 \qquad (15)$$

onde r_k^2 corresponde ao coeficiente de determinação, quando a variável X_k é relacionada, através de um modelo de regressão linear, com as restantes (p-2) variáveis independentes.

Valores elevado do FIV são indicadores de multicolinearidade, considerando-se valores superiores a 10 influenciadores das estimativas dos coeficientes de regressão. No Quadro n.º 13 apresentam-se os FIV para as variáveis utilizadas nos dois modelos, cujos valores mais elevados não indiciam, de algum modo, a existência de multicolinearidade.

Quadro n.º 13. Factor de inflacção da variância

Modelo 1 - VED				Modelo 2 - GMGS			
Variável	*FIV*	*Variável*	*FIV*	*Variável*	*FIV*	*Variável*	*FIV*
Const.		CFC	1,408	Const.		CFC	1,262
RELV	1,277	CFC^2	1,206	RELV	1,279	CP	1,211
ESPEC	1,272	CP	1,244	ESPEC	1,258	*CATA1*	1,738
CCA	1,421	*CATA1*	2,505	TJOGO	1,369	*CATA2*	2,336
CCV	1,259	*CATA2*	2,303	CCA	1,399	CAS	1,617
FCAV	1,246	CAS	1,280	CCV	1,225	CAS^2	1,459
CDEF	1,848	FFC	1,309	FCAV	1,307	FFC	1,237
FDEF	1,777	*FATA1*	2,312	CSC	1,313	*FATA1*	1,695
POSSE	2,554	*FATA2*	2,195	FSC	1,255	*FATA2*	2,287
CSC	1,333	FAS	1,203			FAS	1,146
FSC	1,290						

b_i e $s(b_i)$ - estimativas do coeficiente e do seu desvio padrão para a variável *i*.

A análise de ambos os modelos construídos permite concluir que podem ser aplicados para os dados estudados, uma vez que cumprem, de um modo geral, os pressupostos analisados, pelo que a sua utilização para a previsão dos resultados dos jogos de futebol pode ser realizada.

5. PREVISÃO DOS RESULTADOS DOS JOGOS

Um dos objectivos dos modelos de regressão é, precisamente, a previsão da variável dependente a partir dos valores das variáveis independentes. Com base nos modelos desenvolvidos é possível cumprir este objectivo, calculando a estimativa do resultado de cada jogo ($\hat{Y}_i$), a partir do comportamento das equipas intervenientes: dados estatísticos, relevantes para o modelo, observados para cada jogo (X_{i1}, X_{i2}, …, $X_{i,p-1}$) e das estimativas dos coeficientes dos modelos de regressão (b_0, b_1, …, b_{p-1}), apresentadas nos quadros n.º 12 e 13.

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_{p-1} X_{i,p-1} \qquad i = 1, 2, ..., n \quad (16)$$

Estas estimativas, dos resultados dos jogos, utilizando os modelos desenvolvidos, podem ser, então, comparadas com os resultados efectivamente observados.

É também premente efectuar uma análise de sensibilidade aos resultados dos jogos previstos pelos modelos, para o que se calculam intervalos de confiança para as

estimativas do previsões, sendo utilizado um nível de confiança de 95%. Sendo X_h o vector coluna constituído pelos valores das variáveis independentes, para uma observação em particular, os limites inferior (LI) e superior (LS) do intervalo de confiança, a 95%, podem ser obtidos através da seguinte expressão.

$$\hat{Y}_i \pm t(0{,}025, n-p) \cdot \sqrt{MSE \cdot \left[X'_h \cdot (X'X)^{-1} \cdot X_h \right]} \qquad (17)$$

Sendo *t* o valor da distribuição *t* de *Student* para uma probabilidade de 0,025 e para *n-p* graus de liberdade e X'_h a matriz transposta de X_h.

Após o cálculo da previsão para o valor da variável dependente, bem como dos seus limites de confiança, que apresentam valores contínuos, importa transformá-los em valores que representam o resultado de um jogo de futebol, em termos de vitória, empate ou derrota de uma das equipas, objectivo atingido através de análise que se apresenta seguidamente.

Nas duas épocas estudadas, foram disputados 612 jogos, tendo ocorrido 304 vitórias da equipa que joga em casa (VC), 174 empates (E) e 134 vitórias das equipas que actuam fora (VF). Esta distribuição dos três resultados possíveis, para o período em análise: 49,7% VC, 28,4% E e 21,9% VF apresenta valores semelhantes aos verificados em outros períodos de tempo. Se calcularmos, por exemplo, a distribuição dos resultados para as seis épocas de futebol, em que o campeonato decorreu em condições semelhantes às actuais (com 18 equipas e três pontos atribuídos à vitória), desde 94/95 até 99/2000, verificam-se os seguintes valores: 50,3% VC, 26,4% E e 23,3% VF. Esta constância da distribuição de resultados consubstancia o passo seguinte: a elaboração de regras de decisão para definir os resultados dos jogos em função dos valores previstos pelos modelos para as variáveis dependentes utilizadas.

As regras de decisão, apresentadas no Quadro n.º 14, consistem em que:

- Para o Modelo 1 – VED, considera-se vitória da equipa da casa quando o valor previsto para a variável dependente é superior a 1,37, vitória da equipa que joga fora quando o valor previsto é inferior a 0,90 e empate nos restantes casos.
- No Modelo 2 – GMGS, a valores previstos, para a variável dependente, superiores a 0,50 corresponde uma vitória caseira, inferiores a –0,25 indicam vitória forasteira e aos restantes está associado o empate.

Deste modo, a distribuição dos três resultados possíveis, previstos pelos modelos desenvolvidos, é idêntica à distribuição observada para os 612 jogos em análise: as diferenças percentuais observadas, nas frequências relativas de cada resultado, são mínimas.

Quadro n.º 14. Regras de decisão para definir os resultados dos jogos ($\hat{Y}_i$)

	Frequência	Modelo 1 - VED		Modelo 2 - GMGS	
Resultado	Observada	Regra de decisão	Frequência	Regra de decisão	Frequência
VC	49,7%	$\hat{Y}_i > 1{,}37$	49,8%	$\hat{Y}_i > 0{,}50$	50,1%
E	28,4%	$0{,}90 < \hat{Y}_i < 1{,}37$	28,2%	$-0{,}25 < \hat{Y}_i < 0{,}50$	28,2%
VF	21,9%	$\hat{Y}_i < 0{,}90$	22,1%	$\hat{Y}_i < -0{,}25$	21,8%

As equipas intervenientes em cada jogo são representadas, em todos os quadros e gráficos seguintes, por uma abreviatura com três dígitos, de acordo com o Quadro n.º 15.

Quadro n.º 15 – Abreviaturas dos nomes das equipas

ACA	Académica	CAM	Campomaiorense	MAR	Marítimo
ALV	Alverca	CHA	Desp. Chaves	POR	Porto
BEI	Beira-mar	EST	Estrela Amadora	RIO	Rio Ave
BEL	Belenenses	FAR	Farense	SAL	Salgueiros
BEN	Benfica	GIL	Gil Vicente	SCL	Santa Clara
BOA	Boavista	GUI	Guimarães	SET	Vit. Setúbal
BRA	Sp. Braga	LEI	U. Leiria	SPO	Sporting

Estabelecidas as regras de decisão, torna-se possível efectuar a previsão do resultado de cada um dos jogos, utilizando os modelos de regressão. No Quadro n.º 16 são apresentadas as previsões calculadas com ambos os modelos, a título de exemplo, para os jogos entre os quatro primeiros classificados de cada uma das épocas estudadas e comparadas com os resultados efectivamente observados. Os jogos são apresentados por ordem cronológica.

Importa explicar que os resultados dos jogos (Res.) são representados pelos pontos atribuídos à equipa que joga em casa, cujos significados são, de acordo com a simbologia já utilizada, "3" – VC, "1" – E e "0" – VF.

Na primeira coluna é identificado o jogo em análise, bem como o resultado verificado, em termos de golos marcados, golos sofridos e pontos obtidos, relativos à equipa que actua em casa. Para os dois modelos desenvolvidos são calculadas as estimativas do valor da variável dependente e dos seus limites de confiança, a 95%, bem como os resultados correspondentes, de acordo com as regras de decisão estabelecidas e a análise da sua concordância (☑) ou não (☒) com o resultado efectivamente observado.

Analise-se, a título de exemplo, o último jogo apresentado da temporada de 1999/2000, que poderia decidir o título: Sporting-Benfica (0-1). Ambos os modelos preconizam uma vitória do Sporting, através da estimativa da variável dependente, resultado do jogo, a partir dos dados estatísticos. Apenas o limite de confiança inferior, para um grau de confiança de 95%, permite chegar à conclusão de que o

resultado poderia ser um empate, mas nenhum dos modelos estima o resultado que realmente se verificou, uma vitória do Benfica.

Quadro n.º 16 Resultados dos jogos: valores observados e previstos

Campeonato de 1998/1999

Valor Observado		Modelo 1 – VED						Modelo 2 - GMGS					
Jogo (GM-GS)	Res.	$\hat{Y}_i$	Res.	LI	Res.	LS	Res.	$\hat{Y}_i$	Res.	LI	Res.	LS	Res.
POR-BOA (0-2)	0	-0,44	0 ☑	-0,67	0 ☑	-0,21	0 ☑	-1,69	0 ☑	-2,06	0 ☑	-1,32	0 ☑
BOA-BEN (2-1)	3	0,03	0 ☒	-0,25	0 ☒	0,31	0 ☒	-1,86	0 ☒	-2,38	0 ☒	-1,34	0 ☒
BOA-SPO (2-2)	1	-0,08	0 ☒	-0,31	0 ☒	0,14	0 ☒	-2,03	0 ☒	-2,47	0 ☒	-1,59	0 ☒
POR-BEN (3-1)	3	2,25	3 ☑	2,03	3 ☑	2,48	3 ☑	2,61	3 ☑	2,21	3 ☑	3,02	3 ☑
POR-SPO (3-2)	3	1,38	3 ☑	1,18	1 ☒	1,58	3 ☑	0,73	3 ☑	0,37	1 ☒	1,09	3 ☑
SPO-BEN (1-2)	0	0,44	0 ☑	0,17	0 ☑	0,71	0 ☑	-0,19	1 ☒	-0,62	0 ☑	0,25	1 ☒
BOA-POR (0-0)	1	1,06	1 ☑	0,89	0 ☒	1,23	1 ☑	-0,36	0 ☒	-0,68	0 ☒	-0,09	1 ☑
BEN-BOA (0-3)	0	1,05	1 ☒	0,71	0 ☑	1,40	3 ☒	0,74	3 ☒	0,15	1 ☒	1,32	3 ☒
BEN-POR (1-1)	1	1,60	3 ☒	1,41	3 ☒	1,78	3 ☒	0,09	1 ☑	-0,24	1 ☑	0,42	1 ☑
SPO-BOA (1-1)	1	1,64	3 ☒	1,44	3 ☒	1,83	3 ☒	0,70	3 ☒	0,42	1 ☑	0,97	3 ☒
SPO-POR (1-1)	1	1,53	3 ☒	1,36	1 ☑	1,69	3 ☒	1,51	3 ☒	1,27	3 ☒	1,75	3 ☒
BEN-SPO (3-3)	1	1,17	1 ☑	0,89	0 ☒	1,44	3 ☒	0,19	1 ☑	-0,31	0 ☒	0,68	3 ☒

Campeonato de 1999/2000

Valor Observado		Modelo 1 - VED						Modelo 2 - GMGS					
Jogo (GM-GS)	Res.	$\hat{Y}_i$	Res.	LI	Res.	LS	Res.	$\hat{Y}_i$	Res.	LI	Res.	LS	Res.
BOA-POR (1-1)	1	1,20	1 ☑	1,07	1 ☑	1,34	1 ☑	0,22	1 ☑	-0,03	1 ☑	0,46	1 ☑
SPO-BOA (2-0)	3	1,80	3 ☑	1,62	3 ☑	1,99	3 ☑	1,30	3 ☑	0,97	3 ☑	1,64	3 ☑
BEN-BOA (1-1)	1	0,97	1 ☑	0,73	0 ☒	1,20	1 ☑	-0,15	1 ☑	-0,54	0 ☒	0,25	1 ☑
POR-SPO (3-0)	3	1,91	3 ☑	1,65	3 ☑	2,17	3 ☑	2,35	3 ☑	1,95	3 ☑	2,75	3 ☑
POR-BEN (2-0)	3	2,06	3 ☑	1,84	3 ☑	2,28	3 ☑	1,04	3 ☑	0,68	3 ☑	1,40	3 ☑
BEN-SPO (0-0)	1	1,14	1 ☑	0,81	0 ☒	1,47	3 ☒	0,68	3 ☒	0,08	1 ☑	1,28	3 ☒
POR-BOA (1-0)	3	1,55	3 ☑	1,41	3 ☑	1,70	3 ☑	1,13	3 ☑	0,89	3 ☑	1,37	3 ☑
BOA-SPO (0-1)	0	0,71	0 ☑	0,53	0 ☑	0,89	0 ☑	-0,29	0 ☑	-0,58	0 ☑	0,00	1 ☒
BOA-BEN (1-1)	1	0,58	0 ☒	0,40	0 ☒	0,76	0 ☒	-0,81	0 ☒	-1,12	0 ☒	-0,50	0 ☒
SPO-POR (2-0)	3	1,96	3 ☑	1,70	3 ☑	2,22	3 ☑	0,40	1 ☒	-0,08	1 ☒	0,87	3 ☑
BEN-POR (1-0)	3	1,15	1 ☒	0,98	1 ☒	1,31	1 ☒	0,21	1 ☒	-0,09	1 ☒	0,51	3 ☑
SPO-BEN (0-1)	0	1,59	3 ☒	1,36	1 ☒	1,82	3 ☒	0,58	3 ☒	0,18	1 ☒	0,99	3 ☒

Os modelos permitem efectuar as previsões da variável dependente, resultado do jogo, para todas as observações, à excepção de um deles, ACA-SAL (0-1), para o qual não estão disponíveis dados estatísticos, tal como já foi referido anteriormente.

Aplicando os modelos a todos os jogos, podem calcular-se as classificações finais das equipas estimadas pelos modelos, através do somatório de pontos conseguidos por cada equipa e compará-las com as classificações efectivamente verificadas, como pode observar-se no Quadro n.º 17.

Para uma análise mais detalhada, dividem-se os pontos totais em pontos obtidos em casa e fora e apresentam-se, para cada modelo, os limites inferiores e superiores da classificação final, tendo por base os limites de confiança, a 95%, para o pior e melhor resultado, respectivamente, de cada equipa, em cada jogo.

A título de informação adicional, verifica-se que o modelo 1 – VED estima resultados diferentes dos observados para 246 jogos e relativamente ao modelo 2 – GMGS, existem 256 jogos nestas condições. Os dois modelos originam previsões diferentes, entre ambos, em 140 jogos.

Quadro n.º 17. Classificações finais observadas e estimadas pelos modelos

Campeonato de 1999/2000

	Pontos em casa			Pontos fora			Total de pontos						
								Mod. 1-VED			Mod. 2-GMGS		
	Obs.	VED	GMGS	Obs.	VED	GMGS	Obs.	LI	Mod.	LS	LI	Mod.	LS
SPO	42	47	43	35	26	24	77	60	73	75	47	67	80
POR	49	45	45	24	17	20	73	55	62	71	55	65	72
BEN	44	38	42	25	20	24	69	43	58	67	59	66	74
BOA	33	22	23	22	20	19	55	33	42	54	33	42	52
GIL	35	34	33	18	15	12	53	39	49	58	38	45	51
MAR	31	39	37	19	25	20	50	47	64	71	46	57	68
GUI	38	33	35	10	12	15	48	38	45	61	38	50	59
EST	25	35	36	20	18	20	45	44	53	67	41	56	63
BRA	26	29	27	17	13	11	43	33	42	52	32	38	55
LEI	27	23	22	15	7	9	42	25	30	48	25	31	46
ALV	31	20	23	10	8	16	41	22	28	45	26	39	54
BEL	26	24	16	14	15	14	40	28	39	55	18	30	44
CAM	27	26	26	9	8	13	36	24	34	45	29	39	53
FAR	24	24	23	11	10	12	35	25	34	40	25	35	41
SAL	20	21	16	14	17	15	34	28	38	49	26	31	46
RIO	27	37	35	6	8	11	33	37	45	51	38	46	55
SET	21	31	25	12	17	16	33	37	48	59	32	41	49
SCL	22	36	30	9	8	14	31	33	44	52	37	44	58

Campeonato de 1998/1999

	Pontos em casa			Pontos fora			Total de pontos						
								Mod. 1-VED			Mod. 2-GMGS		
	Obs.	VED	GMGS	Obs.	VED	GMGS	Obs.	LI	Mod.	LS	LI	Mod.	LS
POR	48	46	48	31	25	25	79	62	71	74	61	73	77
BOA	42	35	32	29	30	29	71	49	65	72	36	61	74
BEN	38	36	39	27	26	25	65	56	62	75	49	64	72
SPO	40	37	42	23	25	18	63	48	62	71	53	60	75
SET	36	40	37	17	23	19	53	50	63	64	49	56	72
LEI	31	26	23	21	20	15	52	26	46	58	25	38	55
GUI	35	37	41	15	15	15	50	42	52	60	42	56	65
EST	32	35	40	13	15	11	45	38	50	54	34	51	61
BRA	28	35	47	14	16	12	42	39	51	64	44	59	71
MAR	26	23	26	15	10	12	41	25	33	48	22	38	50
FAR	26	24	27	13	15	15	39	32	39	55	29	42	55
SAL	28	18	24	10	9	10	38	22	27	36	22	34	45
CAM	26	24	23	11	11	11	37	27	35	50	21	34	45
ALV	25	18	21	10	14	14	35	24	32	45	26	35	45
RIO	20	24	18	15	17	13	35	29	41	51	23	31	50
BEI	24	27	24	9	22	17	33	34	49	67	33	41	56
CHA	19	20	22	6	8	7	25	20	28	33	20	29	44
ACA	14	15	19	7	15	21	21	28	30	49	24	40	48

De modo a sistematizar a análise de como a classificação estimada pelos modelos difere da observada, no Quadro n.º 18 indicam-se as alterações classificativas resultantes das estimativas obtidas pelos modelos, com a indicação de manutenção (=), subida (↗*n*) ou descida (↘*n*) de *n* posições na tabela.

Quadro n.º 18. Comparação das classificações estimadas com as observadas

Classificações da época de 1999/2000							Classificações da época de 1998/1999						
Observada		Mod. VED			Mod. GMGS			Observada		Mod. VED			Mod. GMGS	
SPO	77	SPO	73	=	SPO	67	=	POR	79	POR	71	=	POR	73 =
POR	73	MAR	64	↗4	BEN	66	↗1	BOA	71	BOA	65	=	BEN	64 ↗1
BEN	69	POR	62	↘1	POR	65	↘1	BEN	65	SET	63	↗2	BOA	61 ↘1
BOA	55	BEN	58	↘1	MAR	57	↗2	SPO	63	BEN	62	↘1	SPO	60 =
GIL	53	EST	53	↗3	EST	56	↗3	SET	53	SPO	62	↘1	BRA	59 ↗4
MAR	50	GIL	49	↘1	GUI	50	↗1	LEI	52	GUI	52	↗1	SET	56 ↘1
GUI	48	SET	48	↗10	RIO	46	↗9	GUI	50	BRA	51	↗2	GUI	56 =
EST	45	GUI	45	↘1	GIL	45	↘3	EST	45	EST	50	=	EST	51 =
BRA	43	RIO	45	↗7	SCL	44	↗9	BRA	42	BEI	49	↗7	FAR	42 ↗2
LEI	42	SCL	44	↗8	BOA	42	↘6	MAR	41	LEI	46	↘4	BEI	41 ↗6
ALV	41	BOA	42	↘7	SET	41	↗6	FAR	39	RIO	41	↗4	ACA	40 ↗7
BEL	40	BRA	42	↘3	ALV	39	↘1	SAL	38	FAR	39	↘1	LEI	38 ↘6
CAM	36	BEL	39	↘1	CAM	39	=	CAM	37	CAM	35	=	MAR	38 ↘3
FAR	35	SAL	38	↗1	BRA	38	↘5	ALV	35	MAR	33	↘4	ALV	35 =
SAL	34	CAM	34	↘2	FAR	35	↘1	RIO	35	ALV	32	↘1	SAL	34 ↘3
RIO	33	FAR	34	↘2	LEI	31	↘6	BEI	33	ACA	30	↗2	CAM	34 ↘3
SET	33	LEI	30	↘7	SAL	31	↘2	CHA	25	CHA	28	=	RIO	31 ↘2
SCL	31	ALV	28	↘7	BEL	30	↘6	ACA	21	SAL	27	↘6	CHA	29 ↘1

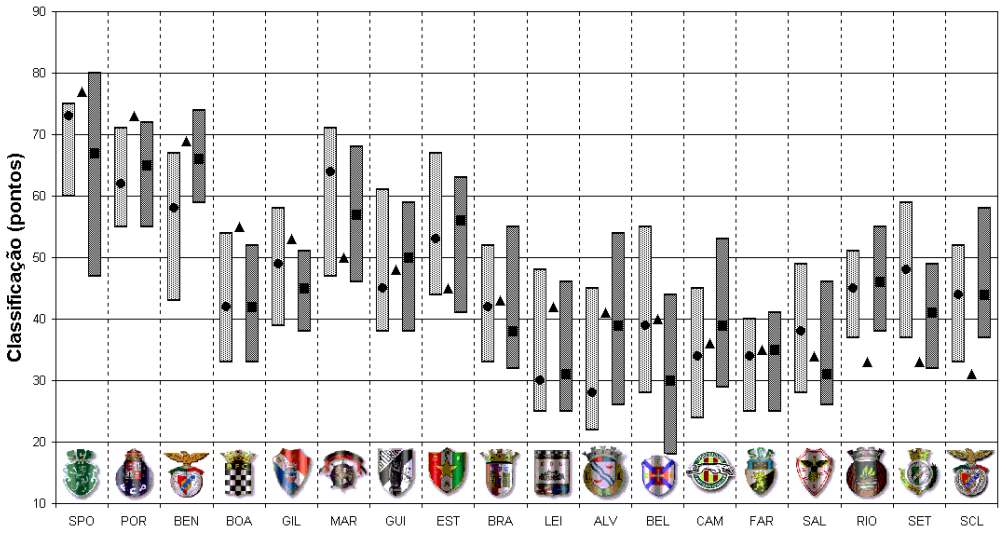
Ambos os modelos confirmam, através dos seus resultados, os vencedores de ambos os campeonatos e indicam subidas significativas, em 1999/2000, para o Marítimo, Est. Amadora, Setúbal, Rio Ave e Santa Clara, que obtiveram uma classificação inferior à estimada pelos modelos e descidas significativas para o Boavista, Braga, Leiria, Alverca e Belenenses, que obtiveram resultados acima dos estimados. Na temporada anterior, em 1998/99, as subidas mais significativas estimadas pelos modelos são do Braga, Beira-mar e Académica, sendo as descidas relevantes para o Leiria, Marítimo e Salgueiros.

No entanto, pode colocar-se uma questão: se a regra de decisão fosse outra ou se o modelo tivesse uma estrutura diferente, qual a influência nos resultados previstos pelo modelo? Uma abordagem à resposta a esta questão pode ser obtida através de uma análise de sensibilidade aos resultados, utilizando os limites dos intervalos de confiança, a 95%, para as previsões pelos modelos, dos valores estimados das classificações finais, para cada temporada, já apresentados no Quadro n.º 17 e ilustrados pelo Gráfico n.º 11. Estes limites dão uma ideia da variação que pode ocorrer nas classificações finais estimadas pelos modelo, por variações pontuais do resultado estimado de cada jogo.

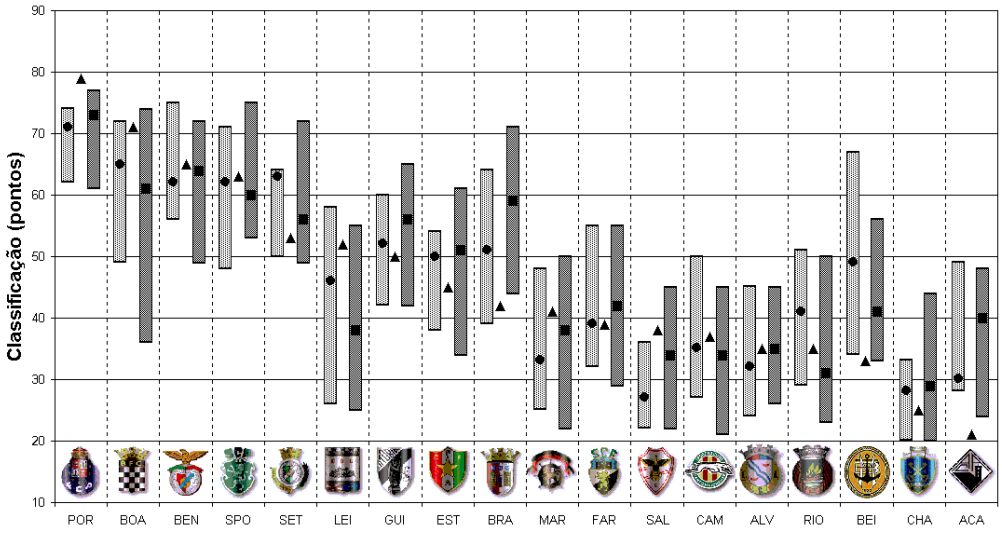
A análise deste gráfico permite visualizar, facilmente, a comparação entre os valores observados para a pontuação obtida na classificação final, por cada equipa e os mesmos valores previstos pelos modelos de regressão linear, já interpretada anteriormente, para as equipas com maiores diferenças. Os intervalos de confiança, previstos por ambos os modelos, apresentam alguma concordância entre si e incluem, na maioria dos casos, os valores observados das classificações.

Gráfico n.º 11. Análise de sensibilidade à classificação final estimada

Época de 1999/2000



Época de 1998/1999



Classificações: ▲ – observada.

● – prevista pelo Modelo 1 – VED – e ▯ - respectivo intervalo de confiança, a 95%.

■ – prevista pelo Modelo 2 – GMGS – e ▯ - respectivo intervalo de confiança, a 95%.

6. CONSIDERAÇÕES FINAIS

Os procedimentos que dão origem aos modelos de regressão permitem exemplificar o desenvolvimento e aplicação de um modelo de regressão para estudos observacionais explicativos, ilustrando os vários passos que vão sendo tomados.

Os modelos desenvolvidos, após a eliminação de *outliers*, respeitam os pressupostos exigidos para modelos de regressão, apresentando a sua aplicação, aos resultados dos jogos de futebol do campeonato nacional, alguma sustentabilidade. No entanto, partindo de outros prismas, no que diz respeito à análise dos resultados dos jogos de futebol, em função das estatísticas disponíveis, poderão ser elaborados outros modelos, com variações mais ou menos relevantes, mas com resultados que também podem ser significativos.

Será importante verificar a aplicação dos modelos a um conjunto superior de observações e as alterações por elas provocadas nas estimativas dos coeficientes dos modelos e até nas variáveis que os integram, através do estudo de futuras épocas do campeonato nacional de futebol.

É de referir que os limites dos intervalos de confiança, a 95%, provocam alterações significativas nos resultados dos modelos, devido a estes apresentarem uma variabilidade considerável, quanto à previsão dos resultados, facto que se deve, certamente à aleatoriedade associada a um desafio de futebol, que não passa de um

jogo: o resultados de muitos partidos pode variar por pequenas alterações nos factores analisados e, certamente, devido a inúmeros factores não estudados.

Existe um grande número de variáveis que influenciam a variável dependente em estudo e que não fazem parte dos modelos desenvolvidos, como sejam o comportamento do árbitro, a estabilidade psicológica, a motivação, o orçamento de cada equipa, os cantos, as grandes penalidades convertidas e falhadas, os momentos em que são marcados os golos, entre muitas outras.

As variáveis incluídas neste modelo são de fácil análise e percepção, pelo que os resultados podem ser facilmente entendidos por um leitor menos familiarizado com estes métodos estatísticos. A extrapolação desta aplicação para o estudo de outros casos é assim de fácil desenvolvimento.

BIBLIOGRAFIA

BARNETT, V.; LEWIS, T., "*Outliers in Statistical Data"*, 3rd Edition, John Wiley & Sons, New York, 1994.

BENNETT, Jay (Editor), "*Statistics in Sport*", *Arnold Applications of Statistics Series*, Oxford University Press, London, 1998.

DRAPER, N.R.; SMITH, H., "*Applied Regression Analysis*", 2.nd Edition, John Wiley & Sons, New York, 1981.

FREEDMAN, D.A., "*A Note on Screening Regression Equations*", *The American Statistician*, 37, pp. 152-155, Alexandria, 1982.

GRAY, Philip K., "*Testing market efficiency: Evidence from the NFL sports betting market*", *The Journal of Finance*, 52 (4), pp. 1725-1737, Cambridge, 1997.

LADANY, S. P. e MACHOL, R. E. (Editores), "*Optimal Strategies in Sports*", North-Holland, New York, 1977.

INFORDESPORTO, *http://www.infordesporto.pt*.

KAHANE, Leo, "*Team roster turnover and attendance in major league baseball*", *Applied Economics*, 29 (4), p. 425, London, 1997.

MACHOL, R. E., LADANY, S. P. e MORRISON, D. G. (Editores), "*Management Science in Sports*", North-Holland, New York, 1976.

MILLER, A.J., "*Subset Selection in Regression"*, Ed. Chapman and Hall, London, 1990.

NETER, J.; KUTNER, M.H.; NACHTSHEIM, C.J.; WASSERMAN, W., "*Applied Linear Statistical Models"*, *Fourth Edition*, Irwin, Chicago, 1996.

PESTANA, M. Helena; GAGEIRO, João N., "*Análise de dados para Ciências Sociais - A complementaridade do SPSS"*, Edição Revista, Edições Sílabo, Lisboa, 2000.

POPE, P.T.; WEBSTER, J.T., "*The Use of an F-Statistic in Stepwise Regression*", *Technometrics*, 14, pp. 327-340, 1972.

ROUSSEEUW, P.J.; LEROY, A.M., "*Robust Regression and Outlier Detection"*, Ed. John Wiley & Sons, New York, 1987.

SMITH, Tyler, "*Can the NCAA Basketball tournament seeding be used to predict margin of victory*?", *The American Statistician*, 53 (2), pp. 94-98, Alexandria, 1999.

SZYMANSKI, S., SMITH, R., "*The English football industry: profit, performance and industrial structure*", *International Reviews of Applied Economics*, 11, pp. 135-154, 1997.