Gestão e Desenvolvimento, 8 (1999), 141-170

DESENVOLVIMENTO E APLICAÇÃO DE MODELOS ESTATÍSTICOS. O ESTUDO DE UM CASO: O CAMPEONATO NACIONAL DE FUTEBOL DE 1998/99

Paulo Almeida Pereira*

Pretende desenvolver-se e aplicar um modelo estatístico de regressão a um estudo observacional explicativo, neste caso, uma competição desportiva: o Campeonato Nacional de Futebol da I Divisão de 1998/99. O modelo serve os objectivos de descrição, controlo e previsão dos resultados dos jogos de futebol (variável dependente) a partir de dados estatísticos para o comportamento das equipas em análise (variáveis independentes). O modelo desenvolvido permite realizar, de um modo significante, uma estimativa e respectivos intervalos de confiança da classificação final e a sua comparação com a realmente observada.

Os procedimentos utilizados, no desenvolvimento do modelo estatístico para a análise deste caso, permitem exemplificar a aplicação de um modelo de regressão para estudos observacionais explicativos, ilustrando os vários passos seguidos, podendo servir de base a outros estudos semelhantes.

1. INTRODUÇÃO À ANÁLISE DE REGRESSÃO

A análise de regressão é um método estatístico que utiliza a relação entre duas ou mais variáveis quantitativas, de modo a que uma variável possa ser estimada a partir de outra, ou de outras. Foi desenvolvida por *Sir* Francis Galton, no final do século XIX, quando estudou a relação entre as alturas de

* Instituto Universitário de Desenvolvimento e Promoção Social - Pólo de Viseu do Centro Regional das Beiras da Universidade Católica Portuguesa.

pais e filhos, concluindo que as alturas dos filhos, tanto de pais altos como baixos, *regrediam* para um valor médio. Por este motivo ainda hoje o termo regressão persiste para designar relações estatísticas entre variáveis.

O Modelo de Regressão Linear, discutido por alguns autores (Neter *et al.*, 1996; Draper e Smith, 1981, por exemplo), é um meio formal de exprimir a relação estatística entre uma ou mais variáveis independentes, de previsão ou explicativas e uma variável dependente, de resposta ou explicada. Deste modo, um modelo de regressão exprime a tendência da variável dependente para variar com as variáveis independentes de um modo sistemático. Quando existe apenas uma variável independente, estamos perante um modelo de regressão linear simples, quando há várias, a regressão linear é múltipla.

Os modelos de regressão têm as mais diversas aplicações nas Ciências Sociais e Comportamentais, na Gestão, nas Ciências Biológicas e em muitas outras áreas. Constatou-se que publicações recentes (Smith, 1999; Gray, 1997, Kahane, 1997, Hamilton, 1997) desenvolvem aplicações destes modelos para competições desportivas.

Com o desenvolvimento das tecnologias e sistemas de informação, temos presentemente uma grande disponibilidade de dados, nomeadamente estatísticos, "à espera de serem trabalhados". A Infordesporto[1] (Infordesporto, *site* na *Internet*) é uma dessas bases de dados, acessível a partir da Internet, que disponibiliza uma informação completa sobre os jogos de futebol do campeonato nacional da primeira divisão.

Conjugando a aplicação dos modelos de regressão a competições desportivas e a informação disponível, surgiu a ideia de realizar um estudo, utilizando a análise de regressão, para desenvolver um modelo de regressão múltipla que relacionasse os dados estatísticos para as equipas integrantes do campeonato nacional de futebol da I divisão, de 1998/99, com os resultados por elas obtidos. Esta análise é bastante interessante, a nível científico, pela vasta informação à qual se pode aplicar e desenvolver um modelo de regressão, permitindo exemplificar e detalhar os passos conducentes à validação e aplicação do 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 existem alguns textos de apoio em português (Pestana e Gageiro, 1998; Bryman e Cramer, 1993).

2. DADOS ESTATÍSTICOS

A Infordesporto disponibiliza uma informação estatística completa sobre os jogos do campeonato nacional de futebol da I divisão, de 1998/99. Para aplicação e desenvolvimento de um modelo de regressão é necessário, em primeiro lugar, sistematizar os dados disponíveis.

O objectivo deste estudo é relacionar o resultado de um jogo de futebol (vitória, empate ou derrota) com o comportamento dos intervenientes (as duas equipas). A variável dependente em análise é o resultado do jogo, para o qual se utiliza uma escala ordinal, em que 2 = vitória, 1 = empate e 0 = derrota. A escala é assim definida, para garantir a simetria da variável. Esta variável é função de várias variáveis independentes: a informação estatística disponível para as duas equipas intervenientes no jogo. As abreviaturas que representam estas variáveis independentes, com determinado valor para cada jogo, são catalogadas no Quadro I, agrupadas em categorias.

Quadro I
VARIÁVEIS INDEPENDENTES

	Gerais		Guarda-redes:
RELV	estado do relvado[2]	DC	defesas completas
ESPEC	nº de espectadores	DI	defesas incompletas
TJOGO	tempo jogado (% do 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
COM	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
CA_2	duplos cartões amarelos	AS	assistências
CV	cartões vermelhos	GM	golos marcados
		GS	golos sofridos

Torna-se premente, desde já, dar uma explicação que será importante para o desenvolvimento do modelo de regressão. Em cada jogo, são recolhidos dados sobre cada uma destas variáveis que, com excepção das variáveis gerais, são duplicadas, uma vez que as informações sobre cada variável existem para

as duas equipas em jogo. Deste modo, em cada jogo será recolhida informação sobre o desempenho da equipa em análise e do adversário, uma vez que o resultado será, como é óbvio, função do comportamento das duas equipas em confronto. Assim, no desenvolvimento do modelo de regressão, as variáveis terão o prefixo "P", quando dizem respeito à equipa em análise e o prefixo "A", quando dizem respeito ao adversário. Este método origina que, em cada jogo, haja duas observações, uma correspondente à equipa que joga em casa e outra correspondente à equipa que joga fora. Cada uma delas inclui dados sobre o comportamento do adversário. Por este motivo o número total de observações (612) é o dobro dos jogos disputados (306).

A partir dos dados, podem calcular-se estatísticas descritivas que representam, de uma forma compreensível, a informação neles contida. No Quadro II são listadas algumas estatísticas: média, desvio padrão, valores mínimo e máximo e observações[3], para todas as variáveis.

Quadro II

MÉDIA ($\overline{x}$), DESVIO PADRÃO (*s*), MÍNIMOS (*min*), MÁXIMOS (*max*) E OBSERVAÇÕES VÁLIDAS (*val*) DE CADA VARIÁVEL

Variável	$\overline{x}$	*s*	*min*	*max*	*val*	Variável	$\overline{x}$	*s*	*min*	*max*	*val*
RELV	3,14	0,99	1	5	612	DI	1,47	1,38	0	9	610
ESPEC	8478	9635	700	70000	608	SC	5,70	3,23	0	20	610
DEF	3,68	1,18	0,57	9,70	604	SI	1,46	1,40	0	9	610
MCD	8,70	1,85	4,48	15,93	604	FC	21,07	6,05	5	42	610
MCO	9,21	2,38	3,25	18,17	604	FS	20,45	6,03	4	42	610
ATA	5,07	1,96	0,80	12,07	604	P	22,83	6,88	4	50	610
TOT	26,65	4,11	16,15	37,71	604	R	64,34	11,95	37	100	610
POSSE	50%	7%	32%	68%	604	AT	38,13	11,25	11	77	610
TJOGO	55%	4%	39%	66%	606	CR	15,82	8,00	1	43	610
CA	3,07	1,68	0	8	610	RE	14,41	5,83	0	31	610
CA_2	0,15	0,38	0	2	612	AS	1,44	1,53	0	12	610
CV	0,09	0,31	0	2	610	GM	1,32	1,24	0	7	612
DC	2,53	1,78	0	11	610	GS	1,32	1,24	0	7	612

Deve salientar-se também que o conjunto de dados disponíveis permite a apresentação de estatísticas que caracterizam o comportamento das variáveis em análise para todas as equipas e para todos os jogadores intervenientes no campeonato nacional de 1998/99. No Quadro seguinte apresentam-se alguns resultados, a título de exemplo, para os cinco valores máximos ou mínimos de algumas das variáveis em análise, para jogadores e equipas, ao longo dos 306 jogos do campeonato. Os resultados apresentados no Quadro III são valores

médios por 90 minutos jogados para os jogadores, com a participação num mínimo de 10 jogos (J) e valores médios por jogo para as equipas.

Quadro III

VALORES MÉDIOS (MÁXIMOS E MÍNIMOS) DE ALGUMAS DAS VARIÁVEIS EM ANÁLISE, PARA JOGADORES E EQUIPAS

Estatísticas para jogadores (médias por 90 minutos jogados)

equipa	nº	jogador	J	GS	equipa	nº	jogador	J	GS
CHA	12	Orlando	10	2,13	POR	99	Vítor Baía	16	0,57
CHA	1	Arteaga	25	2,03	BEN	12	Ovchinnikov	14	0,67
CAM	1	Paulo Sérgio	15	1,92	POR	1	Rui Correia	11	0,73
ACA	12	Pedro Roma	25	1,76	BOA	12	William	29	0,84
SAL	12	Jorge Silva	34	1,62	SPO	1	Tiago Ferreira	28	0,86
equipa	nº	jogador	J	T^4	equipa	nº	jogador	J	AT
BEI	3	Gila	34	3060	POR	21	Capucho	33	14,0
SAL	12	Jorge Silva	34	3060	SPO	13	Krpan	27	12,7
SAL	7	Abílio	34	3034	ACA	14	Luís Filipe	19	12,6
EST	22	Rebelo	34	3031	RIO	22	Gama	34	12,2
SET	4	Quim	34	2984	SPO	10	Edmilson	24	11,8
equipa	nº	jogador	J	RE	equipa	nº	jogador	J	CR
EST	6	Assis	12	6,40	POR	21	Capucho	33	6,67
BRA	7	Formoso	21	5,77	POR	11	Drulovic	32	6,22
LEI	10	Dinda	32	5,30	GUI	17	Riva	30	5,94
LEI	27	Konadu	12	5,29	ACA	14	Luís Filipe	19	5,91
POR	16	Jardel	32	5,23	BEN	7	Poborsky	27	5,69
equipa	nº	jogador	J	AS	equipa	nº	jogador	J	GM
BEN	17	Kandaurov	22	0,94	BOA	26	Atelkin	10	1,20
POR	21	Capucho	33	0,86	POR	16	Jardel	32	1,18
POR	11	Drulovic	32	0,83	POR	28	Quinzinho	12	1,14
SAL	29	Deco	12	0,80	LEI	27	Konadu	12	0,96
SPO	8	Pedro Barbosa	18	0,73	BEN	21	Nuno Gomes	34	0,93

Estatísticas para equipas (médias por jogo)

		RELV		ESPEC		ESPEC			
		equipa	casa	equipa	casa	equipa	fora		
		POR	4,06	POR	27764	BEN	12676		
		RIO	4,06	BEN	27235	SPO	12617		
		SPO	3,88	SPO	20794	POR	11794		
		ALV	3,38	GUI	9441	BOA	11312		
		SET	3,35	BRA	8353	ACA	9111		
equipa	PAT	equipa	PRE	equipa	PCR	equipa	PAS	equipa	GM
SPO	50,08	SPO	19,03	POR	24,12	POR	3,06	POR	2,50
BEN	47,18	POR	17,62	SPO	24,09	BEN	2,56	BEN	2,09
POR	47,03	BEN	17,09	BEN	20,97	SPO	1,94	SPO	1,88
GUI	41,06	GUI	16,56	GUI	17,56	FAR	1,79	BOA	1,68
BOA	38,88	BOA	15,71	SAL	17,06	SET	1,71	GUI	1,59
equipa	AAT	equipa	ARE	equipa	ACR	equipa	AAS	equipa	GS
SPO	27,94	SPO	8,79	SPO	8,24	SPO	0,85	POR	0,76
BRA	32,12	BOA	11,15	BOA	12,26	POR	1,09	LEI	0,85
BOA	33,06	POR	11,79	BRA	12,56	EST	1,15	BOA	0,85
POR	33,21	BRA	12,74	POR	12,88	BEN	1,18	BEN	0,85
GUI	35,03	BEM	13,35	LEI	13,15	LEI	1,21	SPO	0,94

As equipas são designadas por abreviaturas, como se indica de seguida:

```
ACA - Académica  ALV - Alverca  BEI - Beira - mar     BEN - Benfica
BOA - Boavista   BRA - Braga    CAM - Campomaiorense  CHA - Chaves
EST - Estrela    FAR - Farense  GUI - Guimarães       LEI - Leiria
MAR - Marítimo   POR - Porto    RIO - Rio Ave         SAL - Salgueiros
SET - Setúbal    SPO - Sporting
```

Analisando as estatísticas para equipas, a variável golos marcados (GM) parece estar relacionada com as variáveis ataques (PAT), cruzamentos (PCR), remates (PRE) e assistências (PAS) da equipa em análise. Do mesmo modo, a variável golos sofridos (GS) poderá ser função das variáveis ataques (AAT), cruzamentos (ACR), remates (ARE) e assistências (AAS) da equipa adversária.

O Modelo de Regressão, como explicado anteriormente, relaciona estatisticamente uma variável dependente com uma ou mais variáveis independentes, exprimindo a tendência da variável dependente para variar com as variáveis independentes de um modo sistemático. Uma possibilidade de análise residiria no estudo das relações estatísticas entre as variáveis referidas no parágrafo anterior.

No entanto, o modelo aqui desenvolvido partirá de outros pressupostos, como veremos de seguida e será diferente dos exemplos dados.

3. CONSTRUÇÃO DE MODELOS DE REGRESSÃO

A hipótese formulada neste estudo consiste em estabelecer uma relação estatística entre o resultado de um jogo de futebol do campeonato nacional de futebol de 1998/99 e as estatísticas disponíveis para variáveis que estão relacionadas com o comportamento das equipas intervenientes nesse jogo. Essa relação estatística será estudada, utilizando a análise de regressão, de modo a servir três objectivos principais:

- descrição, através do desenvolvimento de um modelo válido e utilizável para descrever a relação entre as variáveis;
- controlo, que permite verificar se os resultados efectivamente obtidos são, ou não, diferentes dos previstos pelo modelo;
- previsão de resultados a partir da definição de novos valores para as variáveis dependentes.

3.1. Formulação do modelo de regressão

A fórmula geral do modelo é a seguinte:

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

A esta equação dá-se o nome de curva de regressão da população, onde:

- n é o número de observações;
- Y_i é a variável dependente. O i representa a observação i em n.
- X_{p-1} são as variáveis independentes ou explicativas. No modelo há $p-1$ variáveis explicativas e p parâmetros;
- β_k são os parâmetros do modelo. O parâmetro β_k indica-nos a variação média do valor esperado de Y, com o aumento de uma unidade de X_k, quando todas as outras variáveis explicativas no modelo permanecem constantes:
- ε_i é o termo aleatório (termo de erro, termo de perturbação, variável aleatória residual ou resíduo), representando todas as variáveis com poder explicativo sobre a variável dependente que foram omitidas pelo modelo. Este termo aleatório vai seguir uma distribuição de probabilidade com média nula e com uma determinada variância.

Um dos princípios básicos do modelo consiste em que uma componente da variação na variável dependente é explicada pelas variáveis independentes incluídas no modelo, que é medida pelo coeficiente de determinação (r^2), mas também existe uma componente que reflecte a nossa ignorância referente a factores explicativos não contemplados.

3.2. Selecção das variáveis independentes

Recolha dos dados

A análise dos dados em causa é um estudo observacional explicativo, em que a informação sobre as variáveis independentes são dados secundários. Uma regra empírica para situações deste tipo é que devem existir 6 a 10 observações por cada variável independente.

Investigações preliminares

Depois de recolhidos os dados, dá-se início ao processo formal de desenvolvimento do modelo, identificando a forma funcional das variáveis independentes, uma vez que a sua relação com a variável dependente pode não ser linear e analisando interacções a incluir no modelo.

Redução do número de variáveis independentes

Num estudo observacional explicativo o número de variáveis independentes é muito grande, existindo normalmente correlação entre elas. Deve reduzir-se o número de variáveis a incluir no modelo final (Miller, 1990), pelas razões que se apresentam de seguida.

A primeira, que é óbvia, traduz-se na dificuldade de trabalhar e compreender um modelo com muitas variáveis. Além disso, muitas delas apresentam forte correlação entre si, o que provoca aumento na variabilidade associada aos coeficientes do modelo diminuindo as capacidades descritivas e de controlo do modelo, uma vez que os valores estimados da variável dependente apresentam grande variância. As capacidades preditivas do modelo são também diminuídas.

Utilizam-se procedimentos para seleccionar e testar subgrupos de variáveis independentes importantes, com o objectivo de desenvolver um modelo com um menor número de variáveis.

3.3. Diagnósticos, medidas correctoras e validação

Depois de definir um ou vários modelos apropriados para os dados em análise, é necessário validá-los, uma vez que os modelos de regressão têm de cumprir os seguintes pressupostos:

- Os termos de erro aleatório (ε_i) seguem uma distribuição normal, com valor esperado nulo e variância constante (homocedasticidade);
- As observações da variável dependente (Y) são independentes umas das outras;

• Os termos de erro, resíduos ou variáveis aleatórias residuais são independentes entre si, não estando auto-correlacionados, tendo, portanto, covariância nula. Não há também correlação entre eles e as variáveis independentes (X_i).

Através da realização de testes estatísticos, que incluem análise gráfica de resíduos, estudo da multicolinearidade (correlação entre variáveis independentes), identificação de *outliers* (casos extremos influentes), análise da homocedasticidade (variância constante dos termos de erro) e medida da auto-correlação aos vários modelos anteriormente encontrados, procede-se à sua validação.

Torna-se importante referir o facto de que o modelo tem determinado alcance, restringido ao intervalo ou região de valores das variáveis independentes, determinado a partir dos dados disponíveis.

4. DESENVOLVIMENTO DO MODELO DE REGRESSÃO PARA O CASO EM ANÁLISE

O objectivo é descrever e controlar a variável dependente (Y), resultado de um jogo de futebol, através da construção de um modelo que a relacione com as 43 variáveis independentes em análise.

A variável dependente, como já referimos anteriormente, é o resultado de um jogo de futebol do campeonato nacional de futebol de 1998/99, para a equipa em análise, em que 2 = vitória, 1 = empate e 0 = derrota, de forma a quantificá-lo de forma simétrica. Neste caso, existem 43 variáveis independentes cuja relação com a variável dependente pode ser estudada. Essas 43 variáveis, reportando-nos ao Quadro I, são as três variáveis gerais, as restantes variáveis, duplicadas, para a equipa em análise (prefixo "P") e para o adversário (prefixo "A"), com excepção da variável POSSE, que só é definida para a equipa em análise. As variáveis GM e GS não são incluídas no modelo, uma vez que a sua relação com as restantes é óbvia, o que originaria correlação entre variáveis, que contraria um dos pressupostos do modelo.

Estão em análise 612 observações, valor 14 vezes superior ao número de variáveis em análise. São conhecidos os valores para todas as variáveis em estudo, excepto os casos em que não estão disponíveis alguns dados (*missing values*), pelo facto dos seus valores serem desconhecidos. Nesta situação de dados indisponíveis são utilizados na construção do modelo, para algumas variáveis, os valores médios[5] 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.

Depois de recolhidos e preparados os dados, o primeiro passo é identificar a forma funcional como as variáveis independentes se relacionam com a variável dependente. Foi estudada, para todas as variáveis, a influência das seguintes formas funcionais: linear, logarítmica, inversa, exponencial, quadrática e cúbica. Utilizou-se como critério para identificar a melhor forma funcional a maximização do coeficiente de determinação do modelo, sem afectar o nível de significância das variáveis. Deste modo, no Quadro IV identificam-se as formas funcionais não lineares da relação entre a variável dependente e as variáveis independentes (X_i), que poderão melhorar o modelo. Para as restantes variáveis independentes a forma funcional que melhor se adequa ao modelo é a relação linear.

Quadro IV
FORMAS FUNCIONAIS NÃO-LINEARES DE RELAÇÃO ENTRE A VARIÁVEL DEPENDENTE E AS VARIÁVEIS INDEPENDENTES

X_i	Forma funcional	X_i	Forma funcional	X_i	Forma funcional
AAS	QUA	ASC	QUA	PFC	INV
AATA	CUB	PAS	QUA	PFS	INV
ACR	CUB	PCA	CUB	PR	QUA
ADI	QUA	PCR	QUA	PRE	CUB
ARE	QUA	PDI	QUA	PSC	QUA

QUA - Quadrática CUB - Cúbica INV – Inversa

De seguida passamos à redução do número de variáveis independentes. Para a selecção de quais as variáveis a incluir no modelo, a situação ideal seria escolher, através de algum critério (como a maximização do coeficiente de determinação ajustado), o melhor modelo de todos os possíveis. Com 43 variáveis (sem incluir as relações não lineares) existem $8{,}8\times10^{12}$ modelos possíveis (com conjuntos ou subconjuntos das variáveis independentes), pelo que é impraticável testar qual o modelo com melhor comportamento.

Nestas condições torna-se bastante útil um procedimento que desenvolva os melhores subconjuntos de variáveis sequencialmente. O procedimento mais utilizado é o *Forward Stepwise* que, essencialmente, desenvolve uma sequência de modelos de regressão, adicionando ou retirando em cada passo uma variável independente, recorrendo ao cálculo de uma estatística F^*, que segue uma distribuição *F de Snedecor.*

4.1. Modelo #1

De modo a construir um modelo com um número reduzido de variáveis independentes, exigiu-se para a estatística F^* um nível de significância de 51% para adicionar uma variável e de 50% para remover uma variável. A escolha destes valores é absolutamente empírica (Freedman, 1983; Pope e Webster, 1972), uma vez que não apresentam um significado probabilístico preciso.

Desenvolveu-se assim um primeiro modelo de regressão, #1, para explicar as relações estatísticas entre as variáveis em análise, para as 610 observações disponíveis. No Quadro V apresentam-se as variáveis que foram seleccionadas para integrar o modelo pelo método *forward stepwise*, bem como alguns resultados significativos para este modelo.

Quadro V
RESULTADOS PARA O MODELO DE REGRESSÃO #1

$r^2 = 0{,}4281$		g.l.	*SS*	*MS*
Desvio padrão = 0,6440	Regressão	27	180,64	6,6903
F = 16,13247 ⇒ Significância F =0,00	Resíduos	582	241,36	0,4147

r^2 - coeficiente de determinação
g.l. - graus de liberdade
SS - somatório dos quadrados
MS - média do somatório dos quadrados

Variável *i*	B_i	$s(b_i)$	Sig. *t*	Variável *i*	b_i	$s(b_i)$	Sig. *t*
AAS	-0,2076	0,0355	0,00	PAS^2	-0,0126	0,0056	0,02
AAS^2	0,0127	0,0055	0,02	PATA	0,0516	0,0272	0,06
AATA	-0,0515	0,0276	0,06	PCA_2	-0,1409	0,0736	0,06
ACA	0,0208	0,0175	0,23	PCR^2	$-1{,}28\cdot10^{-4}$	$1{,}26\cdot10^{-4}$	0,31
ACR^2	$1{,}37\cdot10^{-4}$	$1{,}27\cdot10^{-4}$	0,28	PCV	-0,2186	0,0882	0,01
ACV	0,2037	0,0880	0,02	PDC	0,0336	0,0175	0,05
ADC	-0,0308	0,0177	0,08	PDEF	0,0773	0,0313	0,01
ADEF	-0,0729	0,0312	0,02	PFS	-0,0095	0,0050	0,06
AFS	0,0099	0,0049	0,04	PP	0,0036	0,0045	0,42
AP	-0,0047	0,0045	0,30	PPOSSE	-3,1116	0,9065	0,00
AR	-0,0154	0,0036	0,00	PR	0,0161	0,0036	0,00
ARE	-0,0198	0,0075	0,01	PRE	0,0196	0,0076	0,01
ASC	-0,0193	0,0097	0,05	PSC	0,0204	0,0097	0,04
PAS	0,2079	0,0357	0,00	Constante	2,4572	0,5620	0,00

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 coeficiente de determinação indica-nos que apenas 42,81% da variação

que ocorre na variável dependente (resultado do jogo) é explicada pelas 25 variáveis incluídas no modelo. O teste *F* à significância global do modelo (que testa a hipótese de todos os coeficientes do modelo apresentarem um valor nulo) apresenta uma significância nula, que o valida. São apresentados também outros resultados estatísticos importantes, como os somatórios e médias dos quadrados da regressão e dos resíduos.

Indicam-se as estimativas dos coeficientes associados a cada variável do modelo, cujo significado se exemplifica para a variável AAS: a variação de uma unidade na variável AAS provoca uma variação média esperada negativa de -0,2076 na variável independente, ou seja, o aumento de uma assistência realizada pelo adversário provoca uma diminuição média esperada de 0,2076 no resultado do jogo. A estimativa da constante do modelo representa a influência de variáveis não incluídas no modelo. A estimativa do desvio padrão de cada variável representa a variabilidade associada à estimativa do coeficiente dessa mesma variável.

A significância do teste *t* de *Student* para cada variável indica-nos a probabilidade dessa variável tomar um valor nulo no modelo desenvolvido e, portanto, não ser significante para o modelo. Existem cinco variáveis (ACA, ACR^2, AP, PCR^2 e PP) para as quais o valor do teste *t* apresenta resultados demasiado elevados, que serão corrigidos posteriormente.

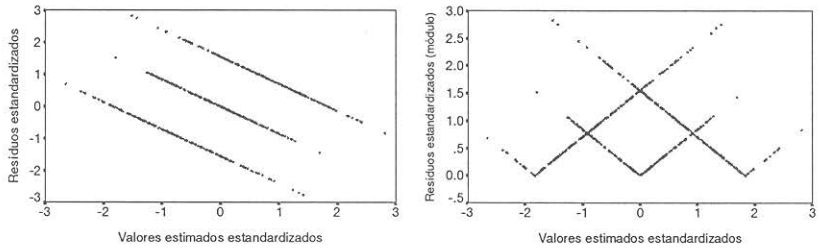
É, de seguida, imperioso analisar os pressupostos do modelo, para verificar a sua validade.

Homocedasticidade

Sabendo que um resíduo ou valor residual resulta da diferença entre os valores previstos pelo modelo e os valores observados, um dos processos alternativos para analisar a homocedasticidade (variância constante) consiste em observar a relação entre os resíduos estandardizados e os valores estimados estandardizados de *Y*. No Gráfico I está ilustrada esta relação, tanto para os resíduos estandardizados, como para o seu valor absoluto, que pretende tornar mais fácil a análise gráfica.

Verifica-se que não ocorre a manutenção de uma amplitude aproximadamente constante em relação ao eixo horizontal, pelo que parece existir uma tendência para uma maior variância quando os valores estimados estandardizados apresentam valores próximos de zero, diminuindo a variância quando estes valores estimados aumentam, em valor absoluto. Este modelo poderá, por este motivo, apresentar variância não constante, ou seja, heterocedasticidade.

Gráfico I
RELAÇÃO ENTRE RESÍDUOS ESTANDARDIZADOS
E VALORES ESTIMADOS ESTANDARDIZADOS DE *Y*



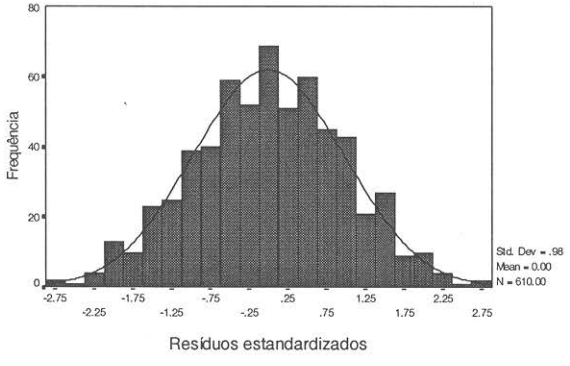
Covariância nula

O teste de Durbin-Watson permite verificar este pressuposto de existência de independência entre as variáveis aleatórias residuais. O resultado deste teste para o modelo #1 é 1,99671, valor próximo do valor de referência (dois), que indica a verificação do pressuposto.

Normalidade

A distribuição normal das variáveis aleatórias residuais pode ser analisada pelo Gráfico II, no qual se observa a diferença entre o histograma da distribuição das variáveis aleatórias residuais e a distribuição normal, que é visível em algumas das classes. Para testar a significância desta diferença utiliza-se o teste Kolmogorov-Smirnov (*K-S*), com a correcção de Lilliefors, apresentado no Quadro VI. Para este teste exige-se, normalmente, um nível de significância de 5% para não rejeitar a hipótese dos resíduos seguirem uma distribuição normal. Neste caso, a significância é superior a 20%, pelo que não se pode rejeitar aquela hipótese.

Gráfico II
HISTOGRAMA DOS RESÍDUOS E DISTRIBUIÇÃO NORMAL

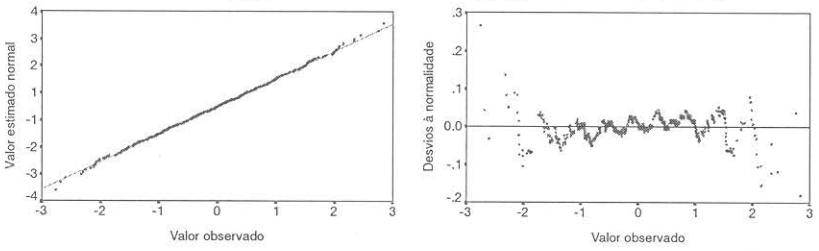


Quadro VI
TESTE À NORMALIDADE DOS RESÍDUOS ESTANDARDIZADOS

Estatística *K-S* (Lilliefors)	Graus de liberdade	Significância
0,0164	610	>0,20

Os desvios à normalidade podem ser observados no Gráfico III, em que se apresentam os gráficos *Q-Q* e *detrended Q-Q*, que ilustram, pelos desvios às linhas oblíqua e horizontal, respectivamente, estes desvios. Os pontos que mais se afastam das linhas são candidatos a *outliers*, pontos extremos com influência no modelo, cuja análise será feita posteriormente.

Gráfico III
GRÁFICOS *Q-Q* E *DETRENDED Q-Q* DOS RESÍDUOS



Multicolinearidade

A intensidade da multicolinearidade é estudada através da análise de:

• Correlação entre as variáveis independentes. Da matriz de correlações, os valores absolutos máximos observados para a correlação são entre as variáveis PAS-PAS^2 (0,86) e AAS-AAS^2 (0,86), o que seria de esperar pois as variáveis indicadas em segundo lugar são o quadrado das primeiras, e entre as variáveis AATA-PPOSSE (-0,75) e PATA-PPOSSE (0,75). Só valores superiores a 0,90 indiciam multicolinearidade, pelo que os valores observados não são indicadores de violação deste pressuposto.

• Tolerância (Tol.) e factor de inflação da variância (VIF). A tolerância mede o grau em que uma variável independente é explicada por todas as outras, sendo o valor de 0,1 o limite abaixo do qual existe multicolinearidade. O factor de inflação da variância (*Variance Inflation Factor*) é o inverso da tolerância, pelo que valores superiores a 10 sugerem multicolineriedade. Pela análise do Quadro VII, verifica-se também a inexistência de multicolinearidade, pois o valor mínimo para a tolerância é de 0,18 e o valor máximo para o VIF é de 5,51. Os valores mais baixos de tolerância e mais altos do VIF ocorrem para as variáveis que apresentavam, na análise anterior,

correlações mais elevadas (PAS, PAS^2, AAS, AAS^2, AATA, PATA e PPOSSE), o que seria de esperar.

• *Condition index* e proporção de variância. Os valores próprios (*eigenvalues*) dão uma indicação de quantas dimensões distintas existem entre as variáveis independentes. O *condition index* é a raiz quadrada do quociente entre o maior valor próprio e cada valor próprio. Valores próprios próximos de zero, *condition index* superiores a trinta e contribuição substancial de uma dimensão (em 90% ou mais) para a variância de duas ou mais variáveis, quando ocorrem simultaneamente, indicam multicolinearidade com elevada intensidade. No Quadro VIII observam-se estes valores para as 28 dimensões em análise do modelo desenvolvido, verificando-se que apenas para as três últimas os valores próprios e o *condition index* assumem valores críticos quanto à multicolinearidade. Mas ao analisarmos a proporção de variância, para nenhuma destas dimensões se observam duas variáveis com valor superior a 0,90, pelo que este método também não sugere multicolinearidade.

Quadro VII
TOLERÂNCIA E VIF

Variável	Tol.	VIF	Variável	Tol.	VIF	Variável	Tol.	VIF
AAS	0,23	4,34	AP	0,73	1,38	PCV	0,94	1,07
AAS^2	0,25	3,98	AR	0,38	2,64	PDC	0,70	1,43
AATA	0,24	4,26	ARE	0,36	2,80	PDEF	0,51	1,98
ACA	0,79	1,27	ASC	0,69	1,44	PFS	0,76	1,32
ACR^2	0,43	2,33	PAS	0,23	4,36	PP	0,72	1,38
ACV	0,94	1,06	PAS^2	0,25	4,01	PPOSSE	0,18	5,51
ADC	0,69	1,45	PATA	0,24	4,16	PR	0,38	2,62
ADEF	0,51	1,97	PCA_2	0,88	1,14	PRE	0,35	2,84
AFS	0,81	1,24	PCR^2	0,43	2,31	PSC	0,70	1,44

Quadro VIII
VALORES PRÓPRIOS (VP) E *CONDITION INDEX* (*CI*)

N	VP	*CI*	n	VP	*CI*	n	VP	*CI*	n	VP	*CI*
1	19,36	1,0	8	0,35	7,4	15	0,10	14,0	22	0,05	20,6
2	1,75	3,3	9	0,28	8,3	16	0,09	14,4	23	0,03	23,6
3	1,10	4,2	10	0,27	8,4	17	0,08	16,0	24	0,03	23,7
4	1,06	4,3	11	0,27	8,5	18	0,06	17,1	25	0,03	25,0
5	0,93	4,6	12	0,20	9,8	19	0,06	18,3	26	0,01	44,3
6	0,77	5,0	13	0,14	11,6	20	0,05	18,9	27	0,01	50,9
7	0,74	5,1	14	0,11	13,2	21	0,05	19,2	28	0,00	118,7

Este modelo, designado por modelo #1, apresenta um bom comportamento geral, tendo sido detectados, no entanto, três tipos de problemas. Em primeiro lugar, o coeficiente de determinação apresenta um valor relativamente baixo, indicando que apenas 42,81% da variação nos resultados dos jogos é explicada pelo modelo. Em segundo lugar, o teste *t* de *Student*, conduzido às estimativas dos coeficientes, resultou em níveis de significância bastante elevados para cinco das variáveis em estudo, pelo que não é garantido que elas devam ser incluídas no modelo. Finalmente, o modelo pode ter problemas ao nível da homocedasticidade, pelo facto de a variância dos valores residuais não ser constante. Os aspectos negativos do comportamento do modelo poderão estar relacionados com a presença de *outliers*, casos extremos e aberrantes com influência na forma como o modelo se comporta. Importa então detectar estes casos e retirá-los para construir um novo modelo.

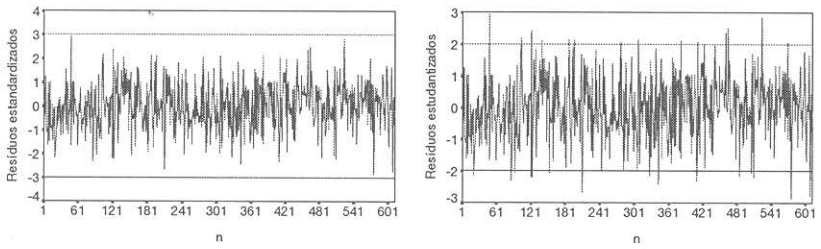
Análise dos *Outliers*

Depois de verificados os pressupostos do modelo, torna-se importante a análise de *outliers* (Barnett e Lewis, 1994; Rousseeuw e Leroy, 1987), casos extremos com influência no modelo, de modo a equacionar se os mesmos deverão ser retirados do modelo. Os *outliers* serão identificados com a ajuda das seguintes estatísticas: resíduos estandardizados, resíduos estudantizados, *leverage*, valores estimados ajustados, distância de *Cook*, *dfBetas* estandardizados e *dfFit* estandardizado.

As condições necessárias para que os resíduos não sugiram *outliers* são que os resíduos estandardizados sejam inferiores a três em valor absoluto e os resíduos estudantizados e estudantizados *deleted* tenham valor absoluto inferior a dois. O gráfico IV exemplifica a identificação de vários *outliers*, que apresentam resíduos estudantizados com valor absoluto superior a dois.

Gráfico IV

RESÍDUOS ESTANDARDIZADOS E ESTUDANTIZADOS



O *leverage* mede a influência de uma observação na qualidade do ajustamento feito. Quando é superior a $2(p+1)/n$, utilizando a nomenclatura

introduzida na formulação do modelo, a observação é considerada influente. O gráfico V ilustra os *outliers* identificados por esta medida.

O valor estimado ajustado é o valor estimado de um caso quando não é incluído no cálculo dos coeficientes de regressão. A diferença entre os valores estimados e os valores estimados ajustados permite a identificação de *outliers*.

A distância de *Cook* é a variação nos resíduos de todas as observações, quando um caso é excluído do cálculo dos coeficientes de regressão. Uma observação é considerada influente quando a distância de *Cook* é superior a $4/(n\text{-}p\text{-}1)$. O gráfico VI permite observar os *outliers* assim identificados.

Gráfico V
LEVERAGE

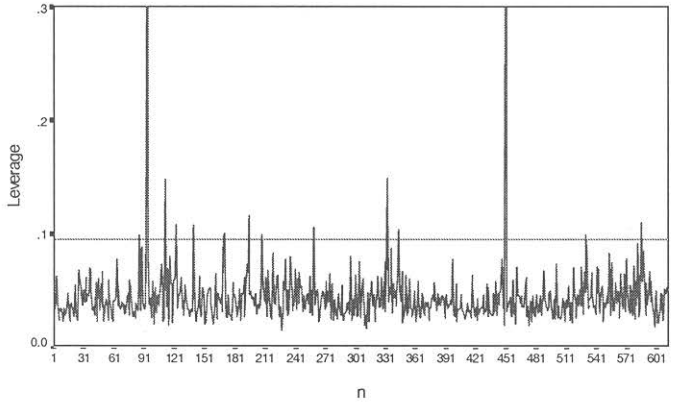
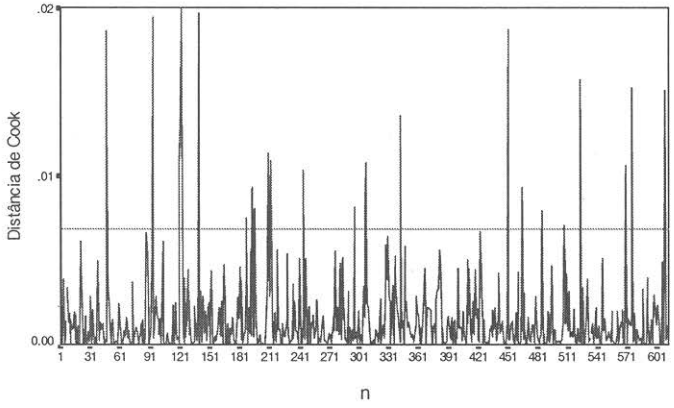
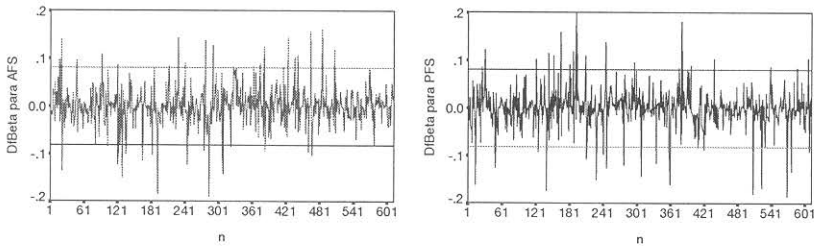


Gráfico VI
DISTÂNCIA DE *COOK*



O *DfBeta*, diferença nos valores de β, coeficientes da regressão, apresenta a mudança nos coeficientes de regressão quando se exclui um caso do modelo. Por este motivo existe um *DfBeta* para cada parâmetro e para a constante do modelo. As observações influentes são sugeridas por valores absolutos de *DfBeta* superiores a $2/\sqrt{n}$. O gráfico VII permite observar os *outliers*, segundo o critério do *DfBeta*, para os coeficientes associados às variáveis AFS e PFS, a título de exemplo.

Gráfico VII
DFBETA



A análise de *outliers* realizada permite identificar os casos extremos considerados influentes para o modelo, que devem ser excluídos para a construção de um novo modelo. Em cada gráfico, os *outliers* são os pontos que ultrapassam os limites horizontais representados: os resíduos estandardizados não apontam para nenhum *outlier*. Os restantes critérios analisados, resíduos estudantizados e estudantizados *deleted*, *leverage*, diferença entre valores estimados e valores estimados ajustados, distância de *Cook* e *DfBeta* para os 28 parâmetros do modelo, sugerem a existência de diversos *outliers*. Foram considerados casos extremos influentes as observações que desrespeitam pelo menos três dos 33 critérios considerados. Deste modo construiu-se um modelo intermédio com 483 observações, tendo sido retiradas 127 *outliers* do modelo #1. Esse modelo intermédio foi analisado, nos mesmos termos do modelo #1, tendo sido retirados mais 83 *outliers*.

4.2 Estimação de um novo modelo, modelo #2

Foi desenvolvido um novo modelo com 400 observações, pois foram excluídos os 210 *outliers* detectados anteriormente. Note-se que estes se distribuem, de modo relativamente uniforme, pelas várias equipas em análise. Este modelo, denominado modelo #2, constituirá o modelo definitivo, que explica as relações estatísticas entre as variáveis em análise, para as 400

observações seleccionadas e servirá o objectivo de previsão. No Quadro IX apresentam-se os resultados obtidos.

O coeficiente de determinação indica-nos que a variação que ocorre na variável dependente (resultado do jogo) passou a ser explicada em 82,09% pelas 25 variáveis incluídas no modelo. Relativamente ao modelo #1, este valor aumentou sensivelmente para o dobro.

Quadro IX
RESULTADOS PARA O MODELO DE REGRESSÃO #2

$r^2 = 0,8209$		g.l.	*SS*	*MS*
Desvio padrão = 0,3466	Regressão	27	204,82	7,5857
F = 63,13667 ⇒ Significância F =0,00	Resíduos	372	44,70	0,1202

r^2 - coeficiente de determinação — g.l. - graus de liberdade

SS - somatório dos quadrados — *MS* - média do somatório dos quadrados

Variável *i*	B_i	$s(b_i)$	Sig. *t*	Variável *i*	b_i	$s(b_i)$	Sig. *t*
AAS	-0,2890	0,0311	0,00	PAS^2	-0,0232	0,0050	0,00
AAS^2	0,0262	0,0062	0,00	PATA	0,0305	0,0190	0,11
AATA	-0,0355	0,0191	0,06	PCA_2	-0,1662	0,0457	0,00
ACA	0,0358	0,0114	0,00	PCR^2	$-2,53 \cdot 10^{-4}$	$8,76 \cdot 10^{-5}$	0,00
ACR^2	$2,53 \cdot 10^{-4}$	$8,55 \cdot 10^{-5}$	0,00	PCV	-0,2389	0,0614	0,00
ACV	0,2243	0,0605	0,00	PDC	0,0505	0,0123	0,00
ADC	-0,0537	0,0121	0,00	PDEF	0,0925	0,0212	0,00
ADEF	-0,0911	0,0210	0,00	PFS	-0,0058	0,0034	0,09
AFS	0,0051	0,0033	0,13	PP	0,0053	0,0034	0,12
AP	-0,0071	0,0032	0,03	PPOSSE	-2,9251	0,6095	0,00
AR	-0,022	0,0026	0,00	PR	0,0250	0,0025	0,00
ARE	-0,0252	0,0050	0,00	PRE	0,0315	0,0051	0,00
ASC	-0,0256	0,0067	0,00	PSC	0,0273	0,0068	0,00
PAS	0,2655	0,0282	0,00	Constante	2,2657	0,3870	0,00

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 teste *F*, com significância nula, valida o modelo, na sua globalidade. O seu valor aumentou de 16,1 para 63,1, do modelo #1 para o modelo #2, pelo que a significância global do último é superior.

A estimativa do desvio padrão do modelo tem o valor de 0,3466, tendo diminuído para aproximadamente metade do valor anterior, o que tornará as previsões mais precisas.

Apresentam-se as estimativas dos coeficientes e respectivos desvios padrão associados a cada variável, cujo significado continua a ser o mesmo que foi explicado anteriormente. Note-se a diminuição generalizada das estimativas do desvio padrão de cada coeficiente, $s(b_i)$, que diminui a imprecisão do modelo.

As estimativas dos coeficientes, b_i, apresentam valores não muito diferentes do modelo #1. Neste modelo, importa analisar o valor das estimativas dos coeficientes de cada variável e verificar se apresentam o sinal esperado. Estimativas com sinal positivo contribuem positivamente para o resultado do jogo, enquanto que estimativas com sinal negativo contribuem negativamente.

Relativamente ao comportamento da equipa em análise:

• as variáveis PAS, PATA, PDC, PDEF, PP, PR, PRE e PSC apresentam estimativas com sinal positivo: as assistências, o tempo de posse de bola no ataque, as defesas completas, as saídas completas, as recuperações e os remates têm uma contribuição positiva para o resultado, como seria de esperar; o tempo de posse de bola na defesa e as perdas de bola deveriam ter, à partida, uma comparticipação negativa no resultado.

• as variáveis PCA_2, PCR, PCV, PFS e PPOSSE, tendo sinal negativo, contribuem desfavoravelmente para o resultado, o que seria de esperar para os cartões amarelos duplos e para os cartões vermelhos, mas não para os cruzamentos, faltas sofridas e percentagem do tempo de posse de bola.

No que diz respeito ao comportamento do adversário, as variáveis:

• com estimativas positivas são ACA, ACR, ACV e AFS: os cartões amarelos e cartões vermelhos do adversário contribuem favoravelmente para um resultado positivo da equipa em análise, o que seria de esperar; já os cruzamentos e faltas sofridas apresentam um sinal contrário ao esperado;

• AAS, AATA, ADC, ADEF, AP, AR, ARE e ASC têm estimativas negativas, pelo que as assistências, o tempo de posse de bola no ataque, as defesas completas, as saídas completas, as recuperações e os remates do adversário contribuem para um resultado negativo da equipa em análise, confirmando as expectativas; esperava-se que o tempo de posse de bola na defesa e as perdas de bola do adversário tivessem um contributo inverso.

As variáveis com contribuições para o modelo com sinal contrário aquele que seria de esperar à partida, podem indicar que as expectativas para essas variáveis não se mantêm ou que haverá problemas ao nível da multicolinearidade. A constante representa as contribuições de outras variáveis não incluídas no modelo.

A significância do teste t de Student para cada variável, que indica a probabilidade dessa variável não ser significante para o modelo, tem valor nulo para a maioria das variáveis e diminuiu consideravelmente naquelas com

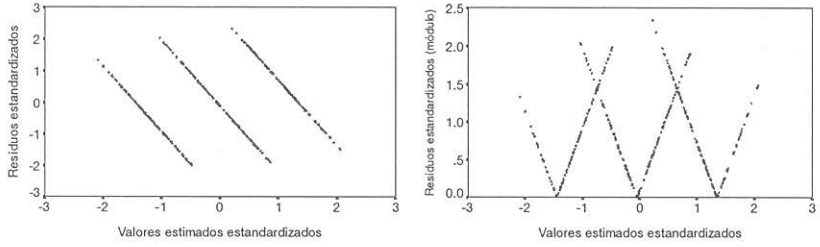
valores mais elevados, passando os quatro coeficientes com significâncias máximas a apresentar valores entre 0,09 e 0,13.

Tal como para o modelo #1, vão ser analisados os pressupostos do modelo #2, para justificar a sua validade.

Homocedasticidade

No gráfico VIII apresenta-se a relação entre os resíduos estandardizados (em valor real e valor absoluto) e os valores estimados estandardizados de *Y*. Verifica-se a diminuição da tendência para uma maior variância quando os valores estimados estandardizados estão próximos de zero: a amplitude apresenta maior constância em relação ao eixo horizontal, sendo as tendências crescentes ou decrescentes menores. Deste modo, a heterocedasticidade diminui relativamente ao modelo #1, não se rejeitando a hipótese de homocedasticidade para este novo modelo.

Gráfico VIII
RELAÇÃO ENTRE RESÍDUOS ESTANDARDIZADOS
E VALORES ESTIMADOS ESTANDARDIZADOS DE *Y*



Covariância nula

O teste de Durbin-Watson tem o valor de 1,88655, valor mais distante, mas ainda próximo, do valor de referência 2 que o obtido para o modelo #1, o que indica a independência entre as variáveis aleatórias residuais.

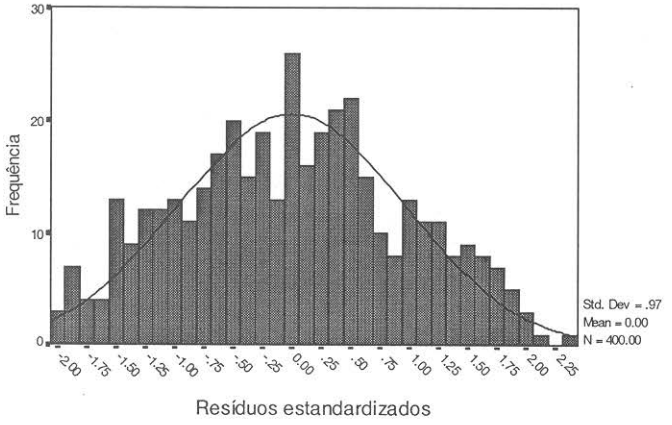
Normalidade

A distribuição normal das variáveis aleatórias residuais pode ser analisada pelo Gráfico IX. O teste Kolmogorov-Smirnov (*K-S*), com a correcção de Lilliefors, apresentado no Quadro X, apresenta uma significância superior a 20%, que permite não rejeitar a hipótese dos resíduos seguirem uma distribuição normal.

O gráfico X (*Q-Q* e *detrended Q-Q*) ilustra os desvios à normalidade. Verifica-se que este segundo modelo apresenta maiores desvios à normalidade

que o modelo #1, mas ainda insuficientes para rejeitar a hipótese de os resíduos seguirem uma distribuição normal.

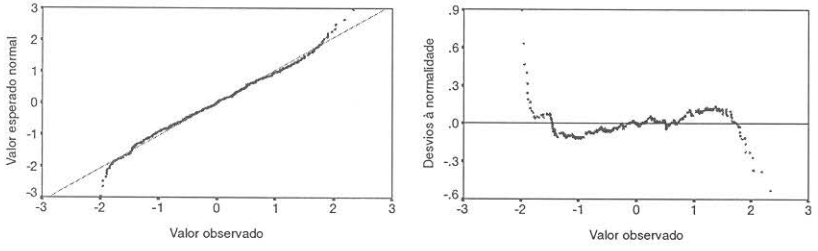
Gráfico IX
HISTOGRAMA DOS RESÍDUOS E DISTRIBUIÇÃO NORMAL



Quadro X
TESTE À NORMALIDADE DOS RESÍDUOS ESTANDARDIZADOS

Estatística *K-S* (Lilliefors)	Graus de liberdade	Significância
0,0315	400	>0,20

Gráfico X
GRÁFICOS *Q-Q* E *DETRENDED Q-Q* DOS RESÍDUOS



Multicolinearidade

• Através da leitura da matriz de correlações, determinam-se os valores absolutos máximos para a correlação entre variáveis. As variáveis AAS-AAS^2 (0,92) e PAS-PAS^2 (0,90) são as que apresentam maior correlação, pois as segundas são o quadrado das primeiras. Entre as variáveis AATA-PPOSSE (-0,75) e PATA-PPOSSE (0,74) ocorre correlação com uma dimensão

inferior. Os valores de correlação superiores a 0,90 podem ter implicações no modelo, pelo que a análise subsequente é importante.

• Tolerância (Tol.) e factor de inflação da variância (VIF). Do Quadro XI verifica-se que o valor mínimo para a tolerância é de 0,13 e o valor máximo para o VIF é de 7,61, mantendo-se ambos dentro dos limites aceitáveis para que não exista multicolinearidade. Novamente os valores mais baixos de tolerância e mais altos do VIF ocorrem para as variáveis que apresentam correlações mais elevadas.

Quadro XI
TOLERÂNCIA E VIF

Variável	Tol.	VIF	Variável	Tol.	VIF	Variável	Tol.	VIF
AAS	0,13	7,61	AP	0,68	1,47	PCV	0,89	1,13
AAS^2	0,14	7,06	AR	0,32	3,09	PDC	0,68	1,47
AATA	0,23	4,43	ARE	0,34	2,91	PDEF	0,48	2,08
ACA	0,77	1,30	ASC	0,71	1,41	PFS	0,74	1,36
ACR^2	0,40	2,48	PAS	0,15	6,58	PP	0,66	1,51
ACV	0,89	1,12	PAS^2	0,17	5,93	PPOSSE	0,18	5,48
ADC	0,68	1,47	PATA	0,22	4,45	PR	0,34	2,98
ADEF	0,47	2,11	PCA_2	0,86	1,17	PRE	0,33	3,04
AFS	0,79	1,27	PCR^2	0,41	2,43	PSC	0,72	1,39

• *Condition index* e proporção de variância. O Quadro XII contém os valores próprios e *condition index* para as n=28 dimensões do modelo. Para as três últimas dimensões do modelo os valores próprios e o *condition index* assumem valores críticos quanto à multicolinearidade, embora para nenhuma destas dimensões se observam duas variáveis com proporção de variância superior a 0,90, pelo que este método não sugere multicolinearidade.

Quadro XII
VALORES PRÓPRIOS (VP) E *CONDITION INDEX* (*CI*)

N	VP	*CI*	*n*	VP	*CI*	*n*	VP	*CI*	*n*	VP	*CI*
1	19,60	1,0	8	0,36	7,4	15	0,09	14,9	22	0,04	22,3
2	1,90	3,2	9	0,28	8,3	16	0,07	16,5	23	0,03	25,0
3	1,12	4,2	10	0,26	8,7	17	0,07	16,9	24	0,03	25,9
4	0,97	4,5	11	0,24	9,0	18	0,06	18,2	25	0,03	27,0
5	0,84	4,8	12	0,19	10,1	19	0,05	19,0	26	0,01	46,4
6	0,76	5,1	13	0,14	11,7	20	0,05	20,8	27	0,01	55,6
7	0,66	5,4	14	0,11	13,2	21	0,04	21,5	28	0,00	122,2

Este modelo, designado por #2, desenvolvido a partir do modelo inicial respeita, tal como foi demonstrado, todos os pressupostos que possam condicionar a sua aplicação. O facto de se retirarem os *outliers* do modelo #1 aumentou consideravelmente, para 82,09%, a variação da variável dependente (resultado do jogo) explicada pelas variáveis independentes incluídas no modelo e diminuiu bastante a variância, facto importante para uma boa capacidade preditiva do modelo.

5. PREVISÃO

Um dos objectivos dos modelos de regressão é, precisamente, a previsão do valor da variável dependente a partir dos valores das variáveis independentes. Devem então ser construídos intervalos de confiança para as previsões, de modo a possibilitar uma análise de sensibilidade ao modelo.

Poderia ser colocada a questão do modelo #2 apresentar resultados, quanto à previsão, bastante diferentes do modelo #1, pela inclusão de um menor número de dados, pois foram retiradas 210 observações do modelo inicial. Pelo facto de terem sido excluídos do modelo apenas *outliers*, verifica-se que as previsões efectuadas com o modelo #1 são semelhantes às do modelo #2, pois nos 306 jogos analisados, as previsões dos dois modelos originam resultados diferentes apenas em 21 jogos, 6,9% do total.

Com este modelo é possível efectuar a previsão do resultado de cada jogo, a partir do comportamento das equipas intervenientes, em termos de dados estatísticos e compará-la com os resultados efectivamente observados. No Quadro XIII são apresentadas esses previsões, a título de exemplo, para os jogos entre os quatro primeiros classificados do campeonato.

Na primeira coluna é identificado o jogo em análise, bem como o respectivo resultado. Nas colunas seguintes apresenta-se o resultado do jogo, em termos de os pontos obtidos por cada uma das equipas, que jogam em casa ou fora. Os valores de Y^*_c / Y^*_f são os valores estimados pelo modelo para o resultado de cada uma das equipas intervenientes, que actuam em casa / fora, a partir do seu desempenho em cada jogo, que é função dos dados estatísticos para cada variável. Os valores anteriores são normalizados, de modo a que o seu somatório seja igual à unidade, representados por P_c / P_f. Coloca-se agora o problema de construir uma regra de decisão que permita definir o resultado de cada jogo em função dos valores estimados normalizados. Depois de testados vários valores, optou-se pela seguinte regra: sempre que Δ (diferença entre valores estimados para as duas equipas) seja superior a 0,300, em valor absoluto, atribui-se a vitória à equipa com valor estimado normalizado superior a 0,650. Quando essa diferença for inferior a 0,300 o resultado do

jogo é o empate. Nas três últimas colunas são apresentados os valores previstos pelo modelo para o resultado do jogo, utilizando a regra de decisão definida, para as equipas que actuam em casa / fora e a sua comparação com o resultado realmente observado.

Quadro XIII
EXEMPLOS DE PREVISÕES DO RESULTADO DO JOGO

Jogo	Y_c	Y_f	Y^*_c	Y^*_f	P_c	P_f	Δ	Y'_c	Y'_f	Obs.
POR-0 : BOA-2	0	3	-0.282	2.514	-0.126	1.126	-1.252	0	3	*ok*
BOA-2 : BEN-1	3	0	1.055	1.118	0.486	0.514	-0.028	1	1	~
BOA-2 : SPO-2	1	1	-0.226	2.202	-0.114	1.114	-1.228	0	3	~
POR-3 : BEN-1	3	0	2.284	-.238	1.117	-0.117	1.234	3	0	*ok*
POR-3 : SPO-2	3	0	1.264	0.699	0.644	0.356	0.288	1	1	~
SPO-1 : BEN-2	0	3	0.718	1.672	0.301	0.699	-0.398	0	3	*ok*
BOA-0 : POR-0	1	1	0.621	1.502	0.292	0.708	-0.416	0	3	~
BEN-0 : BOA-3	0	3	0.566	1.399	0.288	0.712	-0.424	0	3	*ok*
SPO-1 : BOA-1	1	1	1.367	0.735	0.651	0.349	0.302	3	0	~
BEN-1 : POR-1	1	1	0.682	1.419	0.325	0.675	-0.350	0	3	~
SPO-1 : POR-1	1	1	1.659	0.414	0.800	0.200	0.600	3	0	~
BEN-3 : SPO-3	1	1	0.658	1.119	0.370	0.630	-0.260	1	1	*ok*

Y_c / Y_f - valor observado da variável dependente (resultado do jogo, em pontos) para as equipas intervenientes que jogam em casa / fora.

Y^*_c / Y^*_f - valor estimado pelo modelo da variável dependente (resultado do jogo, em pontos) para as equipas intervenientes que jogam em casa / fora.

P_c / P_f - valor estimado normalizado de modo a que a soma dos valores estimados normalizados para as equipas intervenientes que jogam em casa / fora seja igual a um.

Δ - diferença entre P_c e P_f.

Y'_c / Y'_f - valor estimado pelo modelo do resultado do jogo (em pontos) para as equipas intervenientes que jogam em casa / fora.

Obs. - Observação relativamente à concordância (*ok*) ou não (~) entre o valor observado e o valor previsto pelo modelo para o resultado do jogo.

A atribuição de vitória num jogo resulta de um valor estimado normalizado superior a 0,65 para a equipa vitoriosa e, por consequência, um valor inferior a 0,35 para a equipa derrotada. A escolha destes valores justifica-se primeiro pelo facto do resultado da equipa vitoriosa previsto pelo modelo ser, pelo menos, quase o dobro (1,86) do do adversário e a segunda prende-se com o facto de que esta regra de decisão origina a previsão, pelo modelo, de vitória de uma das equipas em 208 jogos e empate nos restantes 98, resultados aproximadamente concordantes com os observados: 212 vitórias e 94 empates.

O modelo permite efectuar as previsões da variável dependente, resultado do jogo, para todas as observações (306 jogos[6]) a partir dos valores estatísticos das variáveis independentes, o que origina os valores estimados pelo modelo para o somatório de pontos conseguidos por cada equipa e consequentemente, a classificação final estimada pelo modelo, que é comparada com a classificação observada na realidade no Quadro XIV. Estão também representados neste Quadro os valores observados e previstos pelo modelo, para cada equipa e a contribuição dos jogos realizados em casa e fora para o total de pontos. Finalmente apresenta-se a classificação final ordenada prevista pelo modelo, podendo observar-se as alterações relativamente à observada ($\nearrow^n$, $\searrow_n$ e = representam subida ou descida de *n* posições ou manutenção de posição classificativa).

Quadro XIV
PREVISÃO DA CLASSIFICAÇÃO FINAL

	Classificação (pontos)						Classificação prevista ordenada (pontos)	
	Total		Casa		Fora			
Equipa	Obs.	Prev.	Obs.	Prev.	Obs.	Prev.	Equipa	Total
POR	79	67	48	42	31	25	POR =	67
BOA	71	61	42	35	29	26	SPO $\nearrow^2$	63
BEN	65	61	38	34	27	27	BOA $\searrow_1$	61
SPO	63	63	40	42	23	21	BEN $\searrow_1$	61
SET	53	48	36	34	17	14	GUI $\nearrow^2$	56
LEI	52	50	31	30	21	20	LEI =	50
GUI	50	56	35	43	15	13	BEI $\nearrow^9$	50
EST	45	48	32	35	13	13	SET $\searrow_3$	48
BRA	42	48	28	37	14	11	EST $\searrow_1$	48
MAR	41	32	26	21	15	11	BRA $\searrow_1$	48
FAR	39	44	26	27	13	17	FAR =	44
SAL	38	26	28	18	10	8	RIO $\nearrow^3$	41
CAM	37	32	26	24	11	8	ALV $\nearrow^1$	37
ALV	35	37	25	23	10	14	MAR $\searrow_4$	32
RIO	35	41	20	24	15	17	CAM $\searrow_2$	32
BEI	33	50	24	32	9	18	ACA $\nearrow^2$	29
CHA	25	27	19	18	6	9	CHA =	27
ACA	21	29	14	17	7	12	SAL $\searrow_6$	26

Obs. - Observado Prev. - Previsto

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? A resposta pode ser obtida através de uma análise de sensibilidade aos resultados.

Com esse propósito, constroem-se os intervalos de confiança a 95% para as previsões dos valores estimados da variável dependente (resultado do jogo), que garantem que as previsões efectuadas pertençam ao intervalo de confiança construído com uma probabilidade de 95%. A partir desses intervalos de confiança determinam-se os valores máximos e mínimos para a classificação final prevista pelo modelo, para cada equipa, que são apresentados no Quadro XV e ilustrados pelo Gráfico XI.

Os valores previstos pelo modelo, para a pontuação de cada equipa na classificação final, em função do seu comportamento estatístico nos jogos efectuados permite verificar que, nalguns casos, a pontuação real é superior à prevista pelo modelo e noutros sucede o inverso.

Quadro XV
ANÁLISE DE SENSIBILIDADE À CLASSIFICAÇÃO FINAL PREVISTA

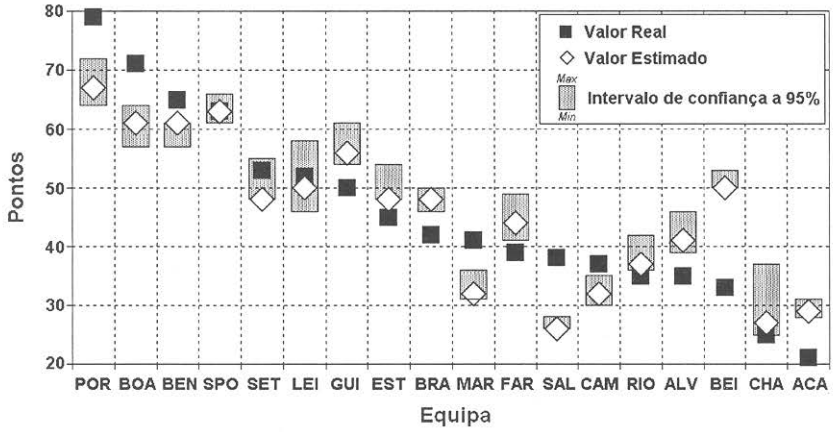
	Classificação observada	Classificação prevista (pontos)		
		Modelo	Mínimo	Máximo
POR	79	67	64	72
BOA	71	61	57	64
BEN	65	61	57	61
SPO	63	63	61	66
SET	53	48	48	55
LEI	52	50	46	58
GUI	50	56	54	61
EST	45	48	48	54
BRA	42	48	46	50
MAR	41	32	31	36
FAR	39	44	41	49
SAL	38	26	26	28
CAM	37	32	30	35
RIO	35	37	36	42
ALV	35	41	39	46
BEI	33	50	50	53
CHA	25	27	25	37
ACA	21	29	28	31

Os intervalos de confiança dão uma ideia da variação que pode ocorrer na classificação final estimada pelo modelo, por variações pontuais do resultado estimado de cada jogo. Através da análise da classificação final prevista pelo modelo para cada equipa e respectivos intervalos de confiança, verifica-se que os intervalos de confiança previstos apenas incluem os valores observados

para algumas equipas (SPO, SET, LEI e CHA), sendo significativamente inferiores aos reais para outras: POR, BOA, BEN, MAR, SAL e CAM; e significativamente superiores para GUI, EST, BRA, FAR, ALV, BEI e ACA. Estas diferenças permitem estabelecer uma classificação final ordenada, prevista pelo modelo, no Quadro XIV, que indica subidas e descidas classificativas, bastante relevantes para o caso de algumas equipas (BEI, MAR e SAL).

Gráfico XI

ANÁLISE DE SENSIBILIDADE À CLASSIFICAÇÃO FINAL PREVISTA



6. CONSIDERAÇÕES FINAIS

O modelo final desenvolvido parece ser de aplicação viável a este caso. No entanto, poderão existir outros com algumas variações, mais ou menos relevantes, com resultados também significantes. Um método que permite verificar a qualidade do modelo desenvolvido consiste em avaliar a possibilidade da sua aplicação a observações futuras, por exemplo, aos dados do próximo campeonato nacional de futebol.

Finalmente, é de referir que as diferenças entre os valores estimados pelo modelo para a variável dependente, resultado de cada jogo, em função dos valores das variáveis independentes, dados estatísticos observados para as equipas intervenientes em jogo, resultam de variáveis não incluídas no modelo e da própria aleatoriedade da variável dependente. Exemplos de variáveis não incluídas no modelo, por não estarem disponíveis, são, entre outras: cantos,

grandes penalidades convertidas e falhadas, comportamento do árbitro, estabilidade psicológica, motivação e orçamento de cada equipa.

Os procedimentos que dão origem ao modelo final obtido para a análise deste caso permitem exemplificar o desenvolvimento e aplicação de um modelo de regressão para estudos observacionais explicativos, ilustrando os vários passos que vão sendo tomados.

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.

NOTAS

[1] A Infordesporto, nomeadamente o seu *site* na Internet, é a fonte de onde foram recolhidos todos os dados estatísticos em que se baseia a análise desenvolvida ao longo deste trabalho.

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

[3] Por vezes, não há informação disponível sobre o valor das variáveis, para determinada observação, pelo que o número de observações válidas para cada variável pode ser inferior ao total de observações.

[4] Tempo total jogado pelo jogador (em minutos).

[5] A substituição dos *missing values* pode provocar alterações no modelo, que são minimizadas pelo facto de serem realizadas para um conjunto reduzido de observações.

[6] O resultado do jogo previsto pelo modelo coincide com o resultado observado em 186 jogos, sendo diferente nos restantes 120, ou seja, 39,2% dos jogos terminam com um resultado previsto pelo modelo diferente do observado.

BIBLIOGRAFIA

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

BRYMAN, Alan; CRAMER, Duncan (1993), *Análise de dados em Ciências Sociais - Introdução às técnicas utilizando SPSS*, Celta Editora, Lisboa.

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

FREEDMAN, D. A. (1982), "A Note on Sreening Regression Equations", *The American Statistician*, 37, p. 152-155, Alexandria.

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

HAMILTON, Barton H. (1997), "Racial Discrimination and professional basketball salaries in the 1990s", *Applied Economics*, 29 (3), p. 287, London.

INFORDESPORTO, "Campeonato nacional I divisão - 1998/99", *site* na Internet: *http://www.infordesporto.pt/ Futebol/idiv/Jogos.*

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

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

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

PESTANA, M. Helena; GAGEIRO, João N. (1998), *Análise de dados para Ciências Sociais - A complementaridade do SPSS*, Edições Sílabo, Lisboa.

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

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

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