ESTUDOS DE ECONOMIA, VOL. XVI–XVII, N.º 4, OUTONO 1997

UTILIZAÇÃO DE UM MODELO LOGLINEAR COM SOBREDISPERSÃO PARA MODELIZAR CASOS DE ACIDENTES DE TRABALHO EM PORTUGAL

Maria Alexandra Abreu Henriques Seco

1 — Introdução

A classe dos modelos lineares generalizados foi inicialmente apresentada num artigo de Nelder e Weddenburn (1972) tendo tido, mais tarde, grande desenvolvimento na primeira edição (1983) do livro de McCullagh e Nelder.

O modelo linear clássico, com erro aditivo de distribuição normal, é um caso particular dos modelos lineares generalizados, bem como os modelos logísticos e os modelos loglineares associados, respectivamente, à distribuição binomial e à de Poisson.

Nelder e Weddenburn (1972), notando entre os vários modelos semelhanças, em vez de diferenças, alargaram as hipóteses da distribuição da variável resposta à família exponencial e introduziram o conceito de função vínculo entre a expressão linear cujos parâmetros são estimados e a média da variável resposta. Estes autores desenvolveram ainda toda uma teoria aplicável a esta classe de modelos, incluindo um algoritmo geral para o cálculo de estimadores da máxima verosimilhança.

Um modelo linear generalizado é composto por três componentes e define-se da forma seguinte: seja Y um vector de N variáveis aleatórias independentes, com a mesma distribuição e vector das médias μ:

1) Componente aleatória, especificando a densidade de probabilidade de Y_i, $i = 1,2,...,N$, que pertence à família exponencial com parâmetro canónico θ_i (1);
2) Componente sistemática: existem p *covariates* $x_1, x_2, \ldots, x_p$, com valores observados $x_{i1}, x_{i2}, \ldots, x_{ip}$, induzindo um preditor linear η_i dado por:

$$\eta_i = \sum_{i=1}^{n} x_{ij}\beta_j, \quad i = 1,2,...,N,$$

onde β_j ($j = 1,...,p$), são parâmetros desconhecidos, ou em notação matricial, $\eta = X\beta$, sendo X uma matriz $N \times p$ chamada matriz do modelo e um vector $p \times 1$ de parâmetros;

(1) Se ϕ é conhecido, diz-se que Y pertence à família exponencial de parâmetro canónico θ se a sua função densidade (ou probabilidade) é da forma $f(y \mid \theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$ com a, b e c funções reais. Exemplos conhecidos são a distribuição normal, a Poisson e a binomial.

3) A componente aleatória do modelo, Y, e a componente sistemática estão relacionadas por $\eta_i = g(\mu_i)$, $i = 1,2,\ldots,N$, onde g é uma função monótona e diferenciável, designada por função vínculo.

Estes modelos generalizam entre outros, o modelo de regressão linear, o modelo loglinear de Poisson e o modelo probit e logit para proporções.

Pode consultar-se Seco (1995) para um estudo desenvolvido dos modelos lineares generalizados.

2 — Estimação de parâmetros num modelo linear generalizado

Para calcular as estimativas dos parâmetros, utiliza-se o seguinte algoritmo iterativo: seja W uma matriz diagonal de elementos $\mathrm{var}(Y_i)$ e $V(\mu) = \frac{\mathrm{var}(Y)}{\phi}$, sendo ϕ o designado parâmetro de dispersão.

Passo 0: Seja $t = 0$. Calcula-se:

$$\widehat{\mu}_i^{(t)} = \bar{y}$$

para todo $i = 1, 2,\ldots,N$.

Passo 1: Calcula-se:

$$\widehat{\eta}_i^{(t)} = g(\widehat{\mu}_i^{(t)})$$

e:

$$\widehat{z}_i^{(t)} = \widehat{\eta}_i^{(t)} + (y_i - \widehat{\mu}_i^{(t)})\left(\frac{d\eta_i}{d\mu_i}\right)^{(t)}$$

$$\widehat{\omega}_{ii}^{(t)} = \frac{1}{a_i(\phi)\, V(\widehat{\mu}_i^{(t)})}\left[\left(\frac{d\mu_i}{d\eta_i}\right)^2\right]^{(t)}$$

$$i = 1,\ldots,N.$$

Passo 2: Determina-se:

$$\widehat{\beta}^{(t+1)} = (X^T \widehat{W}^{(t)} X)^{-1} X^T \widehat{W}^{(t)} \widehat{z}^{(t)}$$

$$\widehat{\eta}^{(t+1)} = X\widehat{\beta}^{(t+1)}$$

e:

$$\widehat{\mu}^{(t+1)} = g^{-1}(\widehat{\eta}^{(t+1)})$$

Passo 3: Compara-se $\widehat{\beta}^{(t+1)}$ com $\widehat{\beta}^{(t)}$. Se a diferença é suficientemente pequena, STOP. Caso contrário faz-se $t = t + 1$, e volta-se ao passo 1.

O algoritmo descrito em geral tem convergência rápida; no entanto a divergência pode ocorrer. O programa GLIM (Generalized Linear Iteractive Modelling), desenvolvido pelo grupo NAG (1987), Numerical Algoritm Group, especialmente concebido para o estudo dos modelos lineares generalizados, pára ao fim de alguns ciclos caso não haja, ainda, convergência. Nesta fase, podem inspeccionar-se as estimativas e, se necessário, pode omitir-se um subconjunto de dados ou modificar o modelo. Pode consultar-se Aitkin (1989) para um estudo pormenorizado do programa.

3 — Análise da *deviance*

No processo de ajustamento do modelo pode analisar-se a discrepância entre os valores observados y e os valores ajustados $\hat{\mu}$. Em geral, os valores ajustados $\hat{\mu}$ não são iguais aos valores observados, surgindo a necessidade de analisar as diferenças entre estes valores. Obviamente, uma grande discrepância entre y e $\hat{\mu}$ não indica tratar-se de um bom modelo; por outro lado, o modelo «exacto» ou saturado onde $y = \hat{\mu}$ contraria o princípio da parcimónia também designado *Ocam's razor*. Tem então de avaliar-se a qualidade do ajustamento do modelo e fazer um balanço entre a vantagem de aumentar o número de parâmetros e a maior complexidade do modelo (ou porventura, alterar de outra forma o modelo).

Um dos caminhos que pode seguir-se é comparar a verosimilhança associada a determinado modelo em estudo, com a verosimilhança associada ao modelo saturado.

Definição 1: Chama-se modelo saturado a um modelo linear generalizado com a mesma distribuição do modelo em estudo, a mesma função vínculo e com um número de parâmetros igual ao número de observações, N (um por observação).

O modelo saturado atribui toda a variação dos y's à componente sistemática e nada à componente aleatória. Repita-se que o modelo saturado é usado como referência, pois reproduz os dados na totalidade não efectuando nenhuma condensação da informação para sublinhar os seus aspectos essenciais.

É mais conveniente trabalhar com a função verosimilhança em termos do valor médio μ, do que com o parâmetro canónico. Assim, seja $l(\hat{\mu}, \phi \mid y)$ a logverosimilhança do modelo em estudo maximizada em para um valor fixo do parâmetro de dispersão e seja $l(y, \phi \mid y)$ a logverosimilhança correspondente ao modelo saturado de N parâmetros.

Definição 2: Chama-se *deviance* à diferença:

$$D(y \mid \hat{\mu}) = 2\,\phi\,[l(y, \phi \mid y) - l(\hat{\mu}, \phi \mid y)]$$

Definição 3: Chama-se *deviance* à escala a:

$$D^*(y \mid \hat{\mu}) = \frac{D(y \mid \hat{\mu})}{\phi}$$

Se ϕ for desconhecido, pode usar-se uma estimativa no seu lugar. Três exemplos onde é conhecido são os casos da distribuição de Poisson, binomial e exponencial; em qualquer dos casos tem-se $\phi = 1$. Verifica-se o seguinte teorema [Dobson (1990)].

Teorema 1: A estatística $D^*(y \mid \hat{\mu})$ tem distribuição assintótica de qui-quadrado com N-p graus de liberdade:

(1) $$D^* \mathrel{\dot{\sim}} \chi^2_{N-p}$$

onde p é o número de parâmetros do modelo em estudo.

A qualidade do ajustamento do modelo pode ser avaliada considerando que se D^* é consistente, ao nível de $100\alpha\,\%$, com a distribuição indicada em **(1)**, então o modelo tem um ajustamento razoável [v. Hinkley e outros (1991) para estudo pormenorizado].

Seja agora:

$$\Delta D^* = D_0^* - D_1^*$$

a diferença entre as *deviances* à escala do «modelo 0» e «modelo 1». São conhecidos os seguintes resultados:

Proposição 1: Sejam dois modelos de Y, respectivamente com p e q parâmetros independentes (admite-se sem perda de generalidade que o espaço dos parâmetros do primeiro modelo incluí o espaço de parâmetros do segundo). Então ΔD^* tem distribuição aproximada de χ^2 com $p-q$ graus de liberdade.

Pode então construir-se um quadro de análise da *deviance* à escala (ANODEV) que generaliza o quadro ANOVA dos modelos de regressão linear, formando uma sucessão de modelos.

Teorema 2: Seja M_0 um modelo com q parâmetros e M_1 um modelo de $p > q$ parâmetros, ambos explicativos de Y_i, $i = 1, 2, \ldots, N$. Seja ainda D_0^* e D_1^* a *deviance* associada, respectivamente, a cada um deles. Na hipótese de independência verifica-se que:

$$\left(\frac{D_0^* - D_1^*}{p-q}\right) \Big/ \left(\frac{D_1^*}{N-p}\right) \;\dot{\sim}\; F_{p-q,\, N-p}$$

4 — Resíduos

A estatística D^* informa se o modelo traduz, ou não, adequadamente os dados mas não indica onde este falha. Para este tipo de análise o estudo dos resíduos é particularmente útil. Pierce e Schafer (1986) e MacCullagh e Nelder (1989) propoem, para os modelos lineares generalizados, resíduos com distribuição aproximadamente normal.

A definição mais usual é a do resíduo de Pearson, que é definido por:

$$r_p = \frac{Y - \hat{\mu}}{\hat{\sigma}_y}$$

onde $\hat{\sigma}_y$ é o desvio padrão estimado. É o resíduo dado pelo programa GLIM. É muitas vezes usado devido à sua simplicidade de interpretação. No entanto, tem desvantagem em distribuições da família exponencial não normais marcadamente assimétricas, pois a sua distribuição insiste em falhar na normalidade assintótica.

Com vista a avaliar o ajustamento do modelo através do estudo dos resíduos, o programa GLIM permite utilizar o seguinte procedimento: podem representar-se graficamente os resíduos padrões ordenados $r_{(i)}$ contra os quantis:

$$\Phi^{-1}\left\{\frac{i-a}{N+1-2a}\right\}$$

da normal estandardizada, onde a é uma constante entre zero e um ($0 \leq a < 1$). O programa GLIM utiliza o valor $a = 0.3175$, uma vez que este permite usar o teste de Filliben (1975) ([2]) para a normalidade. Curvatura sistemática ou observações individuais muito afastadas de uma linha recta indicam falhas na especificação da distribuição de Y ou a presença de *outliers.*

5 — Modelos loglineares

Sejam Y_i, $i = 1, 2,...,N$, N variáveis aleatórias independentes seguindo a lei de Poisson com média μ_i. Chama-se modelo loglinear de Poisson a um modelo linear generalizado da forma:

$$\log \mu_i = \eta_i = \beta^T x_i$$
$$i = 1,...,N,$$

onde a função vínculo é a logarítmica. Como habitualmente, β é o vector de parâmetros e x_1 o vector de variáveis explicativas.

A expressão da *deviance,* neste caso, é:

$$D(y \mid \hat{\mu}) = 2 \sum_{i=1}^{N} [y_i \log (y_i / \hat{\mu}_i) - (y_i - \hat{\mu}_i)]$$

5.1 — Sobredispersão

Segundo a lei de Poisson, a média e a variância coincidem; no entanto, em casos práticos verifica-se frequentemente que $\mathrm{var}(Y) > E(Y)$ ou seja $\phi > 1$, fenómeno que se designa por sobredispersão. Hinkley e outros (1991) afirmam mesmo que a sobredispersão é a norma na prática e que $\phi = 1$ é a excepção, parecendo mais sensato partir do princípio de que a sobredispersão está presente até se mostrar o contrário. Pode também acontecer que $\phi < 1$, situação a que se chama sobdispersão, mas que ocorre com pouca frequência.

([2]) Este autor combina o uso de um coeficiente de correlação com um estudo gráfico, para testar a normalidade. Sugere a construção de um gráfico com os resíduos ordenados $r_{(i)}$, contra as estatísticas medianas ordenadas $M_{(i)}$ de uma normal, $N(0,1)$. As medianas ordenadas $M_{(i)}$ estão relacionadas com as medianas ordenadas $m_{(i)}$ de uma distribuição uniforme em [0,1], por $M_{(i)} = \Phi^{-1}[m_{(i)}]$. Na prática é recomendado o uso de $m_{(i)} = (i - 0.3175)/(N + 0.365)$, $i = 2,3,..., N-1$. Se a amostra provém de uma normal, o gráfico deve apresentar um padrão linear. O teste estatístico que acompanha este gráfico é o coeficiente de correlação ρ ($0 \leq \rho \leq 1$), $\rho = \mathrm{Corr}(R,M) = \frac{\sum_i (r_i - \bar{r})(M_i - \bar{M})}{\sqrt{\sum_i (r_i - \bar{r})^2 \sum_i (M_i - \bar{M})^2}}$.

Uma forma de ultrapassar esta situação é recorrer à quase-verosimilhança, conceito introduzido por Wedderburn (1974). Supõe-se que as componentes do vector resposta Y são independentes com média μ e matriz de covariância:

$$\phi V(\mu)$$

onde ϕ é constante e não depende de β, mas pode ser desconhecido, e $V(\mu)$ é uma matriz de funções conhecidas.

Como as componentes de Y são independentes, assume-se que a matriz $V(\mu)$ é diagonal e escreve-se:

$$V(\mu) = \text{diag}\{V_1(\mu), \dots, V_N(\mu)\}$$

Admite-se ainda que as funções $V_i(\mu)$ dependem apenas da componente i de μ. Ao integral:

$$Q_i(\mu_i \mid y_i) = \int_{y_i}^{\mu_i} \frac{y_i + t}{\phi V(t)} dt$$

caso exista, chama-se quase-verosimilhança ou log-quase-verosimilhança para μ_i baseada em y_i.

Como por hipótese, Y tem N componentes independentes, a função quase-verosi-milhança tem a expressão:

$$(Q_i)(\mu \mid y) = \sum_{i=1}^{N} Q_i(\mu_i \mid y_i)$$

Nos casos de sobredispersão admite-se apenas que:

$$\text{var}(Y) = \phi\mu \ (^3)$$

com $\phi > 1$ (no caso de sobdispersão, $\phi < 1$) e recorre-se à quase-verosimilhança; parte-se do princípio de que a variância é igual à média afectada por um factor desconhecido ϕ. Os métodos descritos no capítulo anterior são ainda válidos, mas obviamente com algumas diferenças.

Para $V(\mu) = \mu$, a quase-verosimilhança corresponde à logverosimilhança associada à distribuição de Poisson. Aitkin e outros (1989) refere ainda o seguinte teorema:

Teorema 3: Os estimadores da máxima quase-verosimilhança para β, com $V(\mu) = \mu$, são os mesmos que se obtêm da distribuição de Poisson e os erros padrões de β podem obter-se multiplicando os erros padrões do modelo de Poisson por um estimador de $\sqrt{\phi}$.

(3) Note-se $V(\mu) = \mu$, para a Poisson.

Pode observar-se depois deste teorema que, para o estudo dos modelos loglineares, o recurso à distribuição de Poisson não é essencial, pois este pode fazer-se da mesma forma partindo apenas de algumas hipóteses baseadas no segundo momento.

O método de estimação da quase-verosimilhança é robusto, no sentido em que é evitada a especificação (completa) da distribuição. Mas, por outro lado, pode haver uma perda de eficiência dos estimadores obtidos desta forma. Cox e Hinkley (1968) verificaram que em algumas condições estes estimadores podem ser tão eficientes como os estimadores da máxima verosimilhança. Sabe-se que o estimador da máxima quase-verosimilhança é ainda muito eficiente na presença de uma «modesta» sobredispersão.

Resta ainda acrescentar que devido à presença de ϕ desconhecido (situação análoga à análise da variância), a distribuição da *deviance* à escala passa a relacionar-se com a distribuição *F-Snedcor* e não com a qui-quadrado. Neste contexto torna-se importante o teorema 2.

O resíduo de Pearson, nesta situação, toma a expressão:

$$r_p = \frac{Y - \hat{\mu}}{\sqrt{\hat{\phi}\hat{\mu}}}$$

6 — Caracterização de casos de acidentes de trabalho em Portugal segundo a actividade económica e sexo

Nesta parte, pretende modelizar-se a frequência de acidentes de trabalho ocorridos em Portugal. O estudo está dividido em duas fases; numa primeira fase modelizam-se dados que se referem ao número de acidentes de trabalho ocorridos segundo os ramos de actividade económica e sexo. Nesta fase procuram conhecer-se as relações existentes entre cada actividade e a perigosidade das mesmas e de que forma o factor sexo influencia esta situação. Numa segunda fase vai estudar-se se há, ou não, alteração significativa da situação desde há 10 anos atrás.

No quadro 1 apresentam-se os casos registados de acidentes de trabalho por ramos de actividade económica e sexo, no ano de 1992. Os dados apresentados são provenientes do Instituto Nacional de Estatística, tendo-se utilizado os mais recentes disponíveis.

A variável resposta Y_{rj} representa o número de acidentes de trabalho do indivíduo j da célula r no ano de 1992 e as variáveis explicativas são factores — ramo de actividade económica e sexo. O factor ramo de actividade económica tem nove níveis e representa-se por *ACT*; assim *ACT(1)* representa «Agricultura, silvicultura, caça e pesca», *ACT(2)* representa «Indústrias extractivas» e assim sucessivamente. O factor sexo tem dois níveis e representa-se pelo mesmo nome. Convencionou-se que *SEXO(1)* representa o masculino e *SEXO(2)* o feminino, notação que se mantém em todos os modelos.

Dada a natureza dos dados em estudo, resultantes de contagens, considerou-se a distribuição de Poisson e optou-se pelos modelos loglineares en-

QUADRO 1

Acidentes de trabalho segundo a actividade económica e sexo no ano de 1992

Ramo de actividade económica	Sexo	Número de empregados	Número de acidentes de trabalho
1) Agricultura, silvicultura, caça e pesca	HM	522 655	15 390
	H	264 938	12 250
	M	257 717	3 140
2) Indústrias extractivas	HM	22 328	2 904
	H	21 328	2 850
	M	1 000	54
3) Indústrias transformadoras	HM	1 072 533	136 260
	H	611 534	111 786
	M	460 999	24 474
4) Electricidade, água e gás	HM	33 692	2 146
	H	25 746	2 035
	M	7 946	111
5) Construção e obras públicas	HM	370 206	53 044
	H	353 834	2 266
	M	16 372	778
6) Comércio por grosso e a retalho, restaurantes e hotéis	HM	774 487	29 471
	H	414 090	20 910
	M	360 397	8 561
7) Transportes, armazenagem e comunicações	HM	219 498	10 772
	H	167 282	9 941
	M	52 216	831
8) Bancos, seguros e outros	HM	286 582	3 703
	H	175 270	2 477
	M	111 312	1 226
9) Serviços prestados à colectividade, serviços sociais e pessoais	HM	1 116 659	16 363
	H	404 436	10 255
	M	712 223	6 108

Fonte: *Inquérito ao Emprego 93, Anuário Estatístico de Portugal 93*, e *Inquérito ao Emprego Açores e Madeira 92* (INE).

quadrados nos modelos lineares generalizados, já retratados aqui. Para a análise utilizou-se o programa GLIM.

Podia parecer que a distribuição adequada fosse a binomial, pois cada acidente teria uma probabilidade p de ocorrer em N observações, no entanto é possível que cada indivíduo tenha mais de um acidente em cada ano, pelo que a distribuição binomial não é adequada, mas sim a de Poisson.

Trata-se ainda de um modelo designado por «agregado» [v. Andrade e Silva (1991)], pois em cada célula r da tabela de contingência regista-se o número total de acidentes ocorridos para n_r indivíduos, sendo n_r conhecido.

As variáveis aleatórias Y_{rj} são independentes e seguem a lei de Poisson com média μ_r, verificando-se que $\ddot{Y}_R = \sum_{j=1}^{n_r} Y_{rj}$, $r = 1, 2, ..., R$ tem distribuição de Poisson com média $\theta_r = n_r \mu_r$. Tem-se então:

$$\begin{aligned} \log \theta_r &= \log n_r + \log \mu_r \\ &= \log n_r + \beta^T \ddot{x}_R \end{aligned}$$

Note-se que as observações têm diferentes pesos conforme o número indicado por n_r, $r = 1, 2, ..., R$, havendo algumas com mais peso do que outras, facto que está já controlado no modelo com a introdução do termo $\log n_r$ no preditor linear.

Posto isto, vai começar por ajustar-se uma sequência de modelos que se resume no quadro 2 ([4]).

QUADRO 2

Análise da *deviance*

Preditor linear	D^*	g.l.	$\hat{\phi}$ (D/g.l.)
1	237 660	17	13 980
1 + SEXO	141 963	16	8 872.69
1 + ACT	56 665	9	6 296.11
1 + SEXO + ACT	1 799.7	8	224.96
SEXO ★ ACT (modelo saturado)	0	0	–

Na coluna da direita indicam-se as estimativas de ϕ. Pode observar-se que essas estimativas são muito superiores ao valor 1, o que indica que há sobredispersão no modelo. Assim, existindo o problema de sobredispersão, vai admitir-se apenas que $E(Y) = \mu$ e $\mathrm{var}(Y) = \phi\mu$ com $\phi > 1$, abandonando a lei de Poisson. As estimativas para os parâmetros do modelo podem obter-se da mesma forma usando o modelo loglinear de Poisson e maximizando a respectiva logverosimilhança. Os erros padrões são iguais aos correspondentes valores obtidos no modelo de Poisson, multiplicados por uma estimativa de $\sqrt{\phi}$ (recorde-se o teorema 3).

Devido à presença de ϕ desconhecido, vai alterar-se o quadro apresentado para se relacionar com a distribuição *F-Snedcor*. Constrói-se o quadro ANODEV adaptado:

QUADRO 3

ANODEV com sobredispersão

Preditor linear	D^*	g.l.	Valor de F	Hipótese a considerar (5 %)
1	237 660	17	–	—
1 + SEXO	141 963	16	10.79	Rejeitar H_0.
1 + ACT	56 665	9	3.59	Rejeitar H_0.
1 + SEXO + ACT	1 799.7	8	77.88	Rejeitar H_0.
SEXO ★ ACT	0	0	–	—

([4]) Manteve-se aqui a linguagem GLIM; «1» representa o *intercept*, «1 + SEXO» significa que se introduz o factor SEXO com dois níveis no preditor linear e «SEXO★ACT» representa o modelo saturado.

Do quadro pode notar-se que ambos os factores *SEXO* e *ACT* são significativos.

Posto isto, o modelo com melhor ajustamento é:

$$1 + SEXO + ACT$$

isto é, o modelo de efeitos principais. As estimativas são:

	Estimativa	e.p.	*Covariates*
1	− 3.118	0.1224	1.
2	1.109	0.3040	ACT(2).
3	1.400	0.1276	ACT(3).
4	0.5388	0.3458	ACT(4).
5	1.206	0.1385	ACT(5).
6	0.2285	0.1492	ACT(6).
7	0.2797	0.1888	ACT(7).
8	− 0.9250	0.2743	ACT(8).
9	− 0.5384	0.1687	ACT(9).
10	− 1.133	0.07968	SEXO(2).

Parâmetro dispersão estimado em 225.0.

Por observação das estimativas e respectivos erros padrões, vai tentar simplificar-se o preditor linear. Há várias estimativas não significativas, 4, 6 e 7 que indicam não serem os respectivos parâmetros significativamente diferentes de zero.

Vai experimentar eliminar-se do preditor linear a *ACT(4)*. A *deviance* aumenta para 2 276.8 com 9 graus de liberdade ($\Delta D^* = +477.1$, Δg.l. = + 1), sendo então $F = 2.12$ para $F_{1,8} = 5.32$ (5 %), aceitando-se o modelo mais simples.

Analisando as estimativas do modelo resultante, desta vez com $\widehat{\phi} = 253$, verifica-se que o coeficiente estimado da actividade 6 não é significativo. Ao igualá-lo a zero a *deviance* aumenta para 2 622.8 com 10 graus de liberdade ($\Delta D^* = +346$, Δg.l. = +1) sendo $F = 1.37$ com $F_{1,9} = 5.12$ (5 %), pelo que se aceita o modelo mais simples.

As estimativas para o presente modelo são:

	Estimativa	e.p.	*Covariates*
1	− 2.958	0.07708	1.
2	− 1.137	0.08598	SEXO(2).
3	0.9494	0.3102	ACT(2).
4	1.240	0.08665	ACT(3).
5	1.046	0.1041	ACT(5).
6	0.1201	0.1734	ACT(7).
7	− 1.084	0.2761	ACT(8).
8	− 0.6970	0.1475	ACT(9).

Parâmetro dispersão estimado em 262.3.

Suprimindo a actividade 7 do preditor linear, obtém-se $D^* = 2\,745.6$ ($\Delta D^* = +122.8$) com 11 graus de liberdade (Δg.l. = + 1), sendo $F = 0.47$ com $F_{1,10} = 4.96$, prosseguindo a simplificação. Todas as estimativas são, agora, significativas.

	Estimativa	e.p.	*Covariates*
1	− 2.936	0.06780	1.
2	− 1.141	0.08370	SEXO(2).
3	0.9274	0.3008	ACT(2).
4	1.219	0.07843	ACT(3).
5	1.024	0.09625	ACT(5).
6	− 1.106	0.2675	ACT(8).
7	− 0.7177	0.1405	ACT(9).

Parâmetro dispersão estimado em 249.6.

Analisando as estimativas do modelo em mão:

$$1 + SEXO + ACT(2) + ACT(3) + ACT(5) + ACT(8) + ACT(9)$$

nota-se que as estimativas dos coeficientes da actividade 5 e actividade 8 são muito semelhantes, mas simétricas. Assim, vai tentar juntar-se o coeficiente de *ACT(5)* ao de *ACT(8)*. Obtém-se:

$D^* = 2764.5$ ($\Delta D^* = +18.9$);
g.l. = 12 (Δg.l. = + 1);
$F = 0.08$;
$F_{1,11} = 4.84$ (5 %);

aceitando-se a simplificação.

O preditor linear fica então:

$$1 + SEXO + ACT(2) + ACT(3) + ACT(58) + ACT(9)$$

cujas estimativas são:

	Estimativa	e.p.	*Covariates*
1	− 2.945	0.05833	1.
2	− 1.140	0.08027	SEXO(2).
3	0.9357	0.2876	ACT(2).
4	1.227	0.06995	ACT(3).
5	− 0.7098	0.1323	ACT(9).
6	1.037	0.08095	ACT(58).

Parâmetro dispersão estimado em 230.4.

Verifica-se que as estimativas dos coeficientes de *ACT(2)* e *ACT(58)* pouco diferem e tenta-se igualá-los. Obtém-se:

$D^* = 2793.5$ $(\Delta D^* = +29)$;
g.l. = 13 (Δg.l. = + 1);
$F = 0.13$;
$F_{1,12} = 4.75$ (5 %);

e aceita-se o modelo:

$$1 + SEXO + ACT(258) + ACT(3) + ACT(9)$$

com:

	Estimativa	e.p.	*Covariates*
1	− 2.945	0.05630	1.
2	− 1.139	0.07752	SEXO(2).
3	1.228	0.06753	ACT(3).
4	− 0.7093	0.1277	ACT(9).
5	1.032	0.07720	ACT(258).

Parâmetro dispersão estimado em 214.9.

Uma vez que os coeficientes estimados de *ACT(3)* e *ACT(258)* têm aproximadamente o mesmo valor tenta-se igualá-los. Obtém-se:

$D^* = 4355.5$ $(D^* = +1562)$;
g.l. = 14 (g.l. = + 1);
$F = 7.27$;
$F_{1,13} = 4.67$ (5 %);

que está na região de rejeição. No entanto pode aceitar-se a simplificação ao nível de 1 % ($F_{1,13} = 9.07$), pelo que vai simplificar-se o modelo.

O preditor linear é, então:

$$1 + SEXO + ACT(2358) + ACT(9)$$

cujas estimativas e resíduos se apresentam em seguida:

	Estimativa	e.p.	*Covariates*
1	− 2.939	0.06814	1.
2	− 1.100	0.09193	SEXO(2).
3	− 0.7296	0.1536	ACT(9).
4	1.153	0.07502	ACT(2358).

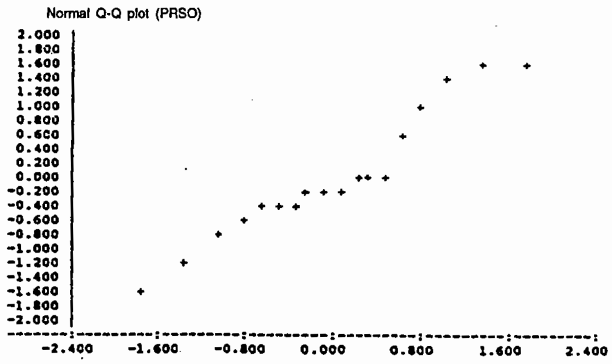
Parâmetro dispersão estimado em 311.1.

Unid.	Observado	Ajustado	Resíduo
1	12250	14020.19	− 0.848
2	2850	3574.69	− 0.687
3	111786	102496.53	1.645
4	2035	1362.45	1.033
5	52266	59304.55	− 1.639
6	20910	21913.13	− 0.384
7	9941	8852.35	0.656
8	2477	2928.45	− 0.473
9	10255	10317.64	− 0.035
10	3140	4537.62	− 1.176
11	54	55.77	− 0.013
12	24474	25707.72	− 0.436
13	111	139.90	− 0.139
14	778	912.99	− 0.253
15	8561	6345.50	1.577
16	831	919.37	− 0.165
17	1226	618.80	1.384
18	6108	6045.35	0.046

Vai fazer-se o teste de normalidade de Filliben aos resíduos, adaptado ao caso de sobredispersão. Obtém-se:

GRÁFICO 1

Representação gráfica dos resíduos padrões ordenados contra os quantis da *N*(0,1) [teste de Filliben (1975)]



O coeficiente de Filliben tem o valor de 0.9727 tendo, para 18 observações, um nível de significância de cerca de 40 %, optando-se por aceitar este modelo.

Os coeficientes das actividades 4, «Electricidade, água e gás», 6, «Comércio por grosso e a retalho, restaurantes e hotéis», e 7, «Transportes, armazenagem e comunicações», podem ser igualados a zero, ficando equivalentes à actividade 1, «Agricultura, silvicultura, caça e pesca», concluindo-se que estas actividades têm comportamentos semelhantes em relação à ocorrência de acidentes de trabalho.

A actividade 9, «Serviços prestados à colectividade, serviços sociais e pessoais», apresenta comportamento diferente, tendo as actividades 2, «Indústrias extractivas», 3, «Indústrias transformadoras», e 5, «Construção e obras públicas», revelado serem muito semelhantes.

Quanto à actividade 8, «Bancos seguros e outros», apesar de se ter juntado às actividades 2, 3 e 5, convém não esquecer que tem efeito simétrico ao destas últimas.

As actividades aparecem no quadro seguinte por ordem, desde as menos perigosas até às mais perigosas.

QUADRO 4

Resumo das estimativas dos parâmetros

Covariates	Parâmetros correspoondentes a	Estimativa
1	Intercept	– 2.939
ACT(8)	Bancos, seguros e outros	– 1.153
ACT(9)	Serviços prestados à colectividade, serviços sociais e pessoais	– 0.7296
ACT(1) ACT(4) ACT(6) ACT(7)	Agricultura, silvicultura, caça e pesca Electricidade, água e gás Comércio por grosso e a retalho, restaurante e hotéis Transortes, armazenagem e comunicações	0.00
ACT(2) ACT(3) ACT(5)	Indústria extractivas Indústrias transormadoras Construção e obras públicas	1.153
SEXO(1)	Homens	0.00
SEXO(2)	Mulheres	– 1.10

Verifica-se que o sexo 2 (feminino), tem muito menos acidentes do que o sexo masculino, em todas as actividades [note-se no preditor linear *0.00SEXO(1)* e *–1.10SEXO(2)*], e o modelo indica que as mulheres têm menos 67 % de acidentes de trabalho do que os homens (note-se $e^{-1.1} = 0.33$).

O *intercept* tem valor negativo, indicando obviamente que há menos acidentes de trabalho do que indivíduos activos.

Observa-se ainda que o parâmetro de dispersão foi estimado em 311.1 representando um valor muito elevado, o que indica que a variância é muito superior ao valor médio.

É interessante comparar esta situação com a de alguns anos atrás, como por exemplo em 1981. Vai analisar-se se há, ou não, alteração da perigosidade das actividades económicas e do factor sexo, desde há cerca de 10 anos. Para isso vai ajustar-se um modelo que represente a evolução que ocorre desde 1981 até 1992.

Os dados de 1981 são os seguintes:

QUADRO 5

Acidentes de trabalho segundo a actividade económica e sexo no ano de 1981

Ramo de actividade económica	Sexo	Número de empregados	Número de acidentes de trabalho
1) Agricultura, silvicultura, caça e pesca	HM	693 423	17 407
	H	438 359	14 476
	M	255 064	2 931
2) Indústrias extractivas	HM	117 915	3 615
	H	77 578	3 539
	M	40 337	76
3) Indústrias transformadoras	HM	880 302	141 245
	H	578 323	127 349
	M	301 979	13 896
4) Electricidade, água e gás	HM	27 688	3 352
	H	24 809	3 167
	M	2 879	185
5) Construção e obras públicas	HM	417 510	49 945
	H	408 345	49 555
	M	9 165	390
6) Comércio por grosso e a retalho, restaurantes e hotéis	HM	494 822	14 150
	H	312 722	10 588
	M	182 100	3 562
7) Transportes, armazenagem e comunicações	HM	181 715	14 049
	H	152 815	13 572
	M	28 900	477
8) Bancos, seguros e outros	HM	98 381	1 365
	H	71 562	1 029
	M	26 819	336
9) Serviços prestados à colectividade, serviços sociais e pessoais	HM	767 956	13 003
	H	355 191	10 143
	M	412 765	2 860

Fonte: *Recenseamento Geral da População 81, Anuário Estatístico de Portugal 87.*

Considerou-se um factor *ANOS*, com dois níveis, onde se convencionou que *ANOS(1)* representa o ano de 1981 e *ANOS(2)*, o ano de 1992.

Há sobredispersão no modelo. A sequência de ajustamento resume-se no quadro seguinte:

QUADRO 6

ANODEV com sobredispersão

Preditor linear	D^*	g.l.	Valor de F	Hipótese a considerar (5 %)
1	486 084	35	–	—
1 + SEXO	302 780	34	20.58	Rejeitar H_0.
1 + ACT	128 285	27	9.41	Rejeitar H_0.
1 + ANOS	483 581	34	0.18	Aceitar H_0.
1 + SEXO + ANOS	302 725	33	0.006	Aceitar H_0.
1 + ACT + ANOS	127 469	26	0.17	Aceitar H_0.
1 + ACT + SEXO	13 986	26	212.48	Rejeitar H_0.
1 + ACT + SEXO + ANOS	13 983	25	0.005	Aceitar H_0.
1 + ACT + SEXO + ACT.ANOS	5 896.8	17	2.59	Rejeitar H_0.
1 + ACT + SEXO + SEXO.ANOS	13 337	24	0.58	Aceitar H_0.
ACT ★ ANOS + SEXO ★ ANOS	5 000.9	16	2.87	Aceitar H_0.

Não há alterações importantes ao longo dos anos, pois o efeito do factor anos foi sempre rejeitado. No entanto, o modelo seleccionado é:

$$1 + ACT + SEXO + ACT.ANOS \text{ (}^5\text{)}$$

existindo uma interacção entre o factor actividade e o factor anos. As suas estimativas são:

	Estimativa	e.p.	*Covariates*
1	– 3.382	0.1416	1.
2	0.1755	0.3399	ACT(2).
3	1.831	0.1496	ACT(3).
4	1.348	0.3514	ACT(4).
5	1.275	0.1643	ACT(5).
6	0.1305	0.2108	ACT(6).
7	0.9422	0.2113	ACT(7).
8	– 0.6804	0.5226	ACT(8).
9	– 0.2161	0.2160	ACT(9).
10	– 1.236	0.07763	SEXO(2).
11	0.2877	0.2061	ACT(1).ANOS(2).
12	1.199	0.4639	ACT(2).ANOS(2).
13	– 0.1486	0.07082	ACT(3).ANOS(2).
14	– 0.5358	0.5149	ACT(4).ANOS(2).
15	0.1966	0.1161	ACT(5).ANOS(2).
16	0.3838	0.1905	ACT(6).ANOS(2).
17	– 0.3894	0.2385	ACT(7).ANOS(2).
18	0.03622	0.5887	ACT(8).ANOS(2).
19	– 0.02232	0.2189	ACT(9).ANOS(2).

Parâmetro dispersão estimado em 346.9.

(5) Podia ter-se analisado o modelo equivalente 1 + ACT + SEXO + ANOS + ACT.ANOS.

Pode observar-se que apenas a interacção com a actividade 2 é significativa, sugerindo que nas outras actividades não há alterações significativas na taxa de acidentes registados desde 1981 até 1992. Tal não é de estranhar, pois a interacção *ACT.ANOS* pode rejeitar-se ao nível de 2,5 %, uma vez que $F_{9,17} = 2.98$. No entanto há uma alteração significativa ao longo destes anos na actividade 2, «Indústrias extractivas».

Como os parâmetros correspondentes às estimativas não significativas são, com grande probabilidade, zero pode ir-se simplificando sucessivamente o modelo eliminando as interacções não significativas.

QUADRO 7

Simplificação do modelo

Preditor linear	D^*	g.l.	Valor de F	Hipótese a considerar (5 %)
1 + ACT + SEXO + ACT.ANOS	5 896.8	17	–	—
1 + ACT + SEXO + ACT.ANOS .... – ACT(8)ANOS(2)	5 898.1	18	0.004	Aceitar H_0.
1 + ACT + SEXO + ACT.ANOS .... – ACT(8)ANOS(2) – ACT(9)ANOS(2)	5 901.7	19	0.01	Aceitar H_0.
1 + ACT + SEXO + ACT.ANOS .... – ACT(8)ANOS(2) – ACT(9)ANOS(2) – ACT(4)ANOS(2)	6 287.8	20	1.24	Aceitar H_0.
1 + ACT + SEXO + ACT.ANOS .... – ACT(8)ANOS(2) – ACT(9)ANOS(2) – ACT(4)ANOS(2) – ACT(5)ANOS(2)	7 282.7	21	3.16	Aceitar H_0.
1 + ACT + SEXO + ACT.ANOS .... – ACT(8)ANOS(2) – ACT(9)ANOS(2) – ACT(4)ANOS(2) – ACT(5)ANOS(2) – ACT(1)ANOS(2)	7 951.9	22	1.93	Aceitar H_0.
1 + ACT + SEXO + ACT.ANOS .... – ACT(8)ANOS(2) – ACT(9)ANOS(2) – ACT(4)ANOS(2) – ACT(5)ANOS(2) – ACT(1)ANOS(2) – ACT(7)ANOS(2)	8 886.8	23	2.58	Aceitar H_0.
1 + ACT + SEXO + ACT.ANOS .... – ACT(8)ANOS(2) – ACT(9)ANOS(2) – ACT(4)ANOS(2) – ACT(5)ANOS(2) – ACT(1)ANOS(2) – ACT(7)ANOS(2) – ACT(6)ANOS(2)	10 350	24	3.79	Aceitar H_0.
1 + ACT + SEXO + ACT.ANOS .... – ACT(8)ANOS(2) – ACT(9)ANOS(2) – ACT(4)ANOS(2) – ACT(5)ANOS(2) – ACT(1)ANOS(2) – ACT(7)ANOS(2) – ACT(6)ANOS(2) – ACT(3)ANOS(2)	11 883	25	3.55	Aceitar H_0.
1 + ACT + SEXO + ACT.ANOS .... – ACT(8)ANOS(2) – ACT(9)ANOS(2) – ACT(4)ANOS(2) – ACT(5)ANOS(2) – ACT(1)ANOS(2) – ACT(7)ANOS(2) – ACT(6)ANOS(2) – ACT(3)ANOS(2) – ACT(2)ANOS(2)	13 986	26	4.42	Rejeitar H_0.

A simplificação resume-se no quadro anterior. A ordem de simplificação corresponde ao menor nível de significância das estimativas.

O presente preditor linear é:

$$1 + ACT + SEXO + ACT(2)ANOS(2)$$

As estimativas são:

	Estimativa	e.p.	*Covariates*
1	– 3.257	0.1214	1.
2	0.05058	0.3815	ACT(2).
3	1.631	0.1273	ACT(3).
4	0.9782	0.3178	ACT(4).
5	1.246	0.1390	ACT(5).
6	0.2482	0.1593	ACT(6).
7	0.6293	0.1837	ACT(7).
8	– 0.7789	0.3285	ACT(8).
9	– 0.3529	0.1755	ACT(9).
10	– 1.239	0.09061	SEXO(2).
11	1.199	0.5430	ACT(2)ANOS(2).

Parâmetro dispersão estimado em 475.3.

Pela análise das estimativas, pode concluir-se que a actividade 2, «Indústrias extractivas», sofreu um aumento muito grande da taxa de acidentes ocorridos em 1992, em relação a 1981 (note-se *0.00ACT(2)ANOS(1)* e *1.199ACT(2)ANOS(2)*). Nas outras actividades não há alterações significativas.

O factor sexo não sofreu alterações nestes 10 anos pois não aparece a interacção *SEXO.ANOS* e note-se que o efeito linha de *ANOS* é sempre rejeitado no modelo de efeitos principais.

Comparando com os dados apresentados pelo INE, de facto, a percentagem de acidentes nas «Indústrias extractivas» aumentou de 3 % em 1981, para 13 % em 1992, não havendo tão grandes diferenças nas restantes actividades.

Note-se que o modelo podia simplificar-se ainda mais, mas o objectivo é apenas estudar a evolução ao longo dos anos, o que já foi feito.

7 — Conclusões

Do estudo realizado verifica-se que há sobredispersão em todos os modelos apresentados com os factores considerados, concluindo-se que o fenómeno ocorrência de acidentes de trabalho é essencialmente aleatório, não sendo explicado unicamente pelos factores actividade e sexo. No entanto há uma certa estabilidade neste fenómeno, pois não há alterações significativas ao longo dos anos segundo estes factores, concluindo-se que, de facto, os factores sexo e actividade económica contribuem de alguma forma para a explicação das taxas registadas de acidentes de trabalho.

Quanto à perigosidade dos factores, concluiu-se que o sexo feminino sofre menos acidentes do que o sexo masculino e que as actividades económicas mais perigosas são as «Indústrias extractivas», «Indústrias transformadoras» e «Construção e obras públicas» e as de menor risco são os «Bancos, seguros e outros» e os «Serviços prestados à colectividade, serviços sociais e pessoais».

REFERÊNCIAS

AITKIN, M., ANDERSON, D., FRANCIS, B., e HINDE, J. (1989), *Statistical Modelling in GLIM*, Claredon Press, Oxford.

ANDRADE E SILVA, J. M. (1991), *Estruturas Tarifárias nos Ramos Reais da Indústria Seguradora*, tese de doutoramento em Economia, Instituto Superior de Economia e Gestão, Lisboa.

COX, D. R. and HINKLEY, D. V. (1968), «A note on the efficiency of least-squares estimates», *Journal of Royal Statistical Society*, B30, pp. 284-289.

FILLIBEN, J. J. (1975), «The Probability Plot Correlation Coefficient Test for Normality», *Technometrics*, **17**, n.º 1, pp. 111-117.

HINKLEY, D. V., REID, N. e SNELL, E. J. (1991), *Statistical Theory and Modelling*, Chapman & Hall, London.

McCULLAGH, P. (1984), «Generalized Linear Models», *European Journal of Operacional Research*, 16, pp. 285-292.

McCULLAGH, P., e NELDER, J. A. (1989),: *Generalized Linear Models*, 2.ª ed., Chapman & Hall, London.

NAG (Numerical Algorithms Group) (1987), *The GLIM System Release 3.77 Manual* (ed. C. D. Payne), 2.ª ed., NAG, Oxford.

NELDER, J. A., e WEDDERBURN, R. W. M. (1972), «Generalized Linear Models», *Journal of Royal Statistical Society*, A 135, pp. 370-384.

PIERCE, D. A., e SCHAFER, D. W. (1986), «Residuals in Generalized Linear Models», *Journal of the American Statistical Association*, **81**, n.º 396, pp. 977-986.

SECO, M. A. (1995), *Modelos loglineares — Uma aplicação a casos de acidentes de trabalho em Portugal*, dissertação de mestrado em Matemática Aplicada à Economia e Gestão, Instituto Superior de Economia e Gestão, Lisboa.

WEDDERBURN, R. W. M. (1974), «Quasilikelihood functions, generalized linear models and the Gauss-Newton method», *Biometrika*, 61, pp. 439-447.

—— (1976), «On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models», *Biometrika*, 63, pp. 27-32.