Epidemiologia de Doenças Transmissíveis – Aulas Práticas

Módulo 18 – Regressão Logística

1. A tabela seguinte classifica 100 homens hospitalizados quanto à idade (≤ 55 anos=0; >55 anos=1) e quanto a evidência de terem tido sinais de doença coronária (CHD) (sim=1; não=0).

índividuo	Idade	CHD	índividuo	Idade	CHD	índividuo	Idade	CHD	índividuo	Idade	CHD	índividuo	Idade	CHD
1	0	0	21	0	1	41	0	0	61	0	0	81	1	1
2	0	0	22	0	1	42	0	0	62	0	0	82	1	1
3	0	0	23	0	0	43	0	0	63	0	0	83	1	1
4	0	0	24	0	0	44	0	0	64	0	0	84	1	1
5	0	0	25	0	1	45	0	1	65	0	1	85	1	0
6	0	1	26	0	0	46	0	0	66	0	1	86	1	1
7	0	1	27	0	1	47	0	0	67	0	0	87	1	1
8	0	0	28	0	0	48	0	0	68	0	0	88	1	0
9	0	1	29	0	0	49	0	0	69	0	0	89	1	0
10	0	0	30	0	0	50	0	0	70	0	0	90	1	1
11	0	0	31	0	1	51	0	1	71	0	0	91	1	1
12	0	0	32	0	1	52	0	0	72	0	0	92	1	1
13	0	1	33	0	0	53	0	1	73	0	0	93	1	1
14	0	0	34	0	0	54	0	0	74	1	1	94	1	0
15	0	0	35	0	0	55	0	0	75	1	0	95	1	1
16	0	1	36	0	0	56	0	0	76	1	1	96	1	1
17	0	1	37	0	0	57	0	1	77	1	1	97	1	1
18	0	0	38	0	1	58	0	0	78	1	1	98	1	1
19	0	0	39	0	0	59	0	0	79	1	1	99	1	1
20	0	1	40	0	1	60	0	1	80	1	0	100	1	1

a) Organize a mesma informação numa tabela de contingência 2 x 2.
b) A tabela seguinte apresenta parte do output computacional com os resultados da análise por regressão logística destes dados, sendo SE o erro padrão:

	b	*SE*
constante	-0.841	0.255
Idade	2.094	0.529

Escreva o modelo de regressão logística, tomando a ocorrência de doença como variável dependente e a idade como factor de risco.

c) Calcule o odds e o risco de ter doença CHD, quer a partir da tabela 2x2, quer a partir do modelo de regressão logística, para um homem hospitalizado com mais de 55 anos.
d) Qual o OR de patologia dos maiores de 55 anos, comparativamente com os mais novos ? Compare com o OR calculado a partir da tabela 2x2.
e) Repita a alínea anterior, para o RR.
f) Construa um IC de 95% para o OR, usando os resultados computacionais
g) Use o teste de Wald para testar a significância do coeficiente b, interpretando o resultado.

2. A tabela seguinte apresenta as frequências de determinado alelo em culturas de quatro estirpes de uma bactéria

alelo ?	estirpe 1	estirpe 2	estirpe 3	estirpe 4	total
presente	20	15	10	5	50
ausente	10	10	10	20	50
	30	25	20	25	100

A tabela seguinte apresenta as estimativas dos coeficientes de uma regressão logística aplicada aos mesmos dados, tomando a estirpe 4 como referência.

variável	*b*	*SE*
constante	-1.386	0.5
estirpe 1	2.079	0.633
estirpe 2	1.792	0.646
estirpe 3	1.386	0.671

a) Escreva o correspondente modelo de regressão logística, tomando a estirpe como variável independente.
b) Qual o odds e o risco da estirpe 1 ter o alelo ? e se for a estirpe 4 ?
c) Qual o OR da estirpe 1? E se for da estirpe 4 ?
d) Qual o odds ratio da estirpe 1 ter o alelo, relativamente à estirpe 2 ?
e) Construa um IC a 95% para o OR da estirpe 1

3. A tabela seguinte apresenta os resultados do seguimento durante 7.7 anos de uma coorte fixa de 4095 homens escoceses de meia-idade que, à partida, não tinham sinal de doença CHD. Foi registada a pressão arterial máxima (SBP) destes homens, bem como o colesterol total no soro sanguíneo. Os dados da tabela registam o número de casos de CHD que entretanto surgiram, relativamente ao total de homens (casos/total), organizados por 5 classes de CHD e de SBP.

	Colesterol total no sangue (mmol/l)				
SBP (mmHg)	<5.42	5.42-6.01	6.02-6.56	6.57-7.31	>7.31
<119	1/190	0/183	4/178	8/157	4/132
119-127	2/203	2/175	6/167	10/166	11/137
128-136	5/173	9/176	9/181	8/167	11/164
137-148	5/139	3/156	10/154	13/174	16/174
>148	5/123	8/123	12/144	13/179	23/180

A tabela seguinte, por outro lado, apresenta as estimativas dos coeficientes do modelo e a análise de deviances do mesmo.

Coeficiente	*b*	*Modelo*	deviance	gdl
constante	-4.5995	constante	84.83	
SBP1	0.0000	constante+SBP	56.73	
SBP2	0.6092	constante+colesterol	49.48	
SBP3	0.8697	constante+SBP+colesterol	18.86	
SBP4	1.0297			
SBP5	1.3425			
Colesterol1	0.0000			
Colesterol2	0.2089			
Colesterol3	0.8229			
Colesterol4	1.0066			
Colesterol5	1.2957			

a) Escreva o modelo completo, com os valores numéricos estimados para os coeficientes.
b) Qual o odds e o risco de CHD para um homem no nível mais alto de tensão e de colesterol ?
c) Qual o odds ratio e o RR para um homem nos níveis mais elevados de SBP e colesterol relativamente a um homem nos níveis mais baixos das mesmas variáveis ?
d) Complete o número de gdl's que faltam na tabela de resultados.
e) O modelo com as duas variáveis ajusta-se satisfatoriamente aos dados ?
f) Teste a significância da pressão arterial (e apenas da pressão) como factor explicativo da doença.
g) Teste a significância do colesterol (e apenas do colesterol) como factor explicativo da doença.
h) Teste a significância da tensão arterial e do colesterol (em conjunto) como factores explicativos da doença.
i) Teste a significância de acrescentar o colesterol a um modelo que já tenha a tensão arterial em consideração.

4. No estudo do exercício anterior foram na realidade usados 4 factores de risco. Desta vez a tensão arterial (SBP) e o colesterol (Col) foram tratados como variáveis contínuas. Foi também usado o índice de massa corporal (BMI, em Kg/m^2) e a idade. O modelo final estimado foi,

Logit = -10.1076 + 0.0171 Idade + 0.3071 Col + 0.0417 BMI + 0.0204 SBP + 0.3225 T

a) Qual o significado do coeficiente da idade b_1=0.0171 ?
b) Qual o odds e o risco de doença de um homem não fumador com 50 anos de idade, 6 mmol/l de colesterol, uma tensão de 125 mmHg e um índice BMI de 25Kg/m^2 ?
c) Considere-se o mesmo homem, mas agora com mais 10 anos e uma tensão de 150. Qual o seu odds de doença relativamente ao da alínea anterior ?
d) É legítimo dizer que o teor em colesterol é a variável "mais importante" na explicação do risco de doença ?
e) Suponha que o desvio-padrão das idades dos homens na amostra é de 15 anos, enquanto desvio-padrão do colesterol é de 3 mmol/l. Compare a importância relativa das duas variáveis.
f) Qual o OR correspondente ao aumento de 1 ano de Idade, ajustado para o efeito de confundimento de todas as outras variáveis ?

5. Durante 12 anos acompanharam-se 200 homens que tinham idade superior a 59 anos, com o objectivo de investigar se o nivel socio-económico (SOC), avaliado por uma variável (0/1), está relacionado com a mortalidade por doença cardiovascular (CVD, também uma variável 0/1). Pretende-se controlar confundimento pelos hábitos tabágicos (SMK, variável 0/1) e pela tensão arterial sistólica (SBP, uma variável contínua). Ao analisar os dados decide-se ajustar dois modelos logísticos, ambos tendo CVD como variável dependente, mas com dois conjuntos diferentes de variáveis independentes. As variáveis de cada modelo e respectivos coeficientes estão nesta tabela:

Modelo 1		Modelo 2	
Variável	Coeficiente	Variável	Coeficiente
Constante	-1,800	Constante	-1,190
SOC	-0,520	SOC	-0,500
SBP	0,040	SBP	0,010
SMK	-0,560	SMK	-0,420
SOC x SBP	-0,033		
SOC x SMK	0,175		

a) Escreva o modelo 1 em termos de Logit: i.e. Logit = ?
b) Usando o Modelo 1, estime o risco de morte (CVD=1) para um individuo de classe social elevada (SOC=1), fumador (SMK=1) com SBP=150.
c) Usando o Modelo 2, estime o risco de morte para as seguintes pessoas:
Pessoa 1: SOC=1, SMK=1, SBP=150; Pessoa 2: SOC=0, SMK=1, SBP=150
d) Comparar o risco da Pessoa 1 do exercício c) com o risco obtido em b)
e) Usando o Modelo 2, calcular o RR da Pessoa 1 em relação à Pessoa 2
f) O estudo descrito foi de coortes. Como seria um estudo caso-control equivalente ? Se fosse caso-controlo poder-se-iam ter efectuado os cálculos das alíneas b) e c) ? porquè?
g) Usando o Modelo 2, estime e interprete o OR do efeito do SOC ajustado para SMK e SBP.
h) Escreva uma equação para representar o efeito do SOC ajustado para SMK e SBP no Modelo 1.

6. Verdadeiro ou falso e, se falso, qual a resposta certa ?
a) O termo constante, b_0, de um modelo de RL pode ser interpretado como o log(odds) de ficar doente, como se as variáveis independentes não existissem.
b) O coeficiente b_1 no modelo de RL pode ser interpretado como a mudança no log(odds) correspondente à mudança de 1 unidade da variável X_1, como se as outras variáveis não existissem.
c) Uma vez ajustado um modelo de RL a dados provenientes de um estudo caso-controlo, pode-se calcular o OR e considerar ser o próprio RR, desde que a doença possa ser considerada rara.
d) Considere um modelo de RL com uma variável independente binária (0/1) e sem efeitos de

interacção. O OR desta variável, ajustado para todas as outras, é apenas exp(b_0) onde b_0 é o termo constante no modelo.

e) Considere um modelo de RL com as variáveis independentes idade, tabaco (0/1) e raça (0/1), estando também presente a interação (tabaco x idade), um OR ajustado para o efeito do tabaco estima-se através de exp(b_{tabaco}), sendo b_{tabaco} o coeficiente do tabaco.
f) Quando só há efeitos directos, a equação OR = exp[$\Sigma\ b_i(X_i^*-X_i^0)$] é uma equação geral para o OR que compara dois grupos de valores das variáveis independentes.

7. Com vista a avaliar a associação entre comportamento e infecção por HIV, foi acompanhado um grupo de homens bisexuais, todos com idade entre 20 e 30 anos e à partida seronegativos para o HIV. Ao fim de 1 ano foram testados para o HIV (1=positivo, 0=negativo). Consideraram-se 4 factores de risco: uso consistente de perservativo (PER, variável 0/1); possuirem um ou mais parceiros sexuais em grupos de alto risco (PAR, variável 0/1), o número de parceiros sexuais (PS) e o número médio de contactos sexuais por mês (NMCS). O principal objectivo era investigar se o uso consistente de perservativo, ajustado para as outras variáveis, evitava a infecção por HIV.
a) Neste contexto, escreva um modelo de RL que cumpra o objectivo pretendido, controlando os factores de risco quer para confundimento quer para efeitos de interacção.
b) Usando o modelo que escreveu, escreva uma equação para o OR que compara um indivíduo que não usa perservativo (PER=1) com um que usa (PER=0)

8. Foram seguidos 609 homens durante 9 anos (Georgia, EUA; Kleinbaum and Klein 2002), tendo-se determinado se tinham sintomas de doença coronária (CHD, 1=doentes, 0=normais). As covariaveis medidas nestes homens e respectivas escalas foram: CAT (nivel de cortisol; 1=alto, 0=baixo), Age (contínua), CHL (colesterol, contínua), ECG (electrocardiograma; 1=anormal, 0=normal), SMK (tabaco, 1=sempre, 0=nunca), HPT (hipertensão, 1=hipertenso, 0=normal). Foi ajustado um modelo de RL aos dados. Neste, a deviance foi 357,05 e o output computacional da análise forneceu a seguinte informação,

Variavel	Coeficiente	SE
Constante	-3,9346	1,2503
CAT	-14,0809	3,1227
AGE	0,0323	0,0162
CHL	-0,0045	0,00413
ECG	0,3577	0,3263
SMK	0,8069	0,3265
HPT	0,6069	0,3025
CC=CATxCHL	0,0683	0,0143

a) Quando o modelo é ajustado sem o termo de interacção, a deviance é 400,39. Fazer o teste da razão de verosimelhança para o efeito do termo de interacção do modelo, controlando para as outras variáveis no modelo.
b) Repetir, usando o teste de Wald. Comparar com os resultados da alínea anterior. Justifica-se a presença do termo de interacção no modelo?
c) Escrever uma fórmula para estimar o OR do efeito de CAT sobre a CHD, ajustada para os efeitos de confundimento de todas as outras covariáveis.
d) Usar a fórmula anterior para estimar o OR quando o colesterol toma os valores CHL=220 e CHL=240. Interprete os resultados.

9) Foi efectuado na Austrália um estudo da prevalência de infecções após operações cirúrgicas, o qual envolveu 12742 doentes operados em 265 hospitais. Para cada doente registou-se:
HT: Tipo de hospital (1=publico, 0= privado)
HS: Tamanho do hospital (1=grande, 0= pequeno)

CT: grau de contaminação do local da cirurgia (1=contaminado; 0=limpo)
AGE: idade do doente, SEX (1=mulher, 0=homem).
Foi ajustado um modelo de RL aos dados com o objectivo de estimar a probabilidade de um doente desenvolver uma infecção pós-operatória durante a hospitalização. Nas questões que se seguem, assuma que o tipo de hospital (HT) é a principal variável de exposição e as restantes são covariáveis de controlo.
a) Escreva um modelo hierarquicamente bem formulado (HBF) no qual tenha em consideração a interacção entre, por um lado, HT e, por outro lado, HS e CT.
b) Suponha que adiciona a interacção HT x AGE x SEX ao modelo. Faça-o de forma a que se mantenha um modelo HBF.
c) Reconsidere o modelo da alínea a). Suponha que efectua um teste de Wald para a hipótese nula de que o efeito directo da variável HS é nulo. Este teste é apropriado ? porquê ?
d) Usando o modelo da alínea a), explique como funciona o método de *hierarchical backward elimination*.
e) Suponha que no modelo da alínea a) se conclui que as interacções HTxHS e HTxCT são significativas. Indique quais as variáveis elegíveis para avaliar confundimento.