Lista 4 - MAE0560

Guilherme NºUSP: 8943160 e Leonardo NºUSP: 9793436

Exercício 1

Os dados exibidos na Tabela 1 são de um estudo sobre doença coronária (CHD) em que CAT = nível de *catecholamine* (0 se baixo e 1 se alto), IDADE (0 se < 55 e 1 se ≥ 55 anos) e ECG = eletrocardiograma (0 se normal e 1 se anormal).

Tabela 1: Estudo sobre doença coronária

CAT	Idade	ECG	CHD		Totais
			Sim	Não	
0	0	0	17	257	274
0	1	0	15	107	122
0	0	1	7	52	59
0	1	1	5	27	32
1	0	0	1	7	8
1	1	0	9	30	39
1	0	1	3	14	17
1	1	1	14	34	58

(a) Ajuste um modelo de regressão logística aos dados desse estudo e apresente conclusões. Avalie o efeito das interações duplas.

Resolução

Ajustando o modelo de ressão logística de forma saturada (com todos os parâmentos possíveis), obtemos a seguinte tabela de diferença de *deviances*:

Tabela 2: Modelos ajustados e diferença de *deviances* entre eles

Modelos	g.l.	*Deviances*	TRV	$\neq$ g.l.	$p - value$	AIC
Nulo	7	21.332				52.043
cat	6	7.201	14.131	1	<0.0001	39.912
idade | cat	5	2.477	4.724	1	0.030	37.188
ecg | cat,idade	4	0.954	1.522	1	0.217	37.666
cat*idade | cat,idade,ecg	3	0.922	0.032	1	0.858	39.634
cat*ecg | cat,idade,ecg,cat*idade	2	0.419	0.504	1	0.478	41.130
idade*ecg |cat,idade,ecg,cat*idade,cat*ecg	1	0.003	0.416	1	0.519	42.714

Com a tabela 2, podemos dizer que ao observarmos os p-valores dos modelos e as medidas AIC, com um nível de significância de 5%, rejeitamos os últimos quatro modelos. O modelo mais apropriado para a análise estatística possui apenas efeito de tratamento de idade e de *catecholamine*.

Assim escolhemos o modelo

$$logito_i = -2.54 + 0.774 * cat_i + 0.62 * idade_i$$

Em que obtemos a seguinte saída com o R:

```
##
## Call:
## glm(formula = as.matrix(data[, c(1, 2)]) ~ cat + idade, family = binomial(link = "logit"),
##     data = data)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7
## -0.72188  -0.15477   1.24039   0.47158  -0.17198  -0.14599   0.34611
##        8
##  0.01182
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.5397     0.2005 -12.664  < 2e-16 ***
## cat           0.7736     0.2955   2.618  0.00884 **
## idade         0.6174     0.2840   2.173  0.02975 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21.3320  on 7  degrees of freedom
## Residual deviance:  2.4768  on 5  degrees of freedom
## AIC: 37.188
##
## Number of Fisher Scoring iterations: 4
```

Para avalaliar a qualidade do ajuste temos o teste de qualidade de ajuste cujas a históteses são:

$$\begin{cases} H_0 : modelo\ ajustado\ e\ satisfatorio \\ H_1 : modelo\ ajustado\ nao\ e\ satisfatorio \end{cases}$$

Em que as estatísticas de teste são:

$$Q_p = \sum_{i.j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_m$$

e

$$Q_L = 2 \sum_{i.j} n_{ij} ln \left(\frac{n_{ij}}{e_{ij}} \right) \sim \chi^2_m$$

com n_{ij} as observação com $i = 1, ...s$ e $j = 1, 2$,

e_{ij} as frequências esperadas sob o modelo ajustado e

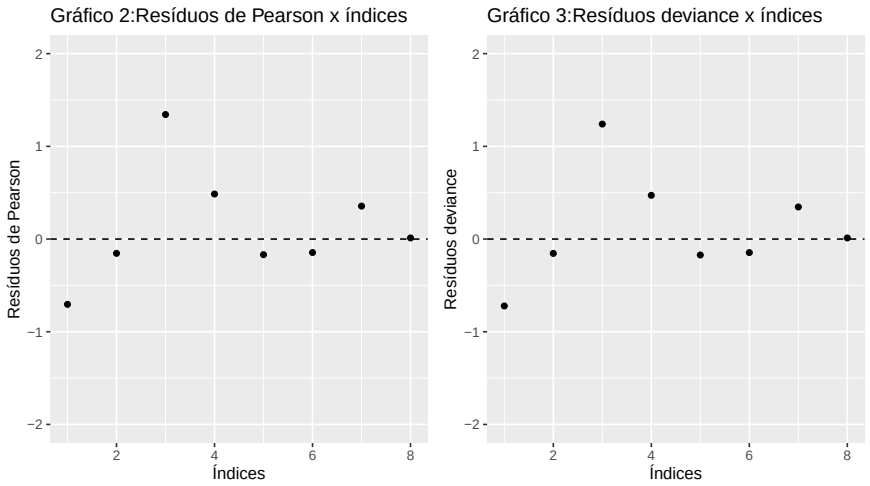
$m =$ nº de subpopulações - nº de parâmetros do modelo ajustado (graus de liberdade)

E obtemos a seguinte tabela:

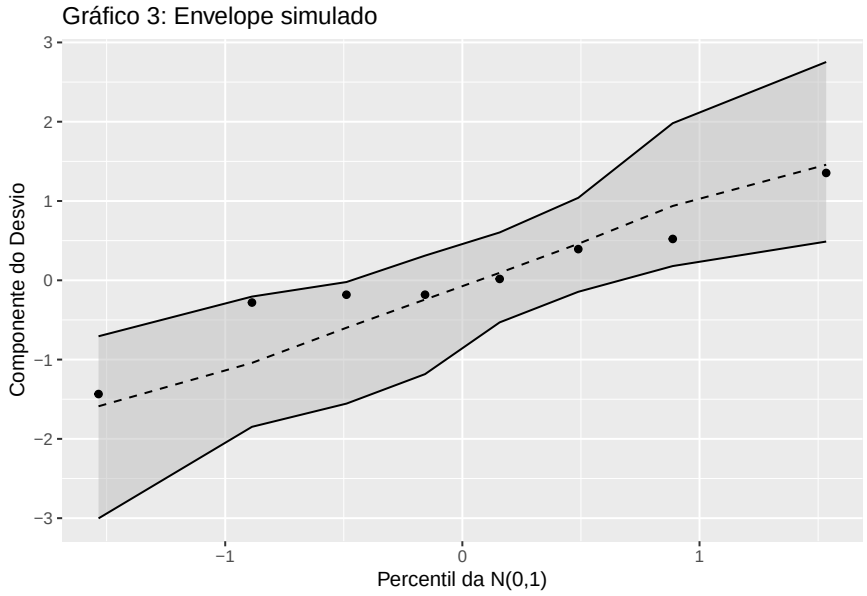
Tabela 3: Teste de qualidade de ajuste

Estatística	Valor	$p - value$
Q_p	2.477	0.780
Q_L	2.736	0.741

Assim, como não rejeitamos H_0, não há evidência de não dizer que o modelo não está bem ajustado e satisfatório.

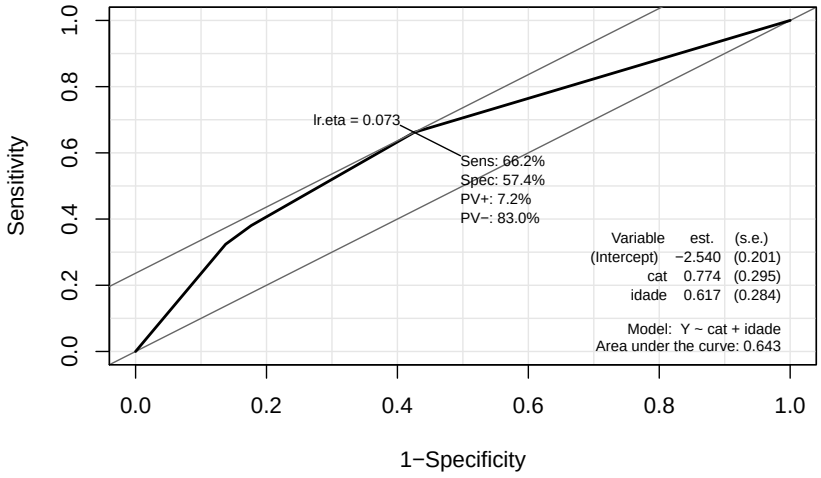


Observando os gráficos 2 e 3, podemos notar que os resíduos estão distribuidos de maneira aleatória como esperado, nos mostrando evidências destes serem independentes.



E a partir do envelope simulado, os pontos estão todos dentro da banda de confiança, logo é possível dizer que o modelo condiz com a distribuição utilizada.

Gráfico 4: Curva ROC



Para finalizar a avaliação da qualidade do ajuste do modelo, pelo gráfico 4, a curva ROC, está acima da reta x=y, e apresentando uma área em baixo da curva ROC de 0.643 o que indica uma classificação melhor que uma classificação aleatória.

Fazendo a interpretação do modelo, temos:

A razão de chances dos pacientes com $cat = 1$ (nível de catecholamine alto) apresentaram chance de doença coronária igual a 2.16 vezes do que a dos pacientes com $cat = 0$ (nível de catecholamine baixo).

$exp(0.773) =$

```
##      cat
## 2.167499
```

A razão de chances dos pacientes com $idade = 1$ (≥ 55 anos) apresentaram chance de doença coronária igual a 1.85 vezes do que a dos pacientes com $idade = 0$ (< 55 anos).

$exp(0.617) =$

```
##    idade
## 1.854041
```

A razão de chances dos pacientes com $cat = 1$ (nível de catecholamine alto) e $idade = 1(>= 55$ anos) apresentaram chance de doença coronária igual a 4 vezes do que a dos pacientes com $cat = 0$ (nível de catecholamine baixo) e $idade = 0(< 55$ anos).

$exp(0.773 + 0.617) = exp(1.39) =$

```
##      cat
## 4.018632
```

(b) Ajuste os modelos probito, clog-log e Cauchy e compare-os em termos de qualidade de ajuste com o modelo de regressão logística.

Resolução

Ajustando o modelo de ressão logística de forma saturada (com todos os parâmentos possíveis) para o modelo com ligação **probito** obtemos a seguinte tabela de diferença de *deviances*:

Tabela 3: Modelos ajustados e diferença de *deviances* entre eles

Modelos	g.l.	*Deviances*	TRV	$\neq$ g.l.	$p-value$	AIC
Nulo	7	21.332				52.043
cat	6	7.201	14.131	1	<0.0001	39.912
idade | cat	5	2.441	4.760	1	0.030	37.152
ecg | cat,idade	4	0.773	1.668	1	0.197	37.485
cat*idade | cat,idade,ecg	3	0.766	0.007	1	0.931	39.477
cat*ecg | cat,idade,ecg,cat*idade	2	0.345	0.421	1	0.517	41.057
idade*ecg |cat,idade,ecg,cat*idade,cat*ecg	1	0.001	0.345	1	0.557	42.712

Agora, ajustando o modelo de ressão logística de forma saturada (com todos os parâmentos possíveis) para o modelo com ligação **cauchy** obtemos a seguinte tabela de diferença de *deviances*:

Tabela 4: Modelos ajustados e diferença de *deviances* entre eles

Modelos	g.l.	*Deviances*	TRV	$\neq$ g.l.	$p-value$	AIC
Nulo	7	21.332				52.043
cat	6	7.201	14.131	1	<0.0001	39.912
idade | cat	5	3.064	4.136	1	0.042	37.776
ecg | cat,idade	4	2.466	0.598	1	0.439	37.178
cat*idade | cat,idade,ecg	3	1.858	0.608	1	0.436	40.570
cat*ecg | cat,idade,ecg,cat*idade	2	1.056	0.802	1	0.370	41.768
idade*ecg |cat,idade,ecg,cat*idade,cat*ecg	1	0.102	0.954	1	0.329	42.813

Agora, ajustando o modelo de ressão logística de forma saturada (com todos os parâmentos possíveis) para o modelo com ligação **C-loglog** obtemos a seguinte tabela de diferença de *deviances*:

Tabela 5: Modelos ajustados e diferença de *deviances* entre eles

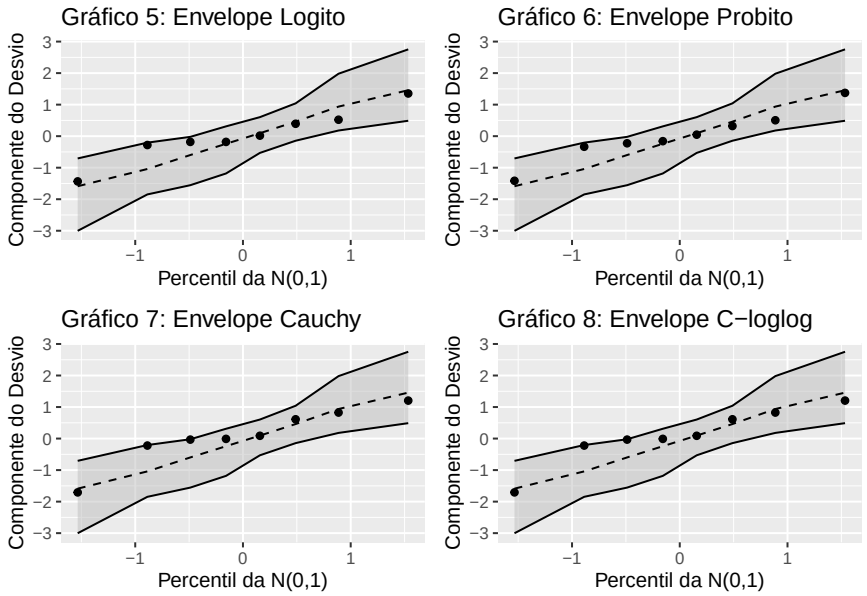
Modelos	g.l.	*Deviances*	TRV	$\neq$ g.l.	$p-value$	AIC
Nulo	7	21.332				52.043
cat	6	7.201	14.131	1	<0.0001	39.912
idade | cat	5	2.502	4.699	1	0.030	37.213
ecg | cat,idade	4	1.059	1.443	1	0.230	37.770
cat*idade | cat,idade,ecg	3	1.005	0.054	1	0.817	39.717
cat*ecg | cat,idade,ecg,cat*idade	2	0.451	0.554	1	0.457	41.163
idade*ecg |cat,idade,ecg,cat*idade,cat*ecg	1	0.005	0.447	1	0.504	42.716

E com as tabelas 3,4 e 5 podemos dizer que a inferência dos parâmetros dos modelos para todas as ligações foi o mesmo escolhido para a ligação logito do item anterior, ao nível de significância de 5%.

Tabela 6: Estatística deviance de qualidade de ajuste e AIC

	Logito	Probito	Clog-log	Cauchy
Q_L	2.477	2.441	2.502	3.064
$p-value$	0.780	0.785	0.776	0.690
AIC	37.188	37.152	37.213	37.776

E com a tabela 6 pode-se dizer que embora as métricas acima sejam muito próximas temos evidências a favor do modelo Binomial com ligação *probito*.



E através dos gráficos 5,6,7 e 8 de envelopes simulados, observa-se que os pontos estão mais próximos da reta pontilhada na ligação probito, confirmando que essa seria a melhor ligação para a análise.

fazendo uma comparação das estimativas e seus erros padrões:

Tabela 7: Comparações entre as estimativas com erro padrão

	Estimativas		
links	$\hat{\beta}_0$ (e.p)	$\hat{\beta}_1$ (e.p)	$\hat{\beta}_2$ (e.p)
Logito	-2.54 (0.20)	0.77 (0.29)	0.62 (0.28)
Probito	-1.46 (0.10)	0.42 (0.16)	0.32 (0.15)
Clog-log	-2.57 (0.19)	0.71 (0.27)	0.58 (0.27)
Cauchy	-4.04 (0.72)	1.41 (0.61)	1.59 (0.84)

Observa-se ainda que os erros padrões são menores na ligação probito.

Exercício 3

Um estudo reuniu informações, entre 1994 e 1995, de 494 indivíduos que sofreram acidente traumático e foram atendidos pelo SIATE (Serviço Integrado de Atendimento ao Trauma em Emergência). A fim de predizer a probabilidade de óbito nas primeiras 24 horas após o acidente, foi ajustado um modelo de regressão logística aos dados do estudo. O modelo final ajustado ficou expresso por

$$ln\left[\frac{\hat{p}(x)}{1-\hat{p}(x)}\right] = 2.211 + 2.607x_1 - 0.52x_2,$$

Em que, x_1 = número de lesões no tórax, que pode variar de 0 a 5, e x_2 = escala de coma de Glasgow (GCS) = total registrado para cada indivíduo no Quadro 1, que pode variar entre 3 e 15.

Quadro 1: Escala de coma de Glasgow

1. Abertura ocular	espontânea	4
	à voz	3
	com dor	2
	ausente	1
2. Resposta verbal	orientada	5
	confusa	4
	desconexa	3
	ininteligível	2
	ausente	1
3. Resposta motora	obedece comandos	6
	apropriada à dor	5
	retirada à dor	4
	flexão à dor	3
	extensão	2
	ausente	1
Total GCS (1+2+3)		

(a) Estime as probabilidades $p(x)$ para todas as possíveis combinações de x_1 e x_2 organizando-as em ordem decrescente a fim de serem identificados os indivíduos que necessitam de encaminhamento hospitalar com muita, moderada ou pouca urgência.

Resolução

As probabilidades estimadas

x_1	x_2	$\hat{p}$	x_1	x_2	$\hat{p}$	x_1	x_2	$\hat{p}$
5	3	1.0000	4	12	0.9983	2	12	0.7658
5	4	1.0000	3	7	0.9983	1	7	0.7646
5	5	1.0000	4	13	0.9972	2	13	0.6604
5	6	1.0000	3	8	0.9972	1	8	0.6588
5	7	1.0000	2	3	0.9972	0	3	0.6572
5	8	1.0000	4	14	0.9953	2	14	0.5362
4	3	1.0000	3	9	0.9953	1	9	0.5344
5	9	1.0000	2	4	0.9953	0	4	0.5327
4	4	1.0000	4	15	0.9921	2	15	0.4073
5	10	1.0000	3	10	0.9921	1	10	0.4056
4	5	1.0000	2	5	0.9920	0	5	0.4040
5	11	0.9999	3	11	0.9868	1	11	0.2886
4	6	0.9999	2	6	0.9867	0	6	0.2872
5	12	0.9999	3	12	0.9779	1	12	0.1943
4	7	0.9999	2	7	0.9778	0	7	0.1933
5	13	0.9998	3	13	0.9635	1	13	0.1254
4	8	0.9998	2	8	0.9632	0	8	0.1247
3	3	0.9998	1	3	0.9630	1	14	0.0786
5	14	0.9997	3	14	0.9400	0	9	0.0781
4	9	0.9997	2	9	0.9396	1	15	0.0482
3	4	0.9996	1	4	0.9392	0	10	0.00479
5	15	0.9994	3	15	0.9031	0	11	0.0291
4	10	0.9994	2	10	0.9025	0	12	0.0175
3	5	0.9994	1	5	0.9019	0	13	0.0105
4	11	0.9990	2	11	0.8462	0	14	0.0062
3	6	0.9990	1	6	0.8453	0	15	0.0037

Observando estas probabilidades, percebemos que quanto maior o valor da variável x_1 maior a probabilidade de um encaminhamento hospitalar com muita urgência. E quanto maior o valor da variável x_2 menor esta probabilidade. De modo geral pode-se dizer q a primeira variável é diretamente proporcional à probabilidade de um encaminhamento hospitalar com muita urgência, enquanto a segunda é inversamente proporcional.