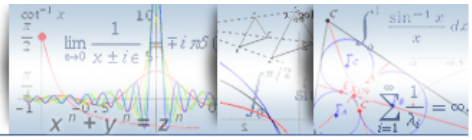
Regressão Logística Aplicada à Pesquisa de Preditores de Morte

Ana Margarida Lopes Gonçalves

DEPARTAMENTO DE MATEMÁTICA
FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA



Regressão Logística Aplicada à Pesquisa de Preditores de Morte

Ana Margarida Lopes Gonçalves

Dissertação para a obtenção do Grau de **Mestre em Matemática**

Área de Especialização em **Estatística, Optimização e Matemática Financeira**

Júri

Presidente:	Carlos Manuel Rebelo Tenreiro da Cruz
Co-Orientador:	Maria Emília Nogueira Mesquita
Co-Orientador:	Adriana Belo
Vogais:	Carlos Manuel Rebelo Tenreiro da Cruz
	Cristina Maria Tavares Martins

Data: Setembro 2013

Resumo

O presente estudo tem como objectivo encontrar preditores de morte em pessoas com Síndrome Coronária Aguda e sem antecedentes cardiovasculares, utilizando para tal um modelo de regressão logística.

Este trabalho está divido em duas partes. Na primeira, de cariz mais teórico, consideram-se variáveis de qualquer tipo. Define-se família exponencial de distribuições e apresentam-se os modelos lineares generalizados, para os quais se especificam as suas componentes (sistemática, aleatória e função de ligação). A escolha da função de ligação depende do problema em questão, e a cada função de ligação corresponde um caso particular dos modelos lineares generalizados, tais como o modelo de regressão linear ou o modelo de regressão logística. Apresenta-se a estimação e a inferência sobre os parâmetros do modelo. Para a estimação dos parâmetros aplica-se o método de máxima verosimilhança, verificando-se que as equações de verosimilhança obtidas para $\boldsymbol{\beta} = [\beta_0\ \beta_1 \dots \beta_p]^T$ são, em geral, não lineares.

Na segunda parte, mais prática, procede-se à análise do caso em que a variável de interesse (resposta) é binária, recorrendo à regressão logística. Faz-se uma análise de regressão logística a dados reais recorrendo ao software estatístico *SPSS* 20.0 (*Statistical Package for the Social Science*). Estes dados foram cedidos pela *Sociedade Portuguesa de Cardiologia* (SPC) no seguimento de um estágio curricular desenvolvido nesta mesma sociedade. Esta análise restringe-se a covariáveis categóricas ordinais e nominais. Começa-se por dividir a amostra em dois conjuntos, um com 60 % dos doentes e outro com os restantes 40%. A partir do primeiro pretende-se encontrar os preditores de morte utilizando regressão logística e analisar o poder discriminatório do modelo. Com os restantes dados faz-se a validação externa do modelo subjacente aos preditores encontrados. Conclui-se que *Sexo*, *Idade*, *Índice massa corporal*, *Frequência cardíaca*, *Pressão arterial sistólica*, *Classe Killip* e *Classificação função VE* são preditores de morte, isto é, influenciam fortemente a ocorrência de morte e que o poder discriminatório é excelente. Quanto à validação externa, os resultados obtidos indicam que o modelo tem um bom desempenho na previsão de ocorrência de morte.

Palavras Chave: Modelos lineares generalizados, logit, regressão logística, odds ratio.

Abstract

This study aims to determine the predictors of death in people with acute coronary syndrome and no history of cardiovascular disease using

a model of logistic regression.

This work is divided into two parts. In the first, more theoretical, we consider variables of any type. We define exponential family of distributions and we present the generalized linear models, for which specify their components (systematic, random and link function). The choice of the link function depends on the problem under study, and every link function corresponds to a particular case of generalized linear models, such as linear regression or logistic regression model. We present the estimation and inference about the model parameters. For the estimation we apply the method of maximum likelihood, verifying that the likelihood equations obtained for $\boldsymbol{\beta} = [\beta_0\ \beta_1 \ldots \beta_p]^T$ are in general nonlinear.

In the second part, more practical, we proceed to consider the case in which response variable (interest) is binary, using logistic regression. Makes it a logistic regression analysis to real data using the statistical software *SPSS 20.0* (*Statistical Package for the Social Sciences*). These data were provided by Portuguese Society of Cardiology (SPC) following a traineeship developed in this same society. This analysis is restricted to ordinal and nominal categorical covariates. We start by dividing the sample in two sets, one with 60 % of patients and the other with the remaining 40 %. From the first we want to find predictors of death using logistic regression and analyze the discriminatory power of the model. With the remaining data we make the external validation of the model underlying the predictors found. We conclude that *Sex*, *Age*, *body mass index*, *Heart rate*, *Systolic*, *Killip Class* and *Rating LV function* are predictors of death, this is, they are strongly influence in the incidence of death and that the discriminatory power is excellent. In respect of external validation, the results indicate that the model as a good performs in predicting the occurrence of death.

Keywords: Generalized linear models, logit, logistic regression, odds ratio.

Agradecimentos

À Prof. Doutora Maria Emília Nogueira Mesquita pela orientação e apoio prestado ao longo da dissertação.

À Mestre Adriana Belo pelos seus sábios conselhos, orientação, amizade e disponibilidade incondicional demonstrada durante todo este trabalho e estágio.

À Sociedade Portuguesa de Cardiologia por ter cedido os dados para a realização da análise apresentada neste trabalho.

Ao Francisco Carvalho pela paciência, apoio e sugestões para uma melhor clareza na apresentação deste texto.

À Dr^a Sandra Corker pela amizade e por dar o exemplo de que a vida pode ser vista sob outro ponto de vista.

Conteúdo

Capítulo 1

Introdução

A estatística é utilizada nas mais diversas áreas do conhecimento científico com a pretensão de responder a problemas subjacentes a estas. Em muitos estudos estatísticos pretende-se saber o efeito que determinadas variáveis provocam na variável resposta (interesse), acreditando que a variabilidade desta é explicada pelas outras.

Para a escolha do modelo estatístico a utilizar na resolução do problema em questão devemos ter em consideração a natureza das variáveis. Na primeira parte deste trabalho de cariz mais teórico, consideramos variáveis de qualquer tipo. Na segunda parte, mais prática, analisamos o caso em que a variável resposta é binária, recorrendo à regressão logística.

Este método de regressão é utilizado na resolução de diversos problemas de resposta binária, isto é, cada indivíduo possui ou não possui determinada característica em estudo. Naturalmente que, a qualquer experiência estão sempre associadas condicionantes aleatórias que não podemos controlar (por exemplo, no caso de aparecimento de uma doença as características genéticas do indivíduo são relevantes) e factores conhecidos, cujo efeito contribui para a presença ou ausência da característica em estudo.

O modelo de regessão logística introduzido por Berkson em 1944 define-se por

$$\log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \boldsymbol{x}^T\boldsymbol{\beta}, \text{ ou equivalentemente, } \pi(\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{x}^T\boldsymbol{\beta}\right)}{1+\exp(\boldsymbol{x}^T\boldsymbol{\beta})},$$

onde $\boldsymbol{x}$ é um vector contituído pelas covariáveis ou variáveis explicativas consideradas, x_1, x_2, $\ldots$, x_p, $\pi(\boldsymbol{x})$ é a probabilidade de um indivíduo com determinados atributos representados pelo vector de covariáveis, apresentar uma determinada característica representada pela variável resposta Y, isto é, $\pi(\boldsymbol{x}) = P\left(Y=1\right)$ e $\boldsymbol{\beta} = [\beta_0\ \beta_1 \ldots \beta_p]^T$ é o vector de parâmetros do modelo.

Para ilustrar este modelo consideremos o seguinte exemplo: se pretendermos analisar a ocorrência de re-enfarte dado um conjunto de covariáveis (por exemplo, sexo, idade, IMC, e etc.) que pensamos explicar a ocorrência de re-enfarte, fazendo uma análise de regressão logística ficamos a conhecer quais as variáveis que são realmente

os preditores de re-enfarte. Além disso, a partir dos coefecientes do modelo encontrado podemos medir o risco de ocorrência do re-enfarte na presença ou ausência de cada uma das características representadas por essas variáveis.

O objectivo central deste trabalho é determinar os preditores de morte em pessoas com Síndrome Coronária Aguda e sem antecedentes cardiovasculares utilizando regressão logística.

No capítulo 2 introduzimos os modelos lineares generalizados (MLG) que têm como caso particular o modelo de regressão logística. Começamos por definir a família exponencial de distribuições, à qual pertencem distribuições como a lei Normal ou a lei Binomial. Sendo esta família a base dos MLG é possível expôr o modelo e as partes que o constituem: componente sistemática, aleatória e função de ligação. A escolha da função de ligação depende do problema em questão, e cada função de ligação origina casos particulares dos MLG, como o modelo de regressão linear, o modelo de regressão logística, o modelo probit e etc. Definido o modelo a utilizar, passamos a estimar os parâmetros do modelo aplicando o método de máxima verosimilhança. As equações de verosimilhança para $\boldsymbol{\beta}$ são em geral não lineares pelo que temos de recorrer a métodos numéricos para as resolver. No final deste capítulo apresentamos a inferência paramétrica destes modelos.

No capítulo 3 definimos regressão logística univariável (uma só covariável) e regressão logística multivariável (mais de uma covariável). Neste capítulo assumimos que a variável resposta é binária e que as covariáveis são categóricas ordinais ou nominais. Introduzimos os conceitos e métodos necessários à realização do estudo apresentado no capítulo 4.

No capítulo 4 fazemos uma análise de regressão logística a dados reais recorrendo ao software estatístico *SPSS* 20.0 (*Statistical Package for the Social Science*). Estes dados foram cedidos pela *Sociedade Portuguesa de Cardiologia* (SPC) no seguimento de um estágio curricular desenvolvido nesta mesma sociedade. Esta análise confina-se a covariáveis categóricas ordinais e nominais. Com vista à construção do modelo de regressão logística dividimos a amostra em dois conjuntos. A partir do primeiro conjunto de dados encontramos os preditores de morte e analisamos o poder discriminatório do modelo constituído pelos preditores encontrados, e com o segundo conjunto fazemos a validação externa do modelo. Por fim, através do valor do *odds ratio*, avaliamos o risco de ocorrência de morte inerente à presença ou ausência de uma determinada característica subjacente a uma covariável.

Capítulo 2

Família exponencial de distribuições e modelos lineares generalizados

Os modelos lineares generalizados pressupõem que a variável resposta tenha uma distribuição pertencente à família exponencial de distribuições. Como tal, começamos por definir a família referida de forma a apresentar os modelos lineares generalizados (MLG). Estes foram introduzidos por Nelder e Wedderburn (1972) com o objectivo de generalizar os modelos lineares clássicos. A generalização incide essencialmente sobre dois aspectos: a distribuição de probabilidade associada à variável resposta Y não se restringe à Normal, podendo ser qualquer distribuição pertencente à família exponencial de distribuições e a função que relaciona a variável resposta e a combinação linear das variáveis independentes deve ser monótona e diferenciável.

2.1. Notação, terminologia e tipo de dados

Para os modelos lineares generalizados, a equação que estabelece a ligação entre cada variável resposta Y, contínua ou discreta, e um conjunto de covariáveis $x_1, x_2, \ldots, x_p$, também de qualquer natureza, tem a forma

$$g\left[E(Y)\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

onde β_0, β_1, $\ldots$ β_p são constantes reais, x_1, x_2, $\ldots$ x_p são variáveis deterministas e g é uma função conhecida. Tal como no modelo clássico, β_0, β_1, $\ldots$, β_p, são designados por parâmetros do modelo e x_1, x_2, $\ldots$ x_p por variáveis explicativas ou covariáveis. A função g é denominada função de ligação.

Um dos problemas que desde logo se levanta é a estimação destes parâmetros. Para tal, necessitamos de dispôr de um conjunto de dados

$$(y_i, x_{i1}, x_{i2}, \ldots, x_{ip}),\ i = 1, 2, \ldots, n.$$

Consideramos y_1, y_2, ..., y_n valores particulares das n variáveis resposta Y_1, Y_2, ..., Y_n, respectivamente, e tais que

$$g\left[E(Y_i)\right] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip},\ i = 1, 2, \ldots, n.$$

Este sistema de equações pode ser escrito na forma matricial, sendo para tal necessário fixar alguma notação. Seja então

$$\boldsymbol{Y} = \left[Y_1\ Y_2 \ldots Y_n\right]^T$$

um vector aleatório real cujas componentes são independentes,

$$\boldsymbol{\beta} = \left[\beta_0\ \beta_1 \ldots \beta_p\right]^T,$$

o vector de parâmetros e

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ 1 & x_{31} & x_{32} & \ldots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

onde $\boldsymbol{x}_i^T = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$ denota a i-ésima linha da matriz $\boldsymbol{X}$, $i = 1, \ldots, n$. Podemos assim escrever

$$\boldsymbol{g}\left[E(\boldsymbol{Y})\right] = \boldsymbol{X}\boldsymbol{\beta}, \quad \text{onde} \quad \boldsymbol{g}\left[E(\boldsymbol{Y})\right] = \begin{bmatrix} g\left[E(Y_1)\right] \\ g\left[E(Y_2)\right] \\ \vdots \\ g\left[E(Y_n)\right] \end{bmatrix}.$$

A construção dos estimadores de $\beta_0, \beta_1, \ldots \beta_p$ será feita no caso das leis de probabilidade das variáveis $Y_1, Y_2, \ldots, Y_n$ pertencerem à família exponencial de distribuições. Assim, antes de definir modelos lineares generalizados, apresentamos na secção seguinte a família exponencial de distribuições.

2.2. Família exponencial de distribuições

Sejam $(\Omega, \mathcal{A}, P)$ um espaço de probabilidade e Y uma variável aleatória real definida sobre Ω cuja lei de probabilidade depende de um parâmetro real desconhecido θ, $\theta \in \Theta$ contido em $\mathbb{R}$. Denotemos por f_θ a função densidade de Y, no caso de Y ser absolutamente contínua, ou função de probabilidade, no caso de Y ser discreta.

Definição 1. *Diz-se que a distribuição de Y pertence à família exponencial de distribuições se f_θ puder ser escrita da forma*

$$f_\theta(y) = \exp\left(a(y)b(\theta) + c(\theta) + d(y)\right),$$

onde $\theta \in \Theta \subset \mathbb{R}$ é um parâmetro escalar e a, b, c, d são funções reais conhecidas, com b diferenciável.

Na definição apresentada, $b(\theta)$ é denominado parâmetro natural e se $a(y) = y$ dizemos que a distribuição está na forma canónica. Se existirem outros parâmetros, para além do parâmetro θ, estes são chamados de parâmetros pertubadores e farão parte das funções a, b, c e d. Supõe-se conhecer estes parâmetros.

Diversas distribuições frequentemente utilizadas, como é o caso, por exemplo, das distribuições Normal, de Poisson e Binomial, pertencem à família exponencial de distribuições. De seguida apresentamos dois exemplos. No primeiro consideramos o caso de uma variável aleatória absolutamente contínua e no segundo o caso de uma v.a. discreta.

Exemplo 1. Distribuição Normal

Se Y segue uma distribuição Normal de média , $\mu \in \mathbb{R}$ e desvio padrão $\sigma \in \mathbb{R}^+$, a função densidade de Y é dada por

$$f_\mu(y) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right),$$

onde consideramos μ como parâmetro de interesse e σ é suposto ser conhecido.

Podemos reescrever $f_\mu(y)$ da seguinte forma,

$$\begin{aligned} f_\mu(y) &= \exp\left[\log\left(\frac{1}{(2\pi\sigma^2)^{1/2}}\right)\right] \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) \\ &= \exp\left(-\frac{1}{2}\log\left(2\pi\sigma^2\right) + \frac{y\mu}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) \end{aligned}$$

onde $a(y) = y$, $b(\mu) = \frac{\mu}{\sigma^2}$, $c(\mu) = -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{\mu^2}{2\sigma^2}$ e $d(y) = -\frac{y^2}{2\sigma^2}$, para $y \in \mathbb{R}$.

△

Exemplo 2. Distribuição Binomial

Seja Y uma variável aleatória real com distribuição binomial de parâmetros n e π, $Y \sim \mathcal{B}(n, \pi)$, $\pi \in]0,1[$, $n \in \mathbb{N}$, onde π é o parâmetro de interesse e assumimos

conhecer n. Uma vez que a função de probabilidade de Y é dada por

$$f_\pi(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}\,,\ y \in \{0,1,2,\ldots,n\},$$

podemos escrever

$$f_\pi(y) = \exp\left(y\log(\pi) - y\log(1-\pi) + n\log(1-\pi) + \log\binom{n}{y}\right),$$

onde $a(y) = y$, o parâmetro natural $b(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$, $c(\pi) = n\log(1-\pi)$ e $d(y) = \log\binom{n}{y}$.

$\triangle$

Apresentamos de seguida condições de regularidade necessárias ao desenvolvimento do estudo, conhecidas por condições de *Cramer-Rao* (cf. Gonçalves & Nazaré, 2003, p. 90).

Seja $(\mathbb{R}^l, \mathcal{B}_l, Q_\theta)_{\theta\in\Theta}$, uma família de espaços de probabilidade associada a uma v.a. Y sobre $\mathbb{R}^l$ de lei Q_θ de suporte S_θ, com $\theta = (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta$ e onde Θ é um aberto de $\mathbb{R}^k$. Seja $f_\theta(y)$, $y \in \mathbb{R}^l$, a função de probabilidade (respectivamente função densidade) de Q_θ. Definamos um conjunto $\mathbb{Y}$ tal que, se Q_θ é absolutamente contínua (resp. discreta), $\mathbb{Y}$ é o menor subconjunto (resp. subconjunto numerável) de $\mathbb{R}^l$, independente de θ que contém S_θ, $\forall\theta \in \Theta$. Diz-se que a família de espaços de probabilidade $(\mathbb{R}^l, \mathcal{B}_l, Q_\theta)_{\theta\in\Theta}$ está nas condições de *Cramer-Rao* se

(i) $\forall y \in \mathbb{Y}\ \forall\theta \in \Theta,\ f_\theta(y) > 0$.

(ii) $\forall y \in \mathbb{Y},\ \forall\theta \in \Theta$, existe $\nabla f_\theta(y) = \left[\frac{\partial f_\theta(y)}{\partial\theta_1}\ \frac{\partial f_\theta(y)}{\partial\theta_2} \cdots \frac{\partial f_\theta(y)}{\partial\theta_k}\right]$.

(iii) Sendo $\mathcal{B}_\mathbb{Y}$ uma σ-álgebra sobre $\mathbb{Y}$ tem-se, $\forall j \in \{1,2,\ldots,k\}\ \forall C \in \mathcal{B}_\mathbb{Y}$,

- $\sum\limits_{y_i\in C} \frac{\partial f_\theta(y_i)}{\partial\theta_j} = \frac{\partial}{\partial\theta_j} \sum\limits_{y_i\in C} f_\theta(y_i)$, se Q_θ é discreta
- $\int_C \frac{\partial f_\theta(y)}{\partial\theta_j} dy = \frac{\partial}{\partial\theta_j} \int_C f_\theta(y) dy$, se Q_θ é absolutamente contínua.

(iv) $\forall\theta \in \Theta$ exitem os momentos de segunda ordem das v.a.r. $\frac{\partial}{\partial\theta_j} log\ f_\theta(Y)$ para todo $j \in \{1,2,\ldots,k\}$, considerando obviamente a restrição de f_θ a $\mathbb{Y}$.

(v) $\forall y \in \mathbb{Y}\ \forall\theta \in \Theta$, existe $\left(\frac{\partial^2 f_\theta(y)}{\partial\theta_j\partial\theta_l},\ j,l \in \{1,2,\ldots,k\}\right)$. Além disso, $\forall C \in \mathcal{B}_\mathbb{Y}$, $\forall\, j,l \in \{1,2,\ldots,k\}$,

- $\sum_{y_i \in C} \frac{\partial^2 f_\theta(y_i)}{\partial\theta_j \partial\theta_l} = \frac{\partial^2}{\partial\theta_j \partial\theta_l} \sum_{y_i \in C} f_\theta(y_i)$, se Q_θ é discreta

- $\int_C \frac{\partial^2 f_\theta(y)}{\partial\theta_j \partial\theta_l} dy = \frac{\partial^2}{\partial\theta_j \partial\theta_l} \int_C f_\theta(y) dy$, se Q_θ é absolutamente contínua.

Nestes modelos, atendendo à condição (i), $\mathbb{Y}$ coincide com o suporte da lei de Q_θ, que é assim, necessariamente, independente de θ. Por outro lado a condição (ii) e o facto de f_θ ser uma função de probabilidade (resp. densidade) asseguram a existência das somas (resp. dos integrais) apresentados na condição (iii). A condição (iii) permite afirmar que se tem $\int_{\mathbb{Y}} \frac{\partial f_\theta(y)}{\partial\theta_j} dy = 0$ ou $\sum_{y_i \in \mathbb{Y}} \frac{\partial f_\theta(y_i)}{\partial\theta_j} = 0$, $j \in \{1, \ldots, k\}$, consoante Q_θ seja absolutamente contínua ou discreta.

Definição 2. *Chamamos vector score do modelo $(\mathbb{R}^l, \mathcal{B}_l, Q_\theta)_{\theta\in\Theta}$ ao vector aleatório $[U_{\theta_1}\ U_{\theta_2} \ldots U_{\theta_k}]^T$, com $U_{\theta_j} = \frac{\partial}{\partial\theta_j} log\ f_\theta(Y)$. Se o parâmetro é real, a variável correspondente é simplesmente designada por score.*

Sob as condições de regularidade, o vector aleatório real $[U_{\theta_1}\ U_{\theta_2} \ldots U_{\theta_k}]^T$ está definido sobre $\mathbb{Y}$, para todo $\theta \in \Theta$. As condições (i) e (ii) garantem a existência do vector e (iv) a existência dos respectivos vector médio e matriz de variâncias-covariâncias. Da condição (iii) decorre que o *vector score* é centrado. Com efeito, por exemplo, no caso em que Q_θ é absolutamente contínua vem

$$\forall j \in \{1, \ldots, k\}\ ,\ E\left(\tfrac{\partial}{\partial\theta_j} log\ f_\theta\,(Y)\right) = E\left(\frac{\partial f_\theta(Y)}{\partial\theta_j}\frac{1}{f_\theta(Y)}\right) = \int_{\mathbb{Y}} \frac{\partial f_\theta(y)}{\partial\theta_j} dy = 0.$$

Este resultado é verificado de modo análogo quando Q_θ é discreta.

A proposição seguinte estabelece, no caso da distribuição de Y pertencer à família exponencial e de $l = 1$, expressões para o valor médio e variância da variável $a(Y)$.

Proposição 1. *Se Y é uma variável cuja distribuição pertence à família exponencial de distribuições, então*

$$E(a(Y)) = -\frac{\frac{d}{d\theta}c(\theta)}{\frac{d}{d\theta}b(\theta)} \quad e \quad Var(a(Y)) = \frac{\frac{d^2}{d\theta^2}b(\theta)\frac{d}{d\theta}c(\theta) - \frac{d^2}{d\theta^2}c(\theta)\frac{d}{d\theta}b(\theta)}{\left(\frac{d}{d\theta}b(\theta)\right)^3}.$$

Demonstração 1. *Façamos a demonstração no caso de Y ser uma variável discreta (no caso absolutamente contínuo, o raciocínio é perfeitamente análogo). Como já referimos da condição (iii) resulta $\sum_{y\in\mathbb{Y}} \frac{d}{d\theta} f_\theta(y) = 0$.*

Por outro lado, como a distribuição de Y pertence à família exponencial, vem

$\frac{d}{d\theta} f_\theta(y) = \left[a(y)\frac{d}{d\theta}b(\theta) + \frac{d}{d\theta}c(\theta)\right] f_\theta(y)$. *Consequentemente*

$$\begin{aligned} 0 &= \sum_{y\epsilon\mathbb{Y}} \left[a(y)\frac{d}{d\theta}b(\theta) + \frac{d}{d\theta}c(\theta)\right] f_\theta(y) \\ &= \frac{d}{d\theta}b(\theta)\sum_{y\epsilon\mathbb{Y}} a(y) f_\theta(y) + \frac{d}{d\theta}c(\theta)\sum_{y\epsilon\mathbb{Y}} f_\theta(y) \\ &= \frac{d}{d\theta}b(\theta)E(a(Y)) + \frac{d}{d\theta}c(\theta). \end{aligned}$$

Logo

$$E(a(Y)) = -\frac{\frac{d}{d\theta}c(\theta)}{\frac{d}{d\theta}b(\theta)}.$$

Como

$$\frac{d^2}{d\theta^2} f_\theta(y) = \left[a(y)\frac{d^2}{d\theta^2}b(\theta) + \frac{d^2}{d\theta^2}c(\theta) + \left(a(y)\frac{d}{d\theta}b(\theta) + \frac{d}{d\theta}c(\theta)\right)^2\right] f_\theta(y),$$

tem-se que

$$\begin{aligned} \sum_{y\epsilon\mathbb{Y}} \frac{d^2}{d\theta^2} f_\theta(y) &= \frac{d^2}{d\theta^2}b(\theta)\sum_{y\epsilon\mathbb{Y}} a(y) f_\theta(y) + \frac{d^2}{d\theta^2}c(\theta)\sum_{y\epsilon\mathbb{Y}} f_\theta(y) + \sum_{y\epsilon\mathbb{Y}} \left(\frac{d}{d\theta}b(\theta)\right)^2 \left[a(y) + \frac{\frac{d}{d\theta}c(\theta)}{\frac{d}{d\theta}b(\theta)}\right]^2 f_\theta(y) \\ &= \frac{d^2}{d\theta^2}b(\theta)E(a(Y)) + \frac{d^2}{d\theta^2}c(\theta) + \left(\frac{d}{d\theta}b(\theta)\right)^2 \sum_{y\epsilon\mathbb{Y}} \left[a(y) - E(a(Y))\right]^2 f_\theta(y) \\ &= \frac{d^2}{d\theta^2}b(\theta)E(a(Y)) + \frac{d^2}{d\theta^2}c(\theta) + \left(\frac{d}{d\theta}b(\theta)\right)^2 Var(a(Y)). \end{aligned}$$

Sendo $\frac{d^2}{d\theta^2}\sum_{y\epsilon\mathbb{Y}} f(y,\theta) = 0$, *tem-se que*

$$\begin{aligned} Var(a(Y)) &= \frac{\frac{d^2}{d\theta^2}b(\theta)\left(-\frac{\frac{d}{d\theta}c(\theta)}{\frac{d}{d\theta}b(\theta)}\right) - \frac{d^2}{d\theta^2}c(\theta)}{\left(\frac{d}{d\theta}b(\theta)\right)^2} \\ &= \frac{\frac{d^2}{d\theta^2}b(\theta)\frac{d}{d\theta}c(\theta) - \frac{d^2}{d\theta^2}c(\theta)\frac{d}{d\theta}b(\theta)}{\left(\frac{d}{d\theta}b(\theta)\right)^3} \end{aligned}$$

□

Definição 3. *Consideremos que o modelo estatístico* $\left(\mathbb{R}^l, \mathcal{B}_l, Q_\theta\right)_{\theta\in\Theta}$ *está nas condições de regularidade de Cramer-Rao, chama-se informação de Fisher de* $\left(\mathbb{R}^l, \mathcal{B}_l, Q_\theta\right)_{\theta\in\Theta}$ *à matriz de variâncias-covariâncias do vector aleatório* $[U_{\theta_1}\, U_{\theta_2} \ldots U_{\theta_k}]^T$. *Denotamos esta matriz por* $\mathcal{I}(\theta)$.

Note-se que se o parâmetro θ é univariado, a informação é um valor real.

É fácil verificar que, se $[U_\theta(\theta)]^2 = \left(\frac{\partial\, log\, f_\theta(Y)}{\partial\theta}\right)^2$, então $\mathcal{I}(\theta) = -E\left[\frac{\partial^2\, log\, f_\theta(Y)}{\partial\theta^2}\right]$, (cf. Gonçalves & Nazaré, 2003, p. 100).

2.3. Modelos lineares generalizados

Consideremos Y_1, Y_2, ..., Y_n variáveis aleatórias independentes cujas leis de probabilidade pertencem à família exponencial de distribuições. Assumimos que a distribuição de cada uma destas variáveis está na forma canónica. Desta forma a lei de $(Y_1, Y_2, \ldots, Y_n)$ é caracterizada pela seguinte função

$$\begin{aligned} f_{(\theta_1,\theta_2,\ldots,\theta_n)}(y_1, y_2, \ldots, y_n) &= \prod_{i=1}^{n} \exp\left(y_i b(\theta_i) + c(\theta_i) + d(y_i)\right) \\ &= \exp\left(\sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)\right) \end{aligned}$$

Os modelos lineares generalizados são definidos por três componentes, a componente aleatória, a componente sistemática e a função de ligação, que passamos a descrever.

1. **Componente aleatória :** as variáveis resposta Y_1, Y_2, ..., Y_n que, tal como referimos, são independentes com distribuição pertencente à família exponencial de distribuições e admitem momento de 1^a ordem finito, $\mu_i = E(Y_i)$, $i = 1, \ldots, n$, onde μ_i é uma função de θ_i.

2. **Componente sistemática:** Dado um conjunto de variáveis explicativas e um conjunto de parâmetros, respectivamente

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ 1 & x_{31} & x_{32} & \ldots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \quad \text{e} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix},$$

   definimos $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$. Este produto é denominado preditor linear.

3. **Função de ligação:** Relaciona as componentes aleatória e sistemática,

$$\eta_i = g(\mu_i).$$

   Esta função é monotona e diferenciável e a sua escolha depende da distribuição de Y.

Note-se que, para determinadas funções de ligação, o preditor linear coincide com o parâmetro canónico, isto é, $\theta_i = \eta_i$. Neste caso a função de ligação denomina-se função de ligação canónica.

De seguida apresentamos dois exemplos, o primeiro diz respeito a um modelo de resposta contínua e o segundo a um modelo de resposta binária.

1. Modelo Normal

Consideremos n respostas independentes $Y_i \sim \mathbb{N}(\mu_i, \sigma^2)$, $i = 1, 2, \ldots, n$, onde

$$E(Y_i) = \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

Este modelo pertence aos modelos lineares generalizados, visto que, as variáveis resposta são independentes, a distribuição pertence à família exponencial de distribuições com $\theta_i = \boldsymbol{x}_i^T \beta$ e a função de ligação é a identidade, $g(\mu_i) = \mu_i$.

Este modelo é usualmente escrito na seguinte forma

$$Y_i = \boldsymbol{x}_i^T \beta + \varepsilon_i,\ i = 1, 2, \ldots, n,$$

onde os ε_i são independentes e identicamente distribuidos, $\varepsilon \sim \mathbb{N}(0, \sigma^2)$. No modelo que estamos a considerar, modelo linear clássico, pressupõe-se que a variância das respostas é constante.

2. Modelo para dados binários ou na forma de proporções

Suponhamos que temos n variáveis resposta independentes $Y_i \sim \mathcal{B}(1, \pi_i)$, $i = 1, 2, \ldots, n$, e que a cada indivíduo ou unidade experimental i está associado um vector de covariáveis $\boldsymbol{x}_i$, $i = 1, 2, \ldots, n$. A função de probabilidade associada a Y_i é

$$f_{y_i}(\pi_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, \quad y_i \epsilon \{0, 1\}.$$

Como vimos anteriormente $\theta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$ e $E(Y_i) = \pi_i$, logo ao fazer

$$\theta_i = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

concluímos que a função de ligação canónica é a função $\ln\left(\frac{\pi_i}{1-\pi_i}\right)$, à qual chamamos *logit*. É fácil de ver que a probabilidade $P(Y_i = 1) = \pi_i$, está relacionada com vector $\boldsymbol{x}_i$ através de

$$\pi_i = \frac{\exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)}{1 + \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)}.$$

2.4. Estimação dos parâmetros do modelo

Tal como referimos anteriormente, os dados são da forma $(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$, $i = 1, 2, \ldots, n$, onde y_i é o valor observado da variável resposta para a i-ésima

unidade experimental (indivíduo) e x_{i1} , x_{i2}, ... x_{in} os correspondentes valores das covariáveis consideradas no estudo. Para simplificar o estudo admitimos que a matriz $\boldsymbol{X}$ tem característica completa. Partimos da suposição que Y_1, Y_2, ..., Y_n são variáveis aleatórias independentes sastisfazendo as propriedades dos modelos lineares generalizados. Nestes modelos os parâmetros de interesse são os β_j, $j = 0, 1, \ldots, p$, que são estimados pelo método da máxima verosimilhança. A lei de $(Y_1, \ldots, Y_n)$ é caracterizada pela seguinte função

$$f_{(\theta_1,\theta_2,\ldots,\theta_n)}(y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} \exp\left(y_i b(\theta_i) + c(\theta_i) + d(y_i)\right).$$

Como foi visto anteriormente θ_i é função de μ_i, sendo

$$E(Y_i) = -\frac{\frac{d}{d\theta_i}c(\theta_i)}{\frac{d}{d\theta_i}b(\theta_i)},$$

$$Var(Y_i) = \frac{\frac{d^2}{d\theta_i^2}b(\theta_i)\frac{d}{d\theta_i}c(\theta_i) - \frac{d^2}{d\theta_i^2}c(\theta_i)\frac{d}{d\theta_i}b(\theta_i)}{\left(\frac{d}{d\theta_i}b(\theta_i)\right)^3}$$

e $g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} = \eta_i$.

Podemos escrever a função de verosimilhança como função de $\boldsymbol{\beta}$, uma vez que μ_i é função de θ_i.

$$L_y(\boldsymbol{\beta}) = \prod_{i=1}^{n} f_{\theta_i}(y_i) = \exp\left(\sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)\right).$$

Portanto a função de *log-verosimilhança*, como função de $\boldsymbol{\beta}$, é dada por

$$log\left(L_y(\boldsymbol{\beta})\right) = l_y(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i b(\theta_i) + c(\theta_i) + d(y_i)\right) = \sum_{i=1}^{n} l_{y_i}(\boldsymbol{\beta}). \tag{2.1}$$

Os estimadores de máxima verosimilhança de $\boldsymbol{\beta}$ são obtidos como solução do sistema de equações de verosimilhança

$$\frac{\partial l_y(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{i=1}^{n}\left(\frac{\partial l_{y_i}(\boldsymbol{\beta})}{\partial \beta_k}\right) = \mathbf{0}, \quad k = 0, 1, \ldots, p,$$

onde

$$\frac{\partial l_{y_i}(\boldsymbol{\beta})}{\partial \beta_k} = \frac{\partial l_{y_i}(\theta_i)}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_k}.$$

Ora

- $$\begin{aligned}\frac{\partial l_{y_i}(\theta_i)}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i}\left(y_i b(\theta_i) + c(\theta_i) + d(y_i)\right)\\ &= y_i\frac{d}{d\theta_i}b(\theta_i) + \frac{d}{d\theta_i}c(\theta_i)\\ &= \left[y_i - \left(-\frac{\frac{d}{d\theta_i}c(\theta_i)}{\frac{d}{d\theta_i}b(\theta_i)}\right)\right]\frac{d}{d\theta_i}b(\theta_i)\\ &= (y_i - \mu_i)\frac{d}{d\theta_i}b(\theta_i)\end{aligned}$$

- $\dfrac{\partial \theta_i}{\partial \mu_i} = \left(\dfrac{\frac{d^2}{d\theta_i^2} b(\theta_i) \frac{d}{d\theta_i} c(\theta_i) - \frac{d^2}{d\theta_i^2} c(\theta_i) \frac{d}{d\theta_i} b(\theta_i)}{\left(\frac{d}{d\theta_i} b(\theta_i) \right)^2} \right)^{-1} = \left(Var(Y_i) \dfrac{d}{d\theta_i} b(\theta_i) \right)^{-1}$

- $\dfrac{\partial \mu_i}{\partial \beta_k} = \dfrac{\partial \mu_i}{\partial \eta_i} \dfrac{\partial \eta_i}{\partial \beta_k} = \dfrac{\partial \mu_i}{\partial \eta_i} \dfrac{\partial \left(\boldsymbol{x}_i^T \boldsymbol{\beta} \right)}{\partial \beta_k} = \dfrac{\partial \mu_i}{\partial \eta_i} x_{ik}.$

  Note-se que $\dfrac{\partial \mu_i}{\partial \eta_i} x_{ik}$ depende da função de ligação escolhida.

As equações de verosimilhança para $\boldsymbol{\beta}$ são

$$\sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ik} \right) = 0, \quad k = 0, \ldots, p.$$

O vector score, $U(\boldsymbol{\beta}) = [U_{\beta_0} \, U_{\beta_1} \ldots U_{\beta_p}]^T$, tem como elemento genérico

$$\begin{aligned} U_{\beta_k} &= \sum_{i=1}^{n} \left[\frac{d}{d\theta_i} b(\theta_i) \left(Y_i - \mu_i \right) \frac{1}{Var(Y_i) \frac{d}{d\theta_i} b(\theta_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ik} \right] \\ &= \sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ik} \right), \quad k = 0, \ldots, p. \end{aligned}$$

Como $E(U_{\beta_k}) = 0$, $k = 0, \ldots, p$, o elemento (k, j) da matriz de covariâncias do vector U, ou seja da matriz de informação de Fisher, é

$$\mathcal{I}_{kj}(\boldsymbol{\beta}) = cov\left(U_{\beta_k}, U_{\beta_j} \right) = E\left(U_{\beta_k} U_{\beta_j} \right), \quad k, j = 0, \ldots, p.$$

Tem-se ainda

$$\begin{aligned} \mathcal{I}_{kj}(\boldsymbol{\beta}) &= E\left\{ \sum_{i=1}^{n} \left(\frac{Y_i - \mu_k}{Var(Y_i)} x_{ik} \frac{\partial \mu_i}{\partial \eta_i} \right) \sum_{l=1}^{n} \left(\frac{Y_l - \mu_l}{Var(Y_l)} x_{lj} \frac{\partial \mu_l}{\partial \eta_l} \right) \right\} \\ &= E\left[\sum_{i=1}^{n} \frac{(Y_i - \mu_i)^2}{Var(Y_i)^2} x_{ik} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \right] \\ &= \sum_{i=1}^{n} \frac{E(Y_i - \mu_k)^2}{Var(Y_i)^2} x_{ik} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \\ &= \sum_{i=1}^{n} \frac{x_{ik} x_{ij}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2, \quad k, j = 0, \ldots, p, \end{aligned}$$

pois Y_1, Y_2, $\ldots$, Y_n são independentes e consequentemente são não correlacionadas. No caso em que a função de ligação é a canónica, a matriz de informação de Fisher coincide com a matriz Hessiana de (2.1) multiplicada por (-1).

Método de *score* de Fisher

As equações de verosimilhança encontradas na secção anterior não têm solução analítica e portanto a sua resolução requer o uso de métodos numéricos. O esquema iterativo para a resolução das equações que se vai apresentar, é baseado no método de *scores* de Fisher. A diferença entre este método e o método de Newton reside na utilização da matriz de informação de Fisher em vez da matriz Hessiana. A vantagem de utilizar a matriz $\mathcal{I}(\boldsymbol{\beta})$ deve-se essencialmente ao facto desta ser mais fácil de calcular.

O desenvolvimento em série de Taylor de $U(\boldsymbol{\beta})$, em torno de uma estimativa inicial de $\boldsymbol{\beta}$, denominada $\boldsymbol{\beta}^{(0)}$, é dado por

$$U(\boldsymbol{\beta}) \approx U\left(\boldsymbol{\beta}^{(0)}\right) + \mathcal{H}(\boldsymbol{\beta})\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}\right).$$

A equação de recorrência obtida a partir deste desenolvimento após a substituição da matriz Hessiana pela matriz de informação de Fisher é a seguinte

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} + \left[\mathcal{I}\left(\hat{\boldsymbol{\beta}}^{(m)}\right)\right]^{-1} U\left(\hat{\boldsymbol{\beta}}^{(m)}\right), \quad m = 0, 1, 2, \ldots$$

onde $\mathcal{I}\left(\hat{\boldsymbol{\beta}}^{(m)}\right)$ é a matriz de informação de Fisher obtida na m-ésima iteração. Multiplicando ambos os membros por $\mathcal{I}\left(\hat{\boldsymbol{\beta}}^{(m)}\right)$ obtemos

$$\mathcal{I}\left(\hat{\boldsymbol{\beta}}^{(m)}\right)\hat{\boldsymbol{\beta}}^{(m+1)} = \mathcal{I}\left(\hat{\boldsymbol{\beta}}^{(m)}\right)\hat{\boldsymbol{\beta}}^{(m)} + U\left(\hat{\boldsymbol{\beta}}^{(m)}\right). \tag{2.2}$$

Recorde-se que $\boldsymbol{X}$ representa a matriz cujas linhas são da forma $(1, x_{i1}, x_{i2}, \ldots, x_{ip})$, $i = 1, \ldots, n$. Seja W uma matriz diagonal de dimensão n cujos elementos da diagonal são

$$w_{ii} = \frac{1}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \tag{2.3}$$

Podemos escrever $\mathcal{I}(\boldsymbol{\beta})$ como

$$\mathcal{I}\left(\hat{\boldsymbol{\beta}}^{(m)}\right) = \boldsymbol{X}^T W \boldsymbol{X}.$$

A prova da igualdade anterior decorre de resultados básicos de álgebra matricial. O segundo membro da equação (2.2) é um vector coluna com $(p+1)$ linhas. A entrada k, $k = 0, \ldots, p$, deste vector resulta da multiplicação da linha k da matriz $\boldsymbol{X}^T W \boldsymbol{X} = \mathcal{I}\left(\hat{\boldsymbol{\beta}}^{(m)}\right)$ pelo vector $\boldsymbol{\beta}^{(m)}$ adicionado por $U\left(\hat{\boldsymbol{\beta}}^{(m)}\right)$, dada por

$$\begin{aligned}
&\sum_{l=0}^{p}\sum_{i=1}^{n}\frac{x_{ik}x_{il}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \hat{\beta}_l^{(m)} + \sum_{i=1}^{n}\left[\frac{y_i - \mu_i}{Var(Y_i)}x_{ik}\frac{\partial \mu_i}{\partial \eta_i}\right] \\
&= \sum_{i=1}^{n}\frac{x_{ik}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\left[\sum_{l=0}^{p} x_{il}\hat{\beta}_l^{(m)} + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}\right].
\end{aligned}$$

Desta forma o segundo membro de (2.2) pode ser escrito, na forma matricial, como

$$\boldsymbol{X}^T W^{(m)} \boldsymbol{z}^{(m)},$$

onde $\boldsymbol{z}^{(m)}$ é um vector coluna, com n linhas, cuja i-ésima componente é dada por

$$z_i^{(m)} = \sum_{l=0}^{p} x_{il} \hat{\beta}_l^{(m)} + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i}, \quad i = 1, \ldots, n. \tag{2.4}$$

A equação recursiva (2.2) na forma matricial é então dada por

$$\boldsymbol{X}^T W^{(m)} \boldsymbol{X} \hat{\boldsymbol{\beta}}^{(m+1)} = \boldsymbol{X}^T W^{(m)} \boldsymbol{z}^{(m)}.$$

Depois de conhecida a equação de recorrência apresentamos o algoritmo para o cálculo das estimativas de máxima verosimilhança de $\boldsymbol{\beta}$:

Escolher uma estimativa inicial, $\hat{\boldsymbol{\beta}}^{(0)}$, para $\boldsymbol{\beta}$.

Para $m = 0, 1, 2, \ldots$

1. Dado $\hat{\boldsymbol{\beta}}^{(m)}$, calcular $W^{(m)}$ e $\boldsymbol{z}^{(m)}$ usando as expressões (2.3) e (2.4) respectivamente (note-se que $\dfrac{\partial \mu_i}{\partial \eta_i} x_{ik}$ depende da função de ligação escolhida).
2. Fazer $\hat{\boldsymbol{\beta}}^{(m+1)} = \left(\boldsymbol{X}^T W^{(m)} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T W^{(m)} \boldsymbol{z}^{(m)}$.

O critério de paragem utilizado é, por exemplo

$$\frac{\| \hat{\boldsymbol{\beta}}^{(m+1)} - \hat{\boldsymbol{\beta}}^{(m)} \|}{\| \hat{\boldsymbol{\beta}}^{(m)} \|} < \epsilon,$$

para algum $\epsilon > 0$, usualmente 10^{-16}.

Note-se que o sucesso deste algoritmo está condicionado pela existência da matriz inversa de $\mathcal{I}(\hat{\boldsymbol{\beta}}^{(m)})$ em cada iteração. Como se assumiu que $\boldsymbol{X}^T \boldsymbol{X}$ tem característica $(p+1)$, a inversa de $\mathcal{I}(\boldsymbol{\beta})$ existe desde que os elementos de $W^{(m)}$ sejam na sua maioria positivos.

2.5. Propriedades assintóticas dos estimadores de máxima verosimilhança

Para inferir sobre o vector de parâmetros $\boldsymbol{\beta}$, nomeadamente, para fazer testes de hipóteses e obter intervalos de confiança, é necessário conhecer a distribuição amostral do estimador.

Se as variáveis resposta forem consideradas normalmente distribuídas, a distribuição amostral usada para a inferência é determinada com exactidão. Porém, para outras distribuições precisamos de recorrer a resultados assimptóticos baseados no

Teorema do Limite Central, que se verificam para grandes amostras quando os modelos em estudo satisfazem certas condições de regularidade. De facto, estas condições são verificadas pelos MLG. Em *Fahrmeir e Kaufmann* (1985) são estabelecidas condições que garantem a consistência e a normalidade assintótica do estimador de máxima verosimilhança, $\hat{\boldsymbol{\beta}}$, dos parâmetros dos MGL.

O estimador de máxima verosimilhança de $\boldsymbol{\beta}$ é obtido como solução de $U(\hat{\boldsymbol{\beta}}) = 0$, onde $U(\boldsymbol{\beta})$ é o vector *score*. Sabemos também que sob as condições de regularidade $E(U(\boldsymbol{\beta})) = 0$ e $Cov(U(\boldsymbol{\beta})) = E(U(\boldsymbol{\beta})U(\boldsymbol{\beta})^T) = \mathcal{I}(\boldsymbol{\beta})$. Considerando uma amostra grande, pelo Teorema do Limite Central temos a garantia de que, pelo menos assintóticamente, $U(\boldsymbol{\beta})$ tem uma distribuição normal multivariada de média $\mathbf{0}$ e matriz variâncias-covariâncias $\mathcal{I}(\boldsymbol{\beta})$. Então, para grandes amostras, a estatística $U(\boldsymbol{\beta})^T \mathcal{I}(\boldsymbol{\beta}) U(\boldsymbol{\beta})$ tem uma distribuição assintótica de um qui-quadrado com $(p+1)$ graus de liberdade, tantos quanto a dimensão de $\boldsymbol{\beta}$, ou seja

$$U(\boldsymbol{\beta})^T \mathcal{I}(\boldsymbol{\beta}) U(\boldsymbol{\beta}) \overset{\bullet}{\sim} \chi^2_{(p+1)}$$

A partir da distribuição assintótica do *vector score*, vamos apresentar a distribuição assintótica do estimador de máxima verosimilhança $\hat{\boldsymbol{\beta}}$. Se desenvolvermos $U(\beta)$ em série de Taylor em torno de $\hat{\boldsymbol{\beta}}$ e retivermos apenas os dois primeiros termos, obtemos

$$U(\boldsymbol{\beta}) \approx U(\hat{\boldsymbol{\beta}}) + \mathcal{H}\left(\hat{\boldsymbol{\beta}}\right)(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).$$

Atendento a que $U(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ e $-\mathcal{H}\left(\hat{\boldsymbol{\beta}}\right) \approx \mathcal{I}(\boldsymbol{\beta})$, o que admitimos ser verdade para grandes amostras (cf. Fahrmeir & Kaufmann, 1985, p. 360), obtemos

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx \mathcal{I}^{-1}(\boldsymbol{\beta}) U(\boldsymbol{\beta}).$$

A partir da expressão anterior podemos deduzir algumas propriedades assintóticas do estimador de máxima verosimilhança de $\boldsymbol{\beta}$.

1. $E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx E\left(\mathcal{I}^{-1}(\boldsymbol{\beta}) U(\boldsymbol{\beta})\right) = 0$, isto é, $\hat{\boldsymbol{\beta}}$ é um estimador de $\boldsymbol{\beta}$ assintóticamente cêntrico.

2. $Cov(\hat{\boldsymbol{\beta}}) \approx E\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\right] = \mathcal{I}^{-1}(\boldsymbol{\beta})$, onde $Cov\left(\hat{\boldsymbol{\beta}}\right)$ denota a matriz de variâncias-covariâncias do vector $\hat{\boldsymbol{\beta}}$.

3. A distribuição assintótica de $\hat{\boldsymbol{\beta}}$ é normal multivariada com vector médio $\boldsymbol{\beta}$ e matriz de variâncias-covariâncias $\mathcal{I}^{-1}(\boldsymbol{\beta})$, isto é

$$\hat{\boldsymbol{\beta}} \overset{\bullet}{\sim} N_{(p+1)}\left(\boldsymbol{\beta}, \mathcal{I}^{-1}(\boldsymbol{\beta})\right).$$

4. É possível verificar que

$$(\hat{\beta} - \beta)^T \mathcal{I}(\boldsymbol{\beta}) (\hat{\beta} - \beta) \overset{\bullet}{\sim} \chi^2_{(p+1)}.$$

   Esta estatística é conhecida por *estatística de Wald*.

5. A distruibuição assintótica de $\hat{\beta}_j$, $j = 0, \ldots, p$, é uma normal de parâmetros $\hat{\beta}_j$ e $\mathcal{I}^{-1}_{jj}(\beta)$, que se representa por

$$\hat{\beta}_j \overset{\bullet}{\sim} N\left(\hat{\beta}_j, \mathcal{I}^{-1}_{jj}(\beta)\right),$$

   onde $\mathcal{I}^{-1}_{jj}(\beta)$ é o elemento (j, j) de $\mathcal{I}^{-1}(\boldsymbol{\beta})$.

Estes resultados são úteis para a construção de intervalos de confiança e testes de hipóteses para $\boldsymbol{\beta}$. A estatística de Wald é uma das estatísticas utilizadas para fazer testes de hipóteses sobre o vector $\boldsymbol{\beta}$. No entanto, este vector é desconhecido e portanto $\mathcal{I}(\boldsymbol{\beta})$ também, visto que depende de $\boldsymbol{\beta}$. De forma a contornar este problema, na prática costuma-se substituir esta matriz por outra conhecida, a matriz de informação de Fisher calculada para a estimativa $\hat{\boldsymbol{\beta}}$.

2.6. Testes de hipóteses e intervalos de confiança

Os problemas de inferência relacionados com testes de hipóteses sobre o vector $\boldsymbol{\beta}$ prendem-se com o facto de se querer testar em simultâneo hipóteses sobre várias combinações lineares dos parâmetros. Como tal, as hipóteses dos testes a q combinações lineares são formulados genericamente da seguinte forma

$$H_0 : C\boldsymbol{\beta} = \boldsymbol{\xi} \quad \text{vs} \quad H_1 : C\boldsymbol{\beta} \neq \boldsymbol{\xi},$$

onde C é uma matriz não aleatória de dimensão $q \times (p + 1)$, com $q \leq p + 1$, de característica q e $\boldsymbol{\xi}$ é um vector de dimensão q.

Quando estamos interessados em testar se as covariáveis são relevantes para o modelo, usualmente utilizamos casos especiais que apresentamos de seguida. Se pretendermos testar cada covariável isoladamente, por exemplo a covariável j, as hipóteses a considerar são

$$H_0' : \beta_j = 0 \quad \text{vs} \quad H_1' : \beta_j \neq 0,$$

sendo, neste caso, a matriz C dada por $C = (0, \ldots, 0, 1, 0, \ldots, 0)$, onde 1 ocupa a j-ésima posição e $\xi = 0$. Caso estejamos interessados em testar a nulidade de um subvector com k componentes de $\boldsymbol{\beta}$, $(\beta_{l_1}, ..., \beta_{l_k})$, com $\{l_1, ..., l_k\} \subset \{0, ..., p\}$, as hipóteses em teste são

$$H_0^{''} : \beta_{l_j} = 0, \forall j \in 1, ..., k \quad \text{vs } H_1^{''} : \exists j \in \{1, ..., k\} : \beta_{l_j} \neq 0.$$

No caso de $\{l_1, ..., l_k\} = 1, ..., k$, a matriz C toma a forma

$$C = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 0 \end{bmatrix} = \left(I_k, \mathbf{0}_{k\times(p+1-k)}\right),$$

onde I_k é a matriz identidade de ordem k e $\mathbf{0}_{k\times(p+1-k)}$ é uma matriz de zeros de dimensão $k \times (p + 1 - k)$. Além disso, $\boldsymbol{\xi} = \mathbf{0}_k$.

A definição das hipóteses utilizando submodelos do modelo constituído por todas as covariáveis consideradas no estudo são importantes para seleccionar as covariáveis significativas para o modelo. Para testar as hipóteses referidas recorre-se usualmente a três estatísticas diferentes que apresentamos de seguida.

Teste Wald

Como foi dito atrás o estimador de máxima verosimilhança de $\boldsymbol{\beta}$ segue assimptóticamente uma lei normal multivariada de média $\boldsymbol{\beta}$ e matriz de variâncias-covariâncias $I(\hat{\boldsymbol{\beta}})$, admitindo que para grandes amostras se tem $I(\boldsymbol{\beta}) \approx I(\hat{\boldsymbol{\beta}})$. Uma vez que $C\hat{\boldsymbol{\beta}}$ é uma transformação linear de $\hat{\boldsymbol{\beta}}$ temos, pelas propriedades da distribuição normal multivariada, que

$$C\hat{\boldsymbol{\beta}} \overset{\bullet}{\sim} N_{(q)}\left(C\boldsymbol{\beta}, C\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})C^T\right),$$

e consequentemente, sob H_0, a *estatística de Wald* é definida por

$$\mathcal{W} = \left(C\hat{\boldsymbol{\beta}} - \xi\right)^T \left[C\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})C^T\right] \left(C\hat{\boldsymbol{\beta}} - \xi\right)$$

e tem uma distribuição assintótica de um χ^2 com q graus de liberdade. Assim, rejeitamos H_0 a um nível de significância α, se o valor observado de $\mathcal{W}$ for superior ao quantil de probabilidade $(1 - \alpha)$ de um qui-quadrado com q graus de liberdade.

Teste Score

Outra estatística para testar H_0, baseada no vector score $U(\boldsymbol{\beta})$, é a *estatística score* $\mathcal{S}$. Atendendo que $\hat{\boldsymbol{\beta}}$ é o estimador de máxima verosimilhança de $\boldsymbol{\beta}$ temos que $U(\hat{\boldsymbol{\beta}}) = 0$. Consideremos $\tilde{\boldsymbol{\beta}}$ o estimador de máxima verosimilhança de $\boldsymbol{\beta}$ sob H_0. Se substituirmos $\hat{\boldsymbol{\beta}}$ por $\tilde{\boldsymbol{\beta}}$ constatamos que o valor de $U(\tilde{\boldsymbol{\beta}})$ se afasta de $U(\hat{\boldsymbol{\beta}})$, caso H_0

não se verifique, pelo que podemos concluir que valores pequenos de $U(\tilde{\boldsymbol{\beta}})$ levam à não rejeição de H_0. A estatística $\mathcal{S}$ determina a diferença entre $U(\tilde{\boldsymbol{\beta}})$ e o vector nulo e é dada por

$$\mathcal{S} = \left[U\left(\tilde{\boldsymbol{\beta}}\right)\right]^T \mathcal{I}^{-1}(\tilde{\boldsymbol{\beta}}) U\left(\tilde{\boldsymbol{\beta}}\right).$$

Usando esta *estatística* rejeitamos H_0 a um nível de significância α se o valor observado da *estatística score* for superior ao quantil de probabilidade $(1-\alpha)$ de um qui-quadrado com q graus de liberdade.

Teste de razão de verosimilhanças

O teste de razão de verosimilhanças é utilizado quando se pretende comparar modelos encaixados, isto é, quando um modelo é submodelo do outro (Turkman & Silva, 2000, p.51). Portanto iremos comparar o modelo sob $H_0 \cup H_1$ com o submodelo restrito às condições de H_0, utilizando a *estatística de razão de verosimilhanças* ou *estatística de Wilks* que é definida por

$$\Lambda = -2\frac{max_{H_0} L(\boldsymbol{\beta})}{max_{H_0 \cup H_1} L(\boldsymbol{\beta})} = -2\{\ell(\tilde{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}})\},$$

onde $\tilde{\boldsymbol{\beta}}$ é o estimador de máxima verosimilhança de $\boldsymbol{\beta}$ restrito a H_0, ou seja, é o valor de $\boldsymbol{\beta}$ que maximiza a função de verosimilhança sujeito às restrições impostas pela hipotese nula, $C\boldsymbol{\beta} = \boldsymbol{\xi}$. A estatística Λ tem, sob H_0, uma distribuição assintótica de um χ^2 sendo o número de graus de liberdade igual à diferença entre o número de parâmetros a estimar sob $H_0 \cup H_1$ (neste caso $p+1$) e o número de parâmetros a estimar sob H_0 $(p+1-q)$. Assim, sob H_0

$$\Lambda = -2\left(l(\tilde{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}})\right) \overset{\bullet}{\sim} \chi^2_q.$$

À semelhança dos outros testes, usando a *estatística de razão de verosimilhanças* rejeitamos a hipótese nula a um nível de significância α se o valor observado de Λ for superior ao quantil de probabilidade $(1-\alpha)$ de um qui-quadrado com q graus de liberdade.

Intervalos de confiança

Para construir um intervalo de confiança assintótico para o parâmetro β_j, $j = 0, \ldots, p$, ao nível de significância α, recorremos à distribuição assintótica de $\hat{\beta}_j$, $N\left(\beta_j, \mathcal{I}^{-1}_{jj}(\beta)\right)$. Tal intervalo é dado por

$$\left]\hat{\beta}_j - z_{(1-\frac{\alpha}{2})}\left(\mathcal{I}^{-1}_{jj}(\beta)\right)^{\frac{1}{2}}, \hat{\beta}_j + z_{(1-\frac{\alpha}{2})}\left(\mathcal{I}^{-1}_{jj}(\beta)\right)^{\frac{1}{2}}\right[,$$

onde $z_{\left(1-\frac{\alpha}{2}\right)}$ é o quantil $(1-\frac{\alpha}{2})$ da lei normal standard.

Capítulo 3

Regressão logística para variáveis de resposta binária

Geralmente nos modelos de regressão a variável resposta Y é uma variável aleatória contínua. No entanto, em determinadas situações a variável Y pode ser discreta, admitindo dois ou mais valores, sendo que cada um destes valores representa uma categoria. As covariáveis dividem-se em dois tipos de variáveis, as ordinais e as nominais. Neste capítulo consideramos a variável resposta binária ou dicotómica, isto é, toma apenas os valores 0 e 1. De forma a modelar este tipo de dados utilizamos o modelo de regressão logística. Foram propostos outros modelos para análise de variáveis binárias (Cox and Snell, citado por Hosmer e Lemeshow 2000), contudo existem duas razões fundamentais para usar o modelo de regressão logística: o facto da função logística ter uma interpretação com significado clínico e ter boas propriedades matemáticas.

Uma análise de regressão logística pode ser univariável ou multivariável. No primeiro caso o modelo é constituído por apenas uma covariável ou variável independente enquanto no segundo caso o modelo é uma generalização do primeiro, o qual é constituído por mais de uma covariável.

3.1. Regressão logística univariável

Consideremos uma variável determinista x e uma variável aleatória real Y binária, tal que

- $P(Y=1) = \pi(x)$, probabilidade de ter sucesso
- $P(Y=0) = 1 - \pi(x)$, probabilidade de ter insucesso

Verifica-se que $Y \sim \mathcal{B}(\pi(x))$.

Em qualquer modelo de regressão pretende-se determinar $E(Y)$. Na regressão linear, esta esperança é dada como uma equação linear em x

$$E(Y) = \beta_0 + \beta_1 x.$$

Nesta expressão, $E(Y)$ e x podem tomar qualquer valor entre $-\infty$ e $+\infty$. Contudo, na regressão logística Y é uma variável binária o que implica que o valor de $E(Y)$ varie no intervalo $[0, 1]$. Da definição de esperança temos

$$E(Y) = \pi(x).$$

Se aplicarmos a transformação *logit* à função $\pi(x)$, onde *logit* é a função de ligação para o modelo de regressão logística que se denota $g(x)$, obtemos

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x.$$

A expressão anterior define o modelo de regressão logística para uma covariável e é equivalente a

$$\pi(x) = \frac{\exp\left(g(x)\right)}{1+\exp\left(g(x)\right)} \quad \text{ou} \quad \pi(x) = \frac{\exp\left(\beta_0+\beta_1 x\right)}{1+\exp(\beta_0+\beta_1 x)}.$$

A importância desta transformação prende-se com o facto de $g(x)$ apresentar algumas propriedades do modelo de regressão linear, nomeadamente a linearidade, a continuidade e o facto de poder tomar qualquer valor entre $-\infty$ e $+\infty$, dependendo do intervalo onde a covariável variar.

Seguidamente fazemos uma breve análise à função $\pi(x)$. Podemos verificar que

$$\lim_{x\to-\infty} \pi(x) = \frac{\exp\left(\beta_0+\beta_1 x\right)}{1+\exp(\beta_0+\beta_1 x)} = 0$$

$$\begin{aligned}\lim_{x\to+\infty} \pi(x) &= \frac{\exp\left(\beta_0+\beta_1 x\right)}{1+\exp(\beta_0+\beta_1 x)} \\ &= \frac{1}{\frac{1}{\exp(\beta_0+\beta_1 x)}+1} \\ &= 1.\end{aligned}$$

Concluímos que qualquer que seja o valor de x a função $\pi(x)$ irá variar no intervalo $]0, 1[$. No que diz respeito à monotonia da função, analisando a derivada de $\pi(x)$,

$$\pi'(x) = \frac{\beta_1 \exp\left(\beta_0+\beta_1 x\right)}{\left[1+\exp\left(\beta_0+\beta_1 x\right)\right]^2},$$

verificamos que $\pi(x)$ é crescente se $\beta_1 < 0$ e decrescente se $\beta_1 > 0$. Se $\beta_1 = 0$ torna-se claro que a variável Y não depende do valor de x.

Se considerarmos $\beta_0 = 0$ e $\beta_1 = -1$, $\pi(x)$ tem a forma da função de distribuição logística de parâmetros $\mu = 0$ e $\sigma = 1$, donde se tira que $\pi(x) = \frac{\exp\left(-x\right)}{1+\exp(-x)}$. A função de distribuição logística é

$$F(x) = \frac{exp(-(x-\mu)/\sigma)}{1+exp(-(x-\mu)/\sigma)}.$$

De seguida vamos estimar pelo método da máxima verosimilhança os parâmetros do modelo, β_0 e β_1. Consideremos dados da forma (x_i, y_i), $i = 1, 2, \ldots, n$, onde as covariáveis associadas a cada indivíduo i são, repectivamente, $x_1, x_2, \ldots, x_n$ e $Y_1, Y_2, \ldots, Y_n$ as variáveis resposta. $Y_i \sim \mathcal{B}(\pi_i)$, com

$$\pi_i = \pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)},\ i = 1, 2, \ldots, n.$$

Sendo Y uma variável de Bernoulli, temos que a sua função de probabilidade é dada por

$$f_{y_i}(\pi_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}, y_i = 0, 1;\ \ i = 1, \ldots, n.$$

A função de máxima verosimilhança é da forma

$$L_y(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i{}^{y_i} (1 - \pi_i)^{1-y_i}, \tag{3.1}$$

onde $y = (y_1, y_2, \ldots, y_n) \in \{0, 1\}^n$ é um valor particular de uma amostra de Y e $\boldsymbol{\beta} = [\beta_0\ \beta_1]^T$. A função Log-Verosimilhança pode ser escrita

$$\begin{aligned}
\ell_y(\beta_0, \beta_1) &= \log\left(\prod_{i=1}^{n} \pi_i{}^{y_i} (1 - \pi_i)^{1-y_i}\right) \\
&= \sum_{i=1}^{n} y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + \log(1 - \pi_i) \\
&= \sum_{i=0}^{n} y_i \exp(\beta_0 + \beta_1 x_i) - [\log(1 + \exp(\beta_0 + \beta_1 x_i))]
\end{aligned}$$

O valor que máximiza a função $\ell_y(\beta_0, \beta_1)$ pode ser obtido resolvendo um sistema de equações. Derivando $\ell_y(\beta_0, \beta_1)$ em ordem aos parâmetros do modelo e igualando a zero obtemos as equações de verosimilhança

$$\begin{cases}
\displaystyle\sum_{i=1}^{n} \left(y_i - \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}\right) = 0 \\
\displaystyle\sum_{i=1}^{n} x_i \left(y_i - \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}\right) = 0
\end{cases}$$

Como as equações são não lineares é necessário recorrer a métodos numéricos para encontrar a solução, obtendo assim uma estimativa de máxima verosimilhança para $\boldsymbol{\beta} = [\beta_0\ \beta_1]^T$, $\hat{\boldsymbol{\beta}}$. Vamos novamente utilizar a adaptação do método de Newton referida anteriormente, em que a matriz $-\mathcal{H}$ é substituída pela matriz de informação de Fisher. Esta matriz para uma covariável é

$$\mathcal{I} = \begin{bmatrix}
\displaystyle\sum_{i=0}^{n} \frac{\exp(\beta_0 + \beta_1 x_i)}{[1 + \exp(\beta_0 + \beta_1 x_i)]^2} & \displaystyle\sum_{i=0}^{n} x_i \frac{\exp(\beta_0 + \beta_1 x_i)}{[1 + \exp(\beta_0 + \beta_1 x_i)]^2} \\
\displaystyle\sum_{i=0}^{n} x_i \frac{\exp(\beta_0 + \beta_1 x_i)}{[1 + \exp(\beta_0 + \beta_1 x_i)]^2} & \displaystyle\sum_{i=0}^{n} x_i^2 \frac{\exp(\beta_0 + \beta_1 x_i)}{[1 + \exp(\beta_0 + \beta_1 x_i)]^2}
\end{bmatrix}.$$

O primeiro passo a efectuar é desenvolver $U(\boldsymbol{\beta})$ em série de Taylor em torno do ponto $\boldsymbol{\beta}^{(0)}$. A expressão que se obtém retendo somente os termos de primeira ordem é

$$U(\boldsymbol{\beta}) \approx U\left(\beta^{(0)}\right) + \mathcal{I}\left(\beta^{(0)}\right)\left(\beta - \beta^{(0)}\right)^{1}.$$

Sendo $\hat{\boldsymbol{\beta}}^{(0)}$ o ponto inicial, a estimativa de $\hat{\boldsymbol{\beta}}$ é obtida a partir do seguinte processo iterativo

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \left(\mathcal{I}\left(\hat{\boldsymbol{\beta}}^{(k)}\right)\right)^{-1} U\left(\hat{\boldsymbol{\beta}}^{(k)}\right), \;\; k = 0, 1, 2 \ldots$$

onde

$$U(\boldsymbol{\beta}) = \begin{bmatrix} \displaystyle\sum_{i=1}^{n}\left(y_i - \frac{\exp\left(\beta_0 + \beta_1 x_i\right)}{1 + \exp(\beta_0 + \beta_1 x_i)}\right) \\ \\ \displaystyle\sum_{i=1}^{n} x_i\left(y_i - \frac{\exp\left(\beta_0 + \beta_1 x_i\right)}{1 + \exp(\beta_0 + \beta_1 x_i)}\right) \end{bmatrix}.$$

Após encontrar estimativa dos parâmetros do modelo devemos testar se a covariável possui uma relação significativa com a variável resposta, isto é, pretendemos saber se a covariável é relevante para o modelo. O teste utilizado é apresentado a seguir

$$H_0 : \beta_1 = 0 \;\; \text{vs} \;\; H_1 : \beta_1 \neq 0$$

A hipótese nula pode ser testada utilizando várias estatíticas de teste, nomeadamente, a *estatística de Wald* e *estatística de Scores* que se apresentam de seguida para uma covariável. Estas estatísticas são uma particularização das apresentadas no capítulo anterior, portanto apenas vamos referir a estatística de teste para uma covariável.

Teste Wald

A distribuição assintótica de $\hat{\beta}$ é normal com média β e matriz de variâncias-covariâncias $\mathcal{I}^{-1}(\beta)$. Consequentemente, sob H_0 temos

$$W = \frac{\hat{\beta}_1^2}{\sigma_{22}} \overset{\bullet}{\sim} \chi_1^2$$

onde σ_{22} é o segundo elemento da diagonal principal de $\mathcal{I}(\hat{\beta})$. À estatística W damos o nome de *estatística de Wald*. Também aqui rejeitamos H_0 a um nível de significância α, se o valor observado de W for superior ao quantil de probabilidade $1 - \alpha$ de um χ_1^2.

Teste Scores

Designando por $\tilde{\beta}$ o estimador de máxima verosimilhança de β sob H_0, a *estatística score* é dada por

$$\mathcal{S} = U(\tilde{\beta})^T \mathcal{I}(\tilde{\beta})^{-1} U(\tilde{\beta}),$$

e tem distribuição assintótica de um χ^2_1. Usando a *estatística score* rejeitamos H_0 a um nível de significância α se o valor observado de S for superior ao quantil de probabilidade $1-\alpha$ de um χ^2_1.

3.2. Regressão logística multivariável

Nesta secção pretende-se apresentar resumidamente a generalização do modelo de regressão logística univariável. Apresentamos o modelo logístico com mais de uma covariável.

Consideremos um conjunto de p covariáveis, $x_1, \ldots, x_p$, e $\boldsymbol{x}^T = (1, x_1, \ldots, x_p)$. Analogamente ao que foi apresentado anteriormente, o modelo de regressão logística multivariável é dado pela expressão que define a probabilidade de que o acontecimento de interesse ocorra

$$\pi(\boldsymbol{x}) = P(Y=1) = \frac{\exp(\beta_0 + \sum_{i=1}^{p} \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{p} \beta_i x_i)},$$

onde β_i é o coeficiente associado à covariável x_i.

Tendo em conta as diversas áreas onde a regressão logística pode ser aplicada, existem várias possibilidades de escolha para as covariáveis, nomeadamente, podem ser sexo, cor dos olhos, etc. Assim, surge a necessidade de atribuir valores numéricos, meramente identificativos, a cada categoria da variável. Segundo Hosmer e Lemeshow (2000) é necessário criar um conjunto de variáveis *dummy* ou codificadoras. Estas variáveis são definidas da seguinte forma

$$D_{jl} = \begin{cases} 1, & \text{se o indivíduo verifica a categoria } l \text{ da covariável } j \\ 0, & \text{caso contrário} \end{cases}.$$

O modelo de regressão logística multivariável com p covariáveis em que a j-ésima covariável é discreta com k_j categorias é definido em termos da função *logit* por

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{l=1}^{k_j - 1} \beta_{jl} D_{jl} + \cdots + \beta_p x_p,$$

onde D_{jl} denota a variável *dummy* e β_{jl} o coeficiente, ambos associados à categoria l da covariável j. A estimação e inferência decorrem dos resultados apresentados

no capítulo 2 fazendo a respectiva particularização, isto é, utilizando a função de ligação *logit*.

3.3. *Odds ratio*

A medida de associação *odds ratio* (OR) é utilizada usualmente na regressão logística univariável para complementar o teste à significância da covariável (x). O facto de existir uma relação entre os parâmetros do modelo logístico e o *odds ratio* constitui a principal vantagem de utilização desta medida. Com vista a apresentar essa relação começamos por assumir que a covariável é binária. O *odds ratio* é dado pelo quociente entre a odds do acontecimento de interesse ocorrer ($Y = 1$) nos individuos com $x = 1$ e a odds desse acontecimento ocorrer nos indivíduos com $x = 0$. A odds do acontecimento de interesse ocorrer nos indivíduos com $x = 1$ é definida por $\frac{\pi(1)}{1-\pi(1)}$. Analogamente, a odds do acontecimento de interesse ocorrer nos indivíduos com $x = 0$ é definida por $\frac{\pi(0)}{1-\pi(0)}$. Assim, o *odds ratio* é uma forma de comparar se a probabilidade do acontecimento de interesse ocorrer é a mesma para os indivíduos com $x = 1$ ou $x = 0$.

As probabilidades do acontecimento de interesse ocorrer para as duas categorias de x, são dadas respectivamente por

$$\pi(1) = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} \quad \text{e} \quad \pi(0) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}.$$

Consequentemente, o valor do *odds ratio* é dado pela expressão

$$OR = \frac{\pi(1)[1 - \pi(0)]}{\pi(0)[1 - \pi(1)]} = \exp(\beta_1),$$

tornando evidente a relação entre o *odds ratio* e o coefeciente do modelo. O valor do OR depende da codificação adoptada para covariável binária x, que pode ser definida por quaisquer dois valores. Considerando a codificação usando os valores genéricos a e b, o valor do *odds ratio* é dado por

$$OR = \frac{\pi(a)[1 - \pi(b)]}{\pi(b)[1 - \pi(a)]} = \exp(\beta_1(a - b)).$$

Constata-se que a interpretação do *odds ratio* não pode ser feita sem antes conhecer a codificação de x. Habitualmente a codificação adoptada é definida em termos de 0 e 1, por permitir uma interpretação trivial dos parâmetros. Na prática o cálculo do *odds ratio* é feito a partir de dados reais geralmente organizados em tabelas contingência. Destas tabelas podemos obter uma estimativa do *odds ratio*, $\hat{OR}$, à

qual podemos aplicar o logaritmo obtendo uma estimativa de β_1 ($\log(\hat{OR}) = \hat{\beta}_1$). O intervalo de confiança a 95% para $\hat{OR}$ obtém-se exponenciando os extremos do intervalo de confiança de $\hat{\beta}_1$. No caso da covariável ter mais de duas categorias ($k > 2$), para determinar o valor do *odds ratio* é necessário utilizar $k - 1$ variáveis codificadoras, denominadas variáveis *dummy*. Usualmente a primeira categoria ($i = 1$) é considerada como classe de referência e toma o valor zero para as $k-1$ variáveis *dummy*. Nas variáveis *dummy* associadas às restantes categorias ($i = 2, \ldots, k$) a entrada i toma o valor 1 e as restantes tomam o valor 0. Após definir estas variáveis, o cálculo do OR é efectuado de forma análoga ao caso em que se consideram duas categorias.

3.4. Selecção das covariáveis

Existem vários algoritmos para a selecção dos preditores. Os métodos disponíveis no *SPSS* mais utilizados são os algoritmos de selecção *forward stepwise* e *backward stepwise*, procedimentos que seguem direcções opostas. O primeiro começa com o modelo mais simples, modelo apenas com a constante, e vai ao encontro de modelos mais complexos. O segundo começa com o modelo completo, com todas as covariáveis, e vai eliminando covariáveis até chegar a um modelo mais simples, onde já nenhuma covariável possa ser eliminada de acordo com uma regra estabelecida previamente para a eliminação das covariáveis. Apresentamos de seguida o algoritmo do método de selecção *Forward Stepwise*.

Algoritmo de selecção *Forward Stepwise*

1. Ajustar o modelo apenas com a constante (modelo nulo);

2. Comparar o modelo nulo com os modelos de regressão logística univariáveis associados a cada uma das covariáveis. O menor p-valor encontrado será comparado com o p-valor de entrada escolhido previamente (P_e =0,05). Se o menor p-valor encontrado for inferior P_e, então a covariável é incluída no modelo e passar ao passo (3). Caso contrário o algoritmo termina e o modelo final é o modelo nulo;

3. Partindo do modelo com a covariável explicativa seleccionada no passo anterior, introduzir individualmente as restantes covariáveis e testar cada um destes novos modelos contra o modelo do passo (2). Se o menor p-valor encontrado

for inferior a P_e, incluir no modelo a respectiva variável e passar ao passo (4). Caso contrário, terminar a selecção e ficar com o modelo encontrado em (2);

4. Comparar o modelo obtido em (3) com os modelos que resultam por exclusão individual de cada uma das covariáveis desse modelo. Se o maior dos p-valores calculados for inferior ao p-valor de saída escolhido previamente (P_s=0,1), a covariável associada a esse p-valor permanece no modelo. Caso contrário, ela é removida. Em qualquer dos casos, ir para o passo (5). Verificar, a partir do modelo ajustado em (3), se existe algum p-valor superior a 0,1 e se existir remover a covariável correspondente a esse p-valor;

5. Ajustar o modelo encontrado no passo anterior e voltar ao passo (3). Repetir o algoritmo até se atingir uma condição de paragem: todas as covariáveis foram incluídas no modelo ou todas as covariáveis incluídas no modelo têm p-valores inferiores a P_s e superiores a P_e .

Os testes mais utilizados na selecção das covariáveis são o teste de *Wald* e o teste de *razão de verosimilhanças*.

3.5. Teste de *Hosmer and Lemeshow*

Hosmer e Lemeshow (2000) propuseram um teste de ajustamento muito utilizado na regressão logística que tem como hipótese nula que o modelo é o adequado. Este teste tem como base a divisão dos dados em g grupos segundo as probabilidades estimadas. Hosmer e Lemeshow definiram duas formas de o fazer. Suponhamos que temos uma amostra com n valores distintos de $\boldsymbol{x}^T = (1, x_1, \ldots, x_p)$, aos quais correspondem n probabilidades estimadas. Os dois tipos de agrupamento são

1. Agrupamento baseado nos percentis das probabilidades estimadas. Fixa-se $g = 10$ em que o primeiro grupo contém os $n'_1 = \dfrac{n}{10}$ indivíduos com as probabilidades de menor valor e o último grupo terá os $n'_{10} = \dfrac{n}{10}$ indivíduos com as probabilidades estimadas mais elevadas.

2. Agrupamento baseado em cut-points pré-fixados. Fixamos 10 grupos construídos segundo cut-points pré-fixados, $\dfrac{k}{10}$, $k = 1, \ldots, 9$. Cada grupo contém todos os indivíduos com probabilidades entre os cut-points dos grupos adjacentes.

Seguidamente determinamos as frequências esperadas para $Y = 1$ obtidas somando as probabilidades estimadas de cada indivíduo do grupo. Para $Y = 0$ estas são

dadas pela soma de (1-probabilidades estimadas) de todos os indivíduos do grupo.

A estatística de teste, $\hat{C}$, é dada por

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)},$$

onde

- $O_k = \sum_{j=1}^{C_k} Y_j$, com C_k o número de valores diferentes do conjunto das p covariáveis observadas.
- n'_k é o número de indivíduos no k-ésimo grupo.
- $\bar{\pi}_k = \sum_{j=1}^{C_k} \frac{m_j \hat{\pi}_j}{n'_k}$, com m_j o número de indivíduos que possuem o mesmo conjunto de valores das covariáveis, $\boldsymbol{x} = \boldsymbol{x_j}$ e $\hat{\pi}_j$ é a probabilidade estimada associada a j.

Segundo Hosmer e Lemeshow, $\hat{C}$ segue assintóticamente um qui-quadradro com $g-2$ graus de liberdade. O método de agrupamento mais utilizado é o dos percentis. Salientamos ainda que o valor de $\hat{C}$ depende dos grupos escolhidos.

3.6. Tabelas de classificação

Um modelo de regressão pode ser estatisticamente significativo e não representar a realidade em estudo. Uma das formas de avaliar a eficiência classificativa do modelo é através de tabelas classificativas.

Para a construção destas tabelas precisamos de calcular as probabilidades estimadas para a ocorrência do endpoint e, de seguida, determinamos o cut-off, c, para estas probabilidades. A partir do cut-off vamos assumir que indivíduos com probabilidades estimadas superiores a c experimentam o endpoint e indivíduos com probabilidades abaixo do cut-off não o experimentam. O valor de cut-off usualmente utilizado é 0.5, contudo nem sempre é o mais adequado. De forma a encontrar o valor mais apropriado usamos gráficos, por exemplo a curva ROC, que nos permitem identificar o valor para o qual a sensibilidade e a especificidade do modelo se equilibram.

A sensibilidade do modelo é definida como a probabilidade de prevermos a ocorrência do enpoint entre os indivíduos em que este foi observado. A sensibilidade dá-nos a proporção de verdadeiros positivos. A especificidade fornece a proporção de falsos negativos, valor este que é determinado pela probabilidade de prevermos a não ocorrência do enpoint entre os indivíduos em que este não foi observado (cf. Braga, 2000). De forma resumida

Tabela 3.1: Tabela de classificação

		Estimados	
		Endpoint=1	Endpoint=0
Observados	Endpoint=1	A	B
	Endpoint=0	C	D
	Desempenho	Sensibilidade $\frac{A}{A+B}$	Especificidade $\frac{D}{C+D}$

Determinado o cut-off, podemos construir a tabela cujas linhas apresentam os valores observados e as colunas os valores estimados para a variável resposta. Num modelo perfeito todos os casos estariam sobre a diagonal principal. Contudo, na prática é muito difícil obter um modelo perfeito e, como tal, teremos de classificar a sua capacidade preditora. Esta é considerada boa se a sensibilidade e a especificidade forem superiores a 80%, razoável se estes dois valores estiverem entre 50% e 80% e medíocre se ambos forem inferiores a 50%.

3.7. Curva ROC

A análise da curva ROC (Receiver Operating Characteristic) pode ser feita por meio de um gráfico que nos permite estudar a variação da sensibilidade e especificidade para cada valor de cut-off. A sensibilidade é apresentada no eixo das ordenadas e (1-especificidade) no eixo das abcissas.

O objectivo desta análise é identificar ou confirmar a qualidade do ajustamento do modelo. Quando observamos o gráfico verifica-se que o ideal seria encontrar uma área sob a curva ROC perto de 1, uma vez que, quanto mais próxima estiver a curva do canto superior esquerdo, mais verdadeiros positivos e menos falsos negativos iremos ter. Por exemplo, se tivermos uma área de 0,5, podemos dizer que o poder discriminatório do modelo é idêntico a lançar uma moeda ao ar para determinar se o indivíduo tem o endpoint ou não. Usualmente utiliza-se o seguinte critério para classificar o poder discriminatório de um modelo de regressão logística

- Se ROC = 0,5 o modelo não faz qualquer discriminação entre os indivíduos com e sem endpoint.
- Se $0{,}6 \leq \text{ROC} < 0{,}7$ o modelo apresenta uma discrimínção limitada.
- Se $0{,}7 \leq \text{ROC} < 0{,}8$ o modelo apresenta uma discrimínção aceitável.
- Se $0{,}8 \leq \text{ROC} < 0{,}9$ o modelo apresenta uma excelente discrimínção.
- Se $\text{ROC} \geq 0{,}9$ o modelo apresenta uma discrimínção quase perfeita.

Capítulo 4

Exemplo prático de aplicação da regressão logística

Neste capítulo apresentamos um estudo no qual se aplica a regressão logística a dados reais fornecidos pela *Sociedade Portuguesa de Cardiologia*. Estes dados foram recolhidos aquando do Registo Nacional de Síndromes Coronários Agudos (RNSCA). O registo iniciou-se em 1 de janeiro de 2002, efectou-se em todos os Serviços ou Departamentos de Cardiologia dos hospitais portugueses e incluiu todos os doentes internados com Síndrome Coronária Aguda que satisfizessem todos os critérios de inclusão e nenhum de exclusão. Estes critérios foram definidos pela instituição referida. O objectivo desta análise é encontrar os preditores de morte em pessoas com Síndrome Coronária Aguda, mas sem antecedentes cardiovasculares num total de 23947 doentes. Para a construção do modelo vamos utilizar 14527 doentes e fazemos a validação externa do modelo com os restantes.

A realização desta dissertação decorreu em simultâneo com um estágio na *Sociedade Portuguesa de Cardiologia*. Assim, o assunto que atribui o título a este trabalho foi aplicado na prática, culminando neste exemplo. Consequentemente, este capítulo tem especial importância nesta dissertação.

4.1. Análise exploratória de dados

Antes da construção do modelo, apresentamos as variáveis intervenientes. A variável resposta é uma variável binária, pode tomar apenas dois valores. Toma o valor 0 caso o doente não tenha sido declarado como morto, ou o valor 1 caso tenha sido declarado o óbito. Vamos dividir a população em estudo em dois grupos: GrupoI e GrupoII. Ao primeiro pertencem os doentes cujo o óbito não se verificou e ao segundo os restantes. Excluímos da análise todos os doentes cujo estado vital não foi identificado (9420). Desta forma, foram incluídos na análise 14527 doentes, dos quais 13813 foram identificados como vivos (95,1%) e 714 foram identificados como óbitos (4,9%). Consideramos 12 covariáveis: *Sexo*, *Índice de massa corporal* (IMC), *Idade*,

Fumador, *Diabetes Mellitus*, *Hipertensão Arterial*(HTA), *Dislipidémia*, *Frequencia Cardíaca*, *Pressao Arterial Sistólica*, *Pressao Arterial Diastólica*, *Classe Killip* e *Classificação Função VE*. De seguida apresentamos a caracterização de cada uma delas quanto à ocorrência de morte. Para cada uma destas foram retirados os valores missing. Note-se que para testar a associação entre cada uma das covariáveis e o endpoint (*Morte*) utilizamos a *estatístiva de Wald* apresentada na secção (3.1). Começamos pela variável *Sexo*.

Tabela 4.1: Sexo

		GrupoI	**GrupoII**	**Total**	**OR(IC 95%)**	**Teste Wald**
Masculino	n (%)	9831 (96,4%)	372 (3,6%)	10203 (100%)	Classe referência	$\boldsymbol{p < 0,001}$
Feminino	n (%)	3919 (92,0%)	339 (8,0%)	4258 (100%)	2,29 (1,96;2,66)	
Total	n (%)	13750 (95,1%)	711 (4,9%)	14461 (100%)		

Nesta amostra 70% dos doentes são do sexo masculino e 30% são do sexo feminino. Quanto à ocorrência de morte verificamos que existe maior proporção de mortes entre os doentes do sexo feminino relativamente aos do sexo masculino (8,0% vs 3,6%). Pretendemos saber se existe associação entre sexo feminino e morte. Pelo teste de Wald concluímos que existe associação entre sexo e a morte ($p < 0,001$), e portanto as diferenças observadas nas proporções são estatisticamente significativas. A estimativa do OR e o respectivo intervalo de confiança são respectivamente 2,29 e $]1,96; 2,66[$. Assim, o risco de ocorrência de morte nos doentes do sexo feminino é 2,29 vezes superior aos do sexo masculino.

Tabela 4.2: Idade

		GrupoI	**GrupoII**	**Total**	**OR(IC 95%)**	**Teste Wald**
<45	n (%)	1078 (9,1%)	10 (0,9%)	1088 (100%)	Classe referência	$\boldsymbol{p < 0,001}$
45-64	n (%)	5525 (98,3%)	95 (1,7%)	5620 (100%)	1,85 (0,96;3,57)	$p = 0,065$
65-74	n (%)	3543 (95,6%)	165 (4,4%)	3708 (100%)	5,02 (2,64;9,54)	$p < 0,001$
>=75	n (%)	3577 (89,0%)	440 (11,0%)	4017 (100%)	13,26 (7,06;24,91)	$p < 0,001$
Total	n (%)	13723 (95,1%)	710 (4,9%)	14433 (100%)		

A idade dos doentes foi categorizada, sendo a classe menos comum constituída pelos indivíduos com menos de 45 anos (7,5%). Quanto ao endpoint, observamos que nesta amostra a maior percentagem de mortes se verifica na faixa etária dos indivíduos com pelo menos 75 anos (11,0%). Consideramos a primeira categoria como classe de referência (< 45 anos), por ser a categoria que teoricamente possuí menor risco de ocorrência de morte. Note-se que o risco de ocorrência de morte nos indivíduos com idades entre os 65 e os 74 anos é 5,02 vezes superior ao dos

indivíduos com menos de 45 anos. Os doentes com pelo menos 75 anos têm um risco de ocorrência de morte 13,3 vezes maior que os doentes pertencentes à classe de referência. No que diz respeito aos doentes com idade entre os 45 e os 64 anos, não existem diferenças significativas, OR=1,85 e IC= $(0,96; 3,57)$. Utilizando o teste de Wald globalmente concluímos que existe associação entre idade e a morte ($p < 0,001$).

Tabela 4.3: Índice de massa corporal

		GrupoI	**GrupoII**	**Total**	**OR(IC 95%)**	**Teste Wald**
Peso baixo	n (%)	76 (91,6%)	7 (8,4%)	83 (100%)	Classe referência	$\mathbf{p < 0,001}$
Peso normal	n (%)	3665 (94,7%)	205 (5,3%)	3870 (100%)	0,61 (0,28;1,33)	$p = 0,214$
Excesso peso	n (%)	5829 (96,4%)	216 (3,6%)	6045 (100%)	0,40 (0,18;0,88)	$p = 0,023$
Obesidade grau I	n (%)	2001 (96,8%)	66 (3,2%)	2067 (100%)	0,36 (0,16;0,81)	$p = 0,013$
Obesidade grau II	n (%)	410 (96,7%)	14 (3,3%)	424 (100%)	0,37 (0,15;0,95)	$p = 0,038$
Obesidade grau III	n (%)	88 (95,7%)	4 (4,3%)	92 (100%)	0,45 (0,14;0,86)	$p = 0,274$
Total	n (%)	12069 (95,9%)	512 (4,1%)	12581 (100%)		

Nesta amostra *Peso baixo* é a categoria que possuí menor percentagem de doentes (0,7%). A maior percentagem diz respeito aos doentes com excesso peso (48%). Quanto ao endpoint, constatamos que nesta amostra, a maior percentagem de mortes se verifica nos doentes com peso baixo, 8,4%, e a menor nos doentes com obesidade de grau I, 3,2%. De acordo com os resulatdos obtidos (OR=0,36; IC= $(0,16; 0,81); p = 0,013$) verificamos que estas diferenças observadas nas proporções são estatisticamente significativas. Pela aplicação do teste de Wald globalmente concluímos que existe associação entre o índice de massa corporal e a morte, $p < 0,001$.

Tabela 4.4: Fumador

		GrupoI	**GrupoII**	**Total**	**OR(IC 95%)**	**Teste Wald**
Não	n (%)	9683 (93,8%)	643 (6,2%)	10326 (100%)	Classe de referência	$\mathbf{p < 0,001}$
Sim	n (%)	4119 (98,3%)	70 (1,7%)	4189 (100%)	0,26 (0,20;0,33)	
Total	n (%)	13802 (95,1%)	713 (4,9%)	14515 (100%)		

Para esta amostra podemos constatar que os doentes não fumadores apresentam maior percentagem de mortes. Pelo teste de Wald podemos concluir que existe associação entre ser fumador e a morte ($p < 0,001$). Dado o valor do OR podemos dizer que o risco de morte nos doentes fumadores é 74,4% inferior ao dos doentes não fumadores.

Tabela 4.5: Hipertensão arterial

		GrupoI	**GrupoII**	**Total**	**OR (IC 95%)**	**Teste Wald**
Não	n (%)	5724 (95,2%)	291 (4,8%)	6015 (100%)	Classe de referência	$p = 0,701$
Sim	n (%)	8036 (95,0%)	421 (5,0%)	8457 (100%)	1,03 (0,88;1,20)	
Total	n (%)	13760 (95,1%)	712 (4,9%)	14472 (100%)		

Verifica-se que não existem diferenças significativas entre os doentes com hipertensão e sem hipertensão arterial, no que diz respeito à morte. Pela utilização do teste de Wald podemos concluir que não existe associação entre a hipertensão e a morte ($p = 0.701$).

Tabela 4.6: Diabetes Mellitus

		GrupoI	**GrupoII**	**Total**	**OR(IC 95%)**	**Teste Wald**
Não	n (%)	10571 (95,6%)	482 (4,4%)	11053 (100%)	Classe de referência	$\boldsymbol{p < 0,001}$
Sim	n (%)	3159 (93,2%)	230 (6,8%)	3389 (100%)	1,60 (1,36;1,88)	
Total	n (%)	213730 (95,1%)	712 (4,9%)	14442 (100%)		

Os doentes com diabetes são os que apresentam maior percentagem de mortes nesta amostra. Analisando o valor de OR podemos dizer que o risco de ocorrência de morte nos indivíduos com diabetes é 59,7% superior ao dos doentes sem diabetes. Pela utilização do teste de Wald podemos concluir que existe associação entre a doença diabetes e a morte ($p < 0,001$).

Tabela 4.7: Dislipidémia

		GrupoI	**GrupoII**	**Total**	**OR(IC 95%)**	**Teste Wald**
Não	n (%)	7857 (93,9%)	509 (6,1%)	8366 (100%)	Classe de referência	$\boldsymbol{p < 0,001}$
Sim	n (%)	5778 (96,7%)	195 (3,3%)	5973 (100%)	0,52 (0,44;0,62)	
Total	n (%)	13635 (95,1%)	704 (4,9%)	14339 (100%)		

Pelos mesmos motivos, também aqui se conclui que existe associação entre a dislipidémia e a morte ($p < 0,001$).

Tabela 4.8: Frequência Cardíaca

		GrupoI	**GrupoII**	**Total**	**OR(IC 95%)**	**Teste Wald**
<60 bpm	n (%)	1678 (94,5%)	97 (5,5%)	1775 (100%)	Classe referência	$\boldsymbol{p < 0,001}$
[60,100[bpm	n (%)	10190 (96,3%)	396 (3,7%)	10586 (100%)	0,67 (0,54;0,84)	$p < 0,001$
>=110 bpm	n (%)	1738 (89,8%)	197 (10,2%)	1935 (100%)	5,02 (2,64;9,54)	$p < 0,001$
Total	n (%)	13606 (95,2%)	690 (4,8%)	14296 (100%)		

Através do teste de Wald podemos concluir que existe associação entre a frequência cardíaca e a morte ($p < 0,001$).

Tabela 4.9: Pressão Arterial Sistólica

		GrupoI	GrupoII	Total	OR (IC 95%)	Teste Wald
<90 mmHg	n (%)	5382 (97,2%)	155 (2,8%)	5537 (100%)	Classe referência	$\boldsymbol{p < 0,001}$
[90,140[mmHg	n (%)	302 (68,3%)	140 (31,7%)	442 (100%)	16,10 (12,46;20,80)	$p < 0,001$
[140,180[mmHg	n (%)	6547 (94,7%)	365 (5,3%)	6912 (100%)	1,94 (1,60;2,34)	$p < 0,001$
>=180 mmHg	n (%)	1485 (97,5%)	38 (2,5%)	1523 (100%)	0,89 (0,62;1,27)	$p = 0,519$
Total	n (%)	13716 (95,2%)	698 (4,8%)	14414 (100%)		

Nesta amostra, observamos que o grupo de doentes com pressão sistólica entre os 90 e os 140 mmHg possui maior percentagem de mortes, enquanto a menor percentegem de mortes se verifica no grupo de doentes com pressão sistólica igual ou superior a 180 mmHg. Note-se que a percentagem de mortes não difere significativamente entre as classes <90 e >=180, uma vez que obtivemos um intervalo de confiança que inclui o valor 1, $(0,62; 1,27)$. No entanto, no geral concluímos pelo teste de Wald que existe associação entre a pressão arterial sistólica e a morte ($p < 0,001$).

Tabela 4.10: Pressão Arterial Diastólica

		GrupoI	GrupoII	Total	OR (IC 95%)	Teste Wald
<50 mmHg	n (%)	300 (78,1%)	84 (21,9%)	384 (100%)	Classe referência	$\boldsymbol{p < 0,001}$
[50,110[mmHg	n (%)	12530 (95,5%)	587 (4,5%)	13117 (100%)	0,17 (0,13;0,22)	$p < 0,001$
>=110 mmHg	n (%)	875 (97,3%)	24 (2,7%)	899 (100%)	0,10 (0,06;0,16)	$p < 0,001$
Total	n (%)	13705 (95,2%)	695 (4,8%)	14400 (100%)		

Observa-se, a partir desta amostra, que é o grupo dos doentes com pressão arterial diastólica inferior a 50 mmHg que possui maior percentagem de mortes e a menor percentegem de mortes verifica-se no grupo de doentes com pressão diastólica superior a 110 mmHg. O valor de OR e seu intervalo de confiança indicam que a diferença observada é significativa. Pelo teste de Wald concluímos que existe associação entre a pressão arterial diastólica e a morte ($p < 0,001$).

Tabela 4.11: Classe Killip

		GrupoI	GrupoII	Total	OR (IC 95%)	Teste Wald
1	n (%)	11393 (97,4%)	303 (2,6%)	11696 (100%)	Classe referência	$\boldsymbol{p < 0,001}$
2	n (%)	1382 (88,6%)	177 (11,4%)	1559 (100%)	4,82 (3,97;5,85)	$p < 0,001$
3	n (%)	479 (85,5%)	81 (14,5%)	560 (100%)	6,36 (4,90;8,26)	$p < 0,001$
4	n (%)	166 (57,0%)	125 (43,0%)	291 (100%)	28,31 (21,86;36,67)	$p < 0,001$
Total	n (%)	13420 (95,1%)	686 (4,9%)	14106 (100%)		

A maior percentagem de mortes verifica-se no grupo de doentes com classe Killip 4, enquanto que a menor percentagem se verifica no grupo de doentes com classe Killip 1. Pela utilização do teste de Wald concluímos que existe associação entre a classe Killip e a morte ($p < 0,001$).

Tabela 4.12: Classificação Função VE

		GrupoI	GrupoII	Total	OR (IC 95%)	Teste Wald
Normal	n (%)	8036 (99,1%)	77 (0,9%)	8113 (100%)	Classe referência	$\mathbf{p < 0,001}$
Ligeiramente deprimida	n (%)	1368 (97,4%)	36 (2,6%)	1404 (100%)	2,75 (1,84;4,10)	$p < 0,001$
Moderadamente deprimida	n (%)	907 (95,5%)	43 (4,5%)	950 (100%)	4,95 (3,39;7,23)	$p < 0,001$
Muito deprimida	n (%)	825 (78,3%)	229 (21,7%)	1054 (100%)	28,97 (22,16;37,87)	$p < 0,001$
Total	n (%)	11136 (96,7%)	385 (3,3%)	11521 (100%)		

A partir dos dados observamos que a maior percentagem de mortes se verifica no grupo de doentes com classificação de função VE muito deprimida, enquanto que a menor percentagem se verifica no grupo de doentes com Classificação da função VE normal. O valor do OR (28,97) e o respectivo intervalo de confiança $]22,16; 37,87[$, confirmam que a diferença entre a classe de referência (normal) e a categoria referida é significativa. De salientar que à medida que a função VE vai agravando o risco de ocorrência de morte aumenta (2,75; 4,95; 28,97). Através do teste de Wald podemos dizer que existe associação entre a classificação da função VE e a morte ($p < 0,001$).

4.2. Construção do modelo de regressão logística

De acordo com os resultados da análise univariável efectuada concluímos que as covariáveis *Sexo*, *IMC*, *Idade*, *Fumador*, *Diabetes Mellitus*, *Dislipidémia*, *Frequência Cardíaca*, *Pressão Arterial Sistólica*, *Pressão Arterial Diastólica*, *Classe Killip* e *Classificação Função VE* são estatisticamente significativas, isto é, influenciam a ocorrência de morte em doentes com Síndrome Coronária Aguda sem antecedentes cardiovasculares. Por outro lado, constata-se que a hipertensão arterial (*HTA*) não é estatisticamente significativa. No entanto, dada a importância atribuída pelos cardiologistas a esta covariável no contexto exposto, vamos incluir *HTA* na análise multivariável. Note-se que, as conclusões neste tipo de análise podem ser diferentes da análise univariável, pelo que a covariável indicada pode ser significativa na presença de outras covariáveis.

Para realizar a análise multivariável indicada utilizamos a regressão logística. Nesta análise excluímos todos casos que tenham valor missing para alguma covariável, procedimento standard do *SPSS* denominado eliminação listwise. Aplicando este procedimento foram incluídos na análise 9619 doentes.

O modelo de regressão logística deve ajustar-se aos dados e ter o menor número de covariáveis possível, pois é favorável usar um modelo com menos covariáveis que consiga prever tão bem a probabilidade de ocorrência de morte como um modelo com todas as covariáveis. Este modelo é usualmente adjectivado de parcimonioso.

Dividimos a análise de regressão logística em 7 passos de forma a esquematizar o estudo.

1. Começamos por codificar as covariáveis categóricas com mais de duas categorias recorrendo às variáveis *dummy*. Na prática é usual considerar a primeira ou a última categoria como classe de referência dependendo da categoria à qual está associado menor risco do endpoint ocorrer. Neste caso, utilizámos como classe de referência a primeira categoria. Na figura (4.1) apresentamos as variáveis *dummy*

Categorical Variables Codings

		Parameter coding				
		(1)	(2)	(3)	(4)	(5)
IMC	Peso baixo	0	0	0	0	0
	Peso normal	1	0	0	0	0
	Excesso peso	0	1	0	0	0
	Obesidade grau I	0	0	1	0	0
	Obesidade grau II	0	0	0	1	0
	Obesidade grau III	0	0	0	0	1
ClassificacaoFuncaoVE	Normal	0	0	0		
	Ligeiramente deprimida	1	0	0		
	Moderadamente deprimida	0	1	0		
	Muito deprimida	0	0	1		
ClasseKillip	1	0	0	0		
	2	1	0	0		
	3	0	1	0		
	4	0	0	1		
Idade	<45	0	0	0		
	45-64	1	0	0		
	65-74	0	1	0		
	>=75	0	0	1		
PressaoArterialSistolica	[140,180[mmHg	0	0	0		
	<90 mmHg	1	0	0		
	[90,140[mmHg	0	1	0		
	>=180 mmHg	0	0	1		
FrequenciaCardiaca	<60 bpm	0	0			
	[60,100[bpm	1	0			
	>=110 bpm	0	1			
PressaoArterialDiastolica	<50 mmHg	0	0			
	[50,110[mmHg	1	0			
	>=110 mmHg	0	1			

Figura 4.1: Codificação das variáveis *dummy*.

2. De acordo com o primeiro passo do algoritmo *Stepwise Forward*, descrito no capítulo 3 usando a *estatística razão de verosimilhanças*, consideramos o modelo apenas com a constante, denominado modelo nulo. De acordo com os resultados apresentados na figura (4.2) o modelo depende apenas da constante

($\beta_0 = -3,511$), que segundo o p-valor do teste Wald ($p < 0,001$) é significativamente diferente de zero. Na figura (4.3) encontram-se as covariáveis que não foram incluídas no modelo e os respectivos p-valores relativos ao teste dos scores.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-3,511	,061	3339,225	1	,000	,030

Figura 4.2: Modelo nulo.

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Idade	223,356	3	,000
		Idade(1)	74,325	1	,000
		Idade(2)	5,165	1	,023
		Idade(3)	214,092	1	,000
		IMC	34,328	5	,000
		IMC(1)	29,174	1	,000
		IMC(2)	15,723	1	,000
		IMC(3)	,835	1	,361
		IMC(4)	2,750	1	,097
		IMC(5)	,004	1	,950
		Sexo	52,730	1	,000
		Fumador	44,643	1	,000
		Dislipidemia	20,171	1	,000
		HTA	,249	1	,618
		DiabetesMellitus	13,550	1	,000
		ClasseKillip	670,404	3	,000
		ClasseKillip(1)	113,225	1	,000
		ClasseKillip(2)	64,673	1	,000
		ClasseKillip(3)	451,005	1	,000
		ClassificacaoFuncaoVE	1003,853	3	,000
		ClassificacaoFuncaoVE(1)	2,952	1	,086
		ClassificacaoFuncaoVE(2)	2,575	1	,109
		ClassificacaoFuncaoVE(3)	977,522	1	,000
		FrequenciaCardiaca	109,709	2	,000
		FrequenciaCardiaca(1)	56,990	1	,000
		FrequenciaCardiaca(2)	108,634	1	,000
		PressaoArterialSistolica	316,875	3	,000
		PressaoArterialSistolica(1)	292,691	1	,000
		PressaoArterialSistolica(2)	4,825	1	,028
		PressaoArterialSistolica(3)	8,439	1	,004
		PressaoArterialDiastolica	88,233	2	,000
		PressaoArterialDiastolica(1)	9,418	1	,002
		PressaoArterialDiastolica(2)	4,552	1	,033
	Overall Statistics		1450,842	26	,000

Figura 4.3: Variáveis não incluidas no modelo.

3. Após sete iterações o algoritmo atinge o critério de paragem. Na figura (4.5) observamos que das covariáveis que não foram incluídas na sétima iteração, o menor p-valor (associado a *Fumador*) é maior que $p_e = 0,1$. Assim, os preditores de morte encontrados são apresentados na primeira coluna de (4.4). Os quadros completos são apresentados em apêndice (figuras A.1 a A.6).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp (B)	95% C.I.for EXP(B) Lower	Upper
Step 7	Idade			56,941	3	,000			
	Idade(1)	,941	,547	2,957	1	,086	2,563	,877	7,495
	Idade(2)	1,301	,545	5,695	1	,017	3,674	1,262	10,695
	Idade(3)	2,136	,536	15,880	1	,000	8,462	2,960	24,192
	IMC			14,102	5	,015			
	IMC(1)	,762	,633	1,447	1	,229	2,142	,619	7,412
	IMC(2)	,271	,634	,183	1	,669	1,311	,378	4,548
	IMC(3)	,689	,651	1,121	1	,290	1,992	,556	7,133
	IMC(4)	-,254	,839	,092	1	,762	,776	,150	4,012
	IMC(5)	,325	1,028	,100	1	,752	1,384	,184	10,388
	Sexo	,494	,146	11,517	1	,001	1,639	1,232	2,179
	ClasseKillip			23,844	3	,000			
	ClasseKillip(1)	,441	,176	6,290	1	,012	1,554	1,101	2,194
	ClasseKillip(2)	,453	,234	3,749	1	,053	1,572	,994	2,486
	ClasseKillip(3)	1,419	,306	21,552	1	,000	4,132	2,270	7,520
	ClassificacaoFuncaoVE			254,719	3	,000			
	ClassificacaoFuncaoVE(1)	,850	,249	11,682	1	,001	2,340	1,437	3,811
	ClassificacaoFuncaoVE(2)	1,226	,242	25,687	1	,000	3,408	2,121	5,476
	ClassificacaoFuncaoVE(3)	2,798	,183	232,577	1	,000	16,41	11,452	23,507
	FrequenciaCardiaca			11,457	2	,003			
	FrequenciaCardiaca(1)	-,515	,222	5,398	1	,020	,597	,387	,923
	FrequenciaCardiaca(2)	-,051	,246	,043	1	,837	,951	,587	1,539
	PressaoArterialSistolica			20,877	3	,000			
	PressaoArterialSistolica(1)	1,222	,310	15,526	1	,000	3,393	1,848	6,231
	PressaoArterialSistolica(2)	,481	,165	8,496	1	,004	1,618	1,171	2,237
	PressaoArterialSistolica(3)	-,144	,311	,216	1	,642	,866	,470	1,592
	Constant	-7,446	,884	70,908	1	,000	,001		

Figura 4.4: Covariáveis incluídas no modelo final.

Variables not in the Equation

			Score	df	Sig.
Step 7	Variables	Fumador	1,101	1	,294
		Dislipidemia	1,043	1	,307
		HTA	,327	1	,568
		DiabetesMellitus	,296	1	,586
		PressaoArterialDiastolica	,608	2	,738
		PressaoArterialDiastolica(1)	,492	1	,483
		PressaoArterialDiastolica(2)	,022	1	,881
	Overall Statistics		3,654	6	,723

Figura 4.5: Covariáveis não incluídas no modelo final.

4. Após a selecção das covariáveis, vamos testar o ajustamento do modelo constituído por estas. Para este fim utilizamos o teste de *Hosmer and Lemeshow*.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8,579	8	,379

Figura 4.6: Teste de *Hosmer and Lemeshow*.

De acordo com o p-valor obtido ($p = 0,379$) aos níveis de significância usuais não rejeitamos a hipótese nula, isto é, não rejeitamos a hipótese de adequação do modelo aos dados.

5. Para a validação do modelo analisamos o seu poder discriminatório, a sensibilidade, a especificidade e a taxa de acertos. Para tal utilizamos o gráfico da curva ROC, nomeadamente o valor da área sob a curva, e a tabela de classificação.

   (a) No caso em estudo, a área sob a curva é 0,891 o que indica que o modelo encontrado é bom. Como o valor da área se situa entre 0,8 e 0,9 dizemos que o modelo tem um poder discriminatório excelente.

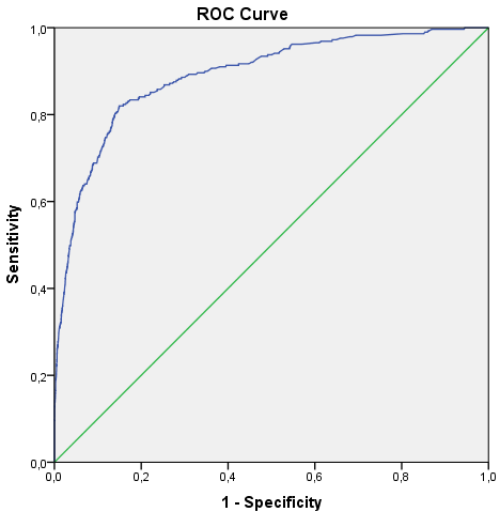


Figura 4.7: Curva ROC.

Test Result Variable(s): Predicted probability

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,891	,010	,000	,870	,911

Figura 4.8: Área sob a curva ROC e o respectivo IC.

(b) Na figura (4.9) apresentamos a tabela de classificação.

	Observed		MortelH		Percentage Correct
			,00	1,00	
Step 1	MortelH	,00	7952	1388	85,1
		1,00	50	229	82,1
	Overall Percentage				85,1

Figura 4.9: Tabela de classificação.

Começamos por determinar o cut-off de forma a transformar os valores preditos numa variável binária. Os valores preditos inferiores ao valor do cut-off tomam o valor 0 e os valores superiores tomam o valor 1.

O output do *SPSS* faculta-nos os pontos coordenados associados ao gráfico da curva ROC que nos permitem descobrir o valor do cut-off, que neste caso é 0,027. De acordo com o cut-off considerado a sensibilidade do modelo é

$$\frac{229}{50+229} \times 100\% = 82,1\%,$$

a especificidade do modelo é dado por

$$\frac{7952}{7952+1388} \times 100\% = 85,1\%,$$

e a taxa de acertos é

$$\frac{7952+229}{7952+1388+50+229} \times 100\% = 85,1\%.$$

A taxa de acertos é boa, pelo que podemos concluir que o modelo consegue fazer boas previsões.

6. Utilizando as estimativas dos coeficientes apresentados na segunda coluna da figura (4.4), construímos o modelo de regressão logística que se apresenta a seguir

$$\pi(\boldsymbol{x}) = \frac{\exp(z)}{1+\exp(z)},$$

onde

$$
\begin{aligned}
z = &-7,446 + 0,4941 \times Sexo + 0,941 \times Idade(1) + 1,301 \times Idade(2) + \\
&2,136 \times Idade(3) + 0,762 \times IMC(1) + 0,271 \times IMC(2) + \\
&0,689 \times IMC(3) + (-0,254 \times IMC(4)) + 0,325 \times IMC(5) + \\
&0,441 \times ClasseKillip(1) + 0,453 \times ClasseKillip(2) + 1,419 \times ClasseKillip(3) + \\
&0,850 \times ClassificacaoVE(1) + 0,226 \times ClassificacaoVE(2) + 2,798 \times ClassificacaoVE(3) + \\
&1,222 \times PressaoArterialSistolica(1) + 0,481 \times PressaoArterialSistolica(2) + \\
&(-0,144 \times PressaoArterialSistolica(3)) + (-0,515 \times FrequenciaCardiaca(1)) + \\
&(-0,051 \times FrequenciaCardiaca(2)). \qquad (4.1)
\end{aligned}
$$

4.3. Validação externa do modelo

Para avaliar o desempenho do modelo fazemos a validação externa do mesmo. Incluímos na análise, após eliminar os valores missings, apenas 6413 doentes. Nestes 6214 tiveram alta e os restantes foram declarados como óbito. Começamos por criar uma variável no *SPSS* constituída pelas probabilidades estimadas para cada doente, pre_sub_40_2, utilizando o modelo (4.1). A curva ROC associada a pre_sub_40_2 é apresentada abaixo. Vamos utilizar as coordenadas do seu gráfico para determinar o valor do cut-off para este conjunto de dados.

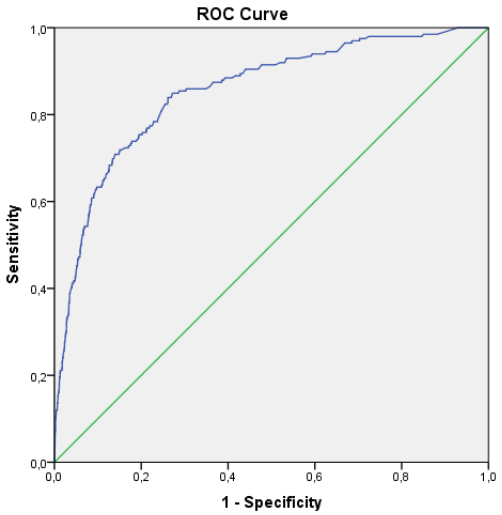


Figura 4.10: Curva ROC.

Test Result Variable(s): pre_sub_40_2

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,854	,014	,000	,825	,882

Figura 4.11: Área sob a curva ROC e o respectivo IC.

Podemos concluir, a partir do valor da área sob a curva ROC e do seu intervalo de confiança, que o modelo tem um poder discriminatório excelente. A partir do cut-off encontrado (0,0098) criamos uma variável, à qual damos o nome de Morte_sub_40, que contém os valores estimados de cada indivíduo para o endpoint, a morte. Esta variável é binária, isto é, toma o valor 1 se a probabilidade estimada é superior ou igual a 0,0098 e o valor 0 caso contrário. Na figura seguinte apresentamos a tabela de classificação

Tabela 4.13: Tabela de classificação

		Morte_sub_40		
		0	1	Total
MorteIH	0	4589	1625	6214
	1	32	167	199
	Total	4621	1792	6413

A sensibilidade, especificidade e taxa de acertos do modelo são, respectivamente, $83,9\%$, $73,8\%$ e $74,2\%$. Dada a taxa de acertos podemos concluir que o modelo tem um bom desempenho na previsão da ocorrência de morte.

4.4. Interpretação do modelo em termos de *Odds ratio*

A partir das estimativas dos coefecientes do modelo podemos determinar estimativas para o OR e para os respectivos intervalos de confiança. Estas estimativas são obtidas aplicando a função exponencial ao valor da estimativa do coeficiente e aos extremos do intervalo de confiança associado a cada covariável. Apresentamos de seguida uma tabela com o valor estimado do OR e o respectivo intervalo de confiança para cada preditor.

Tabela 4.14: Valor do *Odds Ratio* e respectivo IC para cada variável

	OR	95% IC	
		Lower	Upper
Idade	Classe de referência		
Idade(1)	2,563	0,877	7,495
Idade(2)	3,674	1,262	10,695
Idade(3)	8,462	2,960	24,192
IMC	Classe de referência		
IMC(1)	2,142	0,619	7,412
IMC(2)	1,311	0,378	4,548
IMC(3)	1,992	0,556	7,133
IMC(4)	0,776	0,150	4,012
IMC(5)	1,384	0,184	10,388
Sexo	1,639	1,232	2,179
ClasseKillip	Classe de referência		
ClasseKillip(1)	1,554	1,101	2,194
ClasseKillip(2)	1,572	0,994	2,486
ClasseKillip(3)	4,132	2,270	7,520
ClassificacaoFuncaoVE	Classe de referência		
ClassificacaoFuncaoVE(1)	2,340	1,437	3,811
ClassificacaoFuncaoVE(2)	3,408	2,121	5,476
ClassificacaoFuncaoVE(3)	16,408	11,452	23,507
FrequenciaCardiaca	Classe de referência		
FrequenciaCardiaca(1)	0,597	0,387	0,923
FrequenciaCardiaca(2)	0,951	0,587	1,539
PressaoArterialSistolica	Classe de referência		
PressaoArterialSistolica(1)	3,393	1,848	6,231
PressaoArterialSistolica(2)	1,618	1,171	2,237
PressaoArterialSistolica(3)	0,866	0,470	1,592

Sempre que o valor do *Odds Ratio* é superior a 1, existe um aumento do risco de ocorrência de morte. Observando as estimativas do valor do *Odds Ratio* obtidas para as covariáveis consideradas no modelo final, verificamos que a maior parte das covariáveis provocam um aumento do risco de ocorrência de morte. De seguida vamos analisar estes valores de uma forma mais promenorizada. Os resultados apresentados na tabela (4.14) para cada covariável, referem-se a indivíduos que diferem na covariável em análise e têm valores iguais nas restantes covariáveis (por exemplo, ao analisar o valor de OR para *Diabetes Mellitus* escolhemos um indivíduo aleatoriamente que seja diabético e de seguida escolhemos outro indivíduo com as mesmas características que o primeiro, mas não diabético).

Observando os resultados associados à covariável *Idade*, constatamos que o risco de morte tem tendência a aumentar à medida que se envelhece. De facto, tomando como classe de referência a classe <45 anos (categoria 0), verificamos que o valor

do OR aumenta à medida que a idade aumenta (2,56 ; 3,67 ; 8,46). No entanto, o intervalo de confiança associado à categoria 1 indica que as diferenças entre a classe de referência e a classe dos doentes com idades entre os 45 e os 64 anos (categoria 1) não são significativas $(]0,877; 7,495[)$. Verificamos também que o risco de ocorrência de morte nos doentes com idades entre os 65 e os 74 anos é 3.62 vezes superior ao dos doentes com idade inferior a 45 anos. Quanto aos doentes com mais de 75 anos o risco de ocorrência de morte é 8,46 vezes superior ao da classe de referência.

Quanto à variável *Sexo*, o valor de OR leva-nos a concluir que os doentes do sexo feminino têm um risco de morrer 64% superior ao dos doentes do sexo masculino. Podemos ainda dizer com 95% de confiança que este aumento é no mínimo 23,2% e no máximo 179,0%.

Considerando como classe de referência a primeira categoria para as covariáveis *Classe Killip* e *Classificação Função VE*, observamos que o valor de OR associado a ambas é superior a 1 e aumenta à medida que o estado de saúde se agrava. Desta forma concluímos que o risco de ocorrência de morte tem uma tendência crescente para as covariáveis *Classe Killip* e *Classificação Função VE*.

No que diz respeito à frequência cardíaca, podemos observar que o risco de ocorrência de morte nos doentes com frequência cardíaca pertencente ao intervalo $[60, 100[$ bpm é 59,7% inferior ao dos doentes com frequência cardíaca abaixo de 60 bpm (classe de referência). Quanto aos doentes pertencentes à quarta categoria (>=110 bpm) o risco de morte não difere significativamente do risco associado aos doentes pertencentes à classe de referência.

Finalmente vamos averiguar o risco de ocorrência de morte no que concerne à variável *Pressão Arterial Sistólica*. Considerando a categoria 0 como classe de referência observamos que o risco de ocorrência de morte tem tendência a decrescer com o aumento da pressão sistólica. Neste estudo os doentes com pressão sistólica pertencentes à classe $[90, 140[$ e $[140, 180[$ mmHg têm um risco de ocorrência de morte 239% e 61,8% superior ao risco dos doentes pertencentes à classe de referência, respectivamente.

4.5. Conclusões

Tal como referimos, a análise realizada tem como objectivo encontrar os principais factores que influenciam a ocorrência de morte (preditores) em pessoas com Síndrome Coronária Aguda, mas sem antecedentes de doenças cardiovasculares.

Com o intuito de obter o modelo mais robusto para a previsão de morte decidimos fazer uma validação externa do modelo. Para tal, dividimos a amostra em duas partes em que a primeira contém 60% dos doentes e a segunda os restantes. Utilizámos a primeira parte para construir o modelo regressão logística multivariável mais parcimonioso e consequentemente encontrar os preditores de morte.

Os preditores encontrados foram *Sexo*, *Idade*, *Índice de massa corporal*, *Frequência cardíaca*, *Pressão arterial sistólica*, *Classe Killip* e *Classificação função VE*. Utilizámos os testes usuais para testar o poder discriminatório e ajustamento do modelo aos dados. Os resultados obtidos pelo teste de *Hosmer and Lemeshow* permitem concluir que o modelo se ajusta aos dados e a área sob a curva ROC indica um excelente poder discriminatório. A partir da tabela de classificação (4.9) verificamos que 85,1% dos doentes foram bem classificados pelo modelo.

A validação externa do modelo realizada nos restantes 40% dos doentes, veio confirmar que o modelo tem um poder discriminatório excelente, dado que a área sob a curva ROC é 0,854, o que significa que escolhendo aleatoriamente um indivíduo que não teve alta e outro que teve alta, temos 85,4% de hipótese da probabilidade de morte estimada pelo modelo associada ao indivíduo que não teve alta ser superior à do indivíduo que teve alta. A taxa de acertos determinada a partir de (4.13) indica que o modelo tem um bom desempenho na previsão da ocorrência de morte. Estes factos confirmam a robustez do modelo de regressão logística apresentado em (4.1).

4.6. Trabalhos futuros

Um trabalho futuro que consideramos importante por se tratar de um grande problema de saúde do século XXI, é determinar os preditores de re-enfarte em doentes sem antecedentes cardiovasculares. Esta questão foi analisada no âmbito desta dissertação, mas devido ao reduzido número de re-enfartes no internamento não foi possível chegar a um modelo robusto. Uma forma de contornar este problema seria, possivelmente, considerar uma amostra de maior dimensão.

Apêndice A

A.1. Glossário de alguns termos usados em Cardiologia

Pressão arterial sistólica / Pressão arterial diastólica

A tensão arterial é a pressão do sangue dentro do coração e das artérias. É descrita por dois valores, tensão arterial sistólica e tensão arterial diastólica, vulgarmente conhecidas como tensão "máxima" e "mínima" respetivamente. A pressão arterial sistólica mede a pressão provocada pela contracção do coração, enquanto a pressão arterial diastólica quantifica a pressão nas artérias quando o coração relaxa entre duas contracções.

Classe killip

A classe Killip-Kimball é uma escala de classificação do grau de insuficiência cardíaca em indivíduos que sofreram enfarte agudo do miocardio. Os doentes podem ser classificados segundo quatro classes: Classe 1, sem evidência clínica de insuficiência cardíaca; Classe 2, insuficiência cardíaca ligeira; Classe 3, insuficiência cardíaca grave ou presença de edema pulmonar; e Classe 4, presença de choque cardiogénico caracterizado por hipotensão (pressão arterial sistólica < 90 mmHg) e evidência de vasoconstricção periférica. Quanto maior a classe killip pior é o prognóstico do doente.

Diabetes *Mellitus*

A maioria dos alimentos que comemos é convertida pelo nosso organismo em glicose. A glicose é um tipo de açucar que depois de absorvida pelas células do organismo serve de fonte de energia. Na presença de diabetes o nosso organismo ou não produz insulina suficiente (diabetes tipo 1) ou não utiliza a sua própria insulina tão bem como devia (diabetes tipo 2), o que leva à presença de glicémia (níveis de açúcar no sangue) elevada. A diabetes é diagnosticada quando a glicémia em jejum é superior ou igual a 126 mg/dL.

Dislipidémia

Manifesta-se quando os valores do colesterol no sangue são superiores aos níveis máximos recomendados em função do risco cardiovascular individual.

Frequência cardíaca

Frequência cardíaca é determinada pelo número de batimentos cardíacos por unidade de tempo, geralmente expressa em batimentos por minuto (bpm). A frequência cardíaca pode variar de acordo com a necessidade de oxigénio do organismo. Durante o exercício físico a frequência cardíaca eleva-se devido a uma elevada necessidade de oxigénio, já durante o sono o seu valor é mais baixo.

Hipertensão (HTA)

A tensão arterial considera-se elevada quando pressão arterial sistólica ≥ 140 mmHg ou pressão arterial diastólica ≥ 90 mmHg. A HTA é um fator de risco para as doenças cardiovasculares, uma vez que as artérias sujeitas a uma tensão excessiva tornam-se mais espessas e rígidas, o que favorece a progressão da aterosclerose.

Índice de Massa Corporal (IMC)

Indicador utilizado para avaliar a relação entre o peso e a estatura. Calcula-se dividindo o peso (em kg) pelo quadrado da estatura (em m^2).

Tabagismo

É um importante fator de risco para doenças pulmonares e cardiovasculares graves. O fumo do tabaco contém mais de 4000 substâncias químicas, várias das quais com efeitos tóxicos, irritantes ou cancerígenos. A nicotina aumenta a tensão arterial, a frequência cardíaca, diminui o débito cardíaco e o fluxo de sangue nas artérias coronárias. O tabaco torna os vasos rígidos e promove a formação de coágulos, favorece o depósito de colesterol que resulta em aterosclerose e trombose aguda.

Classificação função VE

A avaliação da função do ventrículo esquerdo é uma das principais indicações para a realização de um ecocardiograma e um dos parâmetros mais importantes dessa avaliação ecocardiográfica, fornecendo informações indispensáveis para o diagnóstico, orientação terapêutica e prognóstico de quase todas as patologias cardíacas. Através do ecocardiograma a função VE é classificada em quatro categorias: classificação função VE normal, ligeiramente deprimida, moderadamente deprimida e muito deprimida.

A.2. Tabelas SPSS

A.2.1. Covariáveis incluídas no modelo

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp (B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	ClassificacaoFuncaoVE			519,376	3	,000			
	ClassificacaoFuncaoVE(1)	,977	,243	16,131	1	,000	2,657	1,649	4,280
	ClassificacaoFuncaoVE(2)	1,586	,230	47,494	1	,000	4,886	3,112	7,672
	ClassificacaoFuncaoVE(3)	3,442	,160	460,081	1	,000	31,25	22,817	42,799
	Constant	-4,812	,135	1263,212	1	,000	,008		
Step 2[b]	ClasseKillip			115,202	3	,000			
	ClasseKillip(1)	,846	,168	25,483	1	,000	2,331	1,678	3,237
	ClasseKillip(2)	,853	,219	15,125	1	,000	2,347	1,527	3,607
	ClasseKillip(3)	2,398	,228	110,480	1	,000	11,00	7,033	17,197
	ClassificacaoFuncaoVE			302,661	3	,000			
	ClassificacaoFuncaoVE(1)	,876	,245	12,769	1	,000	2,402	1,485	3,885
	ClassificacaoFuncaoVE(2)	1,298	,237	29,979	1	,000	3,661	2,301	5,827
	ClassificacaoFuncaoVE(3)	2,921	,176	274,319	1	,000	18,55	13,131	26,211
	Constant	-4,969	,139	1276,501	1	,000	,007		
Step 3[c]	Idade			74,385	3	,000			
	Idade(1)	,822	,541	2,309	1	,129	2,276	,788	6,576
	Idade(2)	1,214	,538	5,083	1	,024	3,366	1,172	9,666
	Idade(3)	2,136	,527	16,422	1	,000	8,469	3,014	23,801
	ClasseKillip			96,666	3	,000			
	ClasseKillip(1)	,578	,170	11,559	1	,001	1,782	1,277	2,486
	ClasseKillip(2)	,533	,224	5,676	1	,017	1,703	1,099	2,640
	ClasseKillip(3)	2,276	,232	96,275	1	,000	9,740	6,181	15,346
	ClassificacaoFuncaoVE			284,576	3	,000			
	ClassificacaoFuncaoVE(1)	,820	,246	11,081	1	,001	2,270	1,401	3,680
	ClassificacaoFuncaoVE(2)	1,214	,239	25,745	1	,000	3,368	2,107	5,384
	ClassificacaoFuncaoVE(3)	2,812	,175	256,928	1	,000	16,64	11,799	23,469
	Constant	-6,291	,531	140,477	1	,000	,002		

a. Variable(s) entered on step 1: ClassificacaoFuncaoVE.

b. Variable(s) entered on step 2: ClasseKillip.

c. Variable(s) entered on step 3: Idade.

Figura A.1: Covariáveis incluídas passo a passo por ordem decrescente de significância(continua).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp (B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 4[d]	Idade			75,481	3	,000			
	Idade(1)	,885	,543	2,656	1	,103	2,423	,836	7,023
	Idade(2)	1,321	,540	5,982	1	,014	3,749	1,300	10,808
	Idade(3)	2,224	,529	17,659	1	,000	9,245	3,276	26,087
	ClasseKillip			31,957	3	,000			
	ClasseKillip(1)	,528	,171	9,489	1	,002	1,696	1,212	2,374
	ClasseKillip(2)	,576	,227	6,464	1	,011	1,779	1,141	2,774
	ClasseKillip(3)	1,586	,305	27,110	1	,000	4,886	2,689	8,878
	ClassificacaoFuncaoVE			259,367	3	,000			
	ClassificacaoFuncaoVE(1)	,824	,247	11,158	1	,001	2,279	1,406	3,696
	ClassificacaoFuncaoVE(2)	1,189	,240	24,610	1	,000	3,284	2,053	5,254
	ClassificacaoFuncaoVE(3)	2,733	,178	236,294	1	,000	15,38	10,855	21,793
	PressaoArterialSistolica			22,469	3	,000			
	PressaoArterialSistolica(1)	1,303	,309	17,810	1	,000	3,679	2,009	6,736
	PressaoArterialSistolica(2)	,495	,164	9,136	1	,003	1,640	1,190	2,260
	PressaoArterialSistolica(3)	-,059	,305	,037	1	,848	,943	,518	1,716
	Constant	-6,650	,547	147,990	1	,000	,001		
Step 5[e]	Idade			56,610	3	,000			
	Idade(1)	,874	,542	2,605	1	,107	2,397	,829	6,930
	Idade(2)	1,228	,540	5,171	1	,023	3,413	1,185	9,834
	Idade(3)	2,058	,531	15,041	1	,000	7,829	2,767	22,148
	Sexo	,449	,143	9,782	1	,002	1,566	1,182	2,074
	ClasseKillip			29,765	3	,000			
	ClasseKillip(1)	,489	,172	8,024	1	,005	1,630	1,162	2,286
	ClasseKillip(2)	,560	,227	6,068	1	,014	1,750	1,121	2,732
	ClasseKillip(3)	1,543	,304	25,710	1	,000	4,679	2,577	8,495
	ClassificacaoFuncaoVE			264,192	3	,000			
	ClassificacaoFuncaoVE(1)	,836	,247	11,453	1	,001	2,306	1,421	3,741
	ClassificacaoFuncaoVE(2)	1,213	,240	25,532	1	,000	3,363	2,101	5,382
	ClassificacaoFuncaoVE(3)	2,769	,178	241,163	1	,000	15,95	11,243	22,619
	PressaoArterialSistolica			23,411	3	,000			
	PressaoArterialSistolica(1)	1,311	,309	18,012	1	,000	3,710	2,025	6,796
	PressaoArterialSistolica(2)	,502	,164	9,375	1	,002	1,652	1,198	2,278
	PressaoArterialSistolica(3)	-,113	,308	,134	1	,714	,893	,489	1,634
	Constant	-7,174	,573	156,842	1	,000	,001		

d. Variable(s) entered on step 4: PressaoArterialSistolica.

e. Variable(s) entered on step 5: Sexo.

Figura A.2: Variáveis incluídas passo a passo por ordem decrescente de significância (continua).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp (B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 6[f]	Idade			58,050	3	,000			
	Idade(1)	,917	,546	2,825	1	,093	2,503	,859	7,293
	Idade(2)	1,259	,544	5,359	1	,021	3,523	1,213	10,230
	Idade(3)	2,108	,535	15,543	1	,000	8,230	2,886	23,470
	Sexo	,473	,144	10,754	1	,001	1,605	1,210	2,130
	ClasseKillip			23,768	3	,000			
	ClasseKillip(1)	,425	,175	5,865	1	,015	1,529	1,084	2,156
	ClasseKillip(2)	,459	,232	3,922	1	,048	1,582	1,005	2,490
	ClasseKillip(3)	1,411	,304	21,615	1	,000	4,100	2,262	7,433
	ClassificacaoFuncaoVE			254,913	3	,000			
	ClassificacaoFuncaoVE(1)	,838	,248	11,403	1	,001	2,312	1,421	3,761
	ClassificacaoFuncaoVE(2)	1,237	,242	26,147	1	,000	3,444	2,144	5,532
	ClassificacaoFuncaoVE(3)	2,793	,183	232,339	1	,000	16,33	11,402	23,383
	FrequenciaCardiaca			11,702	2	,003			
	FrequenciaCardiaca(1)	-,530	,221	5,751	1	,016	,589	,382	,908
	FrequenciaCardiaca(2)	-,069	,245	,078	1	,779	,934	,577	1,510
	PressaoArterialSistolica			22,301	3	,000			
	PressaoArterialSistolica(1)	1,253	,307	16,638	1	,000	3,502	1,918	6,396
	PressaoArterialSistolica(2)	,499	,165	9,207	1	,002	1,648	1,193	2,275
	PressaoArterialSistolica(3)	-,141	,310	,206	1	,650	,869	,473	1,595
	Constant	-6,868	,596	132,806	1	,000	,001		
Step 7[g]	Idade			56,941	3	,000			
	Idade(1)	,941	,547	2,957	1	,086	2,563	,877	7,495
	Idade(2)	1,301	,545	5,695	1	,017	3,674	1,262	10,695
	Idade(3)	2,136	,536	15,880	1	,000	8,462	2,960	24,192
	IMC			14,102	5	,015			
	IMC(1)	,762	,633	1,447	1	,229	2,142	,619	7,412
	IMC(2)	,271	,634	,183	1	,669	1,311	,378	4,548
	IMC(3)	,689	,651	1,121	1	,290	1,992	,556	7,133
	IMC(4)	-,254	,839	,092	1	,762	,776	,150	4,012
	IMC(5)	,325	1,028	,100	1	,752	1,384	,184	10,388
	Sexo	,494	,146	11,517	1	,001	1,639	1,232	2,179
	ClasseKillip			23,844	3	,000			
	ClasseKillip(1)	,441	,176	6,290	1	,012	1,554	1,101	2,194
	ClasseKillip(2)	,453	,234	3,749	1	,053	1,572	,994	2,486
	ClasseKillip(3)	1,419	,306	21,552	1	,000	4,132	2,270	7,520
	ClassificacaoFuncaoVE			254,719	3	,000			
	ClassificacaoFuncaoVE(1)	,850	,249	11,682	1	,001	2,340	1,437	3,811
	ClassificacaoFuncaoVE(2)	1,226	,242	25,687	1	,000	3,408	2,121	5,476
	ClassificacaoFuncaoVE(3)	2,798	,183	232,577	1	,000	16,41	11,452	23,507
	FrequenciaCardiaca			11,457	2	,003			
	FrequenciaCardiaca(1)	-,515	,222	5,398	1	,020	,597	,387	,923
	FrequenciaCardiaca(2)	-,051	,246	,043	1	,837	,951	,587	1,539
	PressaoArterialSistolica			20,877	3	,000			
	PressaoArterialSistolica(1)	1,222	,310	15,526	1	,000	3,393	1,848	6,231
	PressaoArterialSistolica(2)	,481	,165	8,496	1	,004	1,618	1,171	2,237
	PressaoArterialSistolica(3)	-,144	,311	,216	1	,642	,866	,470	1,592
	Constant	-7,446	,884	70,908	1	,000	,001		

f. Variable(s) entered on step 6: FrequenciaCardiaca.

g. Variable(s) entered on step 7: IMC.

Figura A.3: Variáveis incluídas passo a passo por ordem decrescente de significância (continuação).

A.2.2. Covariáveis não incluídas no modelo

Variables not in the Equation

			Score	df	Sig.
Step 1	Variables	Idade	102,568	3	,000
		Idade(1)	34,327	1	,000
		Idade(2)	7,158	1	,007
		Idade(3)	97,130	1	,000
		IMC	18,831	5	,002
		IMC(1)	15,118	1	,000
		IMC(2)	13,437	1	,000
		IMC(3)	,080	1	,777
		IMC(4)	1,412	1	,235
		IMC(5)	,195	1	,659
		Sexo	41,689	1	,000
		Fumador	28,684	1	,000
		Dislipidemia	6,077	1	,014
		HTA	2,146	1	,143
		DiabetesMellitus	2,714	1	,099
		ClasseKillip	132,119	3	,000
		ClasseKillip(1)	5,684	1	,017
		ClasseKillip(2)	2,060	1	,151
		ClasseKillip(3)	102,240	1	,000
		FrequenciaCardiaca	24,778	2	,000
		FrequenciaCardiaca(1)	24,623	1	,000
		FrequenciaCardiaca(2)	14,450	1	,000
		PressaoArterialSistolica	102,997	3	,000
		PressaoArterialSistolica(1)	94,187	1	,000
		PressaoArterialSistolica(2)	,011	1	,918
		PressaoArterialSistolica(3)	1,394	1	,238
		PressaoArterialDiastolica	35,159	2	,000
		PressaoArterialDiastolica(1)	7,046	1	,008
		PressaoArterialDiastolica(2)	2,479	1	,115
	Overall Statistics		272,508	23	,000

Figura A.4: Variáveis que não foram incluídas no modelo(continua).

Variables not in the Equation

			Score	df	Sig.
Step 2	Variables	Idade	81,588	3	,000
		Idade(1)	26,480	1	,000
		Idade(2)	6,121	1	,013
		Idade(3)	75,919	1	,000
		IMC	17,080	5	,004
		IMC(1)	14,100	1	,000
		IMC(2)	10,464	1	,001
		IMC(3)	,036	1	,849
		IMC(4)	2,255	1	,133
		IMC(5)	,010	1	,920
		Sexo	31,063	1	,000
		Fumador	22,044	1	,000
		Dislipidemia	4,027	1	,045
		HTA	1,736	1	,188
		DiabetesMellitus	,604	1	,437
		FrequenciaCardiaca	8,755	2	,013
		FrequenciaCardiaca(1)	8,139	1	,004
		FrequenciaCardiaca(2)	3,332	1	,068
		PressaoArterialSistolica	21,287	3	,000
		PressaoArterialSistolica(1)	13,189	1	,000
		PressaoArterialSistolica(2)	2,460	1	,117
		PressaoArterialSistolica(3)	1,204	1	,272
		PressaoArterialDiastolica	4,288	2	,117
		PressaoArterialDiastolica(1)	,031	1	,861
		PressaoArterialDiastolica(2)	2,461	1	,117
	Overall Statistics		143,778	20	,000
Step 3	Variables	IMC	15,582	5	,008
		IMC(1)	10,417	1	,001
		IMC(2)	11,750	1	,001
		IMC(3)	1,415	1	,234
		IMC(4)	1,279	1	,258
		IMC(5)	,004	1	,949
		Sexo	8,793	1	,003
		Fumador	1,325	1	,250
		Dislipidemia	1,097	1	,295
		HTA	,001	1	,982
		DiabetesMellitus	,223	1	,637
		FrequenciaCardiaca	11,838	2	,003
		FrequenciaCardiaca(1)	11,284	1	,001
		FrequenciaCardiaca(2)	5,050	1	,025
		PressaoArterialSistolica	22,871	3	,000
		PressaoArterialSistolica(1)	11,754	1	,001
		PressaoArterialSistolica(2)	4,603	1	,032
		PressaoArterialSistolica(3)	2,324	1	,127
		PressaoArterialDiastolica	2,526	2	,283
		PressaoArterialDiastolica(1)	,116	1	,734
		PressaoArterialDiastolica(2)	1,845	1	,174
	Overall Statistics		62,786	17	,000

Figura A.5: Variáveis que não foram incluídas no modelo(continua).

Figura A.6: Variáveis que não foram incluídas no modelo(continuação).

Variables not in the Equation

			Score	df	Sig.
Step 4	Variables	IMC	14,046	5	,015
		IMC(1)	9,172	1	,002
		IMC(2)	10,695	1	,001
		IMC(3)	1,387	1	,239
		IMC(4)	1,025	1	,311
		IMC(5)	,000	1	,989
		Sexo	9,864	1	,002
		Fumador	1,686	1	,194
		Dislipidemia	1,140	1	,286
		HTA	,303	1	,582
		DiabetesMellitus	,410	1	,522
		FrequenciaCardiaca	10,795	2	,005
		FrequenciaCardiaca(1)	10,785	1	,001
		FrequenciaCardiaca(2)	6,396	1	,011
		PressaoArterialDiastolica	,447	2	,800
		PressaoArterialDiastolica(1)	,435	1	,510
		PressaoArterialDiastolica(2)	,102	1	,750
	Overall Statistics		39,793	14	,000
Step 5	Variables	IMC	14,667	5	,012
		IMC(1)	9,894	1	,002
		IMC(2)	9,782	1	,002
		IMC(3)	1,098	1	,295
		IMC(4)	1,579	1	,209
		IMC(5)	,023	1	,879
		Fumador	,596	1	,440
		Dislipidemia	1,580	1	,209
		HTA	,086	1	,770
		DiabetesMellitus	,094	1	,759
		FrequenciaCardiaca	11,842	2	,003
		FrequenciaCardiaca(1)	11,720	1	,001
		FrequenciaCardiaca(2)	6,280	1	,012
		PressaoArterialDiastolica	,446	2	,800
		PressaoArterialDiastolica(1)	,407	1	,524
		PressaoArterialDiastolica(2)	,053	1	,819
	Overall Statistics		29,852	13	,005
Step 6	Variables	IMC	14,353	5	,014
		IMC(1)	9,246	1	,002
		IMC(2)	8,988	1	,003
		IMC(3)	1,200	1	,273
		IMC(4)	1,960	1	,161
		IMC(5)	,056	1	,812
		Fumador	,648	1	,421
		Dislipidemia	1,472	1	,225
		HTA	,156	1	,692
		DiabetesMellitus	,114	1	,736
		PressaoArterialDiastolica	,580	2	,748
		PressaoArterialDiastolica(1)	,524	1	,469
		PressaoArterialDiastolica(2)	,063	1	,801
	Overall Statistics		18,016	11	,081
Step 7	Variables	Fumador	1,101	1	,294
		Dislipidemia	1,043	1	,307
		HTA	,327	1	,568
		DiabetesMellitus	,296	1	,586
		PressaoArterialDiastolica	,608	2	,738
		PressaoArterialDiastolica(1)	,492	1	,483
		PressaoArterialDiastolica(2)	,022	1	,881
	Overall Statistics		3,654	6	,723

Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	ClassificacaoFuncaoVE	-1262,653	547,838	3	,000
Step 2	ClasseKillip	-988,734	104,935	3	,000
	ClassificacaoFuncaoVE	-1093,514	314,496	3	,000
Step 3	Idade	-936,266	80,989	3	,000
	ClasseKillip	-938,555	85,567	3	,000
	ClassificacaoFuncaoVE	-1043,948	296,353	3	,000
Step 4	Idade	-926,108	82,912	3	,000
	ClasseKillip	-900,130	30,957	3	,000
	ClassificacaoFuncaoVE	-1019,775	270,247	3	,000
	PressaoArterialSistolica	-895,772	22,240	3	,000
Step 5	Idade	-910,378	61,149	3	,000
	Sexo	-884,652	9,696	1	,002
	ClasseKillip	-894,220	28,833	3	,000
	ClassificacaoFuncaoVE	-1017,799	275,991	3	,000
	PressaoArterialSistolica	-891,445	23,282	3	,000
Step 6	Idade	-905,523	62,835	3	,000
	Sexo	-879,435	10,659	1	,001
	ClasseKillip	-885,669	23,126	3	,000
	ClassificacaoFuncaoVE	-1008,089	267,967	3	,000
	FrequenciaCardiaca	-879,804	11,397	2	,003
	PressaoArterialSistolica	-885,241	22,271	3	,000
Step 7	Idade	-897,811	62,031	3	,000
	IMC	-874,106	14,619	5	,012
	Sexo	-872,501	11,409	1	,001
	ClasseKillip	-878,411	23,230	3	,000
	ClassificacaoFuncaoVE	-1000,756	267,919	3	,000
	FrequenciaCardiaca	-872,382	11,172	2	,004
	PressaoArterialSistolica	-877,236	20,880	3	,000

Figura A.7: Teste de razão de verosimilhança para testar a inclusão/exclusão de variáveis em cada passo.

Bibliografia

[1] Braga, A.C.S. (2000). *Curvas ROC: aspectos funcionais e aplicações* (Dissertação submetida à Universidade do Minho para obtenção do grau de doutor no ramo de engenharia de produção e sistemas, área de métodos numéricos e estatísticos).

[2] Dobson, A.J. (2002). *An introduction to generalized linear models* (2^a ed.) Boca Raton: Chapman & Hall/CRC.

[3] Fahrmeir, L. & Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics* , 13 (1), 342-368.

[4] Gonçalves, E. & Lopes, N. M. (2003). *Estatística: teoria matemática e aplicações*. Lisboa: Escolar editora.

[5] Gourieroux, C. & Monfort, A.(1981). Asymptotic properties of the maximum likelihood estimator in dicothomous logit model. *Journal of Econometrics*, 17, 83-97.

[6] Hosmer, D.W. & Lemeshow, S. (2000). *Applied logistic regression* (2^a ed.) New York: John Wiley & Sons.

[7] McCullagh, P. & Nelder, J.A. (1989). *Generalized linear model* (2^a ed.). Monographs on statistics and applied probability (37). London: Chapman & Hall.

[8] Silva, G.L. (1992). *Modelos logísticos para dados binários* (Dissertação apresentada ao Instituto de Matemática e Estatística da Universidade de São Paulo para obtenção do grau de mestre em estatística).

[9] Turkman, M. A. A. & Silva, G. L. (2000). *Modelos lineares generalizados: da teoria à prática*. Lisboa: DEIO/FC e CEAUL; DM/IST e CMA. (Trabalho parcialmente financiado por FCT PRAXIS XXI e FEDER).