Paula Susana Pereira Martins

Análise estatística da performance de um conjunto de testes auditivos

Paula Susana Pereira Martins

Análise estatística da performance de um conjunto de testes auditivos

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requesitos necessários à obtenção do grau de Meste em Matemática e Aplicações, realizada sob a orientação científica da Doutora Andreia Oliveira Hall, Professora Associada do Departamento de Matemática da Universidade de Aveiro e do Doutor António Joaquim da Silva Teixeira, Professor Auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

Dedicatória Aos utentes da CERCIAV

o júri

presidente

Adelaide de Fátima Baptista Valente Freitas
Professora auxiliar da Universidade de Aveiro

vogais

Ana Cristina da Silva Braga
Professora auxiliar da Universidade do Minho

Andreia Oliveira Hall
Professora associada da Universidade de Aveiro

António Joaquim da Silva Teixeira
Professor auxiliar da Universidade de Aveiro

Agradecimentos

À Doutora Andreia Hall e ao Doutor António Teixeira, pelo apoio e incentivo.

Às pessoas que me são mais próximas, pela ajuda e paciência.

Aos professores que me ensinaram a gostar de estatística.

À colega Sandra Pinho, pelo que aprendemos juntas no mestrado.

Palavras-chave Estatística multivariada, regressão logística, discriminação e classificação.

Resumo

Este trabalho incidiu na análise estatística da performance do primeiro conjunto de testes de processamento auditivo central adaptados ao português europeu. Os dados em análise resultaram da aplicação desses testes a um conjunto restrito de indivíduos, tendo por objectivo diagnosticar uma patologia específica.

Dada a novidade e complexidade do estudo em causa, foi necessário recorrer a métodos estatísticos adequados para avaliar a capacidade de diagnóstico dos testes mencionados e caracterizar a forma como o fazem. Este último ponto engloba dois aspectos: a identificação de um conjunto reduzido de testes que contribuem significativamente para efectuar o diagnóstico e a construção de um modelo estatístico adequado para classificar novos elementos.

A regressão logística foi o método escolhido para resolver este problema, sem prejuízo da aplicação de métodos complementares de análise. Para a amostra em estudo, identificou-se um conjunto de três variáveis, num total de dez, que satisfaz as condições pretendidas. Foram seleccionados dois modelos, um com duas e outros com três das variáveis mais importantes, cuja performance preditiva foi comparada. O primeiro permitiu separar correctamente todos os elementos e revelou melhor performance preditiva, mas evidenciou sobreajuste. No segundo não se verificou este problema, mas os seus resultados são menos satisfatórios tanto na separação como na classificação de elementos. Antes de usar a bateria de testes como meio de diagnóstico, recomenda-se a sua aplicação a um conjunto mais vasto de indivíduos.

Key-words Multivariate statistics, logistic regression, discrimination and classification.

Abstract

The present work is about the statistical analysis of the performance of the first central auditory processing test set aimed to European Portuguese. The analyzed data are the outcomes of that test set applied to a small group of persons targeting the diagnosis of a specific pathology.

Given the novelty and complexity of this study, adequate statistical methods were demanded to evaluate the diagnosis capability of the aforementioned tests and characterize the way they do it. The later point brings up two distinct aspects: the identification of a small test set that contribute for the diagnostic significantly and the construction of an adequate statistic model useful for new elements classification.

Logistic regression was the chosen method to solve this problem, although other complementary analysis methods can be used. For the studied sample, a set of three variables out of ten was identified as satisfying the requested conditions. Two models have been selected, one with two and another with three of the most important variables, to compare their predictive performance. The first one separated correctly every single element and presented better predictive performance, but overfits. In the second one, this problem does not occur, but its results were not so good both separating and classifying elements. Before using the battery of tests for making diagnostics, is recommended to apply it to a broader group of persons.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Enquadramento e motivação

A maioria dos problemas auditivos são diagnosticados através da aplicação de um conjunto apropriado de testes audiológicos. Se, por um lado, um maior número de testes pode levar a um diagnóstico mais correcto, por outro lado, o acréscimo em tempo e em custos levam a uma procura de um conjunto reduzido de testes. Num estudo realizado recentemente, por Martins (2007), foi criada uma bateria de testes adaptada ao português europeu, para diagnosticar uma patologia do sistema auditivo. Por ser o primeiro conjunto de testes desta natureza, houve necessidade de criar um conjunto alargado para garantir uma análise abrangente. Os 10 testes que compõe essa bateria são: (1) *Teste Padrão de Frequência*, (2) *Teste de Fala no Ruído*, (3) *Teste de Fala Filtrada*, (4) *Teste de Fusão Binaural* e (5) *Teste SSW* (*Straggered Spondaic Word*), aplicados a cada um dos ouvidos. Este trabalho pretende contribuir, através da utilização de métodos estatísticos adequados, para a escolha de um conjunto reduzido de testes, subconjunto dos anteriores, capaz de diagnosticar a patologia em causa.

O estudo do sistema auditivo e a descrição pormenorizada dos testes audiológicos saem foram do âmbito deste trabalho. No entanto, é fundamental compreender o problema em causa para traçar os objectivos e conduzir a análise de forma adequada. De um modo muito simplificado, pode dizer-se que o sistema auditivo é constituído pelo sistema auditivo central e pelo sistema auditivo periférico. O primeiro encontra-se no cérebro e o segundo compreende o ouvido externo, o ouvido médio, o ouvido interno e o nervo auditivo. Um indivíduo

pode não ter qualquer problema no sistema auditivo periférico e apresentar perturbações do processamento auditivo, ao nível do sistema auditivo central. Estas podem manifestar-se de diversas formas, sendo uma das mais frequentes a dificuldade de discriminação auditiva em ambiente de ruído. A bateria de testes anteriormente referida pretende ser um meio de diagnóstico eficaz desta perturbação quando aplicada a indivíduos com limiares auditivos dentro dos parâmetros de normalidade.

Os testes foram aplicados a um conjunto de indivíduos, em ambos os ouvidos, primeiro no direito e depois no esquerdo. Todos os elementos da amostra foram sujeitos a uma avaliação auditiva preliminar para avaliar a ausência de problemas auditivos ao nível do sistema auditivo periférico. Foi ainda realizada a divisão da amostra em dois grupos, consoante se verificasse a ausência ou a presença de queixas da dificuldades mencionada. O primeiro constitui o grupo sem queixas, ou grupo de controlo, e o segundo o grupo com queixas. No sentido de fundamentar a inclusão dos elementos da amostra num ou noutro grupo, foi feita a análise dos reflexos acústicos e realizado um inquérito, relativamente ao qual foram consideradas apenas as respostas às quatro questões transcritas a seguir.

Tem dificuldade de comunicação em:

- *Ambiente silencioso com uma pessoa a falar?*
- *Ambiente silencioso com várias pessoas a falar?*
- *Ambiente ruidoso com uma pessoa a falar?*
- *Ambiente ruidoso com várias pessoas a falar?*

Relativamente a estas questões, considerou-se que um indivíduo do grupo de controlo não pode responder afirmativamente a mais do que uma. No que concerne aos reflexos ipsi e contra-laterais, devem situar-se entre 80 e 90dB HL para os indivíduos que não têm perturbações do processamento auditivo, podendo estar alterados ou ausentes para os indivíduos com perturbações.

1.2 Definição do problema

Os objectivos deste trabalho decorrem naturalmente da motivação que justificou a criação do conjunto de testes auditivos. O primeiro consiste em avaliar se as dez variáveis, correspondentes aos cinco testes aplicados a ambos os ouvidos, permitem discriminar os elementos dos dois grupos. Satisfeita esta condição, é conveniente identificar um conjunto reduzido de testes capaz de o fazer, avaliando ainda se há necessidade de aplicar um mesmo teste aos dois ouvidos. Como meio de diagnóstico, os testes devem permitir a classificação de novos elementos, sendo necessário escolher um regra ou um modelo que o permita fazer.

Em suma, pretende-se com este trabalho:

- aferir se as 10 variáveis permitem separar os elementos dos dois grupos;
- identificar um conjunto reduzido de variáveis capaz de efectuar essa separação;
- definir uma forma de classificar novos indivíduos, conhecidos os valores das variáveis seleccionadas no ponto anterior.

1.3 Dificuldades e limitações do estudo

A amostra em análise é de reduzida dimensão, limitando o uso de alguns métodos estatísticas e condicionando a generalização de alguns resultados. Algumas técnicas pressupõe a utilização de uma maior quantidade de elementos, sobretudo quando se pretende fazer a análise conjunta das variáveis. Mesmo que se utilizem metodologias de análise adequadas para amostras pequenas, a previsão do comportamento de um conjunto mais vasto de elementos pode vir afectado de erros muito grandes.

A classificação dos elementos desta amostra é conhecida, no entanto há possibilidade de existirem falhas nessa classificação. De facto, a análise dos reflexos acústicos nem sempre é conclusiva e as questões do inquérito são algo subjectivas, pelo que se trata de meios auxiliares de classificação que não são infalíveis. Caso assim não fosse, este trabalho não fazia sentido, uma vez que não havia necessidade de aplicar testes para diagnosticar perturbações do processamento auditivo. A investigadora que construiu e aplicou os testes teve necessidade de

alterar a classificação de alguns indivíduos depois de analisar as informações recolhidas. Há que ter em conta esta fragilidade ao longo de todo o estudo, analisando com alguma precaução os dados desses indivíduos. Depois das dessas alterações, os elementos 12, 20 e 24 passaram a integrar o grupo de controlo e os elementos 7 e 8 mudaram para o grupo com queixas.

1.4 Estrutura da tese

No capítulo seguinte faz-se uma breve descrição de alguns métodos estatísticos que podem ser úteis na análise de problemas desta natureza. Depois de ponderadas as vantagens e desvantagens de cada método específico e tendo em conta os objectivos deste trabalho e a especificidade dos dados, foi definida a análise a efectuar. Como se verá ao longo deste trabalho, a regressão logística parece ser um método adequado para modelar problemas desta natureza. O Capítulo 3 é dedicado ao estudo teórico deste método, explorando os casos univariado e multivariado.

A primeira etapa de análise dos dados é apresentada no Capítulo 4 e tem um carácter exploratória. Procurou-se organizar e resumir a informação para conhecer as especificidades dos valores observados para cada grupos. A escolha do conjunto de variáveis para separar os elementos dos dois grupos e a definição de um modelo para classificar novos elementos é o objectivo da aplicação da regressão logística, no Capítulo 5.

No Capítulo 6, os resultados da aplicação da regressão logística são objecto de um estudo complementar. Trata-se de uma discussão centrada na aplicação e avaliação de alguns modelos e numa análise crítica da construção dos modelos e da classificação dos elementos. Este trabalho termina com algumas conclusões e recomendações para trabalhos futuros.

Capítulo 2

Escolha da metodologia de análise

A selecção dos métodos de análise de um conjunto de dados é de vital importância para a concretização dos objectivos definidos, no contexto de um problema específico. O número de variáveis envolvidas, as respectivas escalas de medida e as relações de dependência ou interdependência entre elas são aspectos relevantes para essa decisão.

Neste trabalho, a amostra é composta por 24 elementos e podem ser considerados dois tipos de variáveis. A primeira é dicotómica e identifica o grupo e as restantes referem-se aos resultados dos testes auditivos. Estes concretizam-se na percentagem de respostas correctas, numa escala de razão com valores discretos.

Para avaliar a relevância e a fiabilidade da aplicação deste conjunto de testes, é necessário analisar se têm capacidade de separar os dois grupos e de efectuar um diagnóstico correcto para novos indivíduos. Há diversas formas de abordar este problema, das quais derivam diferentes modos de análise. Pode-se considerar que os elementos dos dois grupos constituem duas amostras aleatórias, para as quais se analisou o mesmo conjunto de variáveis. Neste caso, procede-se a uma análise comparativa dos valores observados dessas variáveis, em duas amostras independentes. Outra abordagem possível é a existência de uma única amostra, tendo-se observado dois conjuntos de variáveis, como referido acima. Deste ponto de vista, analisam-se os dados considerando uma relação de dependência entre as variáveis.

A seguir é apresentada uma breve descrição de vários métodos estatísticos, indicando os

pressupostos e as vantagens e desvantagens da sua aplicação no contexto deste estudo. Serve este inventário para escolher as técnicas consideradas mais adequadas para atingir os objectivos indicados.

2.1 Análise exploratória de dados

A primeira etapa da análise de dados será de carácter exploratório, recorrendo a medidas de estatística descritiva e a representações gráficas. Pretende-se estudar o comportamento das variáveis, identificar semelhanças e diferenças entre elas e eventuais valores atípicos. Esta forma de análise não é suficiente para responder às questões levantadas, mas é considerada um bom ponto de partida para a exploração de outros métodos mais apropriados no contexto do problema em estudo.

2.2 Análise de variância múltipla - MANOVA

Os dois grupos em análise neste trabalho são formados por elementos que se distinguem por uma característica em particular. A aplicação dos testes auditivos só tem sentido se os resultados neles obtidos explicarem a diferença entre os grupos. Dito de outro modo: se a presença ou ausência de dificuldades de discriminação auditiva se traduzir em diferenças significativas nos resultados da bateria de testes. Esta é a formulação típica de um problema de análise de variância múltipla com um factor e 10 variáveis. O factor é a variável grupo, que toma apenas dois valores, designados níveis. O objectivo é analisar a existência de diferenças estatisticamente significativas dos vectores de médias dos dois níveis.

Numa MANOVA, se houver motivo para rejeitar a hipótese da igualdade dos vectores de médias, pode ainda ser necessário identificar quais as variáveis responsáveis por essas diferenças. Neste ponto, a análise de variância univariada (ANOVA) pode ser aplicada sucessivamente para avaliar a diferença de médias de cada variável. Na presença de um único factor com dois níveis, podem ainda ser aplicados testes univariados para avaliar essas diferenças.

Segundo Sharma (2006), a ANOVA e os testes univariados podem ser um complemento à MANOVA mas não a devem substituir nem preceder. A aplicação sucessiva de testes univariados ou da ANOVA aumenta consideravelmente a probabilidade de rejeitar a hipótese nula quando ela é verdadeira. Outra desvantagem dessas análises univariadas é o facto de não terem em conta as relações que existem entre as variáveis. Com este procedimento não é considerada informação que pode ser relevante, sobretudo se as variáveis forem fortemente correlacionadas.

O estudo de Martins (2007) incluiu a aplicação da MANOVA para comparar os vectores de médias dos dois grupos e a aplicação de testes estatísticos univariados para avaliar as diferenças de médias de cada variável. Os resultados obtidos com estas metodologias serão referidos mais à frente e comparados com a análise efectuada neste trabalho. A aplicação da MANOVA seguida da realização de testes univariados para comparar os dois grupos estão de acordo com os objectivos deste trabalho. No entanto, desconhece-se o comportamento das variáveis em estudo nas populações de onde provêm os elementos dos dois grupos. Acresce ainda o facto do número de elementos observados em cada grupo ser reduzido. Dadas estas condicionantes e de acordo com o que se refere a seguir, considera-se que a MANOVA não é o melhor método para analisar os dados deste estudo.

A MANOVA pressupõe as seguintes condições:

- as observações são independentes;
- os grupos são amostras aleatórias de populações com distribuição normal multivariada;
- as populações têm matrizes de covariâncias iguais.

Existem ainda outros factores a ter em conta antes de realizar uma MANOVA, nomeadamente a dimensão da amostra e a existência de *outliers*. Segundo Hair (1998), o número de elementos de cada grupo deve ser maior ou igual ao número de variáveis e nunca inferior a 20.

As 10 variáveis em estudo neste trabalho tomam valores discretos que são conhecidos. No entanto, a distribuição das variáveis e a matriz de covariâncias de cada população são desco-

nhecidas. Acresce ainda o reduzido número de elementos observados, 14 de um grupo e 10 do outro. Com estas condicionantes, é conveniente procurar métodos alternativos à MANOVA.

2.3 Análise Discriminante

A análise discriminante é muito utilizada em problemas estatísticos que visam a separação de elementos em dois ou mais grupos e a posterior classificação de novos elementos. Estas duas etapas distinguem-se não só pelos seus objectivos, mas também pelos procedimentos. Dadas as especificidades de cada uma, Huberty (2006) divide este método em análise discriminante descritiva e análise discriminante preditiva.

A análise discriminante descritiva só pode ser aplicada se os grupos forem conhecidos à partida, tal como acontece neste estudo. O objectivo é identificar uma forma de diferenciar os elementos dos vários grupos de acordo com os valores observados de um conjunto de variáveis independentes. O processo consiste em encontrar combinações lineares dessas variáveis, as funções discriminantes, que minimizem a probabilidade de incorrecta classificação dos indivíduos. Nos casos em que a análise discriminante revela que as variáveis permitem separar os elementos, é possível avaliar quais as variáveis que mais contribuem para essa separação e encontrar novas funções discriminantes. Este procedimento pode ser realizado automaticamente em vários programas de estatística.

Ao contrário do que acontecia na MANOVA, a variável grupo é a variável dependente e os testes auditivos constituem as variáveis independentes.

A análise discriminante preditiva usa os resultados da fase anterior para identificar o grupo de novos indivíduos, sendo conhecidos os valores das variáveis independentes. De acordo com as funções discriminantes, o espaço de resultados é dividido em tantas regiões quantos os grupos e a alocação de novos elementos respeita essa divisão. Contudo, esta classificação não é isenta de erros, podendo verificar-se a sobreposição de regiões.

A aplicação da análise discriminante requer os pressupostos que foram referidos para a

MANOVA. Segundo Reis (1997), a desigualdade da dispersão entre os grupos afecta os testes realizados na etapa da separação e as regras de classificação da segunda etapa. A violação do pressuposto da normalidade também pode levar a decisões erradas na definição das funções discriminantes, sobretudo em amostras de pequena dimensão.

2.4 Regressão Logística

Os modelos de regressão são muitas vezes usados para estudar fenómenos que podem ser descritos por meio de uma relação de dependência de uma variável relativamente a outras. A primeira é designada variável resposta ou dependente e as segundas são as variáveis explanatórias ou independentes. Na aplicação de um método de regressão procura-se construir um modelo que defina a variável resposta em função das variáveis explanatórias e de um erro aleatório. Quando a variável dependente é dicotómica, o método de regressão mais adequado para descrever a relação de dependência é a regressão logística.

A aplicação da regressão logística a um conjunto de dados resulta na construção de um modelo que pretende descrever o melhor possível a variável dependente em função dos valores das variáveis independentes. Depois de avaliado, este modelo pode ser usado para prever o comportamento de novos elementos, sendo conhecidos os respectivos valores das variáveis explanatórias. As etapas acabadas de descrever correspondem à construção de um modelo para separar os elementos de dois grupos disjuntos e a sua aplicação na classificação ou alocação de novos elementos. Sempre que for considerado necessário, a regressão logística pode ser aplicada na selecção das variáveis que mais contribuem para a separação dos elementos.

A análise discriminante e a regressão logística têm muitos aspectos em comum, mas o segundo método é menos exigente no que se refere aos propostos. Para além de poder ser aplicada com variáveis explanatórias numéricas, categóricas ou ambas, a regressão logística não assume nenhuma distribuição específica dos dados. Em relação à dimensão da amostra, deve ser grande, sobretudo se o número de variáveis for elevado. No caso da dimensão ser pequena, há risco de sobreajuste do modelo e a classificação de novos elementos pode ficar comprometida.

Pohar, Blas e Turk (2004) realizaram um estudo comparativo da performance da análise discriminante e da regressão logística. No caso de se verificar a normalidade dos dados, concluiu-se que a análise discriminante apresenta melhores resultados, sendo essas diferenças menos acentuadas para amostras de grande dimensão. Pelo contrário, se a distribuição dos dados se afastar da normalidade, a regressão logística é mais adequada, mesmo em amostras de pequena dimensão.

A distribuição da população em estudo não é conhecida, a dimensão da amostra é pequena e não é possível validar algumas das condições requeridas por outros métodos. Optou-se por centrar a análise de dados na aplicação da regressão logística, por ser um método que vai ao encontro dos objectivos deste estudo e por não requerer muitos pressupostos. No capítulo seguinte apresenta-se uma descrição mais detalhada deste método dada a sua importância na realização deste trabalho.

2.5 Análise de *Clusters*

Existem diversos métodos para realizar uma análise de *clusters* ou de agrupamentos[1], mas todos eles visam a separação dos elementos de uma amostra em grupos disjuntos. A agregação dos elementos pode ser feita com base em diversos critérios de semelhança escolhidos de acordo com o problema em análise. Nalguns métodos a escolha do número de grupos é feita antes da construção dos mesmos. Já nos métodos hierárquicos, os elementos vão sendo agregados sucessivamente, formando grupos cada vez maiores. Estes últimos sejam mais flexíveis por permitirem analisar diferentes níveis de agregação, mas não existem regras rigorosas para decidir quais os grupos que devem ser considerados.

No problema em análise pretende-se encontrar uma forma diferenciar dois grupos conhecidos em função dos valores de um conjunto de variáveis. Na análise de aglomerados, mesmo que o número de grupos seja decidido inicialmente, os seus elementos não podem ser escolhidos. A

[1]De acordo com o glossário estatístico Inglês-Português elaborado pela Sociedade Portuguesa de Estatística e pela Associação Brasileira de Estatística

ser usada, esta técnica deveria visar apenas confrontar a classificação dos elementos da amostra, uma vez que podem existir falhas a esse nível. No entanto, o contributo de cada variável para a separação dos dois grupos não é conhecido e é possível que algumas importantes. A análise de *clusters* não permite fazer essa ponderação e cria os grupos independentemente da relevância de cada variável para descrever as diferenças entre eles.

2.6 Análise de Componentes Principais

O método da análise de componentes principais é frequentemente utilizado para reduzir o número de variáveis em estudo. As variáveis iniciais, possivelmente correlacionadas, são transformadas num outro conjunto de variáveis não correlacionadas, as componentes principais. Estas são escritas como combinações lineares das primeiras, usando os vectores próprios, unitários, associados aos valores próprios da matriz de covariâncias. Se as variáveis iniciais forem medidas em escalas amplamente diferentes ou se as unidades de medida não forem comensuráveis, pode ser útil realizar a análise de componentes principais usando os dados estandardizados. Neste caso, as componentes principais são escritas à custa dos vectores próprios assocciados aos valores próprios da matriz de correlações.

Os valores próprios são considerados por ordem decrescente, sendo os coeficientes das variáveis da i-ésima componente principal os vectores próprios associados ao i-ésimo valor próprio. Desta forma as componentes principais são definidas por ordem decrescente da variabilidade dos dados que explicam. A percentagem de variância explicada por uma componente principal é dada pelo quociente do respectivo valor próprio e da soma de todos os valores próprios. O coeficiente de correlação de Pearson entre uma variável e uma componente principal permite avaliar a importância da variável na componente principal. As demonstrações destes resultados podem ser encontradas em Johnson e Wichern (1998) e em Reis (1997).

O número de componentes principais a reter para efectuar a análise deve ser escolhido por forma a minimizar a perda de informação contida nos dados. Existem algumas regras empíricas que são muitas vezes usadas para decidir o número de componentes a reter, ver por exemplo Reis (2001).

A redução do número de variáveis necessárias para separar os elementos dos dois grupos é um dos objectivos deste trabalho. O que acontece na análise de componentes principais é uma transformação das variáveis originais e a utilização de um número mais reduzido de variáveis, as primeiras componentes principais. Estas concentram grande parte da informação contida nos dados, mas não eliminam a necessidade de observar todas as variáveis. O contributo das variáveis para as componentes principais pode ser muito variável, mas todas estão presentes.

Este método de análise não será usado por não servir os interesses deste trabalho. Em trabalhos futuros, depois de identificar e validar um conjunto reduzido de variáveis, a análise de componentes principais pode ser um método complementar de análise.

2.7 *Software* utilizado

Actualmente existem muitas ferramentas computacionais que permitem realizar o tratamente sitemático de grandes quantidades de informação. Também na estatística esses avanços se fazem sentir e são cada vez mais os programas que facilitam a análise de dados. Todos os métodos de análise aqui referidos estão implementados em vários desses programas.

Embora a autora deste trabalho estivesse mais familiarizada com o *SPSS* e o *S-PLUS*, optou por utilizar o *R*. O uso deste programa foi facilitado, não só pelo recurso à ajuda do próprio programa, mas também pela consulta de alguns manuais disponiveis na internet, concretamente Venables, et. al (2008) e R Development Core Team (2008).

A escolha do *R* baseou-se na necessidade de automatizar alguns procedimentos e de definir novas metodologias de análise. A criação de pequenos programas permitiu realizar várias experiências e comparar resultados, tornando o processo de análise mais eficaz. Um desses programas permitiu definir um processo de selecção das variáveis a incluir no modelo de regressão logística que não estava implementado nos programas mencionados. A justificação da escolha desse método é apresentada no Capítulo 5.

Capítulo 3

Regressão Logística

3.1 Modelo de regressão logística univariado

Os modelos de regressão são utilizados na análise de dados com o intuito de descrever a relação entre uma ou mais variáveis explanatórias e uma variável resposta. O modelo clássico de regressão linear, embora seja o mais utilizado, não é aplicável em muitas situações, dado assumir alguns pressupostos como a normalidade da variável resposta. O método de regressão logística é o mais adequado para modelar um problema de regressão com uma variável resposta dicotómica. A análise apresentada neste capítulo baseia-se essencialmente no trabalho de Hosmer e Lemeshow (1989).

O método empregue em qualquer modelo de regressão consiste em estimar o valor esperado da variável resposta, Y, dado o valor do vector das variáveis explanatórias, $\mathbf{x}$, $E[Y|\mathbf{x}]$. O modelo de regressão é definido à custa de $E[Y|\mathbf{x}]$ e de um erro ε que expressa o desvio das observações em relação a esse valor: $Y = E[Y|\mathbf{x}] + \varepsilon$. No caso do modelo de regressão logística, é usual denotar-se $E[Y|\mathbf{x}]$ por $\pi(\mathbf{x})$.

A variável Y tem distribuição de Bernoulli, sendo: $Y = \begin{cases} 1 & \text{se sucesso} \\ 0 & \text{se insucesso} \end{cases}$

O valor esperado de uma variável de Bernoulli é igual à probabilidade de sucesso. Logo, $\pi(\mathbf{x}) = E[Y|\mathbf{x}] = P(Y = 1|\mathbf{x})$, donde se conclui que a distribuição condicional da variável resposta é descrita pela distribuição de Bernoulli com probabilidade $\pi(\mathbf{x})$. Os erros têm

distribuição de Bernoulli deslocada, com média zero e variância $\pi(\mathbf{x})(1-\pi(\mathbf{x}))$:

$$\varepsilon = y - \pi(x) = \begin{cases} 1-\pi(x) & \text{se y=1, com probabilidade } \pi(x) \\ -\pi(x) & \text{se y=0, com probabilidade } 1-\pi(x) \end{cases}$$

Neste trabalho, existem dez variáveis explanatórias, exigindo a utilização do modelo de regressão logística multivariado. Contudo, apresenta-se primeiro uma análise do modelo univariado no que se refere à estimação dos parâmetros e aos testes estatísticos para avaliar o significado dos mesmos.

Considere-se uma amostra de n observações independentes do par (x_i,y_i), onde x_i *e* y_i representam, respectivamente, o valor da variável explanatória e da variável resposta do i-ésimo elemento. A função de regressão logística univariada é dada pela esperança de Y dado x, como se segue:

$$\pi(x) = E[Y|\mathbf{x}] = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{3.1}$$

Os parâmetros são estimados pelo método da máxima verosimilhança, que consiste em determinar os valores dos parâmetros que maximizam a probabilidade de obter o conjunto de valores observados. Uma vez que Y tem distribuição de Bernoulli, $f(y_i) = P(Y = y_i) = \pi(x_i)^{y_i}\,[1-\pi(x_i)]^{1-y_i}$. Dada a independência das observações, a função de verosimilhança é dada por:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\,[1-\pi(x_i)]^{1-y_i}$$

Em lugar da função anterior, é habitual usar-se o seu logaritmo, por se tornar mais fácil encontrar os valores pretendidos:

$$L(\ \beta\ ) = ln(\ l(\ \beta\ )) = \Sigma_{i=1}^{n}\,[y_i\ ln\,[\pi(x_i)] + (1-y_i)\ ln\,[1-\pi(x_i)]]$$

Substitui-se $\pi(x_i)$, usando a equação (3.1) e simplifica-se:

$$L(\ \beta\ ) = \Sigma_{i=1}^{n} \left[y_i\ (\beta_0 + \beta_1\ x_i) - ln(1 + e^{\beta_0 + \beta_1 x_i}) \right] \tag{3.2}$$

As equações de verosimilhança consistem em igualar a zero as derivadas parciais da função (3.2). As respectivas soluções são as estimativas dos parâmetros desconhecidos e determinam-se por processos iterativos.

Equações de verosimilhança:

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

Voltando à formulação inicial, obtém-se:

$$\begin{cases} \sum_{i=1}^{n} \left[y_i - \pi(x_i) \right] = 0 \\ \sum_{i=1}^{n}\ x_i \left[y_i - \pi(x_i) \right] = 0 \end{cases} \tag{3.3}$$

Os estimadores de máxima verosimilhança de β e $\pi(x_i)$ são representados por $\widehat{\beta}$ e $\widehat{\pi}(x_i)$, respectivamente. Este último é designado valor predito do modelo de regressão. Por uma questão de simplificação da notação, em vez de $\pi(x_i)$ ou $\widehat{\pi}(x_i)$, pode escrever-se π_i ou $\widehat{\pi}_i$, respectivamente.

3.1.1 Função *logit*

A função *logit* resulta de uma transformação logarítmica da função de regressão logística, como se segue:

$$g(x) = \ln \left[\frac{\pi(x)}{1 - \pi(x)} \right] = \beta_0 + \beta_1 x \tag{3.4}$$

A função assim definda é linear nos parâmetros, pode ter valores contínuos e variar entre $-\infty$ e $+\infty$, tendo muitas propriedades do modelo de regressão linear clássico. O quociente

$\frac{\pi(x)}{1-\pi(x)}$ é denominado *odds* e o seu valor pode ser interpretado como as chances de sucesso da variável Y. À razão de duas *odds* dá-se o nome de *odds ratio* ou razão das chances[1].

Num modelo univariado, verifica-se que $\beta_1 = g(x+1) - g(x)$, donde se conclui que $\widehat{\beta_1}$ representa a variação da função *logit* pelo aumento de uma unidade do valor da variável explanatória. Usando as propriedades dos logarítmos, verifica-se que $g(x+1) - g(x)$ é o logaritmo neperiano de uma *odds ratio*. Considerando $\widehat{\psi} = e^{(\widehat{\beta_1})}$, o valor de $\widehat{\psi}$ representa a razão das chances de sucesso relativas ao aumento de uma unidade da variável explanatória. Se $\widehat{\beta_1}$ for maior do que zero, o aumento do valor da variável explanatória provoca um aumento da probabilidade de sucesso, enquanto um valor negativo diminui essa probabilidade. Mais detalhes sobre a interpretação dos coeficientes do modelo de regressão logística (univariado e multivariado) podem ser encontrados em Hosmer e Lemeshow (1989) ou Manning (2007).

3.1.2 Avaliar a importância da variável

A comparação entre os valores observados e os valores preditos é feita com base na razão das verosimilhanças do modelo em análise e do modelo saturado. Este é o modelo de regressão logística que contém tantos parâmetros quantos os valores observados. Dado o interesse em utilizar um teste estatístico para avaliar aquela razão de verosimilhanças, usa-se o seu logaritmo multiplicado por menos dois, cuja distribuição é conhecida. Este valor é designado por D e o teste utilizado é o teste da razão de verosimilhanças.

$$D = -2\ln\left[\frac{\text{verosimilhança do modelo univariado}}{\text{verosimilhança do modelo saturado}}\right]$$

Para avaliar o significado de uma variável independente, utiliza-se a diferença dos valores de D com e sem a variável no modelo, estatística G. Uma vez que a verosimilhança do modelo saturado é comum às duas parcelas, usando as propriedades dos logaritmos pode escrever-se G da seguinte forma:

[1] De acordo com o glossário estatístico Inglês-Português elaborado pela Sociedade Portuguesa de Estatística e pela Associação Brasileira de Estatística

$$G = -2\ln\left[\frac{\text{verosimilhança do modelo sem a variável}}{\text{verosimilhança do modelo com a variável}}\right]$$

Da resolução da primeira equação de (3.3), obtém-se a estimativa de β_0. Quando a variável independente não está no modelo, o seu valor é dado por: $\widehat{\beta}_0 = \ln\left(\frac{n_1}{n_0}\right)$, onde $n_1 = \sum_{i=1}^{n} y_i$ *e* $n_0 = \sum_{i=1}^{n}(1-y_i)$. Neste caso $\widehat{y}_i = \frac{n_1}{n} = \overline{y}$.

Sob a hipótese de $\beta_1 = 0$, $G \sim \chi^2_{(1)}$. Rejeita-se a hipótese nula para valores elevados da estatística de teste G, cujo valor é dado por:

$$G = 2\left[\sum_{i=1}^{n}[y_i\ln(\widehat{\pi}_i) + (1-y_i)\ln(1-\widehat{\pi}_i)] - [n_1\ln(n_1) + n_0\ln(n_0) - n\ln(n)]\right] \tag{3.5}$$

Outros testes que podem ser usados para avaliar o significado da variável independente são o teste de Wald e o Score test. Sob a hipótese de $\beta_1 = 0$, as estatísticas destes testes têm distribuição normal centrada e reduzida. A estatística do primeiro é dada pelo quociente da estimativa de β_1 e da estimativa do seu erro: $W = \frac{\widehat{\beta}_1}{\widehat{SE}(\widehat{\beta}_1)}$. A estatística do Score test é:

$$ST = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sqrt{\overline{y}(1-\overline{y})\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

Tanto no teste de Wald como no Score test, rejeita-se a hipótese nula para valores absolutos elevados da correspondente estatística de teste, $|W|$ ou $|ST|$.

O teste da razão de verosimilhanças e o teste de Wald requerem o cálculo da estimativa de máxima verosimilhança de β_1. Já o Score test não requer esse cálculo, o que constitui uma vantagem, embora pouco relevante uma vez que dispomos de meios computacionais com grande capacidade de cálculo.

Selecção de variáveis

Num estudo que envolva várias variáveis independentes, pode começar-se por construir os modelos de regressão logística univariados e verificar se algum deles apresenta bom ajuste.

Caso isso não aconteça, é necessário seleccionar as variáveis a incluir num modelo multivariado. Tal pode ser realizado por aplicação do teste univariado de Wald. Este permite ponderar se cada variável, independentemente das outras, é ou não relevante para modelar a relação de dependência em causa. Este procedimento tem a desvantagem de não ter em conta as relações entre as variáveis explanatórias. De facto, pode acontecer que uma variável, por si só, não seja muito relevante na detecção da característica em estudo, mas associação a outra(s) passe a ser significativa. O contrário também é válido, pois uma variável pode ter um contributo significativo no modelo univariado e tornar-se dispensável quando analisada em conjunto com outra(s). Deste modo, para evitar excluir variáveis eventualmente importantes aquando da aplicação do teste de Wald, é conveniente escolher um nível de significância elevado. Hosmer e Lemeshow (1989) consideram que as variáveis cujos modelos univariados apresentam p-value< 0.25 no teste de Wald devem ser consideradas candidatas para o modelo multivariado, cumulativamente com as variáveis cuja importância biológica seja conhecida.

3.2 Modelo de regressão logística multivariado

A regressão logística pode ser utilizada, com as necessárias adaptações, para modelar situações com mais do que uma variável explanatória. Considerem-se n observações independentes do par $(\mathbf{x}_i, y_i)$, onde $\mathbf{x}_i$ é um vector de p variáveis explanatórias e y_i uma variável dicotómica. A função logística usada para modelar esta situação é idêntica à do modelo univariado (3.1), envolvendo as p variáveis explanatórias:

$$\pi(\mathbf{x}_i) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} \tag{3.6}$$

Os p+1 parâmetros desconhecidos são estimados pelo método da máxima verosimilhança, por processos iterativos, sendo as equações de verosimilhança dadas por:

$$\begin{cases} \frac{\delta L}{\delta \beta_0} = 0 \\ \frac{\delta L}{\delta \beta_j} = 0 \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^{n}[y_i - \pi(x_i)] = 0 \\ \\ \sum_{i=1}^{n} x_{ij}[y_i - \pi(x_i)] = 0 \end{cases}, \ j = 1, ..., p$$

Independentemente do número de variáveis usadas para definir o modelo de regressão logística, pretende-se que permita distinguir dois grupos disjuntos de indivíduos, consoante apresentem ou não determinada característica. Como foi referido anteriormente, no presente estudo importa reduzir o número de variáveis a incluir no modelo. De acordo com Hosmer e Lemeshow (1989), essa redução constitui uma mais valia em termos estatísticos pois o aumento do número de variáveis incluídas tende a aumentar o risco de sobreajuste do modelo, sobretudo em amostras de pequena dimensão. Esta situação traduz-se, regra geral, em valores excessivamente elevados das estimativas dos coeficientes e/ou dos erros padrão.

Para verificar se as variáveis explanatórias permitem identificar correctamente os elementos que pertencem a cada grupo, constrói-se o modelo de regressão logística com todas as variáveis e avalia-se a qualidade do seu ajuste. Os valores preditos são então comparados com os da variável resposta, que toma os valores 0 ou 1. Os indivíduos são bem classificados se o valor absoluto da diferença entre o valor predito e o da variável resposta for menor do que 0.5. Se uma grande percentagem de indivíduos for bem classificada, o ideal seria que tal se verificasse para todos, pode ser conveniente encontrar um modelo com menos variáveis que permita separar os elementos dos dois grupos. Antes de passar à discussão dos vários métodos de selecção de variáveis, discute-se a aplicação do teste da razão de verosimilhanças multivariado.

3.2.1 Teste da razão de verosimilhanças multivariado

A aplicação do teste da razão de verosimilhanças para um modelo multivariado permite averiguar se os coeficientes das variáveis explanatórias são todos nulos, $\beta_1 = \beta_2 = ... = \beta_p = 0$. A estatística deste teste tem a formulação apresentada em (3.5), com a particularidade de $\hat{\pi}_i$ ser escrita em função dos p+1 parâmetros estimados. Sob a hipótese dos referidos coeficientes serem todos nulos, a estatística de teste, G, tem distribuição aproximada de um $\chi^2_{(p)}$. Para valores pequenos de G não se rejeita a hipótese nula e considera-se que nenhuma das variáveis

é importante para classificar os indivíduos relativamente à característica de interesse. Na situação contrária, apenas se pode concluir que pelo menos um dos coeficientes, $\beta_j,\ j = 1, ..., p$, é diferente de zero, não sendo possível, com este teste, identificar o(s) coeficiente(s) em questão.

O teste da razão de verosimilhanças também pode ser utilizado para comparar a qualidade de ajuste de dois modelos, sempre que o conjunto das variáveis explanatórias de um deles seja um subconjunto das variáveis usadas para construir o outro (modelo maior). Sob a hipótese dos coeficientes das variáveis excluídas do modelo maior serem todos nulos, a estatística de teste tem distribuição aproximada de um χ^2 com um número de graus de liberdade igual ao número de variáveis retiradas. A hipótese da qualidade de ajuste dos dois modelos ser semelhante deve ser rejeitada para valores elevados da estatística de teste.

3.2.2 Métodos de selecção de variáveis

Um procedimento de selecção de variáveis, baseado na aplicação do teste de Wald, foi referido no contexto da construção dos modelos univariados, na secção (3.1.2). Ao incluir todas as variáveis cujos modelos univariados apresentam p-value inferior a 0.25 no teste de Wald, corre-se o risco de usar mais variáveis do que as estritamente necessárias. Após a construção do modelo com as variáveis seleccionadas, compara-se a qualidade do seu ajuste com a do modelo com todas as variáveis. Se não houver motivo para rejeitar a hipótese dos coeficientes das variáveis excluídas serem todos nulos, é ainda conveniente estudar outros modelos com menos variáveis.

Existem vários métodos automáticos de selecção de variáveis para encontrar o melhor modelo de regressão logística, entre os quais se destacam os métodos passo a passo, por serem muito utilizados e estarem implementados em vários programas de estatística. Uma opção consiste em construir o modelo com todas as variáveis e em seguida retira-se, a cada passo, a variável menos significativa. Também é possível começar por construir o modelo com a variável que mais contribui para a separação dos dois grupos e incluir, a cada passo, a variável mais importante ainda não presente no modelo. Após a introdução ou remoção de uma variável, compara-se a qualidade de ajuste do novo modelo com a do anterior, termi-

nando o processo quando essa qualidade for semelhante ou inferior, respectivamente. Estes dois processos podem ser combinados, começando pela introdução das variáveis seguida da eliminação ou vice-versa. Apesar da larga difusão destes métodos, têm-lhes sido apontados alguns problemas, como as referidos por Whitaker (1987) e King (2003): o erro na determinação do número de graus de liberdade e falha na selecção do melhor subconjunto de variáveis de uma dada dimensão.

Nos casos em que o número de variáveis envolvidas são seja demasiado elevado, pode aplicar-se o método exaustivo de selecção de variáveis. Deste modo, evita-se que algum bom modelo não seja analisado e a construção de cada modelo deixa de estar dependente de decisões anteriores. São construídos e comparados, relativamente à qualidade de ajuste, todos os modelos com uma variável, depois com duas e assim sucessivamente, até ao número total de variáveis. Em alternativa, o processo pode terminar quando for encontrado um modelo com boa qualidade de ajuste e que satisfaça os objectivos pretendidos.

3.3 Avaliar o ajuste do modelo

A avaliação da qualidade do ajuste de um modelo de regressão logística envolve a análise de medidas das diferenças entre os valores observados da variável resposta, y, e os valores preditos $\widehat{\pi}_i$, denominadas resíduos. Antes de prosseguir com o estudo dessas medidas, é necessário fazer algumas considerações e apresentar alguma notação.

Num modelo de regressão logística com p variáveis explanatórias e n indivíduos, podem existir indivíduos diferentes que apresentem os mesmos valores para o conjunto das p variáveis. Neste caso, denota-se por J o número de valores diferentes de $\mathbf{x}$, sendo $\mathbf{x} = (x_1, x_2, ..., x_p)$, e por $m_j, j = 1, 2, ..., J$, o número de indivíduos com $\mathbf{x} = \mathbf{x_j}$.

Quando se pretende avaliar o ajuste de um modelo, podem ser usadas representações gráficas dos valores dos resíduos e testes baseados em estatísticas sumárias desses valores. No primeiro caso é possível comparar os resíduos dos vários elementos, enquanto que os testes se baseiam apenas no valor da estatística de teste, avaliando a qualidade de ajuste do modelo

em termos globais. Após a aplicação de um teste de análise de resíduos, se não houver motivo para rejeitar a hipótese da qualidade de ajuste do modelo, tal não significa que essa qualidade se verifique para todos os elementos. Neste caso, é conveniente verificar se existem elementos com valores absolutos de resíduos elevados, comparativamente aos resíduos dos restantes elementos.

De entre as medidas das diferenças dos valores observados e preditos usadas em regressão logística, destacam-se os resíduos de Pearson, r, e os *deviance residuals*, d:

$$r_j = r(y_j, \widehat{\pi}_j) = \frac{y_j - m_j\widehat{\pi}_j}{\sqrt{m_j\widehat{\pi}_j(1-\widehat{\pi}_j)}} \tag{3.7}$$

$$d_j = d(y_j, x_j) = \pm\sqrt{2\left[y_j \ln\left(\frac{y_j}{m_j\widehat{\pi}_j}\right) + (m_j - y_j)\ln\left(\frac{m_j - y_j}{m_j(1-\widehat{\pi}_j)}\right)\right]} \tag{3.8}$$

As estatísticas sumárias X^2 e D são dadas pela soma dos quadrados dos resíduos de Pearson e dos *deviance residuals*, respectivamente:

$$X^2 = \sum_{j=1}^{J} r(y_j, \widehat{\pi}_j)^2 \qquad e \qquad D = \sum_{j=1}^{J} d(y_j, \widehat{\pi}_j)^2$$

Sob a hipótese do modelo ser adequado, ambas as estatísticas têm uma distribuição aproximada de um $\chi^2_{J-(p+1)}$, devendo rejeitar-se a hipótese nula para valores elevados da estatística de teste. Segundo Kuss (2002), aquela aproximação só é válida se os valores de $m_j, j = 1, 2, ..., J$ forem elevados, pelo que estes testes não devem ser usados se os dados são esparsos.

Hosmer e Lemeshow (1989) propuseram uma estatística da qualidade de ajuste de um modelo de regressão logística que pressupõe que os dados sejam agrupados em k grupos segundo as probabilidades estimadas, nomeadamente considerando os seus percentis ou decis. Para cada grupo k, denota-se por n_k o número de indivíduos e por c_k o número de valores diferentes do conjunto das p variáveis explanatórias. A soma dos valores da variável resposta e a média das probabilidades estimadas para o grupo k denotam-se por o_k *e* $\overline{\pi}_k$, *com* $o_k = \sum_{j=1}^{c_k} y_i \quad e \quad \overline{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j\widehat{\pi}_j}{n_k}$. A estatística de Hosmer-Lemeshow, C, tem uma distribuição

aproximada de um χ^2_{k-2} sob a hipótese do modelo ser adequado. A hipótese nula deve ser rejeitada para valores elevados da estatística de teste, a qual tem a seguinte formulação:

$$C = \sum_{k=1}^{g} \frac{o_k - n_k\overline{\pi}_k}{n_k\overline{\pi}_k(1-\overline{\pi}_k)}$$

Uma das desvantagens deste teste consiste no facto do seu resultado poder estar dependente dos grupos escolhidos.

Vários têm sido os testes de qualidade de ajuste desenvolvidos por diversos autores, sendo conveniente analisar, para cada caso concreto, quais os mais apropriados. Para o efeito pode recorrer-se a estudos comparativos da performance dos mesmos, tais como os sugeridos por Kuss (2002) e Hosmer et. al (1997).

Capítulo 4

Análise preliminar de dados

Daqui em diante, considera-se que a variável grupo toma os valores 0 ou 1, consoante o indivíduo pertença ao grupo de controlo ou com queixas. O primeiro tem 14 elementos e o segundo 10 e passam a ser designados G0 e G1, respectivamente.

O vector das variáveis referentes aos testes auditivos escreve-se da seguinte forma: $X = [X_{1d}, X_{1e}, X_{2d}, X_{2e}, X_{3d}, X_{3e}, X_{4d}, X_{4e}, X_{5d}, X_{5e}]$. A numeração identifica o teste correspondente, conforme apresentado na introdução, e as letras d e e designam os ouvidos direito e esquerdo, respectivamente.

Neste capítulo procede-se a uma breve análise descritiva dos dados recorrendo a métodos gráficos e a medidas de estatística descritiva. Sem perder de vista o objectivo de identificar as variáveis que melhor diferenciam os indivíduos do grupo com queixas daqueles que as não apresentam, é conveniente comparar, para cada variável, os dados obtidos em ambos os grupos. Antes de finalizar, analisam-se os candidatos a *outliers*.

4.1 Análise comparativa dos dois grupos

As caixas de bigodes são diagramas que permitem ter uma ideia da natureza dos valores observados, no que respeita à localização central, aos quartis, aos extremos, à dispersão, à simetria e à existência ou não de candidatos a *outliers*. Na Figura 4.1 estão representadas as caixas de bigodes dos valores observados das 10 variáveis para G0 e G1.

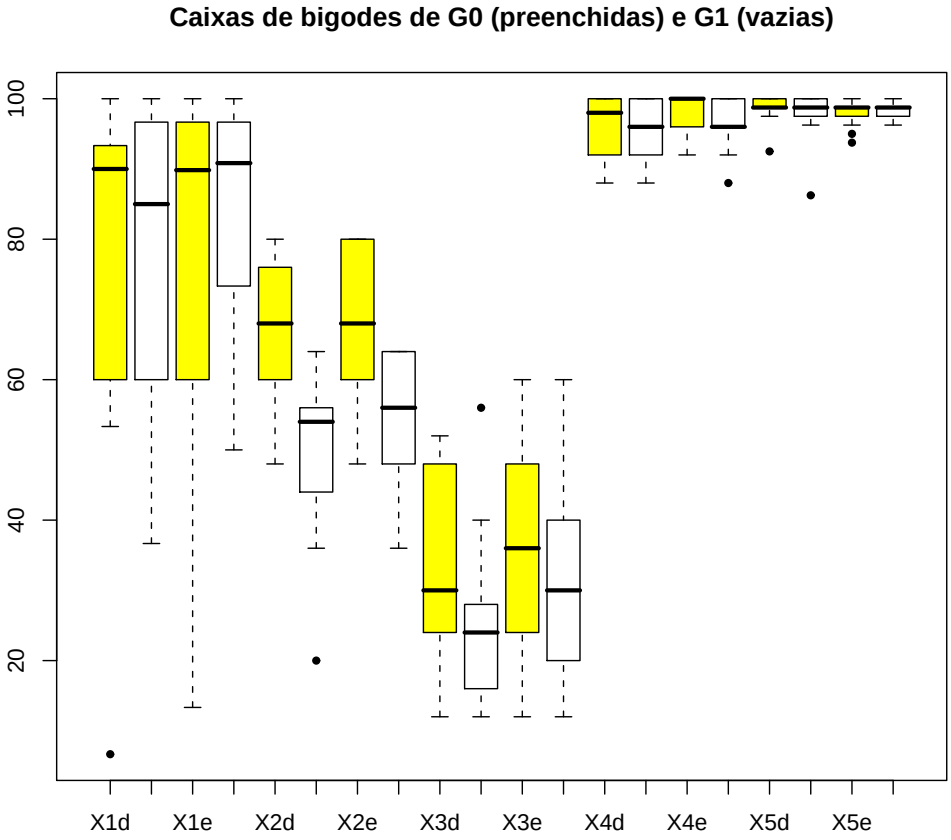

Figura 4.1: Caixas de bigodes das 10 variáveis para os grupos G0 e G1.

Tanto em G0 como em G1, as variáveis apresentam diferenças assinaláveis na gama de valores observados, na dispersão dos dados e no valor da mediana amostral, entre outras. Pese embora a reduzida dimensão da amostra, tais diferenças indiciam que o grau de dificuldade dos testes audiológicos não é semelhante.

O uso da mesma escala para todas as caixas de bigodes permite ter uma ideia global das diferenças entre as variáveis, mas dificulta a análise comparativa dos dois grupos relativamente aos valores observados de cada variável. Esta figura não é a mais indicada para analisar as variáveis representadas nas últimas caixas de bigodes, uma vez que os seus valores

são muito elevados e a amplitude amostral é muito inferior à das restantes. Na Figura 4.2 estão representados os mesmos pares de caixas de bigodes, mas com a escala adaptada a cada variável.

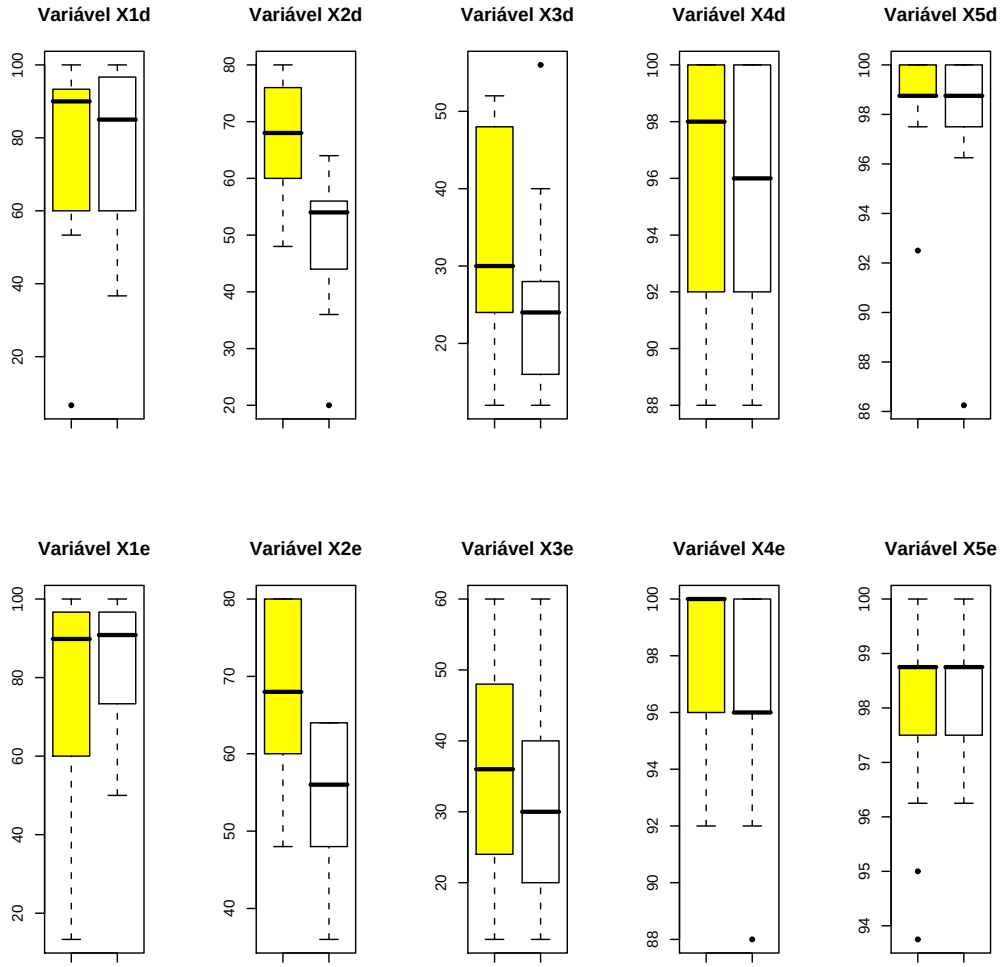

Figura 4.2: Pares de caixas de bigodes de G0 (preenchidas) e G1 (vazias); ouvido direito em cima e ouvido esquerdo em baixo.

Se uma determinada variável contribui significativamente para detectar a característica em estudo, é de esperar que os indivíduos do grupo sem queixas apresentem, regra geral, valores mais elevados. Assim, as variáveis X_{2d}, X_{2e} e X_{3d} parecem revestir-se de particular interesse na diferenciação dos indivíduos de G0 e de G1. As medianas das variáveis X_{3e} e

X_{4d} são superiores para os indivíduos de G0, no entanto as caixas de bigodes não evidenciam diferenças muito acentuadas na distribuição empírica das amostras de ambos os grupos. Os dois últimos pares de caixas de bigodes, referentes ao teste SSW, apresentam valores muito elevados e diferem pouco nos dois grupos, pelo que o referido teste não aparenta ser um bom candidato para atingir os objectivos pretendidos. A variável X_{4e} também apresenta valores muito elevados, no entanto a distribuição dos dados de ambos os grupos tem características claramente distintas. Em G0 pelo menos 50% dos indivíduos responderam correctamente a todas as questões do teste correspondente à variável referida. Por fim, as caixas de bigodes de X_{1d} e X_{1e} não são muito diferentes para os dois grupos, apresentando contudo alguns valores de G0 muito baixos, sendo um deles candidato a *outlier*.

Em G0, o menor valor observado das variáveis X_{1d} e X_{1e} pertence ao elemento 12. O segundo menor valor observado de cada uma dessas variáveis é muito superior ao mínimo, com uma diferença de pelo menos 40 unidades. Na Figura 4.3 estão representadas as caixas de bigodes dos valores observados dessas variáveis, excluindo o elemento 12. Verifica-se que as diferenças entre os dois grupos se mantêm pouco acentuadas, sendo visível o aumento da mediana da variável X_{1e} que passa a ser superior à mediana da mesma variável de G1. Por conseguinte, independentemente do elemento 12, as variáveis X_{1d} e X_{1e} não parecem contribuir muito para a diferenciação dos dois grupos.

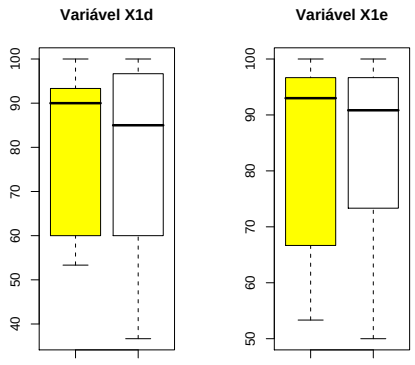

Figura 4.3: Caixas de bigodes de X_{1d} e X_{1e}, sem o elemento 12, para G0 e G1.

Na Tabela 4.1 estão registadas algumas medidas de estatística descritiva das diferentes variáveis, a média amostral, o desvio padrão, a amplitude amostral e o coeficiente de dispersão. Este obtém-se dividindo o desvio padrão pela média e tem a vantagem de não depender da ordem de grandeza das variáveis. A comparação dessas medidas para ambos os grupos visa complementar a análise efectuada com as caixas de bigodes.

	$\bar{x}$		amp. amos.		dp		coef. disp.	
	G 0	G 1	G 0	G 1	G 0	G 1	G 0	G 1
X_{1d}	77.12	79	93.33	63.33	26.16	22.56	0.34	0.29
X_{1e}	78.31	83.17	86.67	50	25.26	18.37	0.32	0.22
X_{2d}	**68**	**49.6**	32	44	10.05	13.49	0.15	0.27
X_{2e}	**67.71**	**54**	32	28	10.58	10.71	0.16	0.20
X_{3d}	**33.43**	**26.4**	40	44	13.73	13.23	0.41	0.50
X_{3e}	36	31.2	48	48	15.92	13.96	0.44	0.45
X_{4d}	96.29	95.6	12	12	4.56	3.98	0.05	0.04
X_{4e}	98	96	8	12	2.60	3.77	0.03	0.04
X_{5d}	98.57	97.62	7.5	13.75	1.95	4.19	0.02	0.04
X_{5e}	97.75	98.5	6.25	3.75	1.87	1.15	0.02	0.01

Tabela 4.1: Medidas de estatística descritiva das 10 variáveis, dos grupos G0 e G1.

O valor médio dos valores observados é muito superior em G0 para as variáveis X_{2d} e X_{2e}, sendo as diferenças de médias de 18.4 e 13.71, respectivamente. Seguem-se as variáveis X_{3d}, X_{3e} e X_{4e}, cujas diferenças de valor médio são de 7.03, 4.8 e 2. Ao analisar as caixas de bigodes, todas estas variáveis foram identificadas como potenciais candidatas à diferenciação dos indivíduos dos dois grupos, com particular relevo para X_{2d}, X_{2e} e X_{3d}. No caso das variáveis X_{4d} e X_{5d} as diferenças dos valores médios de ambos os grupos são inferiores à unidade. O valor médio é superior em G1 para as variáveis X_{5e}, X_{1d} e X_{1e}, sendo as diferenças de médias de G0 e G1 de -0.75, - 1.88 e -4.86, respectivamente.

Em ambos os grupos, as seis primeiras variáveis, correspondentes a três testes auditivos diferentes, são as que apresentam maior variabilidade dos dados. As maiores diferenças do coeficiente de dispersão dos dois grupos registam-se nas variáveis X_{2d}, X_{1e} e X_{3d}. A ampli-

tude amostral e o desvio padrão têm valores mais elevada para as variáveis X_{1d} e X_{1e}, de G0.

Em G0 o elemento 12 exerce grande influência na determinação dos valores das medidas analisadas, no caso das variáveis X_{1d} e X_{1e}. Como se verifica na Tabela 4.2, se o elemento 12 for excluído da análise, a média dos valores das variáveis referidas é superior em G0, contrariamente ao que acontecia quando eram usados todos os elementos. Sem o elemento 12, em G0 a amplitude amostral e o desvio padrão dessas variáveis sofrem uma grande redução.

	$\bar{x}$		amp. amos.		dp		coef. disp.	
	G 0	G 1	G 0	G 1	G 0	G 1	G 0	G 1
X_{1d}	82.54	79	46.67	63.33	17.68	22.56	0.38	0.29
X_{1e}	83.81	83.17	46.67	50	17.67	18.37	0.38	0.22

Tabela 4.2: Medidas de estatística descritiva das variáveis X_{1d} e X_{1e}, sem o elemento 12.

A análise comparativa dos dois grupos evidenciou que existem algumas variáveis que apresentam características que apontam no sentido de melhor fazerem a separação dos dois grupos, são elas X_{2d}, X_{2e} e X_{3d}. Pelo contrário, as variáveis X_{1d}, X_{1e}, X_{5d} e X_{5e} apresentam resultados muito semelhantes nos dois grupos.

4.2 Candidatos a *outliers*

Na Figura 4.2 destacam-se 8 candidatos a *outlier*, sendo alguns referentes ao mesmo elemento. A seguir enumeram-se os elementos que apresentam valores candidatos a *outlier* e as variáveis para as quais tais valores ocorrem: 12 para X_{1d}, 8 para X_{2d}, X_{4e} e X_{5d}, 7 para X_{3d}, 24 para X_{5d}, 3 e 20 para X_{5e}.

Tanto em G0 como em G1, as variáveis X_{5d} e X_{5e} apresentam valores muito elevados. Embora os menores valores observados de G0 para essas variáveis surjam como candidatos a *outliers*, são superiores a 90 não havendo particular interesse em analisá-los.

Elemento 12

O elemento 12 já havia sido referido por apresentar valores muito inferiores aos valores observados dos restantes elementos de G0 para as variáveis X_{1d} e X_{1e}. No grupo com queixas, os valores mínimos observados para essas variáveis são também muito superiores aos registados para o indivíduo 12. Considera-se que há motivo para suspeitar que tenha ocorrido alguma falha durante a aplicação dos testes a este indivíduo, nomeadamente falta de atenção ou outros factores não relacionados com os testes auditivos. Outro aspecto a ter em consta consiste no facto da classificação deste elemento ter sido alterada de G1 para G0. Tendo em vista a clarificação destas dúvidas, faz-se uma breve análise dos valores observados deste indivíduo nas restantes oito variáveis. A Tabela 4.3 sintetiza a comparação desses valores com os valores observados de ambos os grupos.

	X_{2d}	X_{2e}	X_{3d}	X_{3e}	X_{4d}	X_{4e}	X_{5d}	X_{5e}
G0	<Q1	<Q1	max	max	$< \widetilde{x}$	min	max	Q1
G1	Q3	$\widetilde{x}$	<max	<max	$\widetilde{x}$	min	max	<Q1

Tabela 4.3: Comparação entre os valores observados do elemento 12 e de ambos os grupos.

Relativamente ao grupo a que pertence, o elemento 12 apresenta valores baixos nas variáveis X_{2d}, X_{2e}, X_{4e} e X_{5e}, mas é o maior valor observado para as variáveis X_{3d}, X_{2e} e X_{5d}. A presença simultânea de valores elevados e baixos nas oito variáveis não é suficiente para considerar que possa ter ocorrido alguma falha na recolha desses valores. No que se refere à possibilidade da classificação deste elemento estar incorrecta, a comparação com os valores observados para os elementos de G1 não é conclusiva. De facto, para cada uma das oito variáveis em análise, o valor observado do elemento 12 encontra-se dentro da gama de valores observados em G1, sendo maior ou igual à mediana em seis dessas variáveis. Pelo exposto, seria conveniente submeter novamente o elemento 12 à realização dos testes audiológicos e rever a sua classificação. Como tal não é possível e não havendo certezas da validade dos valores observados, é importante fazer uma análise crítica da influência deste elemento nos resultados obtidos ao longo deste trabalho.

Elemento 8

O elemento 8 apresenta sempre valor inferior à média das observações efectuadas nesse grupo, sendo o mínimo da amostra em seis das dez variáveis. Os valores apresentados por este indivíduo são coerentes, não havendo motivo para suspeitar de qualquer falha e é de prever que este elemento seja muito relevante na caracterização dos indivíduos com queixas. Ressalve-se no entanto a possibilidade dessa influência ser excessiva e não corresponder à realidade, dada a reduzida dimensão da amostra.

Elemento 7

O candidato a *outlier* para a variável X_{3d} é o valor observado do elemento 7, o qual apresenta também o maior valor para as variáveis X_{2e} e X_{3e}. Em oito das variáveis em estudo, este elemento apresenta valor superior à média dos valores observados de G1. Quando comparado com os elementos de G0, apresenta um valor superior à média para cinco variáveis, X_{1d}, X_{1e}, X_{3d}, X_{3e} e X_{5e}. No entanto, no caso das variáveis X_{2d}, X_{2e} e X_{4e} os valores deste indivíduo são inferiores à média, sendo as diferenças de 16, 3.7 e 6, respectivamente. Pelas características apresentadas e por ser um dos elementos da amostra que sofreu uma alteração da sua classificação durante a investigação, passando a integrar o grupo com queixas, é conveniente prestar-lhe particular atenção no decurso do estudo a efectuar.

Capítulo 5

Regressão logística aplicada aos dados

Neste capítulo são construidos vários modelos de regressão logística com os dados em análise. Procura-se encontrar um modelo com bom ajuste e com um número reduzido de variáveis.

Foram criadas algumas funções e realizados pequenos programas para facilitar o tratamento dos dados no *R*. A função *logistica* (ver Anexo I) constrói um modelo de regressão logística, recorrendo à função *lrm* pré-definida no *R*, dados o vector dos valores observados da variável resposta, a matriz dos dados das variáveis explanatórias e outros parâmetros genéricos. Da construção de um modelo resultam as estimativas dos coeficientes e dos respectivos erros padrão, as estatísticas de teste dos testes de Wald e da razão de verosimilhanças e os correspondentes valores de p-value, entre outros.

A função *lrm* não devolve directamente os valores preditos, mas os valores da função *logit*, usando a instrução *lin<− modelo$linear.predictors*. Para determinar os valores preditos, usa-se *pred<− 1-1/(1+exp(lin))*.

5.1 Modelo de regressão com todas as variáveis

No sentido de verificar se as 10 variáveis explanatórias permitem discriminar os elementos dos grupos G0 e G1, construiu-se o modelo de regressão logística envolvendo a totalidade das variáveis, modelo *Total*. A aplicação do teste da razão de verosimilhanças (TRV) apenas permite avaliar se os coeficientes são todos nulos contra a hipótese alternativa de pelo menos um ser diferente de zero. Neste caso, G=32.6 e o p-value≈0.0003, havendo motivo para rejeitar a hipótese dos coeficientes serem todos nulos, a um nível de significância de 5%.

Comparando os valores preditos, $\pi(\widehat{x_i})$, com os da variável resposta, verifica-se que todos os elementos da amostra são correctamente classificados, pois o valor predito é inferior a 0.5 para os elementos de G0 e superior a esse valor para os de G1. Mais, os valores de $\pi(\widehat{x_i})$ são muito próximos de 0 ou de 1, como se verifica na Figura 5.1.

O modelo de regressão logística com a totalidade das variáveis permite separar correctamente os elementos dos dois grupos, contudo as estimativas dos erros dos coeficientes estimados são muito elevadas, como se pode constatar na Tabela 5.1. Este é um indicador de sobreajuste do modelo naturalmente relacionado com a reduzida dimensão da amostra face ao número de variáveis em análise.

X_i	X_{1d}	X_{1e}	X_{2d}	X_{2e}	X_{3d}	X_{3e}	X_{4d}	X_{4e}	X_{5d}	X_{5e}
$\widehat{\beta_i}$	0.2092	-0.3110	-1.8771	-0.7071	-2.1154	0.7133	-1.4849	-4.0738	-4.3964	15.4738
$\widehat{SE}(\widehat{\beta_i})$	366.8	315.7	937.8	1073.1	376.3	525.8	923.5	1311.9	2127.1	3036.6

Tabela 5.1: Modelo de regressão logística com as 10 variáveis.

Tal como foi referido no Capítulo 2, não é possível validar alguns pressupostos requeridos pela MANOVA. No entanto, a aplicação deste método permite obter resultados concordantes com os do modelo *Total* da regressão logística. Aplicando a MANOVA a este conjunto de dados, MARTINS (2007) concluiu que havia motivo para rejeitar a hipótese do vector de médias das variáveis em estudo ser igual para os dois grupos.

Na secção seguinte procura-se encontrar um modelo com menos variáveis que permita dife-

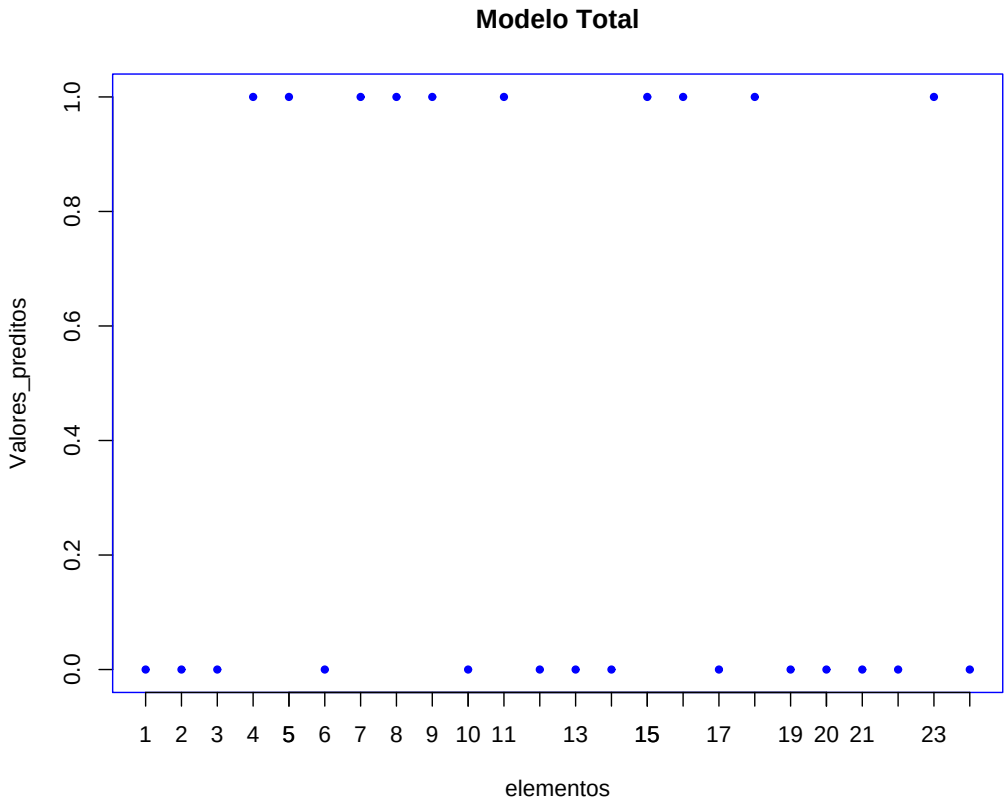

Figura 5.1: Valores preditos do modelo de regressão logística com as 10 variáveis (Todos os elementos são bem classificados).

renciar os indivíduos dos dois grupos sem o efeito de sobreajuste verificado no modelo anterior.

De agora em diante, todos os modelos são designados pelas variáveis explanatórias que os definem, usando o número do teste, de 1 a 5, seguido da letra que identifica o ouvido, *d* ou *e*.

5.2 Selecção de variáveis: método exaustivo

Uma vez que o número de variáveis não é muito elevado, optou-se por aplicar o método exaustivo para construir os vários modelos de regressão no sentido de evitar que algum bom modelo possa não ser analisado. O programa *exaustivo* (Ver Anexo II) foi realizado no *R*

para construir todos os modelos de regressão logística até 4 variáveis[1].

Modelos univariados

Na Tabela 5.2 estão registadas as estimativas dos parâmetros e dos correspondentes erros padrão, bem como as estatísticas de teste e os p-values dos testes de Wald e da razão de verosimilhanças dos modelos de regressão logística univariados.

	$\widehat{\beta}_i$	$\widehat{SE}(\widehat{\beta}_i)$	W	p-value	G	p-value
X_{1d}	0.0034	0.01767	0.19	0.8482	0.0370	0.84750
X_{1e}	0.0106	0.01995	0.53	0.5947	0.2956	0.58665
X_{2d}	-0.1712	0.07298	-2.35	0.0190	12.6683	0.00037
X_{2e}	-0.1355	0.06236	-2.17	0.0298	8.6765	0.00322
X_{3d}	-0.0416	0.03368	-1.24	0.2164	1.6579	0.19788
X_{3e}	-0.227	0.02904	-0.78	0.4342	0.6295	0.42755
X_{4d}	-0.0394	0.09928	-0.40	0.6915	0.1577	0.69131
X_{4e}	-0.2164	0.1506	-1.44	0.1508	2.4057	0.12089
X_{5d}	-0.1089	0.1486	-0.73	0.4637	0.5880	0.44321
X_{5e}	0.2416	0.2899	0.83	0.4046	0.7551	0.38488

Tabela 5.2: Valores dos modelos de regressão logística univariados

De acordo com os valores de p-value dos dois testes, verifica-se que as variáveis que mais contribuem para separar os indivíduos de G0 e G1, quando consideradas individualmente, são X_{2d} e X_{2e}. Ao nível de significância de 5%, há motivo para rejeitar a hipótese $\beta_1 = 0$ nos modelos das referidas variáveis.

A aplicação do TRV para comparar a qualidade de ajuste de dois modelos no *R* é feita à custa da função *lrtest*, cujos parâmetros são os nomes dos modelos a comparar. Usando essa função para comparar o modelo *Total* com os modelos univariados 2_d ou 2_e, resultam os valores aproximados de p-value de 0.0183 e 0.0044. Deste modelo e ao nível de significância de 5%,

[1]Os modelos com mais de 4 variáveis não foram construídos por existirem modelos com menos variáveis com bom ajuste.

há motivo para rejeitar a hipótese dos coeficientes das nove variáveis excluídas serem todos nulos. De facto, no modelo 2_d os valores preditos dos elementos 10, 12, 15 e 16 não estão de acordo com a sua classificação. O modelo 2_e, classifica mal os elementos já referidos, o 5 e o 7.

No que se refere aos valores dos coeficientes da variável explanatória, é de salientar que nos modelos 1_d, 1_e e 5_e esse valor é positivo. Da interpretação da estimativa do coeficiente da variável explanatória de um modelo de regressão logística univariado, resulta que a probabilidade da presença da característica em estudo aumenta com o aumento do valor da variável explanatória. Contudo, no âmbito deste estudo, é de esperar que os indivíduos que não tenham dificuldades de discriminação auditiva em ambiente de ruído apresentem um desempenho superior aos restantes nos testes que contribuem para separar os dois grupos.

Se os modelos univariados forem construídos sem ter em conta o elemento 12, a estimativa do coeficiente da variável explanatória do modelos 1_d e 1_e é um valor negativo, Tabela 5.3. Em ambos os casos, os valores de p-value permanecem muito elevados, donde se depreende que tais variáveis, quando consideradas individualmente, não contribuem significativamente para a separação dos dois modelos.

	$\widehat{\beta}_i$	$\widehat{SE}(\widehat{\beta}_i)$	W	p-value	G	p-value
X_{1d}	-0.0099	0.0223	-0.44	0.6572	0.1977	0.6566
X_{1e}	-0.0005	0.0245	-0.02	0.9845	0.0004	0.9846

Tabela 5.3: Valores dos modelos 1_d e 1_e de G0, sem o elemento 12

Martins (2007) aplicou o teste não paramétrico U de Mann-Whitney para avaliar a importância das variáveis na separação dos dois grupos. A um nível de significância de 5%, apenas o Teste de Fala no Ruído revelou a presença de diferenças nos dois grupos que contribuem para discriminar os seus elementos. Neste estudo, as variáveis que correspondem a esse teste são X_{2d} e X_{2e}. Os modelos de regressão logística univariados destas variáveis foram os únicos a apresentar p-value< 0.05 no teste de Wald.

Modelos com duas variáveis

Os modelos bivariados com melhor ajuste são 2_d5_e e 2_d3_d e os valores de p-value resultantes da aplicação do TRV para comparar a qualidade do seu ajuste com a do modelo *Total* são aproximadamente iguais a 0.124 e 0.080, respectivamente. Ao nível de significância de 5%, não há motivo para rejeitar a hipótese dos parâmetros excluídos do modelo *Total* serem nulos. Ao contrário do que acontecia com o modelo com as 10 variáveis, as estimativas dos erros padrão dos coeficientes estimados destes modelos são pequenos, como se verifica na Tabela 5.4.

	$\widehat{\beta}_i$	$\widehat{SE}(\widehat{\beta}_i)$
X_{2d}	-0.3263	0.1717
X_{3d}	-0.1419	0.0810
X_{2d}	-0.2231	0.1468
X_{2e}	0.0586	0.1328

Tabela 5.4: Estimativas dos coeficientes e dos respectivos erros padrão dos modelos 2_d3_d e 2_d3_e

Apesar dos valores de p-value do TRV serem superiores a 0.05 para os dois modelos, não são muito elevados, pelo que é necessário ter alguma cautela ao assumir que a qualidade do ajuste é semelhante à do modelo *Total*. Da análise dos valores preditos dos dois modelos verifica-se que 2_d5_e e 2_d3_d classificam incorrectamente os elementos $\{4, 10, 13, 16\}$ e $\{7, 10, 13\}$, respectivamente. Na Figura 5.2 estão representados os valores preditos dos modelos 2_d3_d e 2_d5_e, verificando-se que existem valores afastados de 0 e de 1.

Embora a comparação da qualidade de ajuste dos modelos de duas variáveis com o modelo *Total* aponte no sentido de 2_d5_e ser o modelo de duas variáveis com melhor ajuste, parece ser mais conveniente considerar o modelo 2_d3_d. O modelo 2_d5_e, para além de classificar incorrectamente mais um elemento do que 2_d3_d, utiliza a variável X_{5e} que individualmente não contribui muito para a separação dos dois grupos e que tem coeficiente positivo no modelo univariado. Já o modelo 2_d3_d é construído à custa de duas variáveis que estão entre as que mais contribuem para a separação dos elementos dos dois grupos e que apresentam coeficiente negativo nos modelos univariados.

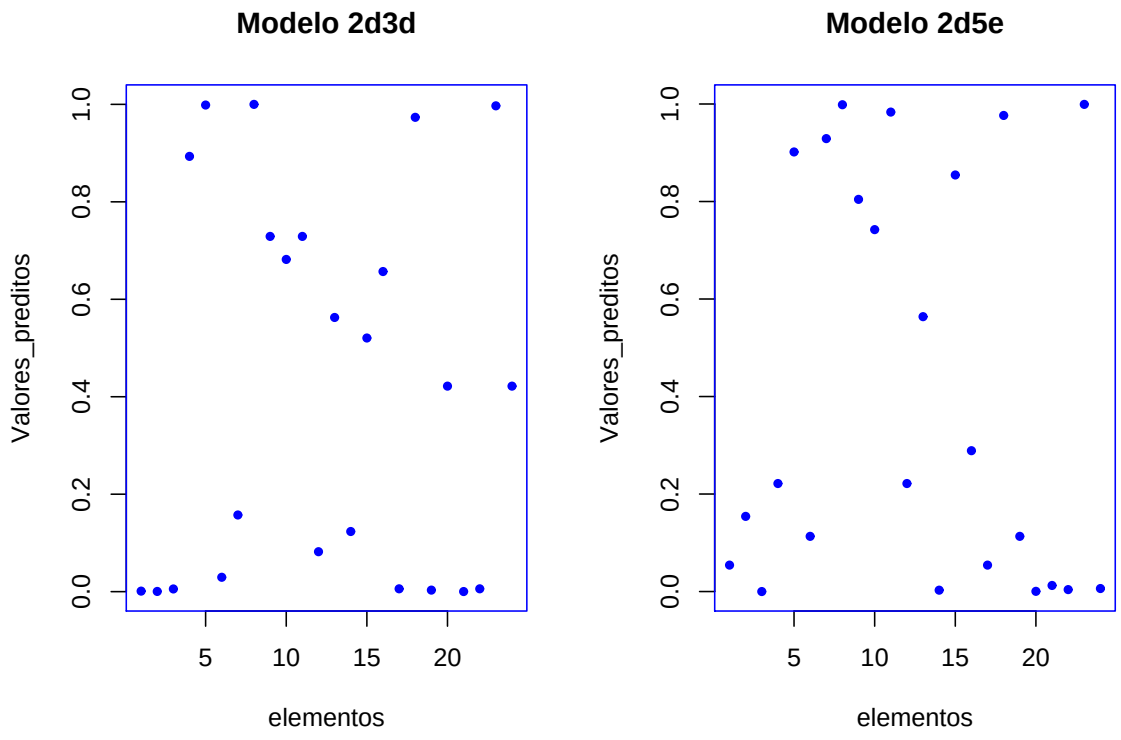

Figura 5.2: Valores preditos dos modelos 2_d3_d (classifica mal os elementos $\{7, 10, 13\}$) e 2_d5_e (classifica mal os elementos $\{4, 10, 13, 16\}$).

Modelos com três variáveis

Os modelos $2_d3_d4_e$ e $2_d2_e3_d$ classificam correctamente todos os indivíduos e têm valores preditos próximos de 0 ou 1, como registado na Figura 5.3. O p-value resultante da aplicação do TRV para comparar a qualidade de ajuste de cada um destes modelos com a do modelo *Total* é aproximadamente igual a 1. O terceiro modelo no que respeita à qualidade de ajuste é o $2_d3_e4_e$, o qual classifica incorrectamente o indivíduo 15 e apresenta valores preditos próximos de 0.5 para os elementos 7, 12, 13 e 15, de acordo com a Figura 5.4.

As estimativas dos coeficientes das variáveis explanatórias dos modelos de três variáveis com melhor ajuste e dos correspondentes erros padrão estão registadas na Tabela 5.5. Em ambos os modelos as estimativas dos erros são muito elevadas, de modo análogo ao que se verificou no modelo *Total*. Regista-se ainda o aumento, em valor absoluto, das estimativas

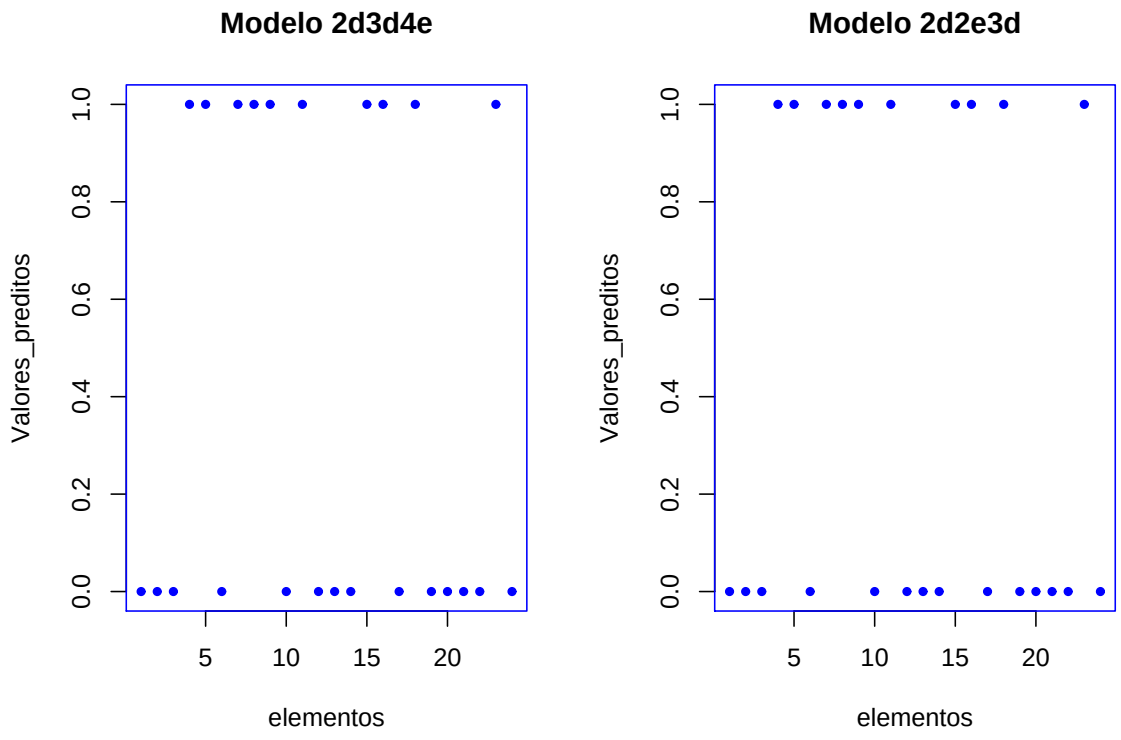

Figura 5.3: Valores preditos dos modelos $2_d3_d4_e$ e $2_d2_e3_d$ (ambos classificam bem todos os elementos).

dos coeficientes, comparativamente aos modelos de uma ou duas variáveis. Estes resultados não significam que as variáveis presentes em cada um dos modelos não sejam adequados para resolver o problema proposto, uma vez que o sobreajuste pode ser devido à reduzida dimensão da amostra. No modelo $2_d2_e3_d$ a estimativa do coeficiente da variável X_{2e} é positivo, podendo tal ser devido a factores como a interacção ou a colinearidade de variáveis.

Ao aplicar o TRV para comparar a qualidade de ajuste dos modelos $2_d3_d4_e$ e 2_d3_d, registou-se um p-value$\approx 1.760e-04$, havendo motivo para rejeitar a hipótese do coeficiente da variável X_{4e} ser nulo.

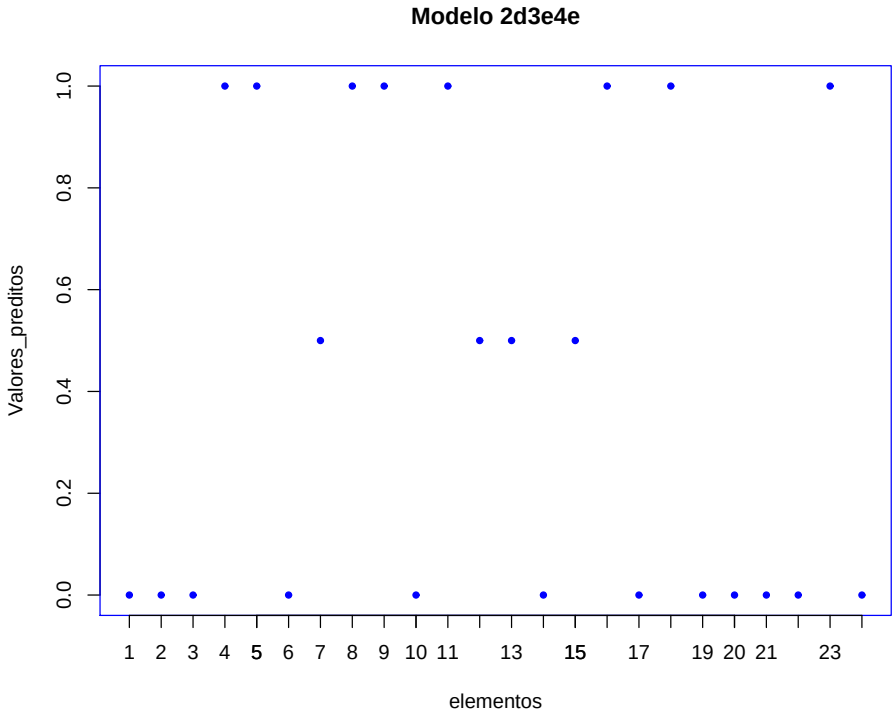

Figura 5.4: Valores preditos do modelo $2_d3_e4_e$ (O elemento 15 é mal classificado).

	$\widehat{\beta}$	$\widehat{SE(\beta)}$
X_{2d}	-20.73	691.9
X_{3d}	-13.75	455.2
X_{4e}	-37.67	1275.3
X_{2d}	-53.79	1445.2
X_{2e}	27.88	750.0
X_{3d}	-16.10	431.2

Tabela 5.5: Estimativas dos coeficientes e dos respectivos erros padrão dos modelos $2_d3_d4_e$ e 2_d2_e3d

Modelos com quatro variáveis

Vários são os modelos de 4 variáveis que classificam correctamente todos os elementos e que apresentam valores preditos próximos de 0 ou 1. No entanto, todos apresentam estimativas dos erros padrão dos coeficientes estimados muito elevadas.

Apesar do sobreajuste registado nos modelos de 4 variáveis, importa fazer algumas considerações. Se a selecção de variáveis fosse realizada com recurso ao teste univariado de Wald, o modelo multivariado seria construído à custa das variáveis X_{2d}, X_{2e}, X_{3d} e X_{4e}, por serem aquelas cujos modelos univariados têm p-value<0.25. No entanto, existem 11 modelos de 4 variáveis com melhor ajuste do que $2_d2_e3_d4_e$. Este é um factor que reforça a importância da escolha do método empregue para seleccionar as variáveis a incluir num modelo multivariado.

5.3 Análise de resíduos

A análise dos valores dos resíduos permite identificar possíveis elementos para os quais se verifique um maior afastamento entre o valor predito e o valor da variável resposta. A existir, tais elementos têm maior influência na definição do modelo do que os restantes, sobretudo em amostras de reduzida dimensão. À medida que a dimensão da amostra aumenta, a influência de observações individuais diminui, o que se traduz na construção de um modelo mais estável e que reflecte mais fielmente as características da população.

A avaliação da qualidade de ajuste de um modelo de regressão deve inclui a análise de medidas sumárias dos valores dos resíduos, por recurso a testes estatísticos. Contudo, no presente estudo optou-se por não aplicar testes de análise de resíduos devido à reduzida dimensão da amostra e ao facto dos dados serem esparsos.

Na exposição teórica da análise de resíduos de um modelo de regressão logística, fez-se referência à possibilidade de existirem vários elementos da amostra que apresentem os mesmos valores para o conjunto das variáveis em análise. Os resíduos seriam calculados apenas para o número de observações diferentes e considerava-se, para cada caso, o número de vezes que esse conjunto de valores havia sido observado. Uma vez que estamos perante uma amostra de dimensão reduzida e que as variáveis explanatórias podem tomar vários valores diferentes, as repetições de observações são raras. Neste caso, não há necessidade de agrupar os elementos e os resíduos são apresentados para os 24 elementos, tal como determinado no *R*. Assim, considera-se $m_j = 1$ nas expressões dos resíduos.

Se $m_j = 1$, a expressão dos *deviance residuals*, apresentada em (3.8), pode ser escrita do seguinte modo:

$$d_j = \begin{cases} -\sqrt{2|\ln(1-\widehat{\pi}_j)|}\ \text{, se } y_j = 0 \\ \sqrt{2|\ln(\widehat{\pi}_j)|}\ \text{, se } y_j = 1 \end{cases}$$

Os *deviance residuals* dos modelos *Total* e $2_d3_d4_e$ encontram-se representados nas Figuras 5.5 e 5.6. Nos dois casos, alguns elementos destacam-se por apresentarem valores de resíduos mais afastados de zero do que os restantes. No modelo *Total* os valores dos resíduos dos elementos $\{13, 15, 16, 17\}$ têm valores absolutos maiores do que 0.001. O mesmo se verifica para os elementos $\{7, 12, 13, 15\}$ no modelo $2_d3_d4_e$.

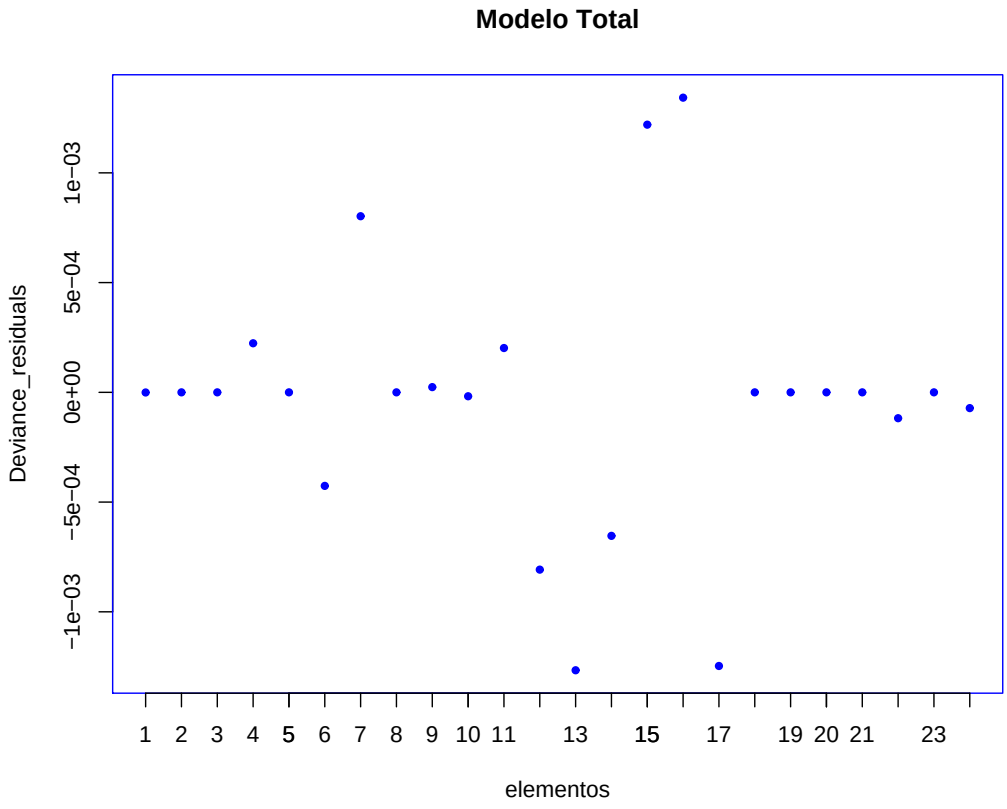

Figura 5.5: *Deviance residuals* do modelo *Total*.

Na secção anterior verificou-se que modelo 2_d3_d apresentava alguns valores preditos afastados de 0 e de 1 e classificava incorrectamente três elementos. Tal como seria de esperar, este

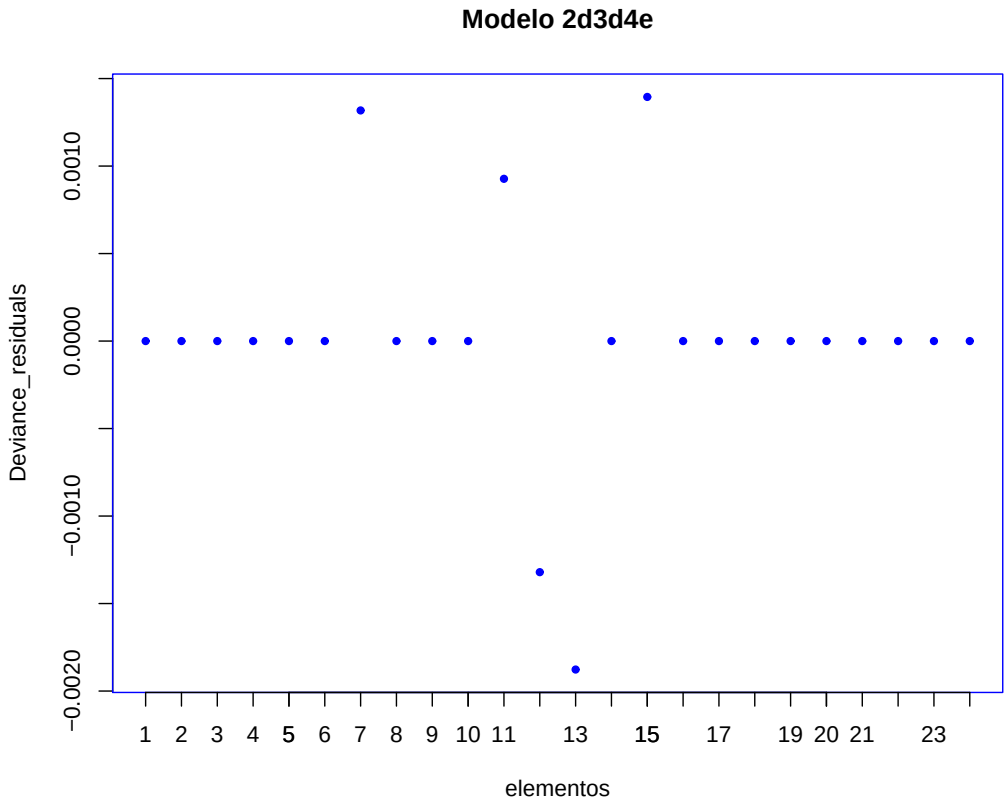

Figura 5.6: *Deviance residuals* do modelo $2_d3_d4_e$.

modelo apresenta valores de resíduos mais afastados de zero do que os modelos *Total* e $2_d3_d4_e$, como se verifica na Figura 5.7. Os elementos $\{7, 10, 13, 15, 20, 24\}$ têm valores absolutos de resíduos superiores a 1.

5.4 Síntese

Ao longo deste capítulo foram construídos vários modelos de regressão logística com diferentes números de variáveis. Se por um lado, o maior número de variáveis permite melhores resultados na separação dos elementos dos dois grupos, por outro lado, a redução do número de variáveis traduz-se na diminuição do sobreajuste dos modelos.

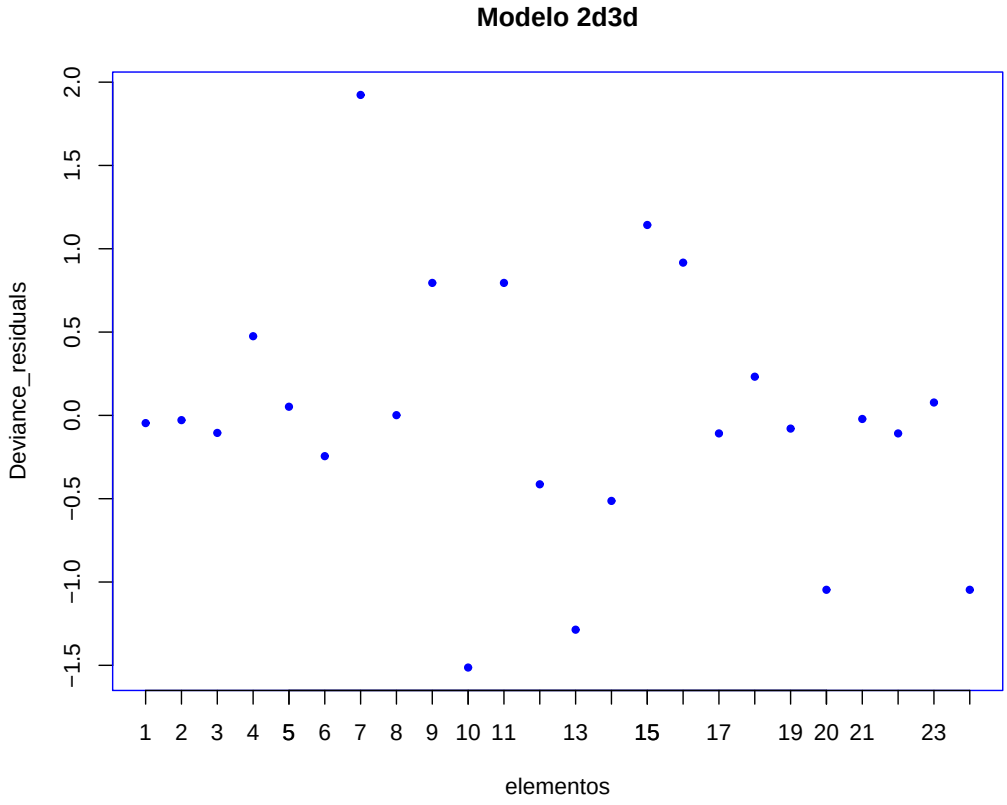

Figura 5.7: Deviance residuals do modelo 2_d3_d.

Principais resultados deste capítulo:

- há motivo para acreditar que os testes audiológicos permitem separar os indivíduos dos dois grupos em estudo;
- para esta amostra, as variáveis X_{2d}, X_{3d} e X_{4e} são suficientes para discriminar os elementos dos dois grupos;
- nos modelos com mais de duas variáveis existem evidências de sobreajuste;
- nenhum modelo com menos de três variáveis classifica correctamente todos os elementos;
- de entre os modelos que não evidenciam sobreajuste, aquele que classifica correctamente mais elementos da amostra em estudo é o 2_d3_d.

Os modelos que classificam bem todos os elementos são demasiado dependentes dos dados utilizados, e têm estimativas dos erros padrão muito elevadas. Por outro lado, os modelos para os quais essas estimativas são baixas não apresentam uma qualidade de ajuste satisfatória. De acordo com estes resultados e sem dispor de novos dados, não é conveniente fazer uma escolha definitiva do modelo mais adequado. Contudo, é de esperar que as variáveis X_{2d}, X_{3d} e X_{4e}, ou apenas X_{2d} e X_{3d} permitam construir um bom modelo de regressão logística para resolver o problema do diagnóstico da dificuldade de discriminação auditiva em ambiente de ruído.

Capítulo 6

Análise crítica e discussão dos resultados

A escolha do modelo a utilizar num determinado problema, deve incluir a avaliação e comparação da performance de diferentes modelos construídos. Quando não de dispõe de novos dados para aplicar e validar modelos previamente seleccionados, como acontece neste estudo, é necessário recorrer a elementos da amostra. A primeira secção deste capítulo é dedicada à aplicação e avaliação de alguns modelos, no que se refere à capacidade preditiva de novos elementos.

A análise realizada nos capítulos anteriores e a escolha dos modelos que melhor respondem às necessidades deste estudo incidiram sobre a totalidade dos elementos da amostra. Contudo, não há garantias de que os modelos seleccionados fossem os mesmos caso não se considerassem todos os elementos da amostra. Na segunda secção deste capítulo é apresentada uma breve análise crítica da selecção de variáveis para a construção dos modelos multivariados.

Outro factor que pode afectar tanto a selecção das variáveis como a construção dos modelos é a possibilidade de existirem elementos mal classificados. Embora com este trabalho não seja possível validar a classificação dos elementos, é importante identificar aqueles cuja classificação possa levantar dúvidas. A última secção incide sobre a classificação de cada elemento e a sua contribuição para a construção dos modelos.

Neste capítulo a avaliação da qualidade de ajuste é realizada por aplicação do teste da razão de verosimilhanças (TRV) para comparar a qualidade de ajuste de dois modelos, sendo um deles o modelo *Total*. Os valores de p-value mencionados referem-se a este teste, salvo menção em contrário.

6.1 Aplicação e validação dos modelos

Quando se pretende avaliar a performance de um modelo que visa a classificação de elementos é conveniente experimentá-lo com novas observações. Quando tal não é possível mas a dimensão da amostra é grande, é usual dividi-la em dois conjuntos, um para construir os modelos e o outro para os aplicar e avaliar. Neste estudo não é conveniente dividir a amostra, uma vez que os elementos em análise são poucos, tendo-se mesmo verificado sobreajuste de alguns modelos. A solução consiste em seleccionar os modelos usando a totalidade dos elementos e só depois proceder à aplicação e avaliação dos mesmos. Segundo Witten (2005), a performance preditiva de um modelo num conjunto de elementos que contribuíram para a sua construção não é um bom indicador da performance num novo conjunto de dados. Assim, os modelos seleccionados devem ser construídos de novo sem ter em conta o(s) elemento(s) escolhidos para avaliar a sua performance preditiva.

No capítulo anterior foram identificados dois modelos de regressão logística, $2_d3_d4_e$ e 2_d3_d, potencialmente apropriados para diagnosticar a dificuldade de discriminação auditiva em ambiente de ruído. Uma vez que a selecção desses modelos se baseou na comparação da qualidade de ajuste com o modelo *Total*, faz-se a comparação dos três modelos no que concerne à capacidade preditiva de novos elementos. Essa comparação é efectuada pela taxa de erro da classificação dos elementos.

Validação cruzada

O método de validação cruzada envolve a selecção de elementos da amostra para aplicar e avaliar os modelos, usando os restantes na construção dos mesmos. Esta é uma técnica muito usada na avaliação de modelos quando os dados são escassos e pode ser realizada aplicando vários procedimentos, nomeadamente *holdout*, *k-fold* e *leve-one-out*.

O procedimento *holdout* consiste em reservar uma parte da amostra, frequentemente um terço, para fazer a avaliação dos modelos. Essa escolha pode ser completamente aleatória ou estratificada, garantindo que cada grupo seja representado de modo análogo ao que se verifica na amostra. No método *k-fold* divide-se a amostra em k subconjuntos, estratificados ou não, reservando um desses grupos para avaliar o modelo. Este procedimento é repetido k vezes, de modo que todos os conjuntos sejam usados uma vez para fazer a avaliação. Um caso particular deste método é o designado *leve-one-out*, considerando k igual à dimensão da amostra. Na prática, retira-se apenas um elemento da amostra para a validação do modelo, repetindo-se o procedimento para todos os elementos.

Ao dividir a amostra em dois subconjuntos, é importante decidir qual a dimensão que cada um deve ter. Neste estudo, deve evitar-se reduzir o número de elementos usados para construir os modelos para não agravar o problema de sobreajuste. O procedimento *leve-one-out* é o que melhor satisfaz este requisito.

Metodologia bootstrap

Para além da validação cruzada, pode também ser usada a metodologia *bootstrap*, que consiste na amostragem aleatória com reposição dos elementos da amostra. Os elementos seleccionados, incluindo as repetições, são usados para construir os modelos e os restantes são reservados para a avaliação. A construção dos modelos será sempre baseada num número de elementos igual à dimensão da amostra, mas a dimensão do conjunto de avaliação é variável. Apesar deste método não reduzir a dimensão da amostra, a diversidade de dados é menor por não contemplar a informação dos elementos que não são seleccionados.

Método *leave-one-out* aplicado aos modelos

Optou-se por usar o método *leve-one-out* para comparar os modelos *Total*, $2_d3_d4_e$ e 2_d3_d, por ser aquele que garante a maior variabilidade possível de dados na construção dos modelos. A taxa de erro registada para o modelo *Total* foi de 25%, uma vez que classificou mal os elementos $\{6, 13, 17\}$ de G0 e $\{7, 15, 16\}$ de G1. O modelo $2_d3_d4_e$ classificou mal os elementos 13 de G0 e $\{7, 15\}$ de G1, registando a taxa de erro mais baixa, 12.5%. No caso

do modelo 2_d3_d a taxa de erro foi aproximadamente igual a 16.7%, com os elementos $\{6, 13\}$ de G0 e $\{7, 15\}$ de G1 a serem mal classificados. A maior taxa de erro do modelo *Total* pode dever-se ao facto de incluir um grande número de variáveis, aumentando a dependência face aos elementos usados para o construir.

Tal como referido anteriormente, o modelo $2_d3_d4_e$ apresenta melhor qualidade de ajuste do que o modelo 2_d3_d. Agora verificou-se que a taxa de erro do primeiro modelo é inferior à do segundo. No entanto, essa diferença não é muito grande e o modelo 2_d3_d tem a vantagem de não evidenciar sobreajuste.

Há ainda a considerar a possibilidade de alguns elementos estarem mal classificados, tendo as taxas de erro sido calculadas assumindo a correcta classificação de todos os elementos. Por exemplo, os elementos 7, 13 e 15 foram mal classificados pelos três modelos e todos eles apresentavam valores de resíduos afastados de zero em vários modelos. A selecção definitiva de um ou outro modelo deve ser feita em estudos posteriores, incluindo a análise de novos elementos.

6.2 Selecção de variáveis

Ao longo deste trabalho, foram destacados alguns elementos por apresentarem valores atípicos, por haver suspeita de estarem mal classificados ou por terem valores de resíduos mais afastados de zero do que os restantes. Dada a reduzida dimensão da amostra, a influência de cada observação na construção dos modelos de é maior do que o desejável, sobretudo nos modelos com mais variáveis. Um único elemento mal classificado ou cujos valores observados se destaquem dos restantes pode determinar a selecção ou eliminação de variáveis e levar à construção de modelos com estimativas dos coeficientes pouco realistas.

A forma mais eficaz de analisar a influência que um elemento exerce na selecção de variáveis e na construção dos modelos consiste em refazer o estudo do capítulo anterior sem esse elemento e comparar os resultados. Para esse efeito, foi aplicado o método exaustivo para construir os modelos de regressão logística considerando todos os subconjuntos de 23 elementos da amostra. Embora se retire um elemento da amostra antes de construir os modelos,

este procedimento difere do *leave-one-out*, uma vez que o que se pretende não é classificar o elemento excluído, mas analisar a selecção de variáveis.

Modelos univariados

No que se refere aos modelos univariados, constatou-se que em 21 dos casos analisados as quatro variáveis mais importantes para a diferenciação dos dois grupos são as mesmas e mantém a mesma ordem: X_{2d}, X_{2e}, X_{4e} e X_{3d}. Noutro caso, as variáveis são as mesmas mas as duas últimas trocam de posições. Se não for considerando o elemento 7 para a construção dos modelos univariados, as variáveis mais importantes são X_{2d}, X_{2e}, X_{3d} e X_{3e}. Se não for considerado o elemento 8, apenas os modelos com uma das variáveis X_{2d}, X_{2e} e X_{5e} têm p-value inferior a 0.25 no teste de Wall.

Modelos bivariados

Na grande maioria, os resultados da construção dos modelos bivariados diferem pouco daqueles que haviam sido obtidos com a totalidade dos elementos. Recorde-se que nenhum modelo, construído com a totalidade da amostra e sem alterar a classificação de nenhum elemento, classificava correctamente todos os elementos e que aqueles que tinham maior valor de p-value no TRV eram 2_d5_e e 2_d3_d. Foi ainda justificado ser mais vantajoso considerar o modelo 2_d3_d.

Apenas quando não foi considerado o elemento 7 foram registadas diferenças assinaláveis na construção dos modelos de duas variáveis. Neste caso, os modelos 2_d3_d e 2_e5_e classificam bem todos os elementos, têm p-value$\approx$1 no TRV e valores preditos próximos de 0 ou 1. Contudo, as estimativas dos coeficientes das variáveis explanatórias do modelo 2_d3_d sofrem grandes alterações e as estimativas dos erros são muito elevadas, como se pode ver na Tabela 6.1. Os valores dessas estimativas passam a ser semelhantes aos obtidos para as mesmas variáveis no modelo $2_d3_d4_e$, considerando todas as variáveis, Tabela 5.5.

Modelos com três variáveis

De entre os modelos de três variáveis construídos com todos os elementos, aquele que apresentava melhor ajuste era o $2_d3_d4_e$. Após retirar um qualquer elemento, o modelo com

	$\widehat{\beta}$	$\widehat{SE(\beta)}$
X_{2d}	-20.48	744.3
X_{3d}	-13.62	496.8

Tabela 6.1: Estimativas dos coeficientes e dos erros padrão do modelo 2_d3_d, sem o elemento 7.

essas variáveis tem sempre p-value≈1 no TRV, classifica correctamente os 23 elementos e tem valores preditos próximos de 0 ou 1. Em todos os casos mantém-se o problema de sobreajuste deste modelo. Apenas quando se retira um dos elementos 7, 12 ou 13, existem outros modelos de três variáveis com melhor ajuste do que $2_d3_d4_e$. No entanto, os valores de p-value são aproximadamente iguais a 1 para todos os modelos, verificando-se a diferença da qualidade de ajuste por ligeiras variações da estatística de teste.

O indivíduo 12 apresenta valores atípicos nalgumas variáveis, nomeadamente X_{1d} e X_{1e}, tendo sido considerada a possibilidade de tais valores resultarem da influência de factores não relacionados com o âmbito deste estudo. Ao efectuar a modelação dos dados sem ter em conta este indivíduo, nenhum modelo de duas variáveis classifica bem todos os elementos, mas existem 5 modelos de três variáveis com bom ajuste. No entanto, em nenhum destes modelos surge qualquer das variáveis X_{1d} e X_{1e}. Portanto, não são os hipotéticos erros nos valores observados deste indivíduo que condicionam a presença dessas variáveis nos modelos de três variáveis com bom ajuste.

Síntese

De um modo geral, os resultado apresentados nesta secção apoiam a análise efectuada no capítulo anterior no que se refere à contribuição individual de cada variável para a separação dos dois grupos e na selecção de variáveis para os modelos de duas e de três variáveis.

As características do modelo $2_d3_d4_e$ construído depois de retirar cada um dos elementos, são semelhantes às do modelo com as mesmas variáveis mas construído com os 24 elementos. O mesmo acontece com o modelo 2_d3_d, excepto quando se retirou o elemento 7. Neste caso, o modelo classifica bem todos os elementos e as estimativas dos coeficientes e dos respectivos erros padrão sofreram grandes alterações, evidenciando sobreajuste.

6.3 Classificação e influência dos elementos

Na primeira secção deste capítulo, com a aplicação do método *leave-one-out* verificou-se que os elementos 7, 13 e 15 são mal classificados pelos modelos *Total*, $2_d3_d4_e$ e 2_d3_d. Embora tal possa dever-se apenas a falhas na capacidade preditiva desses modelos, é importante ponderar a hipótese de haver erro na classificação dos mesmos. O elemento 12 tem sido também destacado ao longo deste estudo por apresentar valores muito baixos nalgumas variáveis. Os quatro elementos referidos são também os que apresentam valores de resíduos mais elevados no modelo $2_d3_d4_e$.

Elemento 7

Os resultados da modelação dos elementos da amostra sem o indivíduo 7 são inesperados, pois existem 2 modelos com 2 variáveis, 2_d3_d e 2_e3_d, e 16 com 3 que classificam correctamente todos os indivíduos. Todos estes modelos classificam o elemento 7 como pertencente a G0, sendo ele de G1. Analisando os modelos de 4 variáveis, verifica-se que existem 59 com p-value≈ 1, dos quais 50 alocam o indivíduo 7 em G0. Os restantes 9 modelos apresentam valores preditos para o elemento 7 próximos de 1. Em 8 destes modelos está presente pelo menos uma das variáveis X_{1d}, X_{1e} ou X_{5e}, que apresentavam coeficiente positivo nos modelos univariados. Há portanto fortes indícios de que este elemento possa estar mal classificado.

Elemento 12

Os valores preditos do elemento 12, de G0, obtidos com vários modelos não são consensuais. Os modelos *Total*, $2_d3_d4_e$, $2_d2_e3_d$ alocam este elemento em G0. Mas existem três modelos de três variáveis com bom ajuste que o alocam em G1. Analisando os modelos de quatro variáveis, verifica-se que existem 34 com bom ajuste, dos quais 19 alocam o elemento 12 em G1 e os restantes 15 em G0. É aconselhável rever a classificação deste indivíduo e submetê-lo novamente à realização dos testes auditivos para despistar eventuais erros nos valores observados.

Elementos 13

Quando se excluiu o elemento 13 da construção dos modelos, foram definidos 4 modelos de 3 variáveis e 35 de 4 variáveis com bom ajuste. O modelo *Total*, os 4 modelos de 3 variáveis

e 23 modelos de 4 variáveis com bom ajuste, alocam o elemento 13 em G1. Os restantes 12 modelos de 4 variáveis alocam esse elemento em G0. Apesar desta análise não ser conclusiva, há motivo para suspeitar que este elemento esteja mal classificado.

Elementos 15

Quando este elemento não é considerado para a construção dos modelos, existem 3 modelos de 3 variáveis e 28 de quatro variáveis com bom ajuste. Tanto o modelo *Total* como os modelos de três variáveis com bom ajuste alocam este elemento em G0, sendo ele de G1. Relativamente aos modelos de 4 variáveis com bom ajuste, 9 alocam o elemento em G0 e os restantes 19 em G1. Apesar dos modelos de 3 variáveis classificarem mal este elemento, a maioria dos modelos de quatro variáveis com bom ajuste classificam-no bem. Desta forma, não há motivo para considerar que o elemento 15 esteja mal classificado.

6.3.1 Perturbar a classificação dos elementos

Para analisar a influência da classificação de um elemento na construção dos modelos, procedeu-se à alteração da sua classificação antes de construir os modelos. Aplicou-se o método exaustivo para definir os modelos de regressão logística com os 24 elementos tendo um deles a classificação alterada. Este procedimento foi repetido para todos os elementos da amostra. Procurou-se analisar o efeito da perturbação da classificação de cada elemento na construção dos modelos, nomeadamente no que se refere ao número mínimo de variáveis necessárias para construir um modelo que classifique bem todos os elementos.

Em todos os casos foi possível construir modelos com bom ajuste, variando o número mínimo de variáveis necessárias para o conseguir, a quantidade de modelos nessas condições e as variáveis envolvidas.

A alteração da classificação de um elemento dos conjuntos $\{1, 2, 21, 22\}$ de G0 ou $\{8, 23\}$ de G1 condiciona visivelmente a construção dos modelos. Nestes casos não foi possível construir nenhum modelo de 4 variáveis com bom ajuste. Estes deverão ser elementos que apresentam uma performance elevada ou baixa na realização dos testes auditivos, consoante sejam ele-

mentos de G0 ou G1, respectivamente.

Após alterar a classificação de um dos elementos dos conjuntos $\{3, 10, 14, 17, 19, 20, 24\}$ de G0 e $\{4, 5, 9, 18\}$ de G1, o número mínimo de variáveis necessárias para construir um modelo com bom ajuste é 4. Nesses modelos surgem com alguma frequência as variáveis que tinham coeficiente positivo nos modelos univariados. Estes 11 elementos também exercem grande influência na construção dos modelos de regressão.

Os resultados apresentados para os 17 elementos referidos apoiam a sua classificação e apontam no sentido de serem elementos que facilmente se diferenciam daqueles que não pertencem ao mesmo grupo.

A alteração da classificação do elemento 7 também influencia fortemente a construção dos modelos, mas de modo contrário ao que se verificou para os elementos já mencionados. Neste caso, existem 2 modelos de duas variáveis e 16 modelos de três variáveis com bom ajuste. Mais uma vez surgem indícios de que este elemento possa estar mal classificado, uma vez que a alteração da sua classificação se traduz numa melhoria da qualidade de ajuste de vários modelos.

No caso dos restantes elementos, $\{6, 12, 13\}$ de G0 e $\{11, 15, 16\}$ de G1, as alterações foram menos significativas. Em todos estes casos foi possível construir modelos de três variáveis com bom ajuste e não existem modelos de duas variáveis que classifiquem bem todos os elementos. A classificação destes elementos não condiciona muito a construção dos modelos, mas tal não significa que estejam mal classificados. De acordo com estes resultados, há motivo para suspeitar que os elementos em causa apresentem valores próximos de elementos do grupo ao qual não pertencem.

Capítulo 7

Conclusões

A necessidade de dispor de meios eficazes de diagnóstico de um problema específico de processamento audiológico, motivou a criação de uma bateria de testes auditivos adaptados ao português europeu. Esse conjunto de testes foi aplicada a um conjunto reduzido de pessoas com o intuito de diagnosticar uma patologia que se manifesta na dificuldade de discriminação auditiva em ambiente de ruído. O objectivo primordial consiste em detectar a presença ou ausência dessa patologia em função dos resultados obtidos nos testes auditivos.

A aplicação do conjunto de testes referido, sendo um meio de diagnóstico da presença ou ausência de uma determinada característica, insere-se no domínio dos problemas de classificação ou separação de elementos em dois conjuntos disjuntos e de alocação de novos elementos de acordo com o desempenho nesses testes. O grupo constitui a variável resposta e as 10 variáveis explanatórias (5 testes aplicados aos dois ouvidos) expressam-se pela percentagem de respostas correctas no teste correspondente.

Neste trabalho fez-se uma análise estatística com vista a avaliar a eficácia da bateria de testes na classificação dos indivíduos, a reduzir o número de variáveis necessárias para efectuar essa separação e a definir um modelo adequado para alocar novos elementos. Inicialmente foram analisadas várias metodologias de estatística multivariada com o intuito de seleccionar as mais pertinentes para alcançar os objectivos definidos. A regressão logística foi o método de eleição para modelar o problema da separação de indivíduos e da alocação de novos elementos, mas antes procedeu-se a uma análise exploratória dos dados. Esta é uma metodologia

genérica que permitiu identificar, no contexto do problema em estudo, semelhanças e diferenças entre as variáveis, detectar valores atípicos e apoiar ou confrontar os resultados da regressão logística.

De acordo com a análise efectuada, há motivo para acreditar que os testes auditivos permitem diagnosticar correctamente o problema de processamento auditivo em análise. Com três ou mais variáveis foi possível construir vários modelos de regressão logística que classificam bem todos os elementos. Considerando apenas os dados em análise, verificou-se que é possível criar um modelo de 3 variáveis, $2_d3_d4_e$, que permite separar os dois grupos de indivíduos tão bem quanto o modelo de 10 variáveis. Os testes auditivos envolvidos na construção desse modelo são o Teste de Fala no Ruído (ouvido direito), o Teste de Fala Filtrada (ouvido direito) e o Teste de Fusão Binaural (ouvido esquerdo).

Todos os modelos que classificam bem a totalidade dos elementos estão afectados de um problema de sobreajuste que se deve à reduzida dimensão da amostra. As estimativas dos erros padrão dos coeficientes estimados desses modelos são muito elevadas. Deste modo e apesar da boa qualidade de ajuste de alguns modelos, a sua aplicação para a alocação de novos elementos não é recomendada.

Existem modelos com duas variáveis que não apresentam o problema de sobreajuste, mas nenhum deles classifica correctamente todos os elementos da amostra. No melhor caso, três elementos são mal classificados, usando para tal o modelo 2_d3_d cujas variáveis correspondem ao Teste de Fala no Ruído (ouvido direito) e ao Teste de Fala Filtrada (ouvido direito).

No sentido de comparar a performance preditiva de dois modelos de regressão logística seleccionados como candidatos a substituir o modelo com as 10 variáveis, $2_d3_d4_e$ e 2_d3_d, foi aplicado o método *leave-one-out* de validação cruzada. Dos 24 elementos em análise, foram classificados incorrectamente 3 com o primeiro modelo e 4 com o segundo, obtendo-se uma taxa de erro de 0.125 e 0.17, respectivamente. Em ambos os casos a taxa de erro é inferior à do modelo *Total* que se situa nos 0.25. Este aumento da taxa de erro ilustra a grande dependência deste modelo face aos elementos usados na sua construção, uma vez que usa um

número elevado de variáveis.

No capítulo anterior procurou-se fazer uma análise crítica da influência de cada elemento na definição dos modelos de regressão logística. Foram identificados alguns elementos que se diferenciam facilmente dos elementos do grupo ao qual não pertencem e que têm uma maior influência na construção dos modelos. Outros elementos são menos influentes na definição dos modelos e apresentam valores mais próximos de alguns elementos do outro grupo. Destacam-se os elementos 7 e 12, o primeiro por existirem fortes indícios da sua classificação estar incorrecta e o segundo por haver suspeita dos valores de algumas variáveis não corresponderem à realidade.

O elemento 7 exerce uma grande influência na construção dos modelos, evidenciada nas alterações decorridas da construção dos modelos sem a sua participação ou com a sua classificação alterada. Nestes casos, contrariamente ao que acontecia quando eram considerados todos os elementos e sem alterar a classificação de nenhum deles, o modelo 2_d3_d classifica bem todos os elementos e apresenta evidências de sobreajuste.

Tudo indica que a bateria de testes permite diagnosticar a patologia em estudo neste trabalho e que é possível fazê-lo usando apenas duas ou três variáveis. A regressão logística revelou ser adequada para modelar o problema em questão e alocar novos elementos. Contudo, devido a limitações relacionadas quer com a reduzida dimensão da amostra quer com a possibilidade de existirem elementos mal classificados, persistem algumas questões sem resposta. No sentido de ultrapassar estas limitações apresentam-se algumas sugestões para trabalhos futuros:

- aplicar a bateria de testes a um conjunto mais vasto de indivíduos e aplicar de novo a regressão logística;
- submeter novamente alguns elementos à realização dos testes auditivos, nomeadamente aqueles cuja classificação é dúbia;
- aplicar os testes a indivíduos com alterações neurológicas documentadas, como sugerido por Martins (2007);

- aplicar a análise discriminante, caso os pressupostos sejam válidos, ao novo conjunto de dados e confrontar com os resultados da regressão logística;
- usar outras metodologias de aplicação e validação dos modelos, nomeadamente o método *holdout* de validação cruzada e a metodologia *bootstrap*, descritos por Witten (2005);
- avaliar o modelo escolhido com elementos que não tenham contribuido para seleccionar as variáveis nem para construir os modelos.

Bibliografia

[1] HAIR, Joseph F., ANDERSON, Rolph E., TATHAM, Ronald L. e BLACK, William C. (1998), *Multivariate Data Analysis*, 5ª ed. Prentice Hall, Upper Saddle River (NJ).

[2] HOSMER, David W. e LEMESHOW, Stanley (1989), *Applied logistic regression*, John Wiley, New-York.

[3] HOSMER, David .W., HOSMER, T., LE CESSIE, S. e LEMESHOW, S. (1997), *A Comparison of goodness-of-fit tests for the logistic regression model*, Statistics in Medicine, 16: 965-980.

[4] HUBERTY, Carl J. e OLEJNIK, Stephen (2006), *Applied MANOVA and discriminant analysis* 2ª ed. Wiley Interscience.

[5] JOHNSON, Richard A. e WICHERN, Dean W. (1997), *Applied multivariate statistical analysis*, 4ª ed. Prentice Hall, Englewood Cliffs (NJ).

[6] KING, Jason E. (2003), *Running a Best-Subsets Logistic Regression: An alternative to Stepwise Methods*, Educational and Psychological Measurement, 63: 392-403.

[7] KUSS, Oliver (2002) *Global goodness-of-fit tests in logistic regression with sparse data*, Statistics in Medicine, 21: 3789-3801.

[8] MANNING, Christopher (2007), *Logistic Regression (with R)*, Apontamentos de Quantitative and Probabilistic Explanation in Linguistics, The Stanford NLP Group, Stanford University.

[9] MARTINS, Elsa M. C. (2007), *Criação de um conjunto de testes para avaliação do processamento auditivo*, Dissertação de Mestrado, Universidade de Aveiro.

[10] POHAR, Maja, BLAS, Mateja e TURK, Sandra (2004), *Comparison of de Logistic Regression and Linear Discriminant Analysis: A Simulation Study*, Metodoloski Zvezki, 1: 143-161.

[11] REIS, Elizabeth (1997), *Estatística Multivariada Aplicada*, Edições Sílabo, Lisboa.

[12] R Development Core Team (2008), *R Language Definition*, Versão 2.6.2, http://cran.r-project.org/

[13] SHARMA, Subhash (1996), *Applied Multivariate Techniques*, John Wiley, New York.

[14] VENABLES, W. N., SMITH, D. M. et. al (2008), *An Introdution to R*, Version 2.6.2, http://cran.r-project.org/

[15] WHITAKER, Jean S. (1997), *Use of Stepwise Methodology in Discriminant Analysis*, annual meeting of the Southwest Educational Research Association, Austin.

[16] WITTEN, Ian H. e FRANK, Eibe (2005), *Data Mining: Practical machine learning tools and techniques*, 2ª ed. Morgan Kaufmann, San Francisco.

Apêndice A

ANEXOS

ANEXO I - Função para criar um modelo de regressão logística

```
# packages necessarios para a regressão logistica.
library(rpart); library(nlme); library(Design);

logistica< −function(var1,vars2,a,b,c) {
    lrm(formula=var1 vars2,tol=a, eps=b,maxit=c,model=TRUE,
    x=TRUE,y=TRUE, linear.predictors=TRUE,se.fit=TRUE);
}

    # var1 e' o vector daos valores da variavel resposta
    # vars2 e' a matriz dos valores das variaveis explanatorias a incluir no modelo;
    # a, b, c são os valores dos parametros tol (criterio de singularidade),
    # eps (criterio de convergencia) e maxit (máximo de iterações) da função lrm.
# restantes parametros do medelo de regressao logistica:
    # x=TRUE e y=TRUE ppara armazenar os valores das variaveis explanatorias
    # e da variavel resposta
    # linear.predictors=TRUE e se.fit=TRUE armazenam os valores da funcao
    #logit e dos erros padrao dos valores preditos
```

ANEXO II - Modelos de regressão logística com um máximo de 4 variáveis

```
# Valores mais usuais de tol, eps e maxit (parâmetros da função logistica)
a1 < −1e − 15; b1 < −0.0001; c1 < −100;
# vR e vP são os valores observados da variável resposta
# e das variaveis explanatorias, respectivamente;
# Definir todos os modelos com duas variáveis p/ identificar os melhores.
# var2test armazena os valores da estat. G e o indice das variaveis
var2test< −NULL
for (i in 1:9)
for (j in (i+1):10) {
    var1< −vP[,i]
    var2< −vP[,j]
    modelo< −logistica(vR,cbind(var1,var2),a1,b1,c1)
    var2test< −rbind(var2test,c(modelo$deviance[2],i,j))
    }
# Ordenar os modelos por ordem crescente da estatistica G;
# Corresponde à ordem decrescente de qualidade de ajuste
var2test< −ordena(var2test,dim(var2test)[1])


# Definir todos os modelos com três variáveis e identificar os melhores.
var3test< −NULL
for (i in 1:8)
for (j in (i+1):9)
for (k in (j+1):10) {
    var1< −vP[,i]
    var2< −vP[,j]
    var3< −vP[,k]
    modelo< −logistica(vR,cbind(var1,var2,var3),a1,b1,c1)
    var3test< −rbind(var3test,c(modelo$deviance[2],i,j,k))
    }
# Ordenar os modelos por ordem decrescente da qualidade de ajuste
```

```
var3test< −ordena(var3test,dim(var3test)[1])

# Definir todos os modelos com quatro variáveis e identificar os melhores.
var4test< −NULL;
for (i in 1:7)
for (j in (i+1):8)
for (k in (j+1):9)
for (l in (k+1):10) {
    var1< −vP[,i]
    var2< −vP[,j]
    var3< −vP[,k]
    var4< −vP[,l]
    modelo< −logistica(vR,cbind(var1,var2,var3,var4),a1,b1,c1);
    var4test< −rbind(var4test,c(modelo$deviance[2],i,j,k,l))
    }
var4test< −ordena(var4test,dim(var4test)[1])
remove(var,var1,var2,var3,var4,i,j,k,l)
```