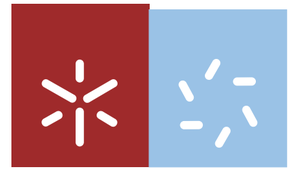


Universidade do Minho
Escola de Ciências

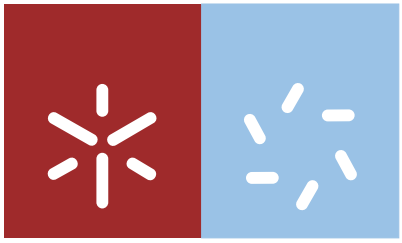
Anabela Costa da Silva

Análise Estatística de Inquéritos *online*

Outubro de 2011

Anabela Costa da Silva **Análise Estatística de Inquéritos** *online*

UMinho | 2011



Universidade do Minho
Escola de Ciências

Anabela Costa da Silva

Análise Estatística de Inquéritos *online*

Relatório de Mestrado
Mestrado em Estatística de Sistemas
- Especialização em Engenharia e Estatística

Trabalho efetuado sob a orientação da
Professora Doutora Ana Cristina Braga

Outubro de 2011

“ Quando recebemos um ensinamento devemos receber como um valioso presente e não como uma dura tarefa. Eis aqui a diferença que transcende”

Alberto Einstein

AGRADECIMENTOS

Durante a realização deste trabalho deparei-me com momentos bons e momentos menos bons, e foi nesses momentos onde as coisas não corriam como esperava que contei com o apoio das pessoas fantásticas que me rodeiam no quotidiano.

Em primeiro lugar cabe-me agradecer a professora Doutora Ana Cristina Braga, pela sua compreensão, disponibilidade, atenção, paciência e força transmitida, assim como todas as dúvidas esclarecidas, e as sugestões dadas. A qualidade deste trabalho deve-se muito a forma como fui orientada.

Agradeço aos meus pais por todo o esforço, sacrifícios e apoio dado ao longo de todo o meu percurso académico.

Um obrigado especial à Empresa GESTA pela oportunidade do estágio, que me possibilitou demonstrar conhecimentos e capacidades adquiridas ao longo deste Mestrado. Agradecendo a compreensão e disponibilidade do orientador externo professor Doutor Paulo Sampaio.

De uma forma geral agradeço a todos os meus familiares que estiveram ao meu lado e acreditaram nas minhas capacidades, e todos os meus amigos que me apoiaram, agradecendo em especial ao meu namorado Bruno pela paciência, compreensão e amizade que disponibilizou sempre e incondicionalmente.

LISTA DE ABREVIATURAS

LME: Lesões Músculo Esqueléticas;

LMELT: Lesões Músculo Esqueléticas Ligadas ao Trabalho;

GESTA: Grupo de Estatística Aplicada;

QWEB: Certificação de Processos de Negócio Electrónico;

JOCALD 2011: XVIII Jornadas de Classificação e Análise de Dados;

ENBIS-11: 11th Annual Conference of European Network for Business and Industrial Statistics;

SNS: Serviço Nacional de Saúde;

SPSS: *Statistical Package for the Social Sciences;*

M.M.V: Método de Máxima Verosimilhança;

ROC: Receiver Operating Characteristic;

RESUMO

As Lesões Músculo Esqueléticas (LME) associadas ao trabalho repetitivo constituem actualmente grande preocupação na generalidade dos países industrializados (Carneiro, 2005).

Com o objectivo de avaliar o risco de Lesões Músculo Esqueléticas Ligadas ao Trabalho (LMELT) nos enfermeiros, elaborou-se um questionário de forma a recolher a opinião de profissionais de enfermagem, que exercem a sua profissão em centros de saúde da região Norte, no que concerne à sintomatologia de Lesões Músculo Esqueléticas associadas ao desempenho das respectivas funções.

A taxa de resposta ao questionário foi de aproximadamente 4,87%, ou seja, entre os 3017 profissionais da área obteve-se 147 respostas completas. Desses 125 prestam apoio ao domicílio.

Após uma análise exploratória das variáveis, verificou-se que a região lombar seria uma região que podia estar relacionada com as LME dos enfermeiros, quando estes prestam apoio ao domicílio. Neste sentido foram calculados os valores de razão de possibilidades para um conjunto de nove regiões (região cervical, ombros, cotovelos, punho/ mão, dorsal, lombar coxas, joelhos e tornozelos).

Apontada a zona lombar como a região com maior evidência para relacionar este tipo de queixas com alguns factores associados com estes profissionais, construiu-se modelos de regressão logística. Neste sentido, filtrando os enfermeiros que prestam apoio ao domicílio, considerou-se como variável dependente "dor lombar" e como variáveis independentes um conjunto de 54 variáveis relacionadas com queixas ou métodos de desempenho do trabalho. (variáveis apresentadas na tabela A1- anexos).

Este processo realizou-se recorrendo a técnicas de selecção de variáveis implementadas no SPSS, selecção passo a passo progressiva ou regressiva (*Forward* e *Backward)*.

Construídos os modelos apresentaram-se as curvas ROC com o objectivo de avaliar em termos de desempenho qual dos modelos exibia um melhor ajuste no que diz respeito à explicação da dor na região lombar nos enfermeiros que prestam apoio ao domicílio. Comparando-se ainda as áreas abaixo da curva ROC para os modelos obtidos (comparação dois a dois), recorrendo-se a região crítica *z*, defendida por *Hanley* e *McNeil* (1983).

ABSTRACT

Musculoskeletal disorders (MSDs) associated with repetitive work is currently great concern in most industrialized countries (Carneiro, 2005).

In order to assess the risk of work related musculoskeletal disorders (WRMSDs) on nurses, who provide home care are prepared a questionnaire to gather the opinion of nurses, who exercise their profession in health centers in the North, regarding the symptoms of musculoskeletal injuries associated with the performance of their duties.

The questionnaire response rate was approximately 4.87%, ie between 3017 professionals we obtained 147 complete responses, of these 125 provide support at home.

After an exploratory analysis of the variables, it was found that the lumbar region would be one of the regions that could be related to MSDs of the nurses, when they provide support at home. In this sense we calculated the odds ratio to a set of nine regions (neck, shoulders, elbows, wrist/ hand, dorsal, lumbar, thighs, knees and ankles).

Considered as the lumbar region with the highest evidence for these complaints relate to some factors associated with these professionals, we constructed logistic regression models. In this sense, by filtering nurses who provide home care, it was considered as the dependent variable complaints in lumbar and a set of 54 variables as independent variables related to complaints or methods of work performance. (variables shown in table A1- Annexes).

This process was carried out using variable selection techniques implemented in SPSS- Forward and Backward stepwise methods.

Built the models, the ROC curves are presented in order to evaluate performance in terms of which of the models showed a better fit with respect to the explanation of lumbar complaints in nurses who provide home support. Finally the areas under the ROC curve obtained for the models are compared (pairwise), using the critical region of z statistic , proposed by Hanley and McNeil (1983).

CONTEÚDOS

AGRADECIMENTOS **..... iii**

LISTA DE ABREVIATURAS **.....v**

RESUMO **.....vii**

ABSTRACT **.....ix**

CAPÍTULO 1- INTRODUÇÃO **.....1**

1.1. LESÕES MÚSCULO ESQUELÉTICAS 1

1.1.1 Definição 1

1.1.2. Lesões Músculo Esqueléticas ligadas ao trabalho 1

1.1.3. Factores de risco 3

1.2. A EMPRESA.....4

1.3. ENQUADRAMENTO E MOTIVAÇÃO 5

1.4. OBJECTIVOS 6

1.5. ESTRUTURA DA TESE.....6

CAPÍTULO 2- ENQUADRAMENTO TEÓRICO **.....7**

2.1. ANÁLISE EXPLORATÓRIA DOS DADOS.....7

2.1.1. Escalas de atitudes e opiniões: 8

2.1.2. Tipo de dados 8

2.2. SOFTWARE UTILIZADO.....9

2.3. REGRESSÃO LOGÍSTICA.....10

2.3.1. Enquadramento 10

2.3.2. Modelo de regressão logística univariado 11

2.3.2.1. Função de verosimilhança 14

2.3.2.2. Teste de *Wald* 15

2.3.3. Modelo de regressão logística multivariado 16

2.3.3.1. Testes de significância estatística 17

2.3.3.2. Métodos de selecção de variáveis 18

2.3.3.2.1. Selecção automática 21

2.3.4. Razão de possibilidades (*odds ratio*) 22
2.3.5. Avaliar o ajuste do modelo 24

2.4. CURVA ROC 26

2.4.1. Perspectiva Histórica 26
2.4.2. Conceitos básicos 27
2.4.3. Gráfico da curva ROC 28
2.4.4. Área abaixo da curva ROC 29
2.4.5. Comparação de modelos com recurso ao teste da área abaixo da curva ROC 30

***CAPÍTULO 3- ANÁLISE DE DADOS* 33**

3.1. ANÁLISE PRELIMINAR DOS DADOS 33

3.2.RAZÃO DE POSSIBILIDADES 35

3.3. MODELOS DE REGRESSÃO LOGÍSTICA PARA A REGIÃO LOMBAR 37

3.4. ANÁLISE DOS RESIDUOS ATRAVÉS DA CURVA ROC 42

3.4.1. Representação da Curva ROC 42
3.4.2. Comparação de modelos com recurso ao teste da área abaixo da curva ROC 43

***CAPÍTULO 4- CONCLUSÃO & TRABALHOS FUTUROS* 49**

4.1. CONCLUSÕES 49

4.2. SUGESTÕES PARA TRABALHOS FUTUROS 50

***ANEXOS* 53**

***BIBLIOGRAFIA* 55**

Páginas da Internet (consultadas no período de Janeiro a Outubro de 2011) 56

ÍNDICE DOS GRÁFICOS:

Gráfico 1- *Curva ROC, para uma dada capacidade de discriminação, com a variação do critério de decisão* 28
Gráfico 2- *Distribuição da percentagem de enfermeiros que prestam apoio ao domicílio segmentado por sexo* 34
Gráfico 3- *Queixas nas regiões corporais por parte dos enfermeiros que prestam apoio ao domicílio (em %)* 34
Gráfico 4- *Queixas nas regiões corporais por parte dos enfermeiros que não prestam apoio ao domicílio (em %)* 35
Gráfico 5- *Curva ROC para os quatro modelos* 43
Gráfico 6- *Curvas ROC dos modelos dois a dois* 45

ÍNDICE DE TABELAS:

Tabela 1-*Razão de possibilidade* 23
Tabela 2- *Representação geral de um teste diagnóstico* 27
Tabela 3- *Valores das estimativas das razões de possibilidade e respectivos I.C.(95%)* 36
Tabela 4- *Valores relevantes da aplicação de regressão logística univariada.* 39
Tabela 5- *Resultados curva ROC* 43
Tabela 6- *Valores de z relativos a comparações de áreas* 46
Tabela 7- *Valores prova associados à região crítica z* 46
Tabela 8: Codificação das variáveis explicativas do modelo 48

CAPÍTULO 1- INTRODUÇÃO

1.1. LESÕES MÚSCULO ESQUELÉTICAS

1.1.1 Definição

Lesões Músculo Esqueléticas (LME) são um conjunto amplo e diversificado de patologias, que se sobrepõem, na sua maioria, às doenças reumáticas, mas que diferem destas por incluírem algumas situações de lesões osteoarticulares e de bolsas sinoviais e por apresentarem, na sua origem, factores de risco de natureza ocupacional (Carneiro, 2005).

Há provas evidentes que em certos factores relacionados com o trabalho estão associados ao elevado risco dos trabalhadores apresentarem lesões no seu sistema músculo-esquelético (Carneiro, 2005).

As Lesões Músculo Esqueléticas associadas ao trabalho repetitivo constituem actualmente grande preocupação na generalidade dos países industrializados. Trata-se efectivamente da expressão de uma hiper-solicitação das articulações, decorrentes de dois tipos de factores, que, associados à sensibilidade individual, condicionam a probabilidade de aparecimento destas patologias (Carneiro, 2005).

Podem evidenciar-se algumas causas de LME de natureza biomecânica, tais como a força e a postura em que os movimentos são realizados, e factores psicossociais associados à representação individual das condições de trabalho e que se traduzem por insatisfação, fadiga e *stress* (Direcção Geral de Saúde, 2008).

1.1.2. Lesões Músculo Esqueléticas ligadas ao trabalho

As Lesões Músculo Esqueléticas ligadas ao trabalho (LMELT) foram, ao longo das últimas décadas, referidas como as alterações de saúde mais frequentes relacionadas com diversos contextos de trabalho (Direcção Geral de Saúde, 2008).

Em Inglaterra segundo *Health & Safety Commission* em 1990, as LMELT são mesmo consideradas as mais frequentes doenças relacionadas como o trabalho, nos Estados Unidos, são consideradas por alguns autores *Maland (1993)* e *Muggleton (1999*) como uma possível epidemia do século XXI (http://www.ensp.unl.pt).

As actividades profissionais com postos de trabalho envolvendo diversos factores de risco de LMELT são muito numerosas, abrangendo designadamente actividades com tarefas repetitivas, aplicações de força e posturas articulares extremas (ou desconfortáveis). Tais características, associadas a outros factores de risco de natureza individual, constituem elementos da matriz etiológica das LMELT, ainda que não sejam bem conhecidas as respectivas relações exposição efeito defendido por *Who* em 1995 referido pela Direcção Geral de Saúde (2008).

Com base nas publicações da Direcção Geral de Saúde (2008) são indicados como factores para o desenvolvimento das LMELT seguintes aspectos:

(1) Uma actividade realizada fundamentalmente por gestos repetitivos que implicam a necessidade de adopção de posições angulares extremas dos membros;

(2) Esforços excessivos;

(3) Elevada receptividade.

Estas lesões afectam principalmente a região dorso-lombar, a zona cervical, os ombros e os membros superiores, mas podem afectar também os membros inferiores. Algumas lesões músculo-esqueléticas, tais como a síndrome do canal cárpico, que afecta o pulso, são lesões específicas que se caracterizam por sinais e sintomas bem definidos. Outras manifestam-se unicamente por dor ou desconforto, sem que existam sinais de uma lesão clara e específica.

A identificação e avaliação dos factores de risco existentes nos postos de trabalho, responsáveis pelo desenvolvimento deste tipo de lesões, e a sua redução através da implementação de medidas de prevenção, deve constituir uma das preocupações dos empregadores, tendo em conta a preservação da saúde e segurança dos seus trabalhadores (Direcção Geral de Saúde, 2008).

1.1.3. Factores de risco

Apoiado no Programa Nacional Contra as doenças reumáticas da Direcção de Saúde (2008), sabe-se que as causas das LMELT são várias, ainda que a "sobrecarga" a nível dos tendões, dos músculos das articulações e dos nervos constitua um importante factor de risco. Essa "sobrecarga" é composta por vários elementos:

(1) Relacionados com a actividade de trabalho;

(2) Individuais, também chamados *co- factores* de risco;

(3) Organizacionais/psicossociais, que, embora sejam igualmente factores de risco profissional, são frequentemente abordados separadamente.

Como factores de risco associados ao trabalho considera-se algo que possa provocar um efeito adverso, como por exemplo as tendinites. A exposição ao factor de risco pode causar doença ou lesão, dependendo de um conjunto de factores adicionais.

Neste sentido, tem-se:

(A) Factores de risco relacionados com a actividade de trabalho:

- Posturas ou posições corporais extremas;
- Aplicação de força;
- Repetitividade;
- Exposições a elementos mecânicos.

(B) Factores de risco individuais:

- Idade;
- Sexo;
- Altura, peso e outras características antropométricas;
- Situação de saúde.

(C) Factores de risco organizacionais/ psicossociais:

- Riscos intensos de trabalho;
- Monotonia das tarefas;
- Insuficiente suporte social;

- Modelo organizacional de produção (Direcção Geral de Saúde, 2008).

1.2. A EMPRESA

No âmbito do Mestrado de Estatística de Sistemas, o estágio decorreu num período de 7 meses no Grupo de Estatística Aplicada (GESTA).
O GESTA pertence ao Departamento de Produção e Sistemas da Universidade do Minho, estando inserido nos *spin-offs* da TecMinho.

A TecMinho é uma associação de direito privado sem fins lucrativos, criada em 1990. Tendo como promotores a Universidade do Minho e a Associação dos Municípios do Vale do Ave e como missão fundamental fazer uma ligação da universidade com a sociedade, contribuindo assim, para o desenvolvimento regional através da melhoria de competitividade das organizações e aumento das competências dos indivíduos (www.tecminho.uminho.pt).

Tendo em conta a política de valorização do conhecimento, a Universidade do Minho incentiva a criação de empresas que visem valorizar o conhecimento resultante das suas actividades de investigação científica e tecnológica (www.tecminho.uminho.pt).

O termo *spin-off* deriva do inglês e utiliza-se para descrever uma nova empresa que nasceu a partir de um grupo de pesquisa, universidade ou centro de pesquisa (público ou privado), tendo normalmente como objectivo explorar um novo produto ou serviço de alta tecnologia. É comum que as *spin-offs* se estabeleçam em incubadoras de empresas ou áreas de concentração de empresas de alta tecnologia
(www.empreendedorismo.uac.pt/spinofss/spinoffs_academicos).

O estatuto de *spin-off* é concebido a projectos com vínculo a departamentos ou centros de investigação que visem criar empresas aptas para valorizarem resultados de investigação gerados no decurso de actividades científicas conduzidas pela sociedade académica, tais como os investigadores, os bolseiros de investigação ou alunos de pós-graduação (www.empreendedorismo.uac.pt/spinofss/spinoffs_academicos).

Integrado no projecto GESTA, foi realizado um trabalho referente à QWEB (Certificação de Processos de Negócios Electrónicos), que teve como objectivo a análise estatística de variáveis provenientes de um inquérito *online*. Esta análise englobou principalmente a parte exploratória e descritiva. Resultou deste trabalho a elaboração de um poster para a JOCALD 2011 (XVIII Jornadas de Classificação e Análise de Dados).

Com o decorrer do trabalho realizou-se um resumo e um *poster* para apresentação na *ENBIS-11* (11th *Annual Conference of the European Network for Business and Industrial Statistics*), onde se focou as LME e um modelo de regressão logística que explicasse a dor na lombar nos profissionais de enfermagem que prestam apoio ao domicílio, assim como a respectiva curva ROC.

1.3. ENQUADRAMENTO E MOTIVAÇÃO

Conhecendo os problemas associados as LME e tendo como base um questionário *online* (desenvolvido por uma aluna de doutoramento do Departamento de Produção e Sistemas), construído com vista a avaliar o risco de LMELT nos profissionais de enfermagem que exercem a sua profissão em centros de saúde da região Norte, no que concerne à sintomatologia de LME associadas ao desempenho das respectivas funções, fez-se uma análise das variáveis do questionário.

Assim, este trabalho focou-se na análise estatística, descritiva e inferencial, das variáveis de forma a encontrar um modelo estatístico mais pertinente e que se melhor se ajustasse a

compreender quais as queixas e factores que poderiam estar relacionadas com as LME nos enfermeiros que prestam apoio ao domicílio.

1.4. OBJECTIVOS

Para desenvolvimento deste trabalho delinearam-se os seguintes objectivos gerais:

1. Análise inicial dos dados;
2. Elaboração de gráficos relevantes ao estudo;
3. Calculo dos valores de razão de possibilidades para as várias sintomatologias;
4. Aplicação de regressão logística:
   (a) Selecção de variáveis candidatas ao modelo;
   (b) Obtenção de modelos explicativos;
5. Analisar os resíduos através da Curva ROC;
6. Comparação dos modelos obtidos.

1.5. ESTRUTURA DA TESE

Este trabalho desenvolveu-se ao longo de 4 capítulos. Os objectivos apresentados na secção anterior traduzem, de uma forma parcial, a forma como o trabalho foi estruturado.

Seguido da introdução, apresenta-se uma parte teórica dividida em 4 subcapítulos incluindo a análise exploratória dos dados, uma breve apresentação do *software* utilizado, a fundamentação da regressão logística, finalizando com a parte da análise dos resíduos com recurso à curva ROC.

No capítulo 3, apresentam-se os resultados obtidos após análise de dados com recurso ao *SPSS*.

Seguindo-se o capítulo 4, referente à análise e discussão de resultados, onde são apresentadas algumas considerações sobre o trabalho realizado, assim como uma orientação para possíveis trabalhos futuros.

CAPÍTULO 2- ENQUADRAMENTO TEÓRICO

Deparados com um conjunto de dados e com o objectivo de efectuar uma análise estatística deve-se fazer uma escolha racional acerca do método mais apropriado a esta.

Neste sentido tem-se que ter em atenção algumas considerações importantes, como é o caso dos objectivos definidos, no contexto de um problema específico, as características matemáticas envolvidas, as hipóteses estatísticas a serem feitas sobre as variáveis em análise e ainda a forma de recolha dos dados (Braga, 1994).

2.1. ANÁLISE EXPLORATÓRIA DOS DADOS

Numa primeira fase, que compreendeu a análise dos dados, procurou entender-se os comportamentos das variáveis em análise, recorrendo-se geralmente a uma análise exploratória. Esta, englobou o cálculo das principais características amostrais, assim como gráficos adequados, de forma a estudar-se os comportamentos das variáveis, e ainda, qual/quais as variáveis de interesse para o estudo.

A estatística descritiva representa as características das unidades observadas ou experimentadas e utiliza-se para descrever esses dados através de estatísticas, como por exemplo a média, a mediana ou desvio padrão.

A estatística indutiva permite tirar conclusões para um domínio mais vasto do que os elementos observados ou experimentados. Essas inferências são realizadas por aplicação em amostras aleatórias de intervalos de confiança e testes paramétricos ou não paramétricos (Pestana e Gageiro, 2005).

2.1.1. Escalas de atitudes e opiniões:

As escalas de opiniões são mecanismos concebidos para medir o grau de intensidade das atitudes e das opiniões de um sujeito a respeito de um fenómeno determinado, visando captação de informação e permitindo ao sujeito diversas opções entre uma série graduada que lhe é proposta. Distinguindo-se a escala unidimensional de *Guttman*, a escala de distância social de *Bogardus,* a escala de intervalo de *Thurstone* e a escala cumutativa de *Likert* (Bessa, 2007).

2.1.2. Tipo de dados

Dados são algoritmos, letras e sinais ou mesmo combinações destes segundo determinadas regras que descrevem uma determinada situação. A descrição e interpretação de dados é uma parte essencial da estatística.

Os dados podem ser de diferentes tipos e, portanto, necessitam ser tratados com métodos estatísticos diferentes. Podendo assim dividir-se os dados em quantitativos e qualitativos (Pestana e Gageiro, 2005).

Os dados quantitativos consistem em números que representem contagens ou medidas, enquanto os dados qualitativos podem ser separados em diferentes categorias que se distinguem por alguma característica não numérica.
Dependendo do tipo de variáveis que constituem os dados estes podem ser expressos em quatro escalas distintas: nominal, ordinal, por intervalos e por rácios.
Os dados qualitativos exprimem-se nas duas primeiras e os dados quantitativos nas duas últimas (Pestana e Gageiro, 2005).

A escala nominal classifica os sujeitos conforme pertençam ou não a uma categoria ou característica. São variáveis categóricas, não podendo avaliar se uma é maior do que a outra. A escala ordinal ordena os sujeitos segundo a ordem que ocupam. Sabe-se que um valor é maior do que outro, mas não há avaliação do intervalo entre dois valores. A escala intervalar atribui valores numéricos a indivíduos, sabendo-se que um valor é maior do que

outro e que os valores diferem em intervalos iguais. A escala de razão ou rácio é uma escala de intervalo, porém possui um zero absoluto (Bessa, 2007).

Os dados quantitativos podem ainda ser descritos pela distinção entre os tipos discretos e contínuos.

Os dados discretos são aqueles que provem de uma variável discreta, isto é, que apenas tomam valores finitos ou numerável de valores distintos. Os dados que não são discretos podem ser provenientes de variáveis contínuas, isto é, tomam valores num conjunto de números reais, possivelmente ilimitado (Athayde, 2005).

2.2. *SOFTWARE* UTILIZADO

Na análise estatística dos dados optou-se pela utilização do *software SPSS* (*Statistical Package for the Social Sciences*) teve origem em 1968, na *Chicago University*, sendo divulgado e com utilização global desde então. Sendo desde 1994, representado em Portugal pela PSE- Produtos e Serviços de Estatística, Lda (www.pse.pt).

O *software SPSS* é um sistema que permite organizar dados e executar análises estatísticas. Tem um ambiente gráfico muito apelativo com o qual, para a maioria das análises a efectuar, basta a selecção das respectivas opções em menus e caixas de diálogos (Laureano e Botelho, 2010).

A utilização do *SPSS* apresenta várias vantagens:

- Flexibilidade para diferentes naturezas de variáveis;
- Facilidade de utilização, sendo um programa muito amigável, com diversos níveis de complexidade, de acordo com as necessidades dos seus utilizadores;
- Participação em todo o processo analítico, desde o planeamento até à recolha de dados para a análise, possibilitando a elaboração de relatórios, quer pelo próprio programa, quer por uma articulação com um processador de texto (Laureano e Botelho, 2010).

2.3. REGRESSÃO LOGÍSTICA

2.3.1. Enquadramento

Através dos recursos matemáticos e estatísticos cedidos pela análise de regressão pode encontrar-se funções que estimem o comportamento de um conjunto de dados que não se dispõem, a partir de dados recolhidos.
O termo de regressão apareceu pela primeira vez na literatura em *Galton* (1885), citado por Braga (1994).

A regressão é um modelo estatístico usado para prever o comportamento de uma variável dependente (*Y*) a partir de uma ou mais variáveis relevantes de natureza essencialmente intervalar ou rácio, as variáveis independentes (*X*), dando informação sobre a margem de erro dessas previsões.

Tal como referido por Hosmer e Lemeshow (1989), nos modelos de regressão linear simples ou múltipla a variável dependente *Y* é uma variável aleatória de natureza contínua, sendo esta em alguns casos qualitativa e expressa em função de duas ou mais variáveis de natureza categórica, isto é, admite dois ou mais valores.

Assim, o que distingue o modelo de regressão logística do da regressão linear é que a variável resultado na regressão logística é usualmente binária (dicotómica). Esta diferença entre regressão logística e linear é reflectida quer na escolha de um modelo paramétrico, quer nas hipóteses a serem consideradas. Desde que esta diferença seja tida em conta, os métodos empregues na análise usando a regressão logística seguem os mesmos princípios usados na regressão linear. Então, as técnicas usadas na análise de regressão linear deverão motivar uma aproximação à regressão logística.

A regressão logística permite o uso de um método de regressão para calcular ou prever a probabilidade de um evento específico. Desta forma, esta usa-se quando se tem uma variável dependente em escala nominal e uma variável independente nominal e/ou contínua,

e serve para descrever a relação entre a variável dependente nominal e o conjunto de variáveis independentes através da função *logit* (Braga, 1994).

Quando a regressão logística é usada deve-se primeiramente achar o modelo que melhor se ajuste aos dados em análise, com o intuito de se obter um modelo moderado e biologicamente razoável, que permita descrever a relação entre a variável resultado e um conjunto de variáveis independentes (Braga, 1994).

2.3.2. 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 independentes e uma variável resposta (Martins, 2008).

A análise apresentada neste capítulo baseia-se essencialmente no trabalho de Hosmer e Lemeshow (1989).

Qualquer problema de regressão passa por estimar o valor esperado da variável resposta, *Y*, dado o valor das variáveis independentes, ***x***.
Na regressão linear assume-se que este valor esperado pode ser expresso como uma equação linear em função de ***x***,
Considerando o modelo de regressão linear simples tem-se

$$E[Y \mid x] = \beta_0 + \beta_1 x \qquad (1)$$

Tendo em conta a expressão anterior, verifica-se, que $E[Y \mid x]$ pode tomar qualquer valor compreendido no intervalo $]-\infty;+\infty[$.

Uma diferença importante entre os modelos de regressão linear e o de regressão logística vai de encontro à distribuição condicional da variável resultado.

Na regressão linear a observação da variável resultado pode ser expressa como

$$y = E[Y \mid x] + \varepsilon, \tag{2}$$

sendo ε o erro associado.

De acordo com (2), ε dá o desvio de uma observação em relação à média condicional. A hipótese mais comum é que este ε segue uma distribuição Normal com média zero e variância constante, ao longo dos níveis da variável independente. Daqui, resulta que esta distribuição condicional da variável resultado dado o valor da variável *x*, segue uma distribuição normal, com média $E[Y \mid x]$ e variância constante.

Contudo, isto, não se verifica quando se tem uma variável resultado dicotómica. Assim, nesta situação deve-se expressar o valor da variável como

$$y = \pi(x) + \varepsilon, \tag{3}$$

considerando-se $\pi(x) = E[Y \mid x]$.

De acordo com Hosmer e Lemeshow (1989), quando se trabalha com dados dicotómicos, a média deverá assumir valores entre 0 e 1. A variação de $E[Y \mid x]$ em função de *x*, é menor consoante a aproximação da média condicional de 0 ou 1. Assim, a curva resultante tem uma forma em S, sendo semelhante ao gráfico de uma distribuição cumulativa de uma variável aleatória. Neste caso, usa-se o modelo de regressão logística.

Foram propostas muitas funções para análise de variáveis dicotómicas, *Cox*, em 1970, (citado por Hosmer e Lemeshow, 1989) apresentou várias razões para a escolha da distribuição logística para a análise de dados, destacando-se:

(1) O ponto de vista matemático, como sendo uma função extremamente flexível e muito usada;

(2) Por si mesma, conduz a uma fácil interpretação dos resultados em termos biológicos.

A resposta esperada é dada pela expressão

$$E[Y \mid x] = \beta_0 + \beta_1 x, \tag{4}$$

sendo *Y* uma variável aleatória que segue uma distribuição de *Bernoulli*, com a seguinte lei de probabilidade:

$$\begin{cases} Y = 1 \Rightarrow P(Y = 1) = \pi(x)\ sucesso \\ Y = 0 \Rightarrow P(Y = 0) = 1 - \pi(x)\ insucesso \end{cases}, \tag{5}$$

Aplicando a definição de valor esperado, obtém-se:

$$E[Y \mid x] = \pi(x) \tag{6}$$

Igualando a expressão obtida em (4) e (6) tem-se

$$E[Y \mid x] = \beta_0 + \beta_1 x = \pi(x) \tag{7}$$

Considere-se uma amostra de *n* observações independentes com o par (x_i, y_i), onde x_i e y_i representam o valor da variável independente e o valor da variável resposta, respectivamente, sendo *i* o iésimo elemento.

A função de regressão logística univariada é dada pela esperança de *Y* dado *x*, ou seja,

$$\pi(x) = E[Y \mid x] = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}, \tag{8}$$

Os parâmetros considerados são estimados pelo método de máxima verosimilhança, que consiste em determinar os valores dos parâmetros que maximizem a probabilidade de obter o conjunto de valores observados.

Uma propriedade interessante que a função logística possui é que pode ser linearizada. Assim, fazendo essa transformação vem

$$g(x) = ln\left(\frac{\pi(x)}{1 - \pi(x)}\right), \tag{9}$$

obtendo-se

$$g(x) = \beta_0 + \beta_1 x, \tag{10}$$

onde $-\infty \le g(x) \le +\infty$; $-\infty \le x \le +\infty$.

Esta transformação é chamada de transformação *logit* de probabilidade $\pi(x)$. A razão

$$\pi(x)\times\left[1-\pi(x)\right], \tag{11}$$

na transformação *logit* é a chamada *odds* ou "*chance*".

A importância desta transformação é que $g(x)$ tem muitas propriedades desejáveis dos modelos de regressão linear. A função *logit*, $g(x)$, é linear nos seus parâmetros, podendo ser contínua, e variar entre valores de $]-\infty;+\infty[$, dependendo do domínio de variação de *x* (Martins, 2008).

2.3.2.1. Função de verosimilhança

O método geral de estimação alternativo ao da função dos mínimos quadrados, para o modelo de regressão linear, é o método de máxima verosimilhança (M.M.V). Este método dá a base para a aproximação de estimação com o modelo de regressão logística (Braga,1994).

Atendendo a Hosmer e Lemeshow 1989, o M.M.V. permite obter valores para os parâmetros desconhecidos, que maximizam a probabilidade de obter o conjunto de observações.

A função de verosimilhança expressa a probabilidade dos dados observados como uma função dos parâmetros desconhecidos. Os estimadores de máxima verosimilhança destes parâmetros, são escolhidos de modo a ser aqueles que maximizam a função de verosimilhança.

Neste caso, em que se tem apenas dois resultados possíveis (sucesso $Y=1$ e o insucesso $Y=0$), e desde que as observações sejam independentes, a função de verosimilhança é dada por:

$$l(\beta)=\prod_{i=1}^{n}\pi(x_i)^{y_i}\cdot\left(1-\pi(x_i)\right)^{1-y_i}, \tag{12}$$

em que $\pi(x_i)$ representa a $P\left[Y=1|\,x\right]$, ou seja, a probabilidade de sucesso.

O princípio de máxima verosimilhança usa para estimativa de β os valores que maximizam a expressão obtida em (12). Contudo, é mais fácil trabalhar com a expressão dos logaritmos da verosimilhança, sendo

$$L\ (\beta) = ln\left(l\left(\beta\right)\right) = \sum_{i=1}^{n}\left\{y_i \cdot ln\left[\pi\left(x\right)\right] + \left(1 - y_i\right) \cdot ln\left[1 - \pi\left(x\right)\right]\right\} \tag{13}$$

e para se obter o valor de β, que maximiza $L\ (\beta)$, deriva-se esta em ordem a cada parâmetro e iguala-se as equações de verosimilhança a zero.

Para regressão logística envolvendo duas variáveis, as equações de verosimilhança são não lineares em β, o que vai requerer métodos especiais para a sua resolução, sendo o método de resolução de equações não lineares usualmente aplicado o método de *Newton-Raphson* (Martins, 2008).

2.3.2.2. Teste de *Wald*

Em regressão logística tem-se variáveis resultado e uma ou mais variáveis explicativas. Para cada variável explicativa do modelo, haverá um parâmetro associado.

O teste de *Wald*, descrito por *Polit* (1996) e *Agresti* (1990) (citado por *Crichton* (2001)), é uma das possíveis formas de testar se os parâmetros associados com um grupo de variáveis explicativas tomam o valor zero.

Segundo *Crichton* (2001), este teste é utilizado para avaliar se o parâmetro é estatisticamente significativo. A estatística teste que se utilizada é obtida através da razão do coeficiente pelo seu respectivo erro padrão, esta estatística de teste segue uma distribuição Normal. A estatística de teste, para avaliar se o parâmetro β é igual a zero e pode-se especificar como sendo:

$$W = \frac{\hat{\beta}}{\sqrt{Var\left(\hat{\beta}\right)}} \tag{14}$$

Todavia, o teste de *Wald*, falha quando se rejeita coeficientes que são estatisticamente significativos (*Hauck* e *Donner*, 1977, citado por *Crichton* (2001). Assim, aconselha-se que os coeficientes, identificados por este teste como sendo estatisticamente não significativos, sejam testados novamente pelo teste da razão de verosimilhança.

2.3.3. Modelo de regressão logística multivariado

A regressão logística pode ser utilizada, fazendo as necessárias adaptações, para modelar situações com mais do que uma variável independente.

Considere-se *n* observações independentes do par $(\boldsymbol{x_i}, y_i)$, em que $\boldsymbol{x_i}$ é um vector de *m* variáveis independentes e y_i uma variável dicotómica. A função logística que se usa para modelar esta situação é semelhante à usada para o modelo univariado apresentado anteriormente, envolvendo as *m* variáveis independentes:

$$\pi(\boldsymbol{x_i}) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_m x_m}}{1-e^{\beta_0+\beta_1 x_1+\ldots+\beta_m x_m}} \tag{15}$$

Os *m+1* parâmetros desconhecidos são estimados pelo método da máxima verosimilhança, aplicando processos iterativos, onde as equações de verosimilhança são dadas por:

$$\begin{cases} \dfrac{\delta L}{\delta \beta_0} = 0 \\ \dfrac{\delta L}{\delta \beta_j} = 0 \end{cases} \Leftrightarrow \begin{cases} \displaystyle\sum_{i=1}^{n} \left[y_i - \pi(\boldsymbol{x_i}) \right] = 0 \\ \displaystyle\sum_{i=1}^{n} x_{ij} \left[y_i - \pi(\boldsymbol{x_i}) \right] = 0 \end{cases}, \; j = 1,\ldots,m \tag{16}$$

Independentemente do número de variáveis usadas para definir o modelo de regressão logística, pretende-se distinguir dois grupos distintos de indivíduos, consoante apresentem ou não determinada característica.

Salienta-se que neste estudo, é importante reduzir o número de variáveis a serem incluídas no modelo. Esta 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, principalmente em amostras de pequena dimensão (Hosmer e Lemeshow, 1989).
Assim, regra geral, esta situação traduz-se em valores extremamente elevados das estimativas dos coeficientes e/ou dos erros padrão.

Com o objectivo de verificar se as variáveis independentes possibilitam identificar correctamente os elementos que pertencem a cada grupo, constrói-se o modelo de regressão logística que inclui todas as variáveis e posteriormente, avalia-se a qualidade do seu ajuste. Assim, os valores preditos são então comparados com os valores da variável resposta, que toma dois valores possíveis, 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 que 0.5. Se a maior percentagem de indivíduos for bem classificada, é conveniente que se tente encontrar um novo modelo com menos variáveis que nos permita separar os elementos de dois grupos (Martins, 2008).

2.3.3.1. Testes de significância estatística

Aquando o modelo ajustado, segue-se a realização de um teste de significância das variáveis que foram incluídas no modelo.
Dado o interesse em se utilizar um teste estatístico de forma a avaliar a razão de verosimilhança, será usado o seu logaritmo, o qual multiplicado por menos dois, resulta numa distribuição conhecida. Este valor é designado por *D*, sendo o teste utilizado o da razão de verosimilhança.
Assim, a estatística *D* tem como objectivo comparar o modelo em análise e o modelo saturado ou seja,

$$D = -2\,ln\left[\frac{verosimilhança\ do\ modelo\ ajustado}{verosimilhança\ do\ modelo\ saturado}\right] \quad , \qquad (17)$$

onde o modelo ajustado corresponde ao modelo que inclui apenas as variáveis desejadas e o modelo saturado corresponde ao modelo com todas as variáveis e interacções. Por outras palavras, o modelo saturado contém tantos parâmetros quanto observações.

Tem-se assim o seguinte teste para testar a significância em que as *m* variáveis são independentes. Para este teste temos as seguintes hipóteses:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_m = 0 \quad vs \quad H_1 : \beta_j \neq 0 \quad \exists_{j=0,...,m \ e \ m=1,...,k} \tag{18}$$

sendo o Teste da Razão de Verosimilhança, o qual se pode definir do seguinte modo:

$$\begin{aligned} G &= D(\textit{verosimilhança sem as m variáveis}) - D(\textit{verosimilhança com as m variáveis}) \\ &= -2\,ln\left[\frac{\textit{verosimilhança}\,(\textit{modelo sem as m variáveis})}{\textit{verosimilhança}\,(\textit{modelo com as m variáveis})}\right] \end{aligned} \tag{19}$$

O teste *G*, segue a distribuição de *Qui-Quadrado* com *m* graus de liberdade, sob a validade da hipótese nula.
Assim, ao rejeitar H_0, pode-se concluir que pelo menos um, ou até os *m* coeficientes poderão ser diferentes de zero (Braga, 1994).

2.3.3.2. Métodos de selecção de variáveis

A inclusão ou a exclusão de uma variável no modelo, pode variar conforme o problema a considerar ou até mesmo a área científica em análise (Braga, 1994).

Quando se minimiza o número de variáveis a incluir no modelo, obtém-se um modelo numericamente mais estável e mais generalizado. As variáveis que não estão correctamente incluídas no modelo podem provocar o aumento dos erros padrão estimados, assim como, uma maior dependência do modelo que se traduz nos dados observados (Braga,1994).

Indo ao encontro de Hosmer e Lemeshow (1989), segue-se alguns passos que podem ajudar quando se tem que seleccionar as variáveis a serem incluídas no modelo de regressão

logística. Este processo é semelhante ao utilizado na construção do modelo de regressão linear.

Deste modo o processo pode ser descrito tendo em conta os seguintes passos:

(A) Deve-se iniciar o processo por uma análise univariada e individual de cada uma das variáveis. Hosmer e Lemeshow (1989) sugeriram que variáveis nominais, ordinais e contínuas com alguns valores inteiros poderão ser tratadas recorrendo-se a tabelas de contingência dos *p* níveis da variável dependente *versus* os *k* níveis da variável independente.
Quando se trata de variáveis independentes e contínuas é desejável que a análise univariada envolva o ajuste de um modelo de regressão logística como o objectivo de se obter estimativas dos coeficientes, estimativas de erro padrão, o teste de razão de verosimilhança para a significância dos coeficientes e estatísticas de *Wald* univariada.
Pode-se ainda usar como alternativa o *teste-t* para duas amostras.

O *teste-t* para duas amostras independentes usa-se quando se pretende comparar as médias de uma variável quantitativa em dois grupos diferentes de indivíduos e se desconhecem as respectivas variâncias populacionais (Pestana e Gageiro, 2005).

A análise baseada neste teste poderá ser útil na determinação da inclusão ou exclusão da variável no modelo.

(B) Quando a análise univariada estiver concluída passa-se para uma análise multivariada. Após sujeitas a um teste univariado selecciona-se as variáveis que apresentarem um valor prova inferior a 0.25, sendo essas variáveis tomadas como candidatas ao modelo multivariado (pode-se ainda incluir no mesmo modelo variáveis consideradas importantes no contexto do estudo ou análise).

A escolha do valor 0.25 como critério de selecção foi feita tendo em conta os trabalhos realizados em regressão linear e regressão logística de *Bendel* e *Afifi*, e, *Mickey* e *Greenland*, citado por Braga (1994).

Segundo estes autores, o valor de 0.05 por vezes falha para algumas das variáveis em análise, por outro lado, quando se consideram níveis elevados podem-se incluir no modelo variáveis com interesse questionável.

Geralmente a decisão começa por ter em conta um modelo multivariado com todas as variáveis possíveis dependente da dimensão e número de elementos que constituem cada grupo de variáveis candidatas ao modelo.
Assim, quando se tem dados adequados para suportar a análise será conveniente começar o modelo multivariado nesse ponto. Caso contrário, esta aproximação pode conduzir a um modelo numericamente instável. Neste ultimo caso, a estatística de *Wald* não deverá ser usada para a selecção das variáveis. Dever-se-á recorrer a uma aproximação para selecção de variáveis baseada no método passo a passo, no qual as variáveis seleccionadas quer por inclusão, quer por exclusão segundo uma ordem sequencial baseada unicamente num critério estatístico (Braga, 1994; Hosmer e Lemeshow, 1989).

(C) Com o modelo multivariado construído, tem-se que verificar a importância de cada variável a ser incluída neste. Para isso, deve-se aplicar o teste de *Wald* para cada variável e comparar o valor de cada coeficiente estimado com o seu valor no modelo univariado contendo somente essa variável.
As variáveis que não contribuam para explicar correctamente o modelo deverão ser eliminadas e ajustar-se um novo modelo. Este novo modelo deverá ser comparado com o antigo aplicando-se o teste da razão de verosimilhança.
O processo de retirar, reajustar, e verificar deve continuar até parecer que as variáveis explicativas do modelo estejam todas incluídas e em oposição às pouco importantes excluídas do modelo.
Se no fim do processo da análise univariada se tiver um número elevado de variáveis candidatas a explicativas ao modelo, será aconselhável utilizar-se a técnica passo a passo (Braga, 1994; Hosmer e Lemeshow, 1989).

(D) Por fim, e após se ter obtido um modelo que pareça conter as variáveis importantes, deve-se fazer uma reanálise de forma a se considerar a necessidade da inclusão de interacção entre variáveis (Braga, 1994; Hosmer e Lemeshow, 1989).

2.3.3.2.1. Selecção automática

Qualquer procedimento para adição ou remoção de variáveis num modelo é baseado num algoritmo que verifica a importância das variáveis, incluindo ou excluindo-as do modelo, baseando-se na regra de decisão.
O critério para adição ou remoção de variáveis, em regressão linear, é geralmente baseado na estatística *F*, comparando os modelos com e sem as variáveis em análise. Em regressão logística, os erros seguem uma distribuição binomial sendo baseado do teste de razão de verosimilhança.

Existem métodos automáticos que podem ser utilizados na decisão de inserir e remover variáveis.
Seguidamente, descrevem-se os métodos implementados no SPSS.

- *Enter*: é um procedimento para a selecção de variáveis em que todas elas em bloco entram no processo uma única vez; (SPSS Inc, 2007).

- *Forward*: Método de selecção *Stepwise*, este procedimento inicia-se com um modelo que não contenha variáveis explicativas. A ideia do método é adicionar uma variável de cada vez, seleccionando em primeiro lugar aquela que apresentar um valor de correlação mais elevado, em módulo, com a variável resposta, e assim consequentemente, até que o processo pára quando o aumento do coeficiente de determinação, devido à inclusão de uma nova variável explicativa no modelo não é mais importante (http://portalaction.com.br).

  - *Forward (condicional)*: baseado na significância da estatística de pontuação e testes de remoção com base na probabilidade de uma estatística de razão de verosimilhança, com base em estimativas de parâmetros condicionais; (SPSS Inc., 2007).

  - *Forward (Likelihood Ratio):* baseado na significância da estatística de pontuação e testes de remoção com base na probabilidade de uma estatística de razão de verosimilhança baseada na máxima verosimilhança parcial das estimativas; (SPSS Inc., 2007).

- *Forward (Wald)*: Método de selecção *Stepwise* baseado na significância da estatística de pontuação e testes de remoção com base na probabilidade da estatística de *Wald*; (SPSS Inc., 2007).

- *Backward*: Enquanto o método *Forward* começa sem nenhuma variável no modelo e adiciona variáveis a cada passo, o método *Backward* faz o oposto. Este incorpora inicialmente todas as variáveis, e ao longo do processo cada uma pode ou não ser eliminada. A primeira variável a ser removida é aquela que apresenta um menor coeficiente de correlação parcial com a variável resposta (http://portalaction.com.br).

- *Eliminação Blackward (Condicional):* baseada na estatística de razão de verosimilhança de probabilidade das estimativas condicionais dos parâmetros (SPSS Inc., 2007).

- *Backward Elimination (Likelihood Ratio):* baseado na probabilidade da estatística de razão de verosimilhança apoiado nas estimativas de probabilidades parciais (SPSS Inc., 2007).

- *Backward Elimination (Wald):* baseado nas probabilidades da estatística de *Wald* (SPSS Inc., 2007).

Stepwise é um dos métodos mais utilizados e consiste na combinação dos dois métodos anteriores (*Forward* e *Backward*). Este, inicia com uma variável (a que apresentar maior correlação com a variável resposta), e a cada passo do *Forward*, depois de incluir uma variável, aplica o *Backward* para ver se será descartada alguma variável. Continua-se o processo até este não incluir ou excluir nenhuma variável (http://portalaction.com.br).

2.3.4. Razão de possibilidades (*odds ratio*)

Actualmente muitos investigadores optam por analisar a relação entre duas variáveis de escala nominal através do rácio de produtos cruzados – *razão de possibilidade,* pois tem uma interpretação mais fácil do que o teste de *Qui Quadrado* (Bessa, 2007).

De acordo com Bessa (2007), a razão de possibilidade é uma medida antiga tendo sido usada por *Snow* no seu trabalho clássico de identificação do factor risco da propagação da cólera em Londres (1853). Sendo utilizado como medida de associação em estudos de "caso- controle" e em estudos transversais controlados.

A razão de possibilidade é a razão entre duas *odds*, onde as *odds* são calculadas da seguinte forma:

$$odds = \frac{"\ Probabilidade\ de\ um\ acontecimento\ ocorrer"}{"\ Probabilidade\ de\ um\ acontecimento\ não\ ocorrer"} \tag{20}$$

Assim, a razão de possibilidade é uma forma de se comparar se a probabilidade de um determinado evento é a mesma para dois grupos (*Wagner e Callegari-Jacques*, 1998*; Rumel in* Revista Saúde Publica, 1986).

Considerando-se a seguinte tabela 2 por 2:

	X^-	X^+	
Y^-	a	b	**a+b**
Y^+	c	d	**c+d**
	a+c	**b+d**	**n=a+b+c+d**

Tabela 1-*Razão de possibilidade*
Fonte: Adaptado de Pestana & Gageiro, 2005

Daqui tira-se que

$$razão\ de\ possibilidades = \frac{a \times d}{b \times c} \quad , \tag{21}$$

e ainda que:

- *razão de possibilidades* $= 1$ implica que o evento é igualmente provável em ambos os grupos;
- *razão de possibilidades* > 1 significa que o evento é mais provável no 1º grupo;
- *razão de possibilidades* < 1 implica que o evento é menos provável no 1º grupo.

Conclui-se que o significado da razão de probabilidade é semelhante ao risco relativo obtido em estudos de *coorte*, e expressa a força de associação o evento e o grupo (Pestana e Gageiro, 2005).

Assim, segundo o que foi referido em Pestana e Gageiro (2005) ,uma medida mais directa comparando as probabilidades em dois grupos é o risco relativo, que também é conhecida como a relação de risco. O risco relativo é simplesmente a razão de duas probabilidades condicionais.

2.3.5. Avaliar o ajuste do modelo

Quando se fala na qualidade do ajuste de um modelo de regressão logística tem que se ter em atenção a análise de medidas das diferenças entre os seus valores observados da variável resposta, *y*, e os resíduos.

Sendo o objectivo avaliar o "bom" ajuste do modelo construído através da regressão logística, pode-se fazê-lo usando representações gráficas dos valores dos resíduos. Este caso permite comparar os resíduos dos vários elementos. Pode-se ainda aplicar testes baseados em estatísticas desses valores, fundamentados no valor da estatística de teste e avaliando a qualidade do ajuste do modelo de uma forma global (Martins, 2008).

Após aplicação de um teste de análise de resíduos e quando a qualidade do modelo não é validada por todos esses elementos, o ideal será verificar a existência de elementos com valores de resíduos elevados (em módulo), comparando-os com os resíduos dos restantes elementos (Martins, 2008).

Relativamente às medidas das diferenças dos valores observados e preditos, usados em regressão logística, destacam-se os resíduos de *Pearson* e os *Deviance residuals*, denotados por **r** e **d**, respectivamente.

Ou seja:

$$r_j = r\left(y_j, \hat{\pi}_j\right) = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j\left(1-\hat{\pi}_j\right)}} \tag{22}$$

e

$$d_j = d\left(y_j, x_j\right) = \mp\sqrt{2\left[y_j\, ln\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + \left(m_j - y_j\right) ln\left(\frac{m_j - y_j}{m_j\left(1-\hat{\pi}_j\right)}\right)\right]} \tag{23}$$

onde $j = 1,2,...,J$ sendo J o número de valores diferentes de $\boldsymbol{x}$, $x = \left(x_1, x_2, ..., x_m\right)$, e m_j o número de indivíduos com $x = x_j$.

Sob a validade do modelo ser o adequado, as estatísticas acima têm aproximadamente uma distribuição $\chi^2_{J-(m+1)}$.

Devendo-se rejeitar a hipótese nula para valores elevados da estatística de teste, essa aproximação só é valida se os valores de m_j forem também elevados (*Kuss* (2002), citado por Martins, (2008)).

Em 1989, Hosmer e Lemeshow, propuseram uma estatística de qualidade de ajuste para um modelo de Regressão Logística, em que os dados devem ser agrupados em *g* grupos com as respectivas probabilidades estimadas.

Denote-se:

n_g : o número de indivíduos;

c_g : o número de valores diferentes do conjunto das p variáveis independentes;

o_g : soma dos valores da variável resposta, com $o_g = \sum_{j=1}^{c_g} y_j$;

$\overline{\pi}_g$: média das probabilidades estimadas para o grupo k , com $\overline{\pi}_g = \sum_{j=1}^{c_g} \frac{m_j\hat{\pi}_j}{n_g}$.

Assim, a estatística de Hosmer-Lemeshow, segue uma distribuição aproximadamente de um *Qui-quadrado* com *g-2* graus de liberdade, segundo uma hipótese de o modelo ser o adequado.

Rejeitando-se a hipótese nula para valores elevados da estatística de teste, *C*, e podendo expressar-se da seguinte forma:

$$C = \sum_{k=1}^{g} \frac{o_g - n_g \bar{\pi}_g}{n_g \bar{\pi}_g \left(1 - \bar{\pi}_g\right)} \tag{24}$$

Note-se que este resultado depende dos grupos que são escolhidos (Martins, 2008).

2.4. CURVA ROC

2.4.1. Perspectiva Histórica

Uma pratica comum, na área relacionada com a medicina, é a forma de se descrever como e quanto uma variável contínua ou categórica ordinal é capaz de classificar materiais ou indivíduos em grupos definidos.

A análise **ROC** (***R**eceiver **O**perating **C**haracteristc*) é uma ferramenta que permite medir e especificar problemas no desempenho do diagnóstico em medicina.

A curva ROC foi usada pela primeira vez durante a segunda Guerra Mundial aplicada à análise de radar antes de ter sido empregue na teoria de detecção de sinais (*Green e Sweets*, citado Braga (2000)). Depois do ataque a *Pearl Harbor*, em 1941, o exército dos Estados Unidos focou-se na investigação vocacionada a aumentar a previsão de detectar correctamente aviões Japoneses através dos sinais de radar.

Nas décadas de 60 e 70, as curvas ROC foram utilizadas na psicologia experimental e em ramos da biomédica, respectivamente. Nesta ultima, o objectivo principal passou basicamente por classificar os indivíduos em “doentes” ou “não doentes”. (Braga, 2000).

2.4.2. Conceitos básicos

A análise da curva ROC pode ser feita por meio de um gráfico simples e robusto, que nos permite estudar a variação da sensibilidade e especificidade, para diferentes valores de corte.

A *sensibilidade* (*Sens.*) é definida como a probabilidade do teste fornecer um resultado positivo, dado que o indivíduo é realmente portador da "doença", enquanto, a *especificidade* (*Esp.*) é definida como a probabilidade do teste fornecer um resultado negativo quando o indivíduo não é portador da "doença" (Margotto).

De outra forma, pode-se dizer que as curvas ROC foram desenvolvidas no ramo das comunicações como uma forma de demonstrar as relações entre sinal-ruído. Neste sentido, podemos interpretar o sinal como os verdadeiros positivos (*sensibilidade*) e o ruído, como os falsos positivos (*1- especificidade*) (Braga, 2000).

A tabela seguinte resume o que foi dito acima:

	Positivos (+)	**Negativos (-)**
Positivo (+)	VP Verdadeiros positivos	FP Falsos positivos
Negativo (-)	FN Falsos negativos	VN Verdadeiros negativos
Total:	VP+FN	FP+VN
Desempenho	***Sensibilidade*** $Sens. = \frac{VP}{VP + FN}$	***Especificidade*** $Esp. = \frac{VN}{FP + VN}$

Tabela 2- *Representação geral de um teste diagnóstico*
Fonte: Adoptado de Braga, 2000

Note-se que a *Sensibilidade* e a *Especificidade* não são calculadas usando os mesmos indivíduos, ou seja, enquanto *Sensibilidade* usa apenas os "doentes", *Especificidade* utiliza os "não doentes", assim, *Sensibilidade* e Especificidade são medidas independentes entre si (Braga, 2000).

2.4.3. Gráfico da curva ROC

Tendo em conta o que foi citado em Braga (2000), a curva ROC é um gráfico de *Sensibilidade* (ou taxa de verdadeiros positivos) *versus* taxa de falsos positivos, ou seja, representa-nos a *Sensibilidade* (ordenadas) e *1 – Especificidade* (abcissas), resultantes da variação de um valor de corte ao longo do eixo de decisão *x*.

Assim, a representação da curva ROC, permite evidenciar os valores para os quais existe optimização da *Sensibilidade* em função da e *Especificidade* correspondente ao ponto que se encontra mais próximo do canto superior esquerdo do diagrama, uma vez que o indício de verdadeiro positivo é 1 e o de falso positivo 0.

Graficamente tem-se:

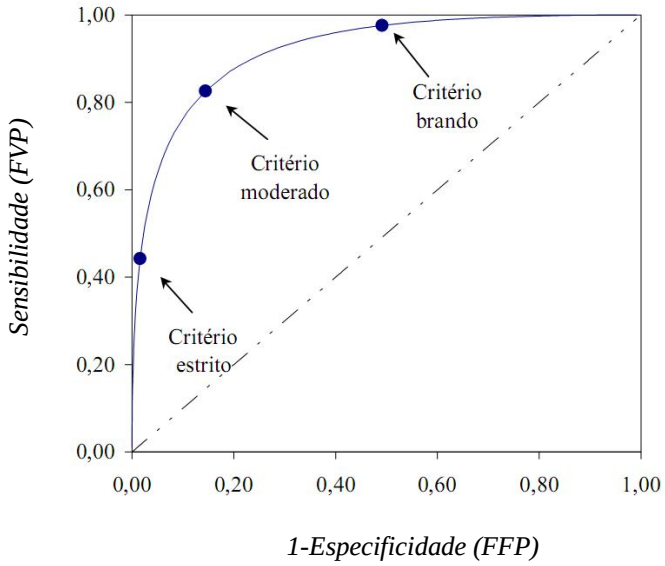


Gráfico 1- *Curva ROC, para uma dada capacidade de discriminação, com a variação do critério de decisão*
Fonte: Braga (2000)

A curva ROC discrimina entre dois estados, onde cada ponto da curva representa um compromisso diferente entre a *Sensibilidade* e o falso positivo que pode ser definido pela adopção de um valor diferente do ponto de corte de anormalidade. Um critério restrito é

aquele que traduz uma pequena fracção de falsos positivos assim como uma pequena fracção de verdadeiros positivos (Braga, 2000).

O valor do ponto de corte é definido com um valor que pode ser seleccionado arbitrariamente pelo pesquisador entre os valores possíveis para a variável de decisão, acima da qual o paciente é classificado positivo e abaixo do qual é classificado como negativo.

De acordo com Braga (2000), para cada ponto de corte são calculados valores de *Sensibilidade* e *Especificidade*, estes valores podem assim ser dispostos no gráfico. Um classificador perfeito corresponderia a uma linha horizontal no topo do gráfico, o que é bastante difícil de se obter. Na prática, curvas consideradas boas estarão entre a linha diagonal e a linha perfeita, onde quanto maior a distância da linha diagonal, melhor o sistema. A linha diagonal indica uma classificação aleatória, ou seja, um sistema que aleatoriamente selecciona saídas como positivas ou negativas. Finalmente, a partir de uma curva ROC, devemos poder seleccionar o melhor limiar de corte para obtermos o melhor desempenho possível.

Se o objectivo for verificar diferenças entre duas ou mais Curvas ROC, a avaliação é feita através da determinação da área abaixo da curva, usando uma modificação do teste da soma de ordens de *Wilcoxon* para esta comparação. Assim é possível quantificar a exactidão de um teste diagnóstico (proporcional à área abaixo da curva), além da possibilidade de comparar testes diagnósticos.

2.4.4. Área abaixo da curva ROC

A área abaixo da curva ROC está associada ao poder discriminante de um teste diagnóstico. Dado um individuo “doente” e outro “não doente”, ambos escolhidos ao acaso, esta medida é interpretada como a probabilidade do individuo “ portador de doença” ter um resultado ao teste diagnóstico de maior magnitude que aquele “não doente” (*Begg*, 1991 citado em Martinez, Neto-Louzada, e Pereira (2003)).

Um teste totalmente incapaz de discriminar indivíduos “doentes” e “não doentes” teria uma área abaixo da curva ROC de cerca de 0.5. Quanto maior a capacidade do teste em

discriminar os indivíduos segundo estes dois grupos, mais a curva se aproxima do canto superior esquerdo do gráfico, e a área abaixo da curva ROC próxima de 1. Para *Pepe* (2000), citado por Martinez, Neto-Louzada, e Pereira (2003), a área abaixo da curva ROC é uma medida não paramétrica da distância entre as distribuições dos resultados dos testes, para indivíduos classificados como "doentes" e "não doentes".

Quando apresentam a curva ROC, alguns autores optam por apresentar para o eixo das abcissas a *Especificidade* em alternativa a *1-Especificidade*, isto não altera a estimativa da área abaixo da curva. Se a curva é ajustada utilizando-se a teoria pertinente à distribuição normal, a área e o seu desvio padrão podem ser obtidos por recurso aos estimadores de máxima verosimilhança (*Begg*, 1987 citado em Martinez, Neto-Louzada, & Pereira, (2003)).

Analiticamente, a área abaixo da curva ROC pode ser determinada através de:

- Métodos de resolução numérica, como por exemplo, a regra do trapézio;
- Métodos estatísticos: Relação com a estatística de Wi*lcoxon-Mann-Witney* (*Hanley*, 1988, citado por Braga (2000)); e estimativa de Máxima Verosimilhança (*Hanley e McNeil*, 1982, citado por Braga (2000)).

2.4.5. Comparação de modelos com recurso ao teste da área abaixo da curva ROC

Numa escala comum, os gráficos que representam duas ou mais curvas ROC associadas a diferentes testes diagnósticos contínuos permitem uma imediata comparação de desempenhos (Martinez, Neto-Louzada, e Pereira, 2003).

Salienta-se que quando se está a comprar duas curvas ROC pode-se encontrar duas situações distintas:

(a) As curvas ROC empíricas são diferentes e não se cruzam, sendo o teste diagnóstico com maior área abaixo da curva aquele que apresenta melhor desempenho;
(b) As curvas ROC cruzam-se, as áreas abaixo da curva são próximas, mas os testes diagnósticos apresentam desempenhos diferentes.

Um método para testar se as diferenças entre duas áreas abaixo das curvas ROC provenientes de amostras independentes são significativas, consiste na utilização da razão crítica *z*, definida por *Hanley* e *McNeil* (1983):

$$z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2}} \sim N(0,1) \tag{25}$$

onde A_1 e A_2 correspondem as áreas e SE_1 e SE_2 correspondem aos erros estimados para a curva ROC, respectivamente para os testes diagnósticos 1 e 2. As áreas e os respectivos erros padrão são obtidos através da aproximação à estatística de *Wilcoxon-Mann-Whitney* (Braga, 2000).

Quando os valores da área abaixo da curva ROC são superiores a 0.5, os erros padrão associados às áreas, podem ser obtidos através da seguinte expressão:

$$SE(A) = \sqrt{\frac{A(1-A) + (n_A - 1)(Q_1 - A^2) + (n_N - 1)(Q_2 - A^2)}{n_A n_N}} \tag{26}$$

onde Q_1 é referente à probabilidade de duas observações anormais, aleatoriamente escolhidas serem classificadas com maior desconfiança do que uma observação normal aleatoriamente escolhida, e Q_2 corresponde à probabilidade de uma observação anormal, aleatoriamente escolhida ser classificada com maior desconfiança do que duas observações normais aleatoriamente escolhidas. E n_A e n_N corresponde, respectivamente a dimensão dos pacientes anormais e normais (Braga, 2000).

CAPÍTULO 3- ANÁLISE DE DADOS

3.1. ANÁLISE PRELIMINAR DOS DADOS

Para iniciar o estudo do questionário *online* começou-se por uma análise inicial de dados, considerando como objectivo principal do trabalho compreender quais as queixas que melhor explicam as lesões músculo esqueléticas (quando é prestado apoio ao domicílio).

O questionário criado estava dividido em quatro partes: A, B, C e D. Após uma análise (principais características amostrais, análise gráficas) destacam-se de seguida alguns resultados obtidos englobando as variáveis que constituem a parte A.

Os enfermeiros inquiridos dividem-se em quatro categorias profissionais, responderam a este inquérito 46% de enfermeiros graduados, 32% de enfermeiros, 16% de enfermeiros especialistas e 6% de enfermeiros chefes.
A idade dos inquiridos varia entre os 24 e os 65 anos, sendo a idade mais comum de 26 anos. Relativamente à antiguidade na profissão a média é de 12.83 anos, sendo o mínimo de anos de trabalho 2 anos e o máximo 42 anos. Na maioria estes trabalham 35 horas semanais, efectuando no máximo 60 horas semanais de trabalho. No que diz respeito ao peso, este varia entre os 45 kg e os 120 kg, sendo o peso médio de 65.63 kg, e a respectiva altura média de 161.85 cm para as mulheres e 174.68 cm para os homens. Tem-se que 15% dos enfermeiros exercem outra actividade em regime de acumulação, sendo maioritariamente ao nível da formação e enfermagem em clínicas privadas. Enquanto 31% já praticou alguma actividade de desporto ou lazer com regularidade, 34% já sofreu uma lesão do foro músculo esquelético.

Uma vez que a variável " Presta cuidados de enfermagem ao domicílio" será considerada como variável independente para a análise, elaborou-se uma análise para esta variável.

Verificou-se que das 147 respostas obtidas 125 enfermeiros prestam cuidados ao domicílio, correspondendo a uma percentagem de 85%.

Efectuando-se uma análise da questão em função do sexo, verificou-se que a percentagem de mulheres é superior a percentagem de homens, quando estes prestam apoio ao domicílio.

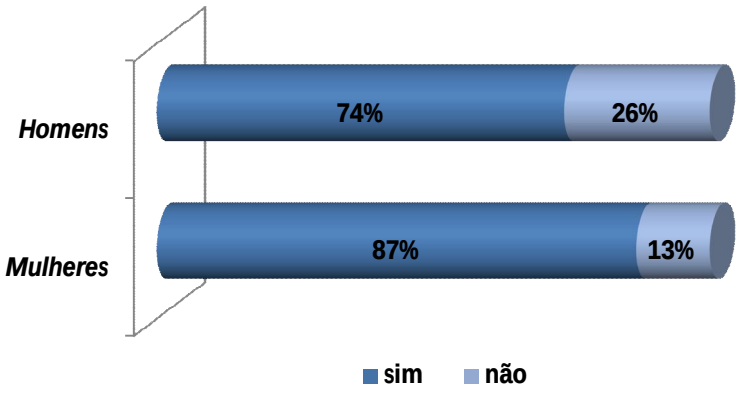


Gráfico 2- *Distribuição da percentagem de enfermeiros que prestam apoio ao domicílio segmentado por sexo*

Apresenta-se, ainda, a análise relativa as regiões do corpo, relativamente ao apoio ao domicílio.

Desta forma, poder-se-á ter uma ideia do comportamento da variável de interesse "presta cuidados de enfermagem ao domicílio" relativamente às sintomatologias nas regiões em causa.

O primeiro gráfico é referente aos enfermeiros que prestam apoio ao domicílio (*n=125*) e o segundo aos enfermeiros que não prestam apoio ao domicílio (*n=25*).

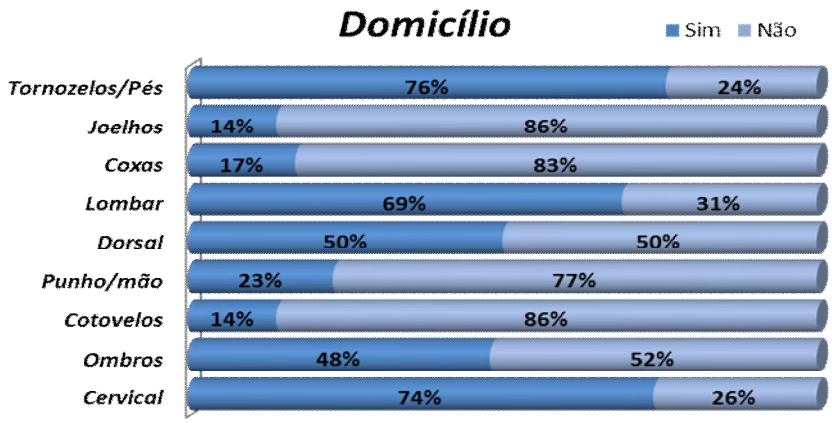


Gráfico 3- *Queixas nas regiões corporais por parte dos enfermeiros que prestam apoio ao domicílio (em %)*

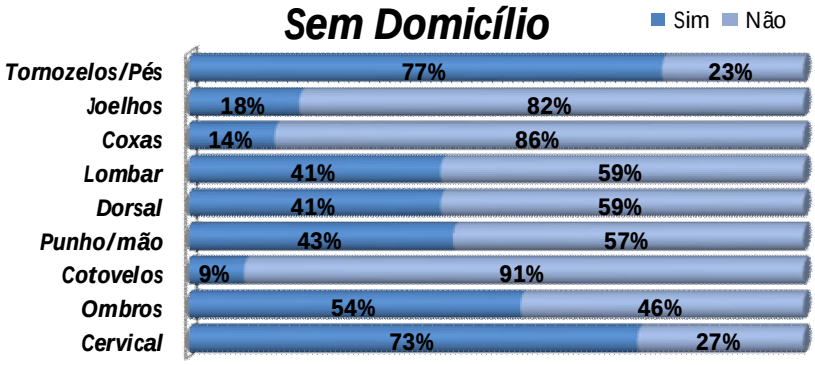


Gráfico 4- *Queixas nas regiões corporais por parte dos enfermeiros que não prestam apoio ao domicílio (em %)*

Após análise gráfica verifica-se que as percentagens de respostas para os dois casos são semelhantes. Contudo, a região lombar apresenta uma percentagem mais elevada quando os enfermeiros prestam apoio ao domicílio (69%) comparando com a percentagem de respostas por parte dos enfermeiros que não prestam apoio ao domicílio (41%). Daqui, pode-se prever que a região lombar está associada como sendo uma causas das LME dos enfermeiros.

3.2.RAZÃO DE POSSIBILIDADES

Nesta secção vai considerar-se como variável independente "presta cuidados de enfermagem ao domicílio" e nove variáveis dependentes.

As variáveis dependentes são referentes à sintomatologia, ou seja, as queixas apontadas pelos 147 enfermeiros, nas várias regiões do corpo. Deste modo tem-se: Região cervical; Ombros; Cotovelos; Punho/ mão; Coluna vertebral – zona dorsal; Coluna vertebral – zona lombar; Coxas; Joelhos; Tornozelos/ pés.

Com o objectivo de compreender qual/quais as variáveis que apresentam um valor significativo para a presença de dor, condicionada com a variável independente, apresenta-se os valores das razões de possibilidade.

A intensidade da relação entre as variáveis qualitativas faz-se usando medidas de associação. A razão de probabilidade mede a ocorrência de um acontecimento em relação a outro. Ou de forma análoga, mede a associação entre duas variáveis nominais, em que uma das variáveis pode ser designada por factor, e a outra por acontecimento.

Está-se assim interessado a testar as seguintes hipóteses:

$$H_0 : "\textit{as var iáveis são independentes, isto é, razão de possibilidade igual a } 1"$$

$$vs$$

$$H_1 : "\textit{Existe uma relação de dependência entre as var iáveis, isto é, razão de possibilidade é diferente de } 1"$$

Apresentam-se de seguida uma tabela para as referidas variáveis, assim como os respectivos intervalos de confiança, a 95%, e os valores de razão de possibilidades.

Região da Queixa	***Razão de possibilidades***	**I.C. 95%**
Região cervical	1.045	[0.377; 2.897]
Ombros	0.769	[0.310; 1.910]
Cotovelos	0.611	[0.156; 2.395]
Punho /mão	0.972	[0.376; 2.573]
Coluna vertebral: zona dorsal	1.468	[0.585; 3.681]
Coluna vertebral: zona lombar	3.185	[1.256; 8.0759]
Coxas	0.782	[0.212; 2.883]
Joelhos	0.601	[0.165; 2.188]
Tornozelos	1.412	[0.426; 4.678]

Tabela 3- *Valores das estimativas das razões de possibilidade e respectivos I.C.(95%)*

A zona que parece mais pertinente para análise, devido ao seu valor da razão de possibilidades, é a zona lombar.

Com 95% de confiança, e com um valor de razão de possibilidade de 3.185 e um respectivo intervalo de confiança de [1.256; 8.0759] pode-se concluir que os enfermeiros que prestam

apoio ao domicílio têm uma possibilidade três vezes maior no que diz respeito a virem a ter queixas de LME na região lombar, comparativamente com os que não prestam apoio ao domicílio.

3.3. MODELOS DE REGRESSÃO LOGÍSTICA PARA A REGIÃO LOMBAR

Na análise de regressão a variável dependente pode ser influenciada pela presença de variáveis quantitativas e qualitativas. As primeiras, podem facilmente ser transformada noutra escala o que não acontece com as variáveis qualitativas.
Um método para se qualificar os atributos é construir variáveis artificiais que assumam valores compreendidos entre 0 e 1, sendo estas variáveis conhecidas como variáveis "*dummy*".

Após se verificar que a zona lombar apresenta maior evidência que relacione este tipo de queixas com alguns factores associados a estes profissionais, pretende-se construir um modelo usando regressão logística.
Neste sentido, e filtrando apenas os enfermeiros que prestam apoio ao domicílio, considerou-se como variável dependente "Dor na Lombar", e como variáveis independentes V1 a V54 (Codificadas em anexo- Tabela A1).

Com objectivo de verificar se as variáveis acima descritas permitem construir um modelo de regressão logística, começou-se pela construção de modelos univariados, como foi referido em *2.3.2.*

Encontram-se de seguida registados os resultados, após aplicação de regressão logística univariada, considerando um intervalo para base [0.20; 0.25], tendo em conta o que foi dito em *2.3.3.2. (B).*

Variável	$\hat{\beta}$	*S.E*	*Wald*	*Sig.*	*Exp(β)*
V1	0.930	0.665	1.959	0.162	2.535
V2	-0.476	0.444	1.147	0.284	0.621
V3	1.046	0.473	4.901	0.027	2.846
V4	0.565	0.682	0.697	0.407	1.760
V5	0.047	0.027	3.124	0.077	1.048
V6	0.051	0.028	3.258	0.071	1.053
V7	0.016	0.015	1.081	0.299	1.016
V8	0.081	0.027	0.002	0.969	1.001
V9	1.4	0.429	10.664	0.001	4.057
V10	1.045	0.409	6.513	0.011	2.842
V11	0.208	0.706	0.086	0.769	1.231
V12	0.997	0.473	4.436	0.035	2.709
V13	0.704	0.394	3.184	0.074	2.021
V14	1.156	0.657	3.097	0.078	3.176
V15	0.787	0.541	2.118	0.146	2.197
V16	1.363	0.78	3.057	0.08	3.908
V17	0	0.028	0	0.99	1
V18	0.139	0.167	0.694	0.405	1.150
V19	0.187	0.208	0.806	0.369	1.206
V20	-0.149	0.138	1.157	0.282	0.862
V21	-0.04	0.144	0.076	0.783	0.961
V22	0.015	0.188	0.007	0.935	1.015
V23	0,919	0.515	3.180	0.075	2.506
V24	0.277	0.388	0.511	0.475	1.315
V25	1.079	0.642	2.824	0.093	2.941
V26	1.119	0.607	3.398	0.065	3.062
V27	-0.125	0.559	0.05	0.823	0.882
V28	0.249	0.185	1.802	0.179	1.282
V29	0.539	0.528	1.044	0.307	1.714
V30	0.077	0.08	0.930	0.335	1.080
V31	1.281	0.667	3.686	0.055	3.6
V32	0.105	0.211	0.25	0.617	1.111
V33	0.965	0.593	2.652	0.103	2.625
V34	0.163	0.29	0.315	0.575	1.177

V35	0.189	0.172	1.206	0.272	1.208
V36	0.178	0.123	2.097	0.148	1.195
V37	0.699	0.572	1.495	0.221	2.012
V38	1.211	0.561	4.657	0.031	3.357
V39	0.325	0.505	0.416	0.519	1.385
V40	-2.254	0.863	6.831	0.009	0.105
V41	0.616	0.701	0.773	0.379	1.851
V42	0.061	0.159	0.145	0.704	1.062
V43	-0.322	0.19	2.874	0.09	0.725
V44	-0.222	0.389	0.327	0.569	0.801
V45	-0.152	1.24	1.511	0.291	0.218
V46	0.335	0.273	1.501	0.221	1.348
V47	0.271	0.296	0,841	0.359	1.311
V48	0.352	0.325	1.178	0.278	1.423
V49	1.227	0.452	7.380	0.007	3.41
V50	0.347	0.465	0.556	0.456	1.415
V51	0.191	0.494	0.149	0.7	1.21
V52	1.387	0.577	5.788	0.016	4.004
V53	-1.161	0.581	3.992	0.046	0.313
V54	-0.283	0.493	0.33	0.565	0.753

Tabela 4- *Valores relevantes da aplicação de regressão logística univariada.*

Na tabela 4 encontram-se destacados os valores prova que permitiu seleccionar as variáveis candidatas e explicativas para o modelo.

Resumindo, tem-se as seguintes variáveis candidatas:

- Outra actividade profissional;
- LME antiga;
- Idade;
- Antiguidade profissão;
- Dor Cervical;
- Dor Ombros;
- Dor Punhos;

- Dor Dorsal;
- Dor Coxas;
- Dor Tornozelo;
- Pulsos: torção lateral;
- Pulsos: postura;
- Pulsos: movimentos repetitivos;
- Pulsos: força;
- Tronco: Torção lateral;
- Pescoço: torção lateral;
- Figura braços;
- Braços: Abdução;
- Braços: rotação;
- Apoio braço;
- Tempo da actividade;
- Auxiliares de movimentação;
- Espaço caracterização;
- Altura cama/sofá;
- Stress;
- Satisfação com o trabalho.

Tendo em conta esta lista de variáveis, o próximo passo consiste em construir um modelo aplicando-se o princípio de regressão logística multivariada, para isso recorreu-se aos métodos implementados no SPSS (e descritos na secção *2.3.3.2.1*).

Em regressão logística não existe um modelo único, neste sentido, e com o objectivo de se encontrar o modelo que mais se ajuste foram construídos quatro modelos possíveis:

(1) Modelo 1 $= 1.763 \times V26 + 0.843 \times V36 - 2.877 \times V40 - 2.464 \times V53 - 0.504 \Rightarrow$ *Forward Stepwise*

(2) Modelo 2 $= 0.336 \times V6 + 1.851 \times V12 + 2.02 \times V14 + 1.702 \times V26 + 0.414 \times V33$
$+ 0.74 \times V36 + 0.618 \times V43 - 4.05 \times V40 + 0.89 \Rightarrow$ *Backward Condicional*

(3) Modelo 3 $= 2.235 \times V12 + 1.899 \times V26 + 1.751 \times V33 - 2.781 \times V40 - 2.089 \Rightarrow$ *Forward Stepwise*

(4) Modelo 4 $= 2.226 \times V3 + 0.006 \times V5 + 0.608 \times V43 + 0.116 \times V45 + 0.983 \times V53 + 3.470 \Rightarrow$ *Forward Stepwise*

No modelo 1 entraram todas a variáveis que foram seleccionadas após aplicação de regressão logística univariada e aplicando-se o método *Forward Stepwise*. Para este modelo obtiveram-se quatro variáveis explicativas: Movimentos repetitivos dos pulsos; Caracterização da figura relativa aos braços; Braços encontram-se em apoio, havendo suporto do seu peso e Satisfação do trabalho.

No modelo 2 usou-se a técnica *Backward Condicional*, que incluiu no modelo 8 variáveis explicativas: Antiguidade na profissão; Dor nos punhos; Dor nas Coxas; Movimentos repetitivos dos pulsos; Torção lateral do pescoço; Caracterização da figura relativa aos braços; Braços encontram-se em apoio, havendo suporto do seu peso e Tempo de desempenho da actividade.

Para os dois últimos modelos apresentados foram retiradas algumas das variáveis candidatas, aplicando-se o método *Forward Stepwise.*
Isto, segundo um critério pessoal, uma vez que se achou interessante verificar o que acontecia se se retirasse da análise algumas variáveis e se considera-se apenas as que parecessem ter relação, como exemplo do modelo 4, em que foram retiradas da análise as variáveis referentes a dores e zonas corporais.

Para o modelo 3 entraram como variáveis candidatas: Idade; Postura dos pulsos; Movimentos repetitivos dos pulsos; Força dos pulsos; Torção lateral do tronco; Torção lateral do pescoço; Figura dos braços; Abdução dos braços; Rotação dos Braços; Apoio dos Braços; Tempo de Actividade; Dor na cervical; Dor nos ombros, Dor nos punhos; Dor na dorsal; Dor nas coxas e Dor nos tornozelos. No final, destas variáveis, apenas se obteve

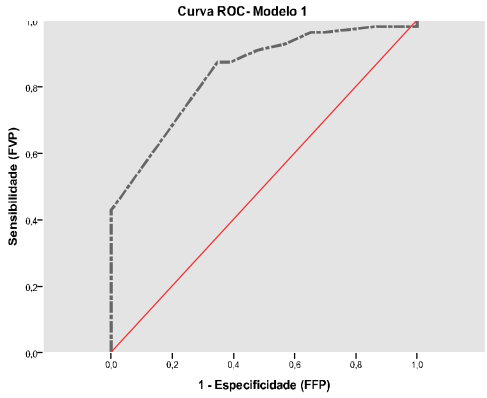
como explicativas para este modelo 4 das variáveis. (Movimentos repetitivos dos pulsos; Torção lateral do pescoço; Braços encontram-se em apoio, havendo suporto do seu peso e Dor nos punhos).

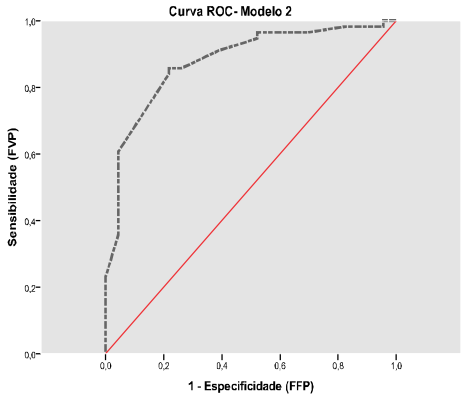
Por último, para construção do modelo 4, as variáveis candidatas utilizadas foram: LME antiga; Outra actividade; Idade; Antiguidade na profissão; Tempo de actividade, Altura da cama/ sofá; Auxiliares de movimentação; *Stress* e Satisfação. Destas, a técnica *Forward Stepwise* seleccionou as LME antigas; Idade; Tempo de actividade; Auxiliares de movimentação e Satisfação do trabalho.

3.4. ANÁLISE DOS RESIDUOS ATRAVÉS DA CURVA ROC

3.4.1. Representação da Curva ROC

Com os modelos de regressão construídos, apresenta-se as respectivas curvas ROC, com o objectivo de avaliar em termos de desempenho, qual dos modelos melhor se ajusta para explicar a Dor lombar nos enfermeiros que prestam apoio ao domicílio.
De referir que a curva ROC será usada como alternativa para avaliar o diagnóstico dos modelos gerados $\hat{\pi}_i$ *vs* y_i.





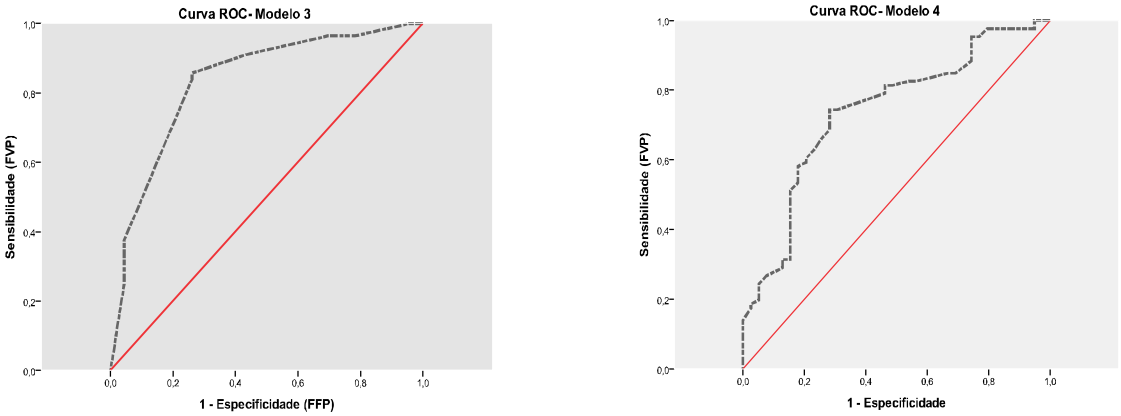


Gráfico 5- *Curva ROC para os quatro modelos*

Para cada uma das curvas apresentadas obteve-se o valor da área e os respectivos intervalos de confiança a 95%.

	Área	***Std. Error***	**I.C.**
Modelo1	0.844	0.045	[0.756; 0.932]
Modelo 2	0.878	0.042	[0.796; 0.961]
Modelo 3	0.834	0.053	[0.731; 0.937]
Modelo 4	0.74	0.048	[0.647; 0.834]

Tabela 5- *Resultados curva ROC*

Através da análise da curva ROC verificou-se que para todos os modelos que o valor da área está acima de 0.7, e com um erro padrão associado não superior a 0.053, o que significa que têm um bom poder discriminante no que concerne à avaliação das queixas referentes à região lombar por parte dos enfermeiros que prestam apoio ao domicílio.

3.4.2. Comparação de modelos com recurso ao teste da área abaixo da curva ROC

Neste ponto, vai-se proceder a comparações dos modelos dois a dois.

Está-se interessado a testar as seguintes hipóteses:

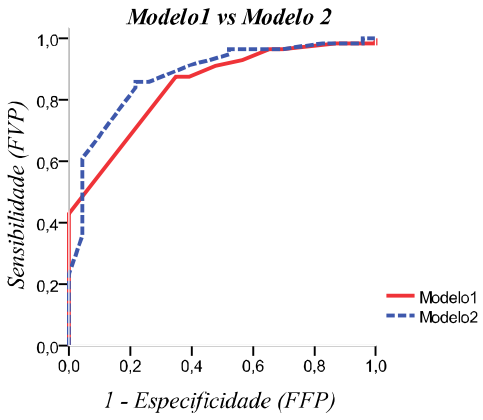
$$H_0 : A_1 - A_2 = 0 \quad vs \quad H_1 : A_1 - A_2 \neq 0 .$$

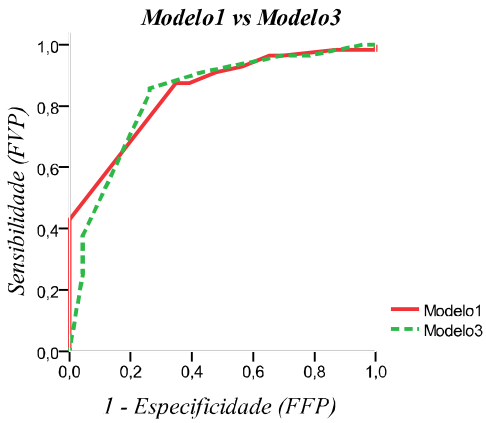
Ou seja, queremos verificar a igualdade das áreas entre os dois modelos. Isto é, vai-se comparar

$$A_1 \textit{ versus } A_2$$
$$A_1 \textit{ versus } A_3$$
$$A_1 \textit{ versus } A_4$$
$$A_2 \textit{ versus } A_3$$
$$A_2 \textit{ versus } A_4$$
$$A_3 \textit{ versus } A_4$$

Os resíduos gerados são independentes, assim a correlação entre as áreas é nula.

Apresenta-se de seguida as representações das curvas ROC dos modelos que serão comprados dois a dois.





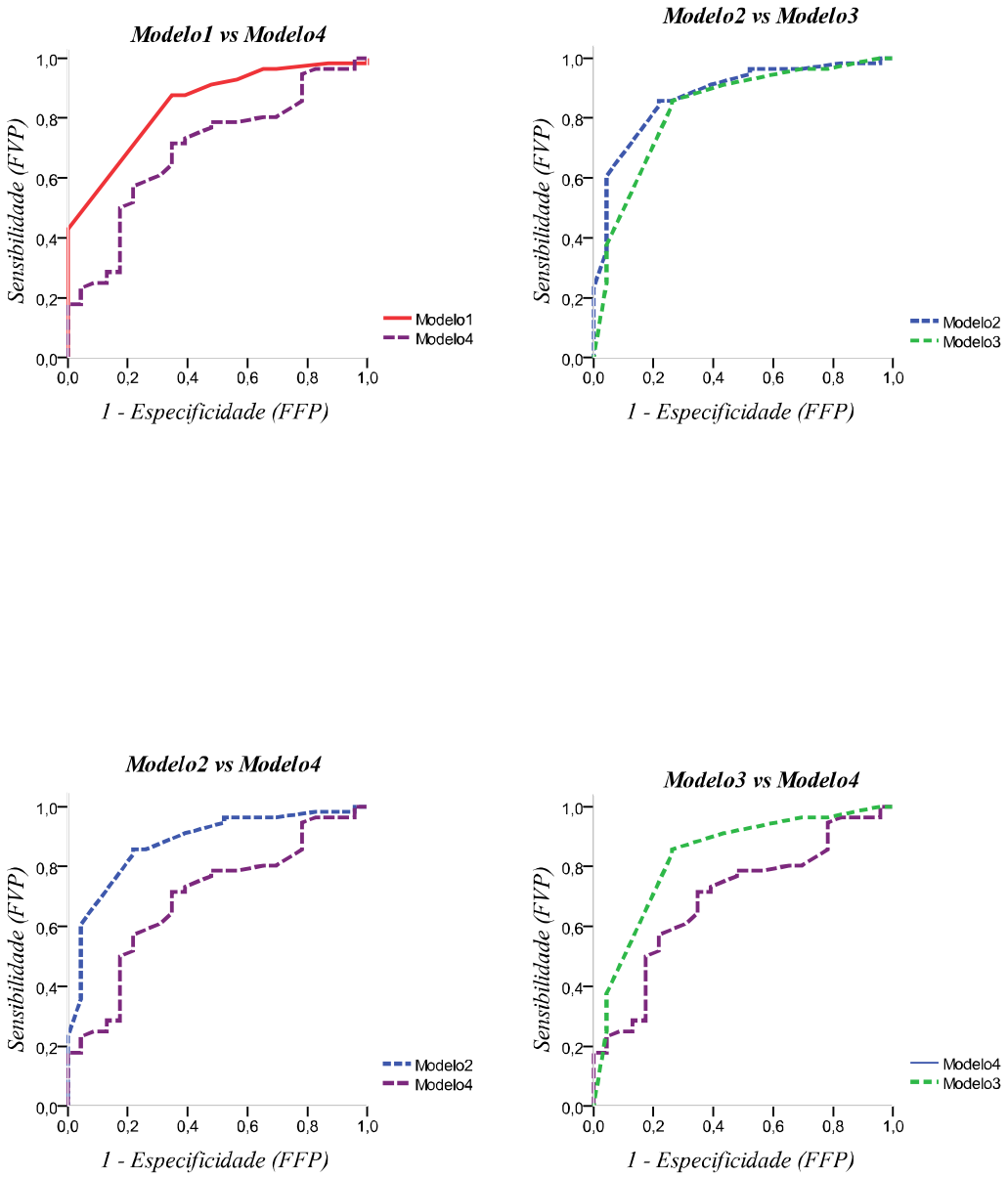


Gráfico 6- *Curvas ROC dos modelos dois a dois*

Indo ao encontro do que foi dito em *2.4.5.*, e aplicando a fórmula (25), tem-se para a comparação entre os quatro modelos a tabela resumo:

	M2	M3	M4
M1	$z_1 = \frac{A_1 - A_2}{\sqrt{SE(A_1)^2 + SE(A_2)^2}} = -0.55$	$z_2 = \frac{A_1 - A_3}{\sqrt{SE(A_1)^2 + SE(A_3)^2}} = 0.14$	$z_3 = \frac{A_1 - A_4}{\sqrt{SE(A_1)^2 + SE(A_4)^2}} = 1.581$
M2		$z_4 = \frac{A_2 - A_3}{\sqrt{SE(A_2)^2 + SE(A_3)^2}} = 0.65$	$z_5 = \frac{A_2 - A_4}{\sqrt{SE(A_2)^2 + SE(A_4)^2}} = 2.164$
M3			$z_6 = \frac{A_3 - A_4}{\sqrt{SE(A_3)^2 + SE(A_4)^2}} = 1.31$

Tabela 6- *Valores de z relativos a comparações de áreas*

Como *z* segue uma distribuição normal padrão recorreu-se ao ***R***, para se obter o valor prova associado a cada *z*.

z_1=1-pnorm(-0.55,lower.tail=T)
z_2=pnorm(0.14,lower.tail=F)
z_3=pnorm(1.581,lower.tail=F)
z_4=pnorm(0.65,lower.tail=F)
z_5=pnorm(2.164,lower.tail=F)
z_6=pnorm(1.31,lower.tail=F)

Obtendo-se os seguintes valores prova, respectivos a cada um dos *z*. Salienta-se os valores obtidos no ***R*** têm de ser multiplicados por 2, uma vez que temos um teste bilateral.

	z_1	z_2	z_3	z_4	z_5	z_6
Valor prova	**0.582**	**0.89**	**0.114**	**0.516**	0.032	**0.19**

Tabela 7- *Valores prova associados à região crítica z*

Por análise dos valores provas associados aos vários valores de *z*, e tomando como base de comparação para regra de decisão 0.05, rejeita-se a hipótese de igualdade de áreas apenas para o valor de z_5.

Como $p-value(z_5)=0.032<0.05$, rejeita a hipótese de igualdades de áreas para z_5, o que significa que se detectaram diferenças significativas entre as áreas do modelo 2 e modelo 4.

Neste sentido, não se detectaram diferenças significativas entre os modelos:

$$Modelo1-Modelo2$$
$$Modelo1-Modelo3$$
$$Modelo1-Modelo4$$
$$Modelo2-Modelo3$$
$$Modelo3-Modelo4$$

Tendo em conta estes resultados e o valor das áreas apresentadas para cada modelo, tomou-se o modelo 2 como o sendo explicativo das LME dos enfermeiros que prestam apoio ao domicílio, associado à dor na lombar.

$$\textbf{\textit{Modelo 2}} = 0.336\times V6+1.851\times V12+2.02\times V14+1.702\times V26+0.414\times V33 + 0.74\times V36+0.618\times V43-4.05\times V40+0.89 \Rightarrow \textit{Backward Condicional}$$

Codificação das variáveis explicativas do modelo:

Variáveis	**Codificação**
V6-Antiguidade na profissão	Expressa em anos
V12-Dor punho	Respondia apenas pelos profissionais que prestam apoio ao domicílio. 0-Não; 1-Sim
V14- Dor nas Coxas	Respondia apenas pelos profissionais que prestam apoio ao domicílio. 0-Não; 1-Sim
V26-Pulsos:movimentos repetitivos	Respondia apenas pelos profissionais que prestam apoio ao domicílio. 0-Não; 1-Sim

V33-Pescoço: torção lateral	Respondia apenas pelos profissionais que prestam apoio ao domicílio. 0-Não; 1-Sim
V36- Figura braços	Respondia apenas pelos profissionais que prestam apoio ao domicílio. 1-20º extensão a 20ºflexao; 2- >20º extensão;3- 20º a 45º flexão; 4-45º a 90º flexão;5- > 90º flexão
V40-Apoio do braço	Respondia apenas pelos profissionais que prestam apoio ao domicílio. 0-Não; 1-Sim
V43- Tempo de actividade	Respondia apenas pelos profissionais que prestam apoio ao domicílio. 0- Nunca; Raramente; 1- Com alguma frequência; Frequentemente; Sempre

Tabela 8: Codificação das variáveis explicativas do modelo

CAPÍTULO 4- CONCLUSÃO & TRABALHOS FUTUROS

4.1. CONCLUSÕES

O tamanho da amostra não permite uma inferência sobre a população, no entanto podemos caracterizá-la em diversos aspectos.

Há uma associação estatisticamente significativa mais elevada entre "queixas músculo esqueléticas na região lombar" e "prestação de cuidados ao domicílio"(com 95% de confiança $[1.256;8.076]$, *razão de possibilidade* $= 3.185\ (p - value < 0.05)$). Ou seja, os enfermeiros que prestam apoio ao domicílio têm cerca de três vezes mais possibilidade de ter queixas músculo-esquelécticas na região lombar do que os outros enfermeiros.

Após se verificar que a zona lombar é a zona que apresenta maior evidência para relacionar este tipo de queixas com alguns factores associados com estes profissionais, construiu-se modelos de regressão logística. Estes modelos foram construídos utilizando os métodos implementados em SPSS, obtendo-se quatro modelos distintos:

- O modelo 1 com quatro variáveis explivativas (Movimentos repetitivos dos pulsos; Caracterização da figura relativa aos braços; Braços encontram-se em apoio, havendo suporto do seu peso e Satisfação do trabalho);
- O modelo 2 com oito variáveis explicativas (Antiguidade na profissão; Dor nos punhos; Dor nas Coxas; Movimentos repetitivos dos pulsos; Torção lateral do pescoço; Caracterização da figura relativa aos braços; Braços encontram-se em apoio, havendo suporto do seu peso e Tempo de desempenho da actividade);
- O modelo 3 com quatro variáveis explicativas (Movimentos repetitivos dos pulsos; Torção lateral do pescoço; Braços encontram-se em apoio, havendo suporto do seu peso e Dor nos punhos);
- O modelo 4 com quatro variáveis explicativas (LME antigas; Idade; Tempo de actividade; Auxiliares de movimentação e Satisfação do trabalho).

Através da análise da curva ROC obteve-se as curvas referentes a cada modelo, assim como as respectivas áreas, erros padrão e intervalos de confiança a 95%.
Uma vez construídos quatro possíveis modelos, através da comparação de áreas abaixo da curva ROC, seleccionou-se o modelo 2.

Verificou-se que o valor abaixo da curva ROC é de 0.844 com um erro padrão associado de 0.042 e com um intervalo de confiança a 95% de [0.796; 0.961] o que indica um bom poder discriminante, no que concerne à avaliação das queixas referentes à região lombar relativamente aos enfermeiros que prestam apoio ao domicílio. O valor 0.844 significa ainda que 84.4% dos casos do modelo *logit* apresentado acerta na predição das queixas de dor na região lombar para esses enfermeiros.
Neste sentido, as variáveis explicativas para as LME dos enfermeiros que prestam apoio ao domicílio são: Antiguidade na profissão; Dor nos punhos; Dor nas Coxas; Movimentos repetitivos dos pulsos; Torção lateral do pescoço; Caracterização da figura relativa aos braços; Braços encontram-se em apoio, havendo suporto do seu peso e Tempo de desempenho da actividade.

4.2. SUGESTÕES PARA TRABALHOS FUTUROS

Nesta secção deixa-se algumas sugestões para possíveis trabalhos futuros.

Com o decorrer da análise dos resultados teve-se a percepção que se poderia ter realizado uma análise mais detalhada, assim como comparação dos resultados utilizando *softwares* diferentes.

Neste sentido deixa-se uma lista de possíveis sugestões:

- Uma vez que em regressão logística a obtenção dos modelos não é única, poderia ter-se obtido uma lista mais elevada de modelos, incluindo ou excluindo novas variáveis;
- Comparação dos resultados utilizando o SPSS e o ***R***, com objectivo de verificar se se obteria resultados semelhantes, e consequentemente, as mesma conclusões;

- Uma análise dos resíduos, para identificar possíveis elementos para os quais se verifique um maior afastamento entre o valor predito e o valor da variável resposta;
- Realizar um estudo das interacções entre as variáveis em análise;

ANEXOS

1. Tabela de codificação das variáveis:

Variáveis	Código
Outra actividade profissional	**V1**
Praticou desporto	**V2**
LME antiga	**V3**
Sexo	**V4**
Idade	**V5**
Antiguidade profissão	**V6**
Peso	**V7**
Altura	**V8**
Dor cervical	**V9**
Dor ombros	**V10**
Dor cotovelos	**V11**
Dor punhos	**V12**
Dor dorsal	**V13**
Dor coxas	**V14**
Dor joelhos	**V15**
Dor tornozelo	**V16**
Horas semanais de apoio ao domicílio	**V17**
Distribuição semanal do apoio ao domicílio	**V18**
Trabalho domiciliário efectuado	**V19**
Actividades ao domicílio	**V20**
Figura antebraço	**V21**
Figura pulsos	**V22**
Pulsos: torção lateral	**V23**
Pulsos: pega	**V24**
Pulsos: postura	**V25**
Pulsos: movimentos repetitivos	**V26**

Pulsos: acções rápidas	**V27**
Pulsos: força	**V28**
Pulsos: choque	**V29**
Figura tronco	**V30**
Tronco: torção lateral	**V31**
Figura pescoço	**V32**
Pescoço: torção lateral	**V33**
Figura pernas	**V34**
Flexão joelhos	**V35**
Figura braços	**V36**
Braços: abdução	**V37**
Braços: rotação	**V38**
Elevação ombro	**V39**
Apoio braço	**V40**
Movimentação paciente	**V41**
Colaboração de um colega	**V42**
Tempo da actividade	**V43**
Pausas	**V44**
Auxiliares de movimentação	**V45**
Caracterização do espaço	**V46**
Espaço disponível	**V47**
Arrumação	**V48**
Altura cama/sofá	**V49**
Dependência do paciente	**V50**
Trato paciente	**V51**
Stress	**V52**
Satisfação com o trabalho	**V53**
Ansioso/ irritavel	**V54**

Tabela A1- *Codificação das variáveis.*

BIBLIOGRAFIA

Athayde, M. (2005). *Estatística. R.* Braga: Publicado pelo Departamento de Matemática da Universidade do Minho.

Bandeira, M. (s.d.). Texto 10: Análise de Dados, cronograma, orçamentos, pertinência, considerações éticas. Departamento de Psicologia- UFSJ.

Bessa, J. (2007). Selecção das Fontes de Dados e Participantes, elaboação do protocolo para colheitas, processamento e análise de dados.

Braga, A. (1994). *Acidente Vascular Cerebral e seus Factores de Risco. Estudo de ocorrência de quatro tipos de AVC.* Tese de Mestrado, Universidade do Minho.

Braga, A. (2000). *Curva ROC: Aspectos fundamentais e Analiação.* Braga: Tese de Doutoramento, Universidade do Minho.

Braga, A. (s.d.). Mini- curso: Curvas ROC. *Tardes de Estatística e Investigação Operacional*.

Carneiro, P. (2005). *Análise ergonómica da postura e dos movimentos na profissão de médicos dentistas.* Universidade do Minho, Tese de Mestrado.

Crichton N. (2001). Wald test; Ó 2001 Blackwell Science Ltd. *Jornal of Clinical Nursin*, www.blackwellpublishing.com/specialarticles/jcn_10_774.pdf.

Direção Geral de Saúde. (2008). Lesões Musculoesqueléctricas Relaciondas com o Trabalho. Programa Nacional contra as doenças reumáticas, Guia de Orientação para a prevenção.

Hosmer, D. J., & Lemeshow, S. (1989). *Applied Logistic Regression.* Copyright by John Wiley & Sons, Inc.

Laureano, M., & Botelho, M. (2010). *SPSS o meu manual de consulta rápida.* Lisboa: Edições Sílabo, Lda, 1ª Edição.

Margotto, P. (s.d.). Curva ROC: Como fazer e Interpretar no SPSS. *Professor do Curso de Medicina da Escola Superior de Ciências da Saúde (ESCS/CES/DF).*

Martinez, E., Neto-Louzada, F., & Pereira, B. (2003). A Curva ROC para testes diagnósticos. Rio de Janeiro, **11 (1):**7-31.

Martins, P. S. (2008). *Análise estatística de performance de um conjunto de testes auditivos.* Tese de Mestrado, Universidade de Aveiro.

Misso, F., & Jacobi, L. F. (2007). *Variáveis dummy:especificações de modelos com parâmetros variáveis.* Brasil: Ciências e Natura, UFSM.

Pestana, M., & Gageiro, J. (2005). *Análise de dados para Ciências Sociais- A Complementaridade do SPSS.* Lisboa: Editora Sílibo, 4ª Edição.

Rumel, D. (1986). Odds ratio-algumas considerações. *Revista de Saúde Publica V.20 n3.* Brasil: Departamento de Epidemiologia da Faculdade de Saúde Publica da Universidade de São Paulo.

SPSS Inc. (2007). *SPSS Regression 17.0.* Obtido em Setembro de 2011, de http://www.helsinki.fi/~komulain/Tilastokirjat/IBM-SPSS-Spec-Regression.pdf

Páginas da Internet ***(consultadas no período de Janeiro a Outubro de 2011)***

www.tecminho.uminho.pt

www.empreendedorismo.uac.pt/spinofss/spinoffs_academicos

www.pse.pt

http://portalaction.com.br

http://www.ensp.unl.pt