Universidade de Lisboa
Faculdade de Ciências
Departamento de Estatística e Investigação Operacional

Aplicação do Modelo de Regressão Logística num Estudo de Mercado

Cleidy Isolete Silva Cabral

Projeto

Mestrado em Matemática Aplicada à Economia e à Gestão

2013

Universidade de Lisboa

Faculdade de Ciências

Departamento de Estatística e Investigação Operacional

Aplicação do Modelo de Regressão Logística num Estudo de Mercado

Cleidy Isolete Silva Cabral

Projeto

Mestrado em Matemática Aplicada à Economia e à Gestão

Orientador:

João José Ferreira Gomes

2013

AGRADECIMENTOS

Agradeço à minha família, especialmente aos meus pais, Aguinaldo Cabral e Eurisanda Silva, que sempre me apoiaram e acreditaram em mim. Aos meus irmãos, Christienne Cabral, Cellice Cabral, Amin Lopes e Diego Cabral, pois são eles as pessoas mais importantes da minha vida.

Ao CEFAR, agradeço pelos dados disponibilizados, sem os quais não seria possível realizar este trabalho. À Dra. Suzete Costa por ter autorizado este estudo.

Agradeço aos meus orientadores de estágio, Zilda Mendes e João Gomes, pelas sugestões, pela disponibilidade e por todo o apoio prestado.

Agradeço também a todos os membros do CEFAR, nomeadamente à Patrícia Ferreira e ao José Pedro Guerreiro que sempre se prontificaram a ajudar e a esclarecer quaisquer dúvidas que eventualmente apareciam. Agradeço à Ana Paula David, Sonia Romano e Paulo Carvalhas pelo carinho e atenção. Agradeço ainda à Marta Gomes por toda a amizade.

Agradeço às minhas colegas de casa, Edilina Pinheiro, Silvana Oliveira e Zelinda Santos, e ao meu namorado, Stivon Silva, que foram a minha segunda família e o meu apoio desde que cá estou.

PREFÁCIO

Este trabalho consiste num relatório de estágio, para obtenção do grau de mestre em Matemática Aplicada à Economia e Gestão, pela Faculdade de Ciências da Universidade de Lisboa. Estágio este realizado no Centro de Estudos e Avaliação em Saúde (CEFAR) sob a orientação de Zilda Mendes.

No presente trabalho será apresentado um estudo de mercado realizado pelo CEFAR, onde o principal objectivo é avaliar a posição da marca em relação à concorrência no mercado de venda de produtos para o tratamento de Eczema/Eczema atópico e Psoríase.

Será aplicado um modelo de regressão logística onde a variável resposta é o consumo, ou não, de produtos de uma determinada marca, identificando o conjunto de variáveis que melhor explicam a utilização desses produtos.

Este trabalho é composto por 4 capítulos distintos. No primeiro capítulo, faremos uma breve introdução ao estudo de mercado, ilustrando o objectivo e as áreas de aplicação desse estudo. Seguidamente, apresentaremos o CEFAR, expondo a metodologia que utilizaram para a recolha de informação, a dimensão da amostra e as variáveis em estudo.

No capítulo seguinte, introduziremos a análise de regressão logística univariada e posteriormente a regressão logística multivariada.

Iniciaremos, em seguida, a análise prática, fazendo a análise descritiva dos dados, comparando os utilizadores dos produtos da concorrência com os utilizadores dos produtos da marca. Após esta análise, aplicar-se-á um modelo de regressão logística utilizando o método stepwise para a seleção das covariáveis de interesse.

Por fim, serão apresentadas as conclusões do estudo.

Palavras-chave: Modelo de Regressão Logística, Estudo de Mercado, Método Stepwise, Eczema Atópico, Psoríase

PREFACE

This internship report was developed in order to obtain the Master Degree in "Matemática Aplicada à Economia e à Gestão", at Faculdade de Ciências, Universidade de Lisboa. This internship was carried out in Centre for Health Evaluation & Research (CEFAR), under the supervision of MSc Zilda Mendes.

Throughout this report it will be described a market research performed by CEFAR, where the main purpose was to assess the market position, of a determined brand, towards competitors in products for the treatment of Eczema/ Atopic Eczema and Psoriasis.

This report focuses on the utilization of Logistic Regression Methods where the outcome variable is binary, consumption or not of products of a determined brand, identifying the best fitting model to describe the relationship between the outcome variable and the set of independent variable.

This report is organized into four different chapters. In the first one, a brief introduction of market research, showing the purpose and application areas of these studies. Afterwards CEFAR was introduced, showing methods used for data collection, sample size estimation and definition of variables under study.

In the second chapter was introduced logistic regression model in the univariate context and then for the multivariate case.

Through the third chapter the practical performance was described, making a descriptive analysis of the data, comparing competitive products users with the brand users. After this analysis, model of logistic regression was applied using the stepwise method for selection of covariates of interest.

Finally, in the fourth and last chapter the study final conclusions were reported.

Keywords: Logistic Regression Model, Market Research, Stepwise Method, Psoriasis, Atopic Eczema

ÍNDICE

ÍNDICE DE TABELAS

ÍNDICE DE FIGURAS

CAPÍTULO I
INTRODUÇÃO

CAPÍTULO I - INTRODUÇÃO

1. ESTUDO DE MERCADO

Qualquer empresa que se orienta numa ótica de marketing necessita de um conjunto de dados e/ou indicadores quer sobre o próprio mercado, quer sobre os consumidores/clientes e a concorrência a fim de planear, controlar e executar todo o processo de gestão e/ou delinear estratégias a prosseguir. A palavra-chave para um bom desempenho e concretização das funções de marketing é a informação. Por isso é imprescindível para qualquer empresa orientada na ótica do marketing, a implementação de um Sistema de Informação de Marketing.

Segundo Kotler, o Sistema de Informação de Marketing (SIM) pode ser definido como:
A ação conjunta de pessoas, equipamentos e métodos visando a recolha, tratamento, análise, avaliação e distribuição em devido tempo de toda a informação atualizada e necessária para a tomada das mais adequadas decisões de marketing. (Lopes, 2007)

Um Sistema de Informação de Marketing concretiza ou materializa as suas funções através de 4 vertentes: dados internos, notícias de marketing, suporte das decisões de marketing e estudos de mercado. Destes, o estudo do mercado é o mais conhecido e mais aplicado pela gestão de marketing.

Define-se estudo de mercado como um processo sistemático e objetivo de recolha e consequente fornecimento da informação necessária e indispensável para a tomada de decisões por parte das gestão/direção de marketing.

Sintetizando, os estudos de mercado auxiliam o marketing não só a detetar e avaliar qualitativa e quantitativamente as necessidades e/ou preferências dos consumidores, o impacto das ações de marketing levadas a cabo e ainda os hábitos e comportamentos dos consumidores face a todas as variáveis do *marketing-mix.*

As áreas de aplicação do estudo de mercado são o Mercado, o Consumidor e a Concorrência.

Este estudo foca tanto o consumidor como a concorrência. No que diz respeito ao consumidor, pretende-se conhece-lo em pormenor de modo a permitir a sua mais adequada segmentação e caracterização. Quanto á concorrência pretende-se responder a questões como: Que implantação têm? Que grau de preferência têm junto do consumidor? Como se comportam relativamente aos nossos? Quais os seus pontos fortes e fracos?

2. CEFAR - CENTRO DE ESTUDOS E AVALIAÇÃO EM SAÚDE

2.1. INTRODUÇÃO

O Centro de Estudos e Avaliação em Saúde (CEFAR) é um departamento da Associação Nacional das Farmácias (ANF) fundada em Janeiro de 1994. As áreas de especialização desse departamento são a investigação e consultoria científica sobre o medicamento e saúde. O CEFAR desenvolve estudos nas áreas de farmacoepidemiologia, estatística aplicada à saúde, economia da saúde, observatório da farmácia e outcome research. Para além destas áreas também realiza estudos de mercado - foi neste âmbito que decidi fazer este relatório de estágio baseando-me num estudo de mercado no qual eu também tomei parte.

O principal objetivo desse estudo foi avaliar o perfil de consumidores de produtos para o tratamento de Eczema/Eczema atópico e Psoríase.

O Eczema/Eczema atópico é uma doença de pele caracterizada pela inflamação e comichão. Manifesta-se habitualmente nos primeiros meses de vida e mais raramente em adultos jovens.

À tendência para este tipo de doenças chama-se atopia. Os atópicos têm a pele seca e áspera. Caracteristicamente sentem prurido (comichão) quando transpiram facto que, aliado à já referida secura cutânea e a uma típica diminuição do seu limiar de sensibilidade para o prurido, os leva a coçar permanentemente o corpo.

A Psoríase é uma doença inflamatória que se manifesta principalmente na pele, mas que também pode atingir outras áreas, como as articulações ou as unhas. É relativamente frequente, afetando cerca de 2% da população. Surge frequentemente no adulto jovem, mas pode aparecer em qualquer idade.

2.2. METODOLOGIA

Para a recolha de informação foi elaborado um questionário a ser aplicado nas farmácias, nos meses de Março a Junho de 2013.

Neste questionário, pretendia-se, antes de mais, obter informações sobre o nome do produto adquirido e o respetivo código (CNP).

Para obter informações acerca do utilizador, foram introduzidas as variáveis sexo, idade, primeira vez, diagnóstico que motivou a compra e o motivo da compra.

Caso o motivo da compra fosse especialidade médica, era pedido ao utilizador que indicasse a especialidade do médico prescritor e a origem da prescrição.

Para os utilizadores que adquiriram o produto anteriormente, pedia-se que indicassem o motivo da compra anterior. Caso esse motivo fosse prescrição médica, pedia-se também a especialidade do médico.

Por último, pretendia-se ainda saber se houve substituição de produtos da marca por produtos da concorrência e caso tal ocorresse, pedia-se o nome do produto.

Enviou-se um convite a 1500 Farmácias filiadas da ANF das quais, 200 enviaram informação que resultou numa amostra de 1786 questionários válidos para análise.

Esse questionário foi dirigido a todos os doentes que se deslocassem à farmácia para adquirir um dos seguintes produtos para o tratamento de Eczema / Eczema atópico ou Psoríase: **Lut Xeramance® , Leti At4®, Dermalex®, Nutratopic®, Atoderm®, A-Derma Exomega® e A-Derma Atopicas®.** Mesmo que o adquirente seja um terceiro, deverá preencher o questionário com as características do doente utilizador final do produto.

2.3. OBJETIVO

O principal objetivo deste estudo é analisar o perfil dos doentes que sofrem de Eczema/Eczema atópico e Psoríase e caracterizar o que motivou a compra de um produto de saúde para o tratamento dessas doenças. No caso de existir prescrição médica identificar a especialidade médica do prescritor e identificar a origem da prescrição.

Pretende-se ainda fazer um estudo comparativo entre os doentes que adquirem produtos de uma marca em relação à concorrência.

Por último, pretende-se ajustar um modelo de regressão logística, em que a variável resposta é dicotómica entre consumir, ou não, produtos de determinada marca.

2.4. VARIÁVEIS EM ESTUDO

A tabela seguinte ilustra as variáveis em estudo e as respectivas categorias.

***TABELA* 1: VARIÁVEIS EM ESTUDO E AS RESPECTIVAS CATEGORIAS**

Variáveis	Categorias
Produto	Nome
	CNP
Sexo	1-Masculino
	2-Feminio
Idade	Anos
Primeira Vez	1 - Sim
	2- Não
Diagnóstico eczema	-1- Sim
	0 - Não
Diagnóstico psoriase	-1- Sim
	0 - Não
Diagnóstico outro	Nome
Motivo	1 - Iniciativa própria
	2- Amigo e ou Familiar
	3 - Publicidade (TV/revistas)
	4 - Recomendação da Farmácia
	5 - Outro profissional de Saúde
	6 - Prescrição médica
Especialidade médica	1 - Pediatra
	2 - Dermatologista
	3 - Médico Clinica Geral Familiar
	4 - Outra

Variáveis	Categorias
Origem	1-Hospital Público
	2- Centro de Saúde
	3 - Hospital Privado
Motivo anterior	1 - Iniciativa própria
	2- Amigo e ou Familiar
	3 - Publicidade (TV/revistas)
	4 - Recomendação da Farmácia
	5 - Outro profissional de Saúde
	6 - Prescrição médica
Especialidade médica anterior	1 - Pediatra
	2 - Dermatologista
	3 - Médico Clinica Geral Familiar
	4 - outra
Substituição da marca por produto da concorrência	1 - Sim
	2- Não

CAPÍTULO II
REGRESSÃO LOGÍSTICA

CAPITULO II – REGRESSÃO LOGÍSTICA

1. INTRODUÇÃO

Os modelos de regressão constituem uma das ferramentas estatísticas mais importantes na análise estatística de dados quando se pretende modelar relações entre variáveis. O principal objetivo destes modelos é explorar a relação entre uma ou mais variáveis explicativas (ou independentes) e uma variável resposta (ou dependente). Um dos casos particulares dos modelos lineares generalizados são os modelos onde a variável resposta apresenta apenas duas categorias ou que de alguma forma foi dicotomizada assumindo valores 0 ou 1 sendo o modelo de regressão logística o mais popular desses modelos.

A regressão logística é uma técnica estatística que tem como objetivo modelar, a partir de um conjunto de observações, a relação "logística" entre uma variável resposta dicotómica e uma serie de variáveis explicativas numéricas (continuas, discretas) e/ou categóricas.

2. REGRESSÃO LOGÍSTICA UNIVARIADA

Na regressão logística, a variável resposta é dicotómica, atribuindo-se o valor 1 ao acontecimento de interesse (sucesso) e 0 ao acontecimento complementar (incucesso).

Em qualquer regressão a quantidade chave é o valor médio da variável resposta dado o valor da variável independente. Esta quantidade é chamada valor médio condicional e é expressa como $E[Y/X]$ onde Y representa a variável resposta e X a variável explicativa. A quantidade $E[Y/X = x]$ é lida como "valor esperado de Y dado X=x"

No modelo de regressão linear, admitimos que o valor médio condicional, pode ser expresso como uma equação linear em x:

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

Note-se que $E[Y/X = x]$ pode assumir valores entre -∞ e +∞.

De modo a simplificar a notação consideremos, $\pi(x) = E[Y/X = x]$l. No modelo de regressão logística, no caso em que a variável resposta toma 2 valores distintos, assume-se que:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \qquad (2)$$

3. TRANSFORMAÇÃO LOGIT

Uma transformação fulcral no estudo dos modelos de regressão logística é a transformação *logit* cujo objetivo é linearizar o modelo, aplicando o logaritmo. Essa transformação define-se como:

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

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

$$g(x) = ln\left(e^{\beta_0+\beta_1 x}\right) = \beta_0 + \beta_1 x$$

Esta transformação assume especial importância pois o modelo com esta transformação possui diversas propriedades do modelo de regressão linear;

- A função *logit*, é linear nos parâmetros;
- Pode ser continua;
- Os seus valores podem variar em $\mathbb{R}$.

Esta transformação é chamada *transformação logit* de $\pi(x)$. A razão $\frac{\pi(x)}{1-\pi(x)}$ é chamada *Odds*.

3.1. ESTIMAÇÃO DOS PARÂMETROS

Na regressão linear a variável resposta pode ser expressa como $Y_x = E[Y/X = x] + \varepsilon_x$, onde a quantidade ε_x é o erro e expressa o desvio de uma observação em relação à média. Parte-se do pressuposto que o erro segue uma distribuição Normal com média zero e variância constante. Mas no caso onde temos uma variável dicotómica isto não acontece. O erro (ε_x) pode assumir apenas dois valores:

- $Y = 1 \rightarrow \varepsilon_x = 1 - \pi(x),\ com\ probabilidade\ \pi(x), onde$

$$\pi(x) = P(Y = 1 | X = x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad (3)$$

- $Y = 0 \rightarrow \varepsilon_x = -\pi(x), com\ probabilidade\ 1 - \pi(x), onde$

$$1 - \pi(x) = P(Y = 0 | X = x) = \frac{1}{1+e^{\beta_0+\beta_1 x}} \quad (4)$$

Então, ε_x tem uma distribuição com média zero e variância $\pi(x)[1-\pi(x)]$. Assim, a distribuição condicional da resposta Y_x é uma distribuição Bernoulli com parâmetro $\pi(x)$.

3.2. AJUSTAMENTO DO MODELO

Considere-se um elemento genérico da População com características $(X = x,\ Y = y)$ e representemo-lo pela variável aleatória, com distribuição de probabilidade Bernoulli, Y_x. A função de probabilidade desta v.a. é

$$f_{Y_x}(y_x) = \pi(x)^{y_x}\left(1-\pi(x)\right)^{1-y_x} com\ y_x \in \{0,1\} \qquad (5)$$

Suponhamos agora que temos uma amostra com n observações independentes do par (x_i, y_i) i=1,2…,n onde y_i é o valor da variável dicotómica e x_i é o i-ésimo valor da variável independente. Para ajustar o modelo é preciso estimar β_0 e β_1, os parâmetros desconhecidos. O método utilizado para a estimação dos parâmetros é o método de máxima verosimilhança.

A função massa de probabilidade Y_{x_i} é dada por:

$$f\left(Y_{x_i}\right) = \pi(x_i)^{y_{xi}}\left(1-\pi(x_i)\right)^{1-y_{xi}} com\ y_i \in \{0,1\} \qquad (6)$$

Assumindo a independência das observações a função de verosimilhança é obtida como o produto dos termos da expressão (6).

$$L(\beta) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i} \qquad (7)$$

A expressão da log- verosimilhança é definida como

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

$$= \sum_{i=1}^{n}\{y_i\ ln[\pi(x_i)] + (1-y_i)ln[1-\pi(x_i)]\}$$

$$= \sum_{i=1}^{n}\{y_i\ ln[\pi(x_i)] + ln(1-y_i) - y_i ln[1-\pi(x_i)]\}$$

$$= \sum_{i=1}^{n}\left\{y_i\ ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] + ln[1-\pi(x_i)]\right\}$$

Substituindo $\pi(x_i)\ e\ (1-\pi(x_i))$ por (1) e (2) respectivamente, obtém-se:

$$l(\beta) = \sum_{i=1}^{n}\left[y_i(\beta_o + \beta_1 x_i) + ln\left[\frac{1}{1+e^{\beta_0+\beta_1 x_i}}\right]\right]$$

$$\sum_{i=1}^{n}\left[y_i(\beta_o + \beta_1 x_i) + \ln(1) - \ln(1+e^{\beta_0+\beta_1 x_i})\right]$$

$$\sum_{i=1}^{n}\left[y_i(\beta_o+\beta_1 x_i)-\ln\left(1+e^{\beta_0+\beta_1 x_i}\right)\right]$$

$$\sum_{i=1}^{n}\left[y_i\beta_o+y_i\beta_1 x_i-\ln\left(1+e^{\beta_0+\beta_1 x_i}\right)\right]$$

O valor β que maximiza $ln[L(\beta)]$ é obtido após derivar $l(\beta)$ em relação aos parâmetros (β_0,β_1).

Derivando em ordem aos parâmetros obtém-se:

$$\frac{\partial ln[L(\beta)]}{\partial \beta_0}=\sum_{i=1}^{n}\left[y_i-\frac{e^{\beta_0+\beta_1 x_i}}{1-e^{\beta_0+\beta_1 x_i}}\right]=\sum_{i=1}^{n}[y_i-\pi(x_i)] \qquad (8)$$

$$\frac{\partial ln[L(\beta)]}{\partial \beta 1}=\sum_{i=1}^{n}\left[y_i x_i-x_i\frac{e^{\beta_0+\beta_1 x_i}}{1-e^{\beta_0+\beta_1 x_i}}\right]=\sum_{i=1}^{n}x_i[y_i-\pi(x_i)] \qquad (9)$$

Essas expressões são equações não lineares nos parâmetros, portanto são necessários métodos iterativos para sua resolução. O método mais utilizado, na maioria dos softwares estatísticos é o método de Newton-Raphson.

4. REGRESSÃO LOGÍSTICA MÚLTIPLA

Na secção anterior introduzimos o modelo de regressão logística univariado, ou seja, para o caso onde temos uma única variável independente. Consideremos agora o caso onde se tem um conjunto de p variáveis independentes expresso pelo vector $\boldsymbol{x}^{\mathbf{T}}\equiv(x_1,x_2,\ldots,x_p)$.

Neste caso, E(Y|**x**)= $\pi(\boldsymbol{x})$ com

$$\pi(\boldsymbol{x})=\frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}} \qquad (10)$$

e o logit da Regressão logística multipla é dado por

$$g(\boldsymbol{x})=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p\,. \qquad (11)$$

4.1. AJUSTAMENTO DO MODELO

Suponhamos que temos uma amostra com n observações independentes do p+1 vetor $(\boldsymbol{x_i},y_i)$ i=1,2,…,n onde y_i é o valor da variável dicotómica e $\boldsymbol{x_i}$ o i-ésimo valor do vetor de variáveis independentes. Para ajustar o modelo é preciso estimar $\boldsymbol{\beta^T}=(\beta_0,\beta_1,\cdots,\beta_p)$

O método utilizado no caso multivariado é novamente o método da Máxima Verosimilhança. Assumindo a independência das observações a função de verosimilhança é expressa por:

$L(\boldsymbol{\beta})=ln[L(\boldsymbol{\beta})]=\sum\{y_i\ \ln[\pi(\boldsymbol{x_i})]\,(1-y_i)ln[1-\pi(\boldsymbol{x_i})]\}$ (12)

=…=

$$\sum_{i=1}^{n}\left[y_i\beta_o + y_i\beta_1 x_{i1} + \cdots + y_i\beta_p x_{ip} - \ln\left(1 + e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}\right)\right]$$

4.2. A MATRIZ DE INFORMAÇÃO DE FISHER

A matriz de covariância dos coeficientes estimados é obtida a partir das derivadas parciais de segunda ordem do logaritmo da função de verosimilhança:

$$\frac{\partial^2 ln[L(\beta)]}{\partial {\beta^2}_j} = -\sum_{i=1}^{n}\left[{x^2}_{ij}\,\pi_i(1-\pi_i)\right] \qquad (13)$$

$$\frac{\partial^2 ln[L(\beta)]}{\partial\beta_j\partial\beta_k} = -\sum_{i=1}^{n}\left[x_{ij}x_{ik}\pi_i(1-\pi_i)\right] \qquad (14)$$

Onde j, k=0,1,2,…,p e π_i representa $\pi(x_i)$.

Se for formada uma matriz quadrada de dimensão (p+1), constituída pelo simétrico dos valores médios dos termos referidos nas duas equações anteriores, obtém-se $I(\beta)$, a chamada **Matriz de Informação**.

As variâncias dos coeficientes e as covariâncias entre os coeficientes estimados são obtidas por inversão desta matriz. Designar-se-á por $\sigma^2(\beta_j)$ o j-ésimo elemento da diagonal principal da matriz, $I^{-1}(\beta) \equiv \Sigma(\boldsymbol{\beta})$ a variância de $\hat{\beta}_j$, e por $\sigma(\beta_j\beta_u)$ a covariância entre β_j *e* β_u.

Os estimadores da variância e da covariância, são obtidos de $\sum(\boldsymbol{\beta})$ quando se substitui $\boldsymbol{\beta}$ pelo seu estimador $\widehat{\boldsymbol{\beta}}$. Serão utilizados $\hat{\sigma}^2(\hat{\beta}_j)$ e $\hat{\sigma}^2(\hat{\beta}_j, \hat{\beta}_u)$, com j, u=0,1,2,…,p, para designar os valores da respetiva matriz.

$\widehat{SE}(\hat{\beta}_J)=\left[\hat{\sigma}^2\hat{\beta}_J\right]^{1/2}$ representa o erro estimado dos coeficientes encontrados.

Usando notação matricial podemos escrever $\hat{I}(\hat{\beta}) = X'VX$, onde X é uma matriz nx(p+1) contendo, além do vetor 1 os valores observados para as variáveis independentes, e V é uma matriz diagonal nxn, de elemento genérico $\hat{\pi}_i(1-\hat{\pi}_i)$:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ 1 & x_{31} & x_{32} & \ldots & x_{3p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}$$

$$V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \dots & 0 \\ 0 & 0 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

4.3. TESTE À SIGNIFICÂNCIA DO MODELO

Uma vez ajustado o modelo, é necessário testar a significância do modelo estimado. Uma medida para testar essa significância é o teste da razão de verosimilhanças.

Com este teste pretende-se testar simultaneamente se os coeficientes de regressão associados a **β** são todos nulos com exceção de β_0.

A comparaçao dos valores observados e dos valores esperados usando o função de verosimilhança é baseada na seguinte expressão:

$$D = -2\ln\left[\frac{\text{Função de máxima verosimilhança do modelo corrente}}{\text{Função de máxima verosimilhança do modelo saturado}}\right]$$

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right] \qquad (15)$$

Onde,

Modelo saturado o corresponde ao modelo com todas as variáveis e interacções e o Modelo corrente corresponde ao modelo com apenas as variáveis desejadas para o estudo

A hipótese a testar é

$$H_0: \beta_1 = \cdots = \beta_p = 0 \text{ v.s } H_1: \exists_{j=1,\dots p}: \beta_j \neq 0$$

Onde a estatística de teste é definida por:

G = D (modelo sem as p variáveis) – D (modelo com as p variáveis)

$$G = -2\ln\left[\frac{\text{modelo corrente}}{\text{modelo saturado}}\right] \cap_{\text{sob } H_0} \chi^2_{(p-1)} \qquad (16)$$

Ao rejeitarmos a hipótese nula podemos concluir que, pelo menos um dos coeficientes é estatisticamente diferente de zero. Assim antes de chegarmos a uma conclusão final, dever-se-á testar se cada um dos coeficientes é significativamente diferente de zero, sendo para isso realizado o Teste de *Wald, que passaremos a descrever.*

TESTE DE *WALD*

O teste de Wald testa se cada coeficiente é significativamente diferente de zero. Deste modo, o teste de *Wald* averigua se uma determinada variável independente apresenta uma relação estatisticamente significativa com a variável dependente.

Então estamos interessados em testar:

$$H_0: \beta_j = 0 \qquad \text{v.s} \qquad H_1: \beta_j \neq 0, j = 0, \dots, p$$

A estatística de teste é dada por:

$$W_j = \frac{\widehat{\beta}_j}{var(\widehat{\beta}_j)} \cap_{sob\ H_0} \chi^2_{(1)} \qquad (17)$$

4.4. INTERPRETAÇÃO DOS COEFICIENTES ESTIMADOS

Nas secções anteriores descreveram-se métodos para ajustar e testar a significância do modelo de regressão logística. Uma vez ajustado o modelo e após avaliar a significância dos coeficientes estimados, é agora necessário interpretar os seus valores.

Para que possamos interpretar os valores associados aos coeficientes do modelo de regressão logística, é conveniente proceder à análise de acordo com a natureza das variáveis independentes.

Abordaremos três situações: variáveis dicotómicas, variáveis policotómicas (nominais ou ordinais com mais de duas categorias) e variáveis contínuas.

4.4.1. VARIÁVEIS INDEPENDENTES DICOTÓMICAS

Suponhamos que X é uma variável explicativa e que está codificada em dois valores distintos o ou 1 levando a que $\pi(x)$ possa apenas assumir os valores π(0) e π(1) e Y tenha a seguinte distribuição de probabilidade:

	X=0	X=1
Y=0	1- π (0)	1- π (1)
Y=1	π (0)	π (1)

Em que:

$$\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}} \qquad \pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$$

$$1\text{-}\pi(0) = \frac{1}{1+e^{\beta_0}} \qquad 1\text{-}\pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$$

Quando a característica X está presente (X=1) o ODDs é $\frac{\pi(1)}{1-\pi(1)}$. Da mesma maneira na ausência da característica, ou seja X=0 esse ODDS é $\frac{\pi(0)}{1-\pi(0)}$.

Aplicando o logaritmo, ou seja a função logit, obtém-se:

$$g(1) = ln\left[\frac{\pi(1)}{1-\pi(1)}\right] \;\; e \;\; g(0) = ln\left[\frac{\pi(0)}{1-\pi(0)}\right] \qquad (18)$$

A medida mais utilizada, em Regressão logística, é a Razão de ODDS, que é designada por OR (Odds Ratio), estimada da seguinte forma:

$$OR = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} \qquad (19)$$

Aplicando o logaritmo obtém-se

$$\ln(OR) = ln\left[\frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}}\right] = g(1) - g(0)$$

A esta diferença dá-se o nome de **diferença *logit.***

Simplificando a expressão OR, substituindo os valores de $\pi(1)$ e $\pi(0)$, de 1 - $\pi(1)$ e de 1 - $\pi(0)$ pelas expressões apresentadas anteriormente, vem

$$OR = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\left(\frac{1}{1+e^{\beta_0}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Para amostras suficientemente grandes, o estimador para β_i, $\hat{\beta}_i$, segue uma distribuição aproximadamente normal. Supondo válida essa aproximação à Normal dos $\hat{\beta}_i$, encontra-se primeiro o intervalo de confiança para β_i,

$$\hat{\beta}_i \pm Z_{1-\frac{\alpha}{2}} \, S\hat{E}\,(\hat{\beta}_i),$$

onde $Z_{1-\frac{\alpha}{2}}$ representa o quantil de probabilidade $\left(1-\frac{\alpha}{2}\right)$ da distribuição Normal de valor médio nulo e variância unitária e seguidamente para e^{β_i}.

Assim, o intervalo de 100(1- α)% de confiança para e^{β_i} é dado por:

$$e^{\left(\hat{\beta}_i \pm Z_{1-\frac{\alpha}{2}} \, S\hat{E}\,(\hat{\beta}_i)\right)} \qquad (20)$$

Se o intervalo de confiança contiver o valor 1, não existe relação significativa entre as variáveis em causa, X (como variável explicativa) e Y (como variável resposta).

4.4.2. VARIÁVEIS INDEPENDENTES POLICOTÓMICAS

Suponhamos agora que uma das variáveis independentes em estudo, tem mais do que duas categorias distintas. Então neste caso deveremos utilizar variáveis auxiliares designadas por variáveis Dummy ou variáveis Indicatrizes.

Este tipo de variáveis assume os valores 0 ou 1. Se a variável nominal em estudo tiver m categorias distintas, então dever-se-ão criar m-1 variáveis Dummy indexadas a essa categoria.

Categoria da variável X	Variáveis Dummy			
	D_1	D_2	⋯	D_{m-1}
C_1	0	0	⋯	0
C_2	1	0	⋯	0
⋮	⋮	⋮	⋱	⋮
C_m	0	0	⋯	1

Neste caso é necessário considerar uma categoria como grupo de referência, C_1. Pode-se então comparar uma qualquer categoria C_k, com k = 2, … , m, com a categoria de referência, C_1.

O cálculo do OR é obtido da mesma forma que no caso das variáveis dicotómicas. O intervalo de confiança é obtido com cálculos idênticos aos já realizados anteriormente.

4.4.3. VARIÁVEL INDEPENDENTE CONTÍNUA

Quando um modelo logístico contém uma variável independente continua, antes de fazer qualquer interpretação do seu coeficiente estimado, é conveniente testar a escala das variáveis contínuas.

4.4.3.1 LINEARIDADE NO LOGIT

Neste caso $g(x) = \beta_0 + \beta_1 x$. β_1 representa o valor da mudança em g(x) quando o valor da variável x aumenta em uma unidade, pois $\beta_1 = g(x+1) - g(x)$, para qualquer valor de x.
Assim, é necessário saber de quantas unidades se deverá aumentar x, de modo que a interpretação seja considerada adequada. Seja k um valor ideal para a realização de uma boa interpretação tal que $k\beta_1 = g(x+k) - g(x)$.. Calculando a exponencial de $k\beta_1$ obteremos o OR (x, x+k) onde o intervalo de 100%(1-α) de confiança vem dado por:

$$e^{\left(k\widehat{\beta}_1 \pm Z_{1-\frac{\alpha}{2}}\, k\, S\hat{E}\ (\widehat{\beta}_1)\right)} \qquad (21)$$

No caso de não se verificar a condição de linearidade no logit ou mesmo no caso em que a escolha de k seja difícil, poderá pôr-se a hipótese de categorizar a variável em estudo, em 2 ou mais categorias.

4.4.3.2 MÉTODO DOS QUARTIS

Hosmer & Lemseshow (Hosmer & Lemeshow, 2000) sugerem um procedimento que permite verificar a hipótese de linearidade no logit, dando-nos a indicação de quantas categorias deverão ser criadas. Este procedimento é designado pelo **Método dos Quartis** e deverá seguir os seguintes passos:

1. Determinar os quartis associados á distribuição de frequências da variável.
2. Criar 3 variáveis *Dummy* baseadas nos quartis obtidos, sendo o primeiro considerado como a categoria de Referência.
3. Determinar os coeficientes estimados para estas 3 variáveis *Dummy*, através do modelo que contém todas as variáveis previamente selecionadas.
4. Proceder á estimação dos OR associada às 3 variáveis *Dummy*.

5. Construção de um gráfico, onde no eixo dos xx, estão representados os pontos médios dos quartis e no eixo dos yy, os valores dos coeficientes estimados das variáveis *Dummy*, no modelo multivariado.

Após construir o gráfico procederemos à análise deste. Se a curva encontrada for "aproximadamente" linear poderemos concluir que a variável é linear no logit. Caso contrario a variável deverá ser categorizada.

4.5. ESTRATÉGIA PARA A CONSTRUÇÃO DO MODELO

Na secção anterior focámos a estimação, testes de significância e interpretação dos coeficientes no modelo de regressão logística.

Nesta secção, apresentaremos um método para a seleção de variáveis a incluir no modelo final.

O principal objetivo de qualquer destes métodos é selecionar as variáveis que resultem no melhor modelo possível. De modo a ser cumprido esse objetivo é necessário:

- Um plano básico para a seleção de variáveis para o modelo.
- Um Método para determinar a adequabilidade dos modelos tanto em termos de cada variável como do ajustamento global.

4.5.1. SELEÇÃO DE VARIÁVEIS

O processo de seleção de variáveis deve começar por uma análise univariada de todas a variáveis. Após essa análise, selecionar-se-ão as variáveis para a análise multivariada. O grau de importância de uma variável é medido pelo p-value de Wald. Quanto menor for este valor tanto mais importante será considerada a variável. Qualquer variável cujo p-value, referente ao teste de Wald, seja inferior ou igual a 0,20 (p_e) deverá ser considerada como candidata ao modelo múltiplo. As variáveis candidatas a sair do modelo, são as que apresentam um p-value superior a 0,25 (p_s). É no entanto possível forçar a entrada de uma variável cuja importância clínica seja relevante.

4.5.2. UM MÉTODO PARA A SELEÇÃO DE VARIÁVEIS

Pretende-se selecionar o subconjunto de variáveis significativas, de entre todas aquelas que estão disponíveis. Para esta selecção usaremos um método de seleção stepwise.

Consideremos que existem p variáveis independentes, todas consideradas importantes para explicar a variável resposta. Passaremos então a descrever o método, com inclusão

progressiva seguido de eliminação regressiva, com base numa regra de decisão critica como já foi referido.

Passo (0): este passo inicia-se por ajustar um modelo contendo apenas os termos independentes, e calcula-se o valor do logaritmo da verosimilhança, que se designará por L_0. De seguida ajusta-se vários modelos univariádos, um para cada uma das p variáveis independentes e calcula-se o valor do logaritmo da verosimilhança para cada um desses modelos. Designa-se por $L_j^{(0)}$, o logaritmo da verosimilhança para o modelo que contém a variável x_j , no passo (0). O subscrito j refere-se à variável incluída no modelo e o sobrescrito refere-se ao passo. Esta notação será utilizada ao longo do processo do stepwise para controlar tanto o número de passo como o número de variáveis no modelo.

Seja $G_j^{(0)} = 2\left(L_j^{(0)} - L_0\right)$ o valor do teste da razão das verosimilhanças, para o modelo contendo a variável x_j versus o modelo que contem apenas os termos constantes. Neste caso o p-value é determinado por $P\left(\chi_v^2 > G_j^{(0)}\right) = p_j^{(0)}$, onde $v = (k-1)$ caso a variável x_j seja policotómica e $v = 1$ caso a variável seja continua.

A variável considerada estatisticamente mais importante é aquela que corresponde a um p-value menor. Seja $p_{e_1}^{(0)} = \min_j\left(p_j^{(0)}\right)$.

Designando-se por p_e, o p-value correspondente á variável mais importante, passa-se para o passo (1) se $p_{e_1}^{(0)} < p_e$, caso contrário paramos por aqui. O subscrito e_1 é utilizado para indicar que a variável é candidata a entrar no passo (1).

Passo(1) Este passo inicia-se com o ajustamento do modelo contendo apenas a variável que corresponde ao menor p-value . Seja x_{e_1}essa variável e seja $L_{e_1}^{(1)}$ o logaritmo da verosimilhança para este modelo.

Seguidamente é necessário determinar quais as variáveis importantes, das p-1, dado que o modelo já contém a variável x_{e_1}. Ajustam-se p-1 variáveis modelos contendo a variável x_{e_1} e x_j, j=1,…, p e $j \neq e_1$. Designa-se por $L_{e_{1j}}^{(1)}$ o valor do logaritmo da verosimilhança que contém x_{e_1} e x_j , e seja a estatística do qui-quadrado do modelo $G_j^{(1)} = 2\left(L_{e_{1j}}^{(1)} - L_{e_1}^{(1)}\right)$.

Considera-se tal como no passo anterior $p_j^{(1)}$, o p-value correspondente à estatística apresentada.

Seja a variável x_{e_2} que deu origem ao menor p-value, neste passo (1). Considere tal como no passo anterior $p_{e_2}^{(1)} = \min_j \left(p_j^{(1)}\right)$. Então prosseguimos ao passo seguinte se $p_{e_2}^{(1)} < p_e$ caso contrário paramos por aqui.

Passo (2) Este passo inicia-se com o ajustamento do modelo contendo as variáveis x_{e_1} e x_{e_2}. Uma vez que estamos perante uma selecção de variáveis progressiva e regressiva, em simultâneo, é possível que ao incrementar a variável x_{e_2}, a variável x_{e_1} possa ter deixado de ser importante. Então esse passo possui a selecção regressiva. Seja $L_{-e_j}^{(2)}$ o valor do logaritmo da verosimilhança, do modelo retirando a variável x_{e_2} e seja a estatística do qui-quadrado do modelo sem a variável x_{e_2} e do modelo com todas as duas variáveis, definida $G_{-e_j}^{(2)} = 2\left(L_{e_1e_2}^{(2)} - L_{-e_j}^{(2)}\right)$ e seja ainda $p_{-e_j}^{(2)}$ o p-value correspondente. Seja x_{s_2}, a variável que foi excluída do modelo cujo valor de p vem dado em $p_{s_2}^{(2)}$ $\max = \left(p_{-e_1}^{(2)}, p_{-e_2}^{(2)}\right)$. Para decidir se a variável x_{s_2}, será ou não removida do modelo, compara-se o valor de p com p_s, sendo p_s o valor de p para saída de variáveis do modelo.

Então se $p_s{}^{(2)} > p_s$, a variável x_{s_2} será removida do modelo. Caso contrário deverá manter-se a variável no modelo.

Na fase de seleção progressiva são ajustados p-2 modelo contendo x_{e_1} , x_{e_2} e x_j, com j=1,2, … , p, $j \neq e_1, e_2$. Calcula-se o logaritmo da verosimilhança para cada um dos modelos encontrados para cada um dos p-2 modelos à semelhança do passo (1), determina-se a estatística de teste da razão de verosimilhança para estes novos modelos versus o modelo contendo apenas as variáveis x_{e_1} e x_{e_2}, determinando-se os respetivos p-value. Suponha-se que a variável x_{e_3} é a variável que corresponde ao menor valor de p encontrado. Se este valor de p for menor que p_e, prossegue-se para o passo (3), senão paramos por aqui

Passo (n) este passo ocorre, se: Todas as p variáveis já entraram no modelo ou se todas as variáveis que constituem o modelo têm p-value inferiores ao valor de p_s e todas as variáveis que não foram incluídas no modelo têm p-value superiores a p_e.

O modelo encontrado neste passo deverá conter todas as variáveis que foram consideradas estatisticamente importantes, de acordo com os valores de p_e e p_s escolhidos.

4.6. DIAGNÓSTICO DO MODELO

Em qualquer modelo de regressão, é necessário proceder à análise dos resíduos para validação da qualidade do modelo estimado. Assim, pretende-se avaliar quais as ”distâncias” entre os valores observados e os valores estimados.

Existem diversas medidas de modo a detetar diferenças significativas entre os valores observados e os valores estimados.

Existem dois tipos de resíduos possíveis que poderão ser utilizados para avaliar a qualidade do ajustamento: os resíduos de Pearson e os resíduos da Deviance.

4.6.1. RESÍDUOS DE PEARSON

O resíduo de Pearson para o j-ésimo individuo é definido por:

$$r(y_j\hat{\pi}_j) = r_j = \frac{y_j - \hat{\pi}_j}{\sqrt{\hat{\pi}_j(1-\hat{\pi}_j)}},\ \text{j=1,2, … n} \qquad (22)$$

A estatística de teste global baseada nos **Resíduos de Pearson** designa-se por estatística de Qui-Quadrado de Pearson e é calculada da seguinte forma:

$$\chi^2 = \sum_{j=1}^{n} r(y_j, \hat{\pi}_j)^2 \cap_{sob\ H_0} \chi^2_{(n-p-1)} \qquad (23)$$

Uma estatística alternativa é obtida à custa dos Residuos da Deviance, ainda sob a mesma hipótese nula, H_0 , onde H_0 significa “O modelo encontrado explica bem os dados”.

4.6.2. RESÍDUOS DE DEVIANCE

O resíduo de Deviance para o j-ésimo individuo é definido da seguinte forma:

$$d(y_j, \hat{\pi}_j) = d_j = \pm\left\{2\left[y_j ln\left(\frac{y_j}{\hat{\pi}_j}\right) + (1-y_j) ln\left(\frac{1-y_j}{1-\hat{\pi}_j}\right)\right]\right\}^{1/2} \qquad (24)$$

A estatística a utilizar é:

$$D = \sum_{j=1}^{n} d(y_j, \hat{\pi}_j)^2 \cap_{sob\ H_0} \chi^2_{(n-p-1)} \qquad (25)$$

4.6.3. RESÍDUDOS DE PEARSON STANDARTIZADOS

Um procedimento mais adequado consiste em dividir os resíduos pelo valor estimado do seu desvio padrão. Este valor é aproximado por $\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)(1-h_j)}$ onde o h_j é o j-ésimo elemento da diagonal da matriz $(\hat{H})$ dada por:

$$\hat{H} = \hat{V}^{1/2}X(X'\hat{V}X)^{-1}X'\hat{V}^{1/2},$$

onde $\hat{V}$ é uma matriz diagonal com n elementos $\hat{\pi}_i(1-\hat{\pi}_i)$, X é uma matriz nxp, X' é a transposta de X e $\hat{V}^{1/2}$ é a matriz diagonal em que os seus elementos são iguais à raiz quadrada dos elementos da matriz $\hat{V}$.

Posto isto os resíduos standartizados são dados por:

$$r_{sj} = \frac{r_j}{\sqrt{1-h_j}} \qquad (26)$$

Uma estatística de diagnóstico extremamente útil é aquela que examina o efeito provocado pela eliminação de uma observação, nos coeficientes estimados, para os coeficientes do modelo e pode ser representada da seguinte forma

$$\Delta\hat{\beta}_j = \frac{{r_{sj}}^2 h_j}{1-h_j} \qquad (27)$$

4.6.4. TESTE DE HOSMER-LEMESHOW

É usual agrupara os valores de X e Y, constituindo g grupos, normalmente g=10, e constitui-los de forma a serem mais ou menos homogéneos nos valores de X.

A hipótese a testar é, H_0: "O modelo encontrado explica bem os dados".

A estatística de teste é obtida pela estatística de Qui-quadrado de Pearson, a partir de uma tabela 2 x g, onde g é o número de grupos criados. Esta tabela contém as frequências observadas e esperadas.

A estatística de teste é dada por

$$\chi^2_{HL} = \sum_{k=1}^{g} \frac{(o_k - n'_k\bar{\pi}_k)}{n'_k\bar{\pi}_k(1-\bar{\pi}_k)} \overset{sob\ H_0}{\bigcap} \chi^2_{g-2}$$

Onde,

n'_k- é o número de indivíduos no k-ésimo grupo

c'_k - é o número de covariáveis padrão no k-ésimo grupo

$o_k = \sum_{i=1}^{c'_k} y_i$ – é o número de respostas ao longo das c'_k classes de variáveis

$$\bar{\pi}_k = \sum_{i=1}^{c'_k} \frac{\bar{\pi}_i m_i}{n'_i}$$

4.6.5. A CURVA ROC

A curva ROC (Receiver Operating Characteristc) é uma ferramenta que permite avaliar o desempenho de um modelo de regressão binária (variável resposta é do tipo 0-1). Pode ser feita por meio de um gráfico simples e robusto, que nos permite estudar a variação da sensibilidade e especificidade, para diferentes pontos de quebra.

Deveremos considerar um ponto de quebra C e comparar cada probabilidade estimada com o valor de C. O valor mais utilizado para C é 0,5 (Hosmer & Lemeshow, 2000).

Neste trabalho o ponto de quebra será o que maximiza ambas as curvas de sensibilidade e especificidade, tal como ilustra o gráfico abaixo.

O ponto onde as retas se cruzam, 0.04, será o ponto de quebra a considerar neste trabalho.

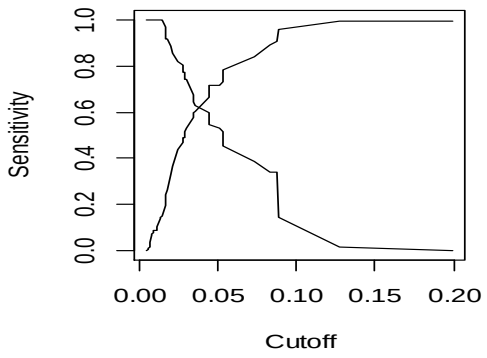

GRÁFICO **1: SENSIBILIDADE VS ESPECIFICADE**

Caso a probabilidade estimada exceda o valor C a variável dicotómica tomará o valor 1, caso contrário tomará o valor 0.

A tabela de contingência 2 x 2 é útil para ilustrar esses valores. Segue uma generalização das tabelas de contingência.

Classificação	Doente		
	Sim(1)	Não	Total
Sim (1)	n_{11}	n_{12}	$n_{1.}$
Não (0)	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

A sensibilidade é definida como a probabilidade do teste fornecer um resultado positivo, dado que o indivíduo é realmente portador da "doença".

Este é dado por $\frac{n_{11}}{n_{.1}}$. A especificidade é definida como a probabilidade do teste fornecer um resultado negativo quando o indivíduo não é portador da "doença". É dado por $\frac{n_{22}}{n_{.2}}$

A percentagem de indivíduos correctamente classificados é dada por $\frac{n_{11}+n_{22}}{\mathrm{n}_{..}} \times 100.$

A curva ROC é um gráfico de Sensibilidade (ou taxa de verdadeiros positivos) versus taxa de falsos positivos, ou seja, representa a Sensibilidade (ordenadas) vs 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 otimização da Sensibilidade em função da 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.

A área abaixo da curva de ROC, fornece-nos uma medida de discriminação, que nos indica a possibilidade de um individuo não doente ter uma probabilidade estimada associada mais elevada do que um individuo doente.

Seja R, o valor que corresponde à área abaixo da curva de ROC, como regra geral temos as seguintes linhas de orientação:

- Se R = 0, 5 não há discriminação
- Se 0,7 ≤ R <0,8 a discriminação é aceitável
- Se 0,8 ≤R <0,9 Discriminação excelente
- Se R ≥ 0,9 Discriminação excepcional

CAPÍTULO III
RESULTADOS

CAPITULO III - RESULTADOS

1. CONSTRUÇÃO DO MODELO

Começamos o trabalho com uma análise descritiva das variáveis, de um modo global. Em seguida procederemos a uma análise bivariada, comparando a concorrência com a marca e ajustando um modelo de regressão logística para cada uma das variáveis em estudo.

Para o modelo de regressão logística múltipla, não serão consideradas todas as variáveis em estudo devido à reduzida dimensão da amostra.

2. METODOLOGIA/SOFTWARE UTILIZADO

O software utilizado, tanto na análise descritiva como na construção do modelo de regressão logística foi o SAS, Enterprise Guide V4.1.

O procedimento utilizado para a construção de modelo de regressão foi o PROC LOGISTIC.

3. ANÁLISE DESCRITIVA

Do total dos utilizadores (n=1784), a maioria era do sexo feminino (54,2%, n=987). Em relação à idade verificou-se que em média os respondentes tinham 23 anos variando dos 0 aos 100 anos. Analisando os resultados por classe etária, verificou-se que 40,7% dos utilizadores têm menos de 5 anos, 18,1% dos utilizadores têm idade compreendida entre os 6 e os 17 anos e 41, 2% dos utilizadores tem mais de 18 anos. Para 53,5% (n=950) dos utilizadores, não era a primeira vez que adquiriam o produto em causa.

No que diz respeito ao diagnóstico, verificou-se que do total de 1766 respostas, Eczema/Eczema atópico foi o diagnóstico mais referido 65,0% (n*=1148), psoríase com apenas 3,5% (n*=61) e outros diagnósticos com 30,9% (n*=542). Desses diagnósticos, destaca-se a pele seca/muito seca (n*=239) e a dermatite (n*=78).

O principal motivo para adquirirem o produto foi prescrição médica (52,2%, n=931), sendo o dermatologista a especialidade médica mais referida (52,7%) seguido da pediatria (33,3%). Relativamente á origem da prescrição verifica-se que a maioria tem origem em consultas privadas (72,3%).

Para os utilizadores que já haviam adquirido o produto anteriormente, verificou-se que prescrição médica foi o motivo principal dessa compra (67,7%, n=618). Dermatologista e pediatria são as especialidades mais referidas, com 53,5% e 36,0% respetivamente.

4. ANÁLISE UNIVARIADA

Como já foi referido anteriormente o objetivo deste estudo foi saber a posição que a marca ocupa no mercado relativamente á concorrência na venda de produtos para o tratamento de eczema e psoríase. Assim fez-se a análise de cada variável, contra a variável resposta (concorrência ou marca) de modo a avaliar o comportamento das mesmas. Para cada variável, para além do cálculo das frequências e da respetiva representação gráfica, será aplicado o teste de Qui-quadrado e um modelo de regressão logística simples, caso seja aplicável.

Variável resposta

A variável resposta será baseada na marca do produto que o respondente está a adquirir, sendo esta categorizada como produto da concorrência ou da marca. Como já se tinha referido, será uma variável categorizada em 0 e 1 em que o sucesso (1) será a aquisição do produto da marca e o insucesso a aquisição do produto da concorrência (0). O gráfico seguinte ilustra os valores absolutos obtidos em cada categoria. Obteve-se um total de 1.786 questionários válidos para análise, n=1721 (96.4%) referente à concorrência e n=65 (3.6%) referentes à marca

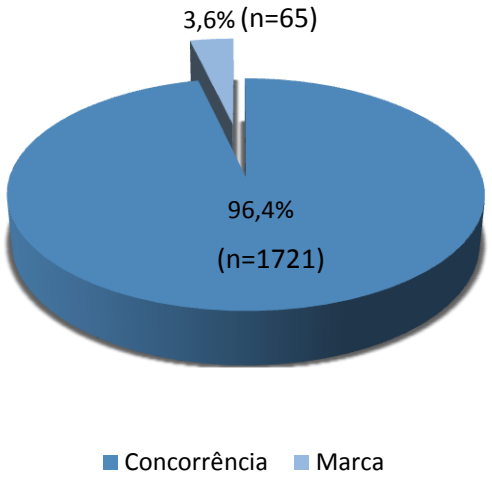

GRÁFICO 2: DISTRIBUIÇÃO DOS UTILIZADORES POR TIPO DE PRODUTO

Distribuição dos respondentes por marca, segundo o sexo

Analisando a distribuição dos respondentes por sexo de acordo com a marca, verificou-se que a maioria dos utilizadores era do sexo feminino (cerca de 54%), em ambas categorias. Não se

encontraram diferenças estatisticamente significativas entre a distribuição de adquirentes da Marca e de produtos da concorrência, segundo o sexo, (χ_1^2 = 0.001; *p-value*= 0,9755).

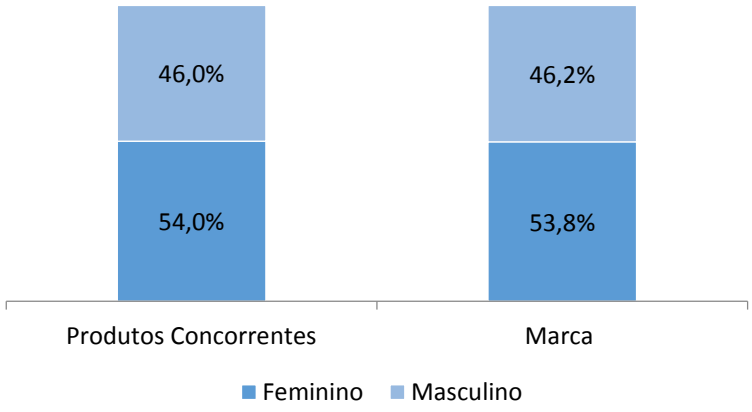

GRÁFICO 3: DISTRIBUIÇÃO DOS RESPONDENTES, POR MARCA, SEGUNDO O SEXO

Classe etária

A variável idade foi dividida em 3 classes, a primeira classe corresponde aos indivíduos com idade inferior a 5 anos, a segunda classe refere-se aos indivíduos com idade compreendida ente os 6 aos 17 anos e a terceira classe corresponde aos indivíduos com mais de 18 anos. Analisando a distribuição dos respondentes por faixa etária, verifica-se que não existem diferenças estatisticamente significativas entre a distribuição de adquirentes da Marca e de produtos da concorrência, (χ_3^2 =2,99; *p_value* = 0,2241).

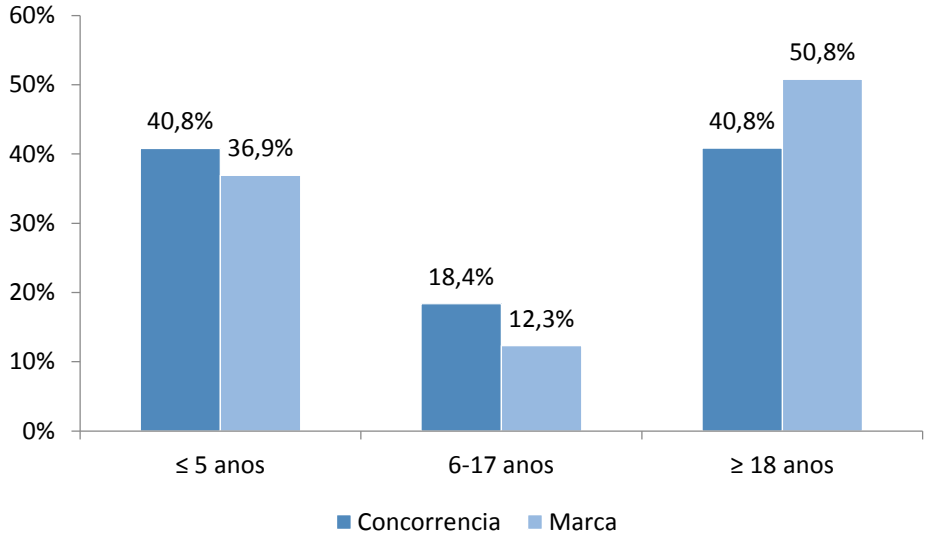

GRÁFICO 4: DISTRIBUIÇÃO DOS RESPONDENTES, POR MARCA, SEGUNDO A CLASSE ETÁRIA

Para o modelo de regressão logística, onde a classe de referência são os utilizadores com mais de 18 anos, obtiveram-se os seguintes resultados:

Dummy 1 - Com um $\hat{\beta}_1$= -0.32, que corresponde a um OR_1= 0.73, que significa que o odds dos utilizadores com idade inferior a 5 anos é menor, 30%, do que o dos indivíduos com mais de 18 anos.

Dummy 2 - Com um $\hat{\beta}_2$= -0.6174, o que corresponde a um OR_1= 0,54 o que significa que o odds dos indivíduos com idade compreendida entre os 6 e os 17 anos reduz para metade quando comparados com o odds dos indivíduos com idade superior a 18 anos .

TABELA 2: REGRESSÃO UNIVARIADA PARA A VARIÁVEL IDADE

Idade	$\hat{\boldsymbol{\beta}}$	OR	Intervalo de confiança do OR (95%)	p-value
≥ 18 anos (Ref).	-3,08			<0,0001
≤ 5 anos	-0,32	0,73	[0,43 ; 1,25]	0,2464
6-17 anos	-0,62	0,54	[0,25 ;1,18]	0,1226

Primeira vez

Para a maioria dos utilizadores da Marca (60.9%), era a primeira vez que utilizavam o produto. Por outro lado, a maioria dos utilizadores da marca concorrente (53,7%) eram prevalentes. Existem diferenças estatisticamente significativas entre a proporção de utilizadores da marca e a proporção de utilizadores de produtos da concorrência, segundo a variável primeira vez (χ_1^2 = 5,33; *p_value*= 0,0209).

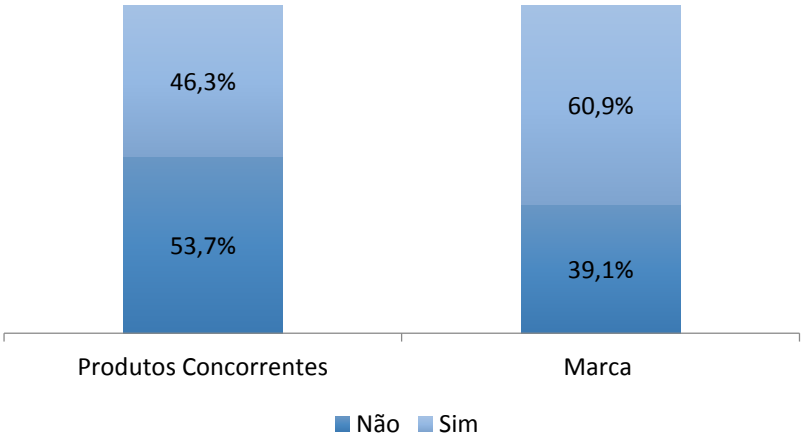

GRÁFICO 5: DISTRIBUIÇÃO DOS RESPONDENTES, POR MARCA, SEGUNDO A PRIMEIRA VEZ

No modelo de regressão logística univariada, obteve-se um $\hat{\beta}_1 = 0.61$, o que corresponde a um $OR_1 = 1.83$ o que nos leva a concluir que o odds dos indivíduos que compram o produto pela primeira vez é duas vezes maior do odds da prevalência.

TABELA 3: REGRESSÃO UNIVARIADA PARA A VARIÁVEL PRIMEIRA VEZ

Primeira Vez	$\hat{\beta}$	OR	Intervalo de confiança do OR (95%)	p-value
Não (Ref.)	-3,61			<0,0001
Sim	0,61	1,83	[1,1 ; 3,06]	0,0201

Diagnóstico da compra por marca

Eczema / Eczema atópico foi o diagnóstico mais referido (82% Marca e 64% Produtos Concorrentes). Para os produtos da Marca, Psoríase foi a segunda causa da compra (16.7%). Já para os Produtos Concorrentes, foram destacados outros diagnósticos (32,2%). Estes diagnósticos estão apresentados na Tabela 4.

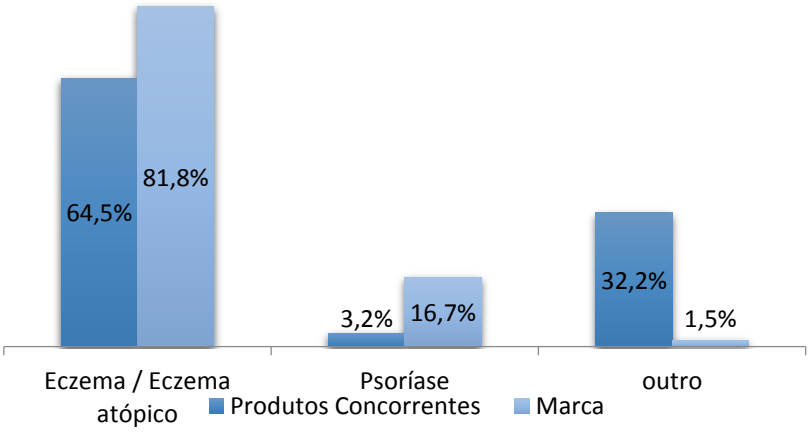

GRÁFICO 6: DISTRIBUIÇÃO DOS RESPONDENTES, POR MARCA, SEGUNDO O DIAGNÓSTICO

A tabela seguinte ilustra para além dos diagnósticos, Eczema / Eczema atópico, Psoríase e outro diagnóstico, a proporção dos utilizadores que tiveram mais do que um diagnóstico em simultâneo. Observa-se que houve 3 indivíduos a registarem Eczema / Eczema atópico e Psoríase como diagnóstico, 7 que registaram Eczema / Eczema atópico e outro e ainda 1 individuo a registar como diagnóstico Psoríase e outro.

TABELA 4: DIAGNÓSTICOS QUE MOTIVARAM A COMPRA

Diagnóstivo que motivou a compra	Produtos Concorrentes		Marca	
	n	%	n	%
Eczema / Eczema atópico	1.095	64,4%	53	81,5%
Eczema / Eczema atópico + Psoríase	3	0,2%	1	1,5%
Eczema / Eczema atópico + Outro	7	0,4%	0	0,0%
Psoríase	51	3,0%	10	15,4%
Psoríase + outro	1	0,1%	0	0,0%
outro	544	32,0%	1	1,5%
Total	1.701	100,0%	65	100,0%
NR	20	-	0	-

Foram ajustados modelos de regressão logística univariada tanto para o diagnóstico de eczema/eczema atópico como para o diagnóstico de psoríase. Os resultados estão apresentados nas tabelas 5 e 6 respetivamente.

Para o diagnóstico eczema obteve-se um $\hat{\beta}_1$= 1,00, o que corresponde a um OR_1= 2,72 o que significa que o odds dos indivíduos com eczema/eczema atópico é 3 vezes maior que o odds dos indivíduos a quem não foi diagnosticado eczema.

TABELA 5: REGRESSÃO UNIVARIADA PARA O DIAGNÓSTICO DE ECZEMA

Diagnóstico Eczema	$\hat{\beta}$	OR	Intervalo de confiança do OR (95%)	p-value
Não(ref)	3,03			<0,0001
sim	1,00	2,72	[1,41 ; 5,23]	0,0028

Para o diagnóstico psoríase obteve-se um $\hat{\beta}_1$= 1,83 , o que corresponde a um OR_1= 6.22 o que nos leva a concluir que o odds dos utilizadores com psoríase é 6 vezes maior, quando comparados com indivíduos sem psoríase.

TABELA 6: REGRESSÃO UNIVARIADA PARA O DIAGNÓSTICO DE PSORÍASE

Diagnóstico psoríase	$\hat{\beta}$	OR	Intervalo de confiança do OR (95%)	p-value
Não(Ref)	-3,44			<0,0001
Sim	1,83	6,22	[3,08 ; 12,53]	<0,0001

Outro diagnóstico

Para os produtos da marca, pele atópica foi o único diagnóstico referido e foi registada por um único individuo. Para os produtos concorrentes, obteve-se um total de 497 respostas. A tabela 7 ilustra os diagnósticos obtidos. Constata-se que pele seca ou muito seca foi o diagnóstico mais referido (n=239) seguido pela dermatite (n=78) e pele atópica (n=60).

TABELA 7: OUTRO DIAGNÓSTICO PARA OS PRODUTOS CONCORRENTES

Outro diagnóstico nos Produtos Concorrentes	n	%
Pele seca ou muito seca	239	48,1%
Dermatite	78	15,7%
Pele atópica	60	12,1%
Hidratação	36	7,2%
Pele sensível	28	5,6%
Sarna/Urticária / Prurido	19	3,8%
Irritação na pele	17	3,4%
Pele Intolerante e reativa	13	2,6%
Acne / borbulhas	10	2,0%
Quimio/Fototerapia	9	1,8%
Aplicação em bebé / recém nascido	8	1,6%
Ptiriase/Manchas na pele	7	1,4%
Rosácea	6	1,2%
Tromboflebite/Edema	3	0,6%
Lúpus	3	0,6%
Pós-operatório	2	0,4%
Queratose	1	0,2%
Escaras	1	0,2%
Artrite Reumatóide	1	0,2%
Total de respondentes	497	-

Distribuição dos resultados pelo motivo da compra

Essa variável corresponde a 6 categorias e foi criada com o objetivo de saber o que levou o respondente a adquirir o produto. Pretende-se saber se foi prescrição médica ou recomendação de outro profissional de saúde, da farmácia ou ainda se foi aconselhado por um amigou/familiar, por iniciativa ou publicidade.

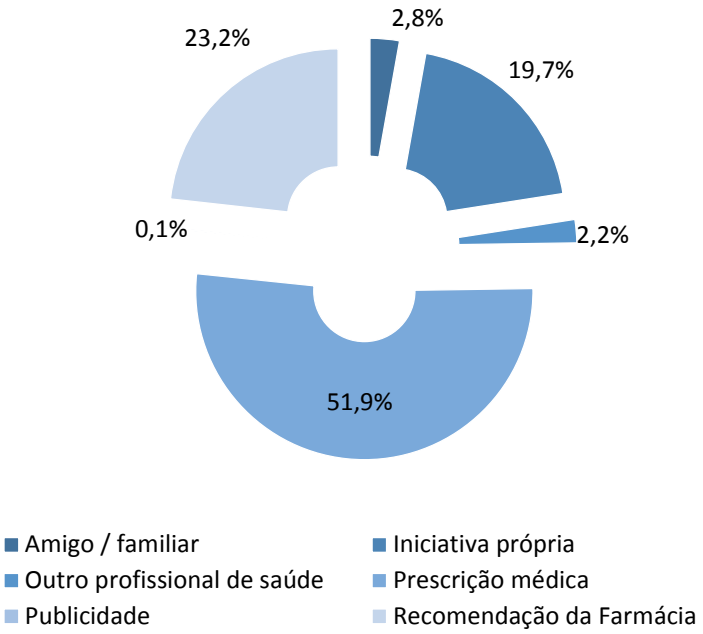

GRÁFICO 7: DISTRIBUIÇÃO DOS RESPONDENTES DA CONCORRÊNCIA, SEGUNDO O MOTIVO DE COMPRA

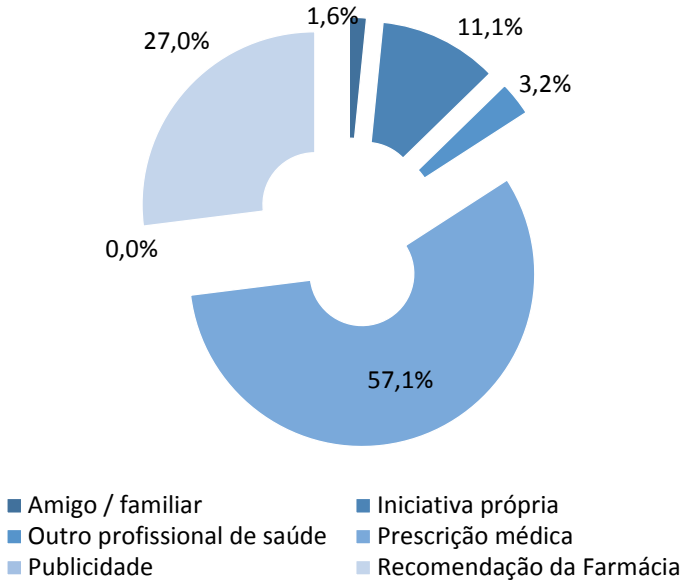

GRÁFICO 8:DISTRIBUIÇÃO DOS RESPONDENTES DA MARCA, SEGUNDO O MOTIVO DE COMPRA

Tanto para a concorrência como para a marca, prescrição médica foi o motivo mais referido, seguido de recomendação da farmácia, embora com maior proporção para marca. Para o teste de hipótese foram consideradas apenas as categorias prescrição médica, recomendação da farmácia e iniciativa própria visto que nas outras categorias o n é inferior a 5 que é o mínimo recomendado para uma tabela de contingência. Não se encontraram diferenças estatisticamente significativas entre a distribuição de adquirentes da Marca e de produtos da concorrência, segundo o motivo, ($\chi_1^2 = 2.71$; p_value $= 0{,}2578$).

Especialidade Médica

Para a especialidade médica, obteve-se um total de 885 respostas para a concorrência e 36 para a marca. O gráfico 10 ilustra a distribuição dessas respostas por tipo de especialidade médica.

Analisando o gráfico, conclui-se que os produtos da concorrência são maioritariamente prescritos pelo dermatologista (53,6%) seguido do pediatra (32.9%). Para os utilizadores da marca a situação inverte-se: Pediatra é o médico prescritor da maioria dos utilizadores (52,8%) seguido do dermatologista (44,4%). Não houve registos de outra especialidade para a marca e médico clinica geral e familiar corresponde a 2,8% (n=1). Comparando as especialidades dermatologista e pediatra obteve-se um $\chi_1^2 = 3{,}72$ que corresponde a um $\text{p\_value} = 0{,}0537$ o que significa que não se encontraram diferenças estatisticamente significativas entre a distribuição de adquirentes da Marca e de produtos da concorrência, segundo a especialidade médica.

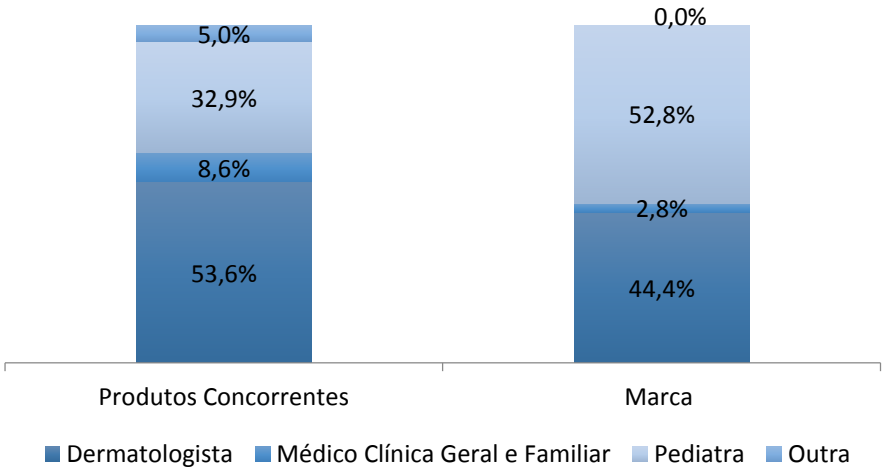

GRÁFICO 9: DISTRIBUIÇÃO DOS RESPONDENTES POR MARCA, POR ESPECIALIDADE MÉDICA

Considerando as especialidades pediatra e dermatologista obteve-se os seguintes resultados no modelo logístico. A classe de referência é a dermatologista.

Com um $\hat{\beta}_1 = 0{,}66$, o que corresponde a um $OR_1 = 1{,}93$ o que significa que o odds dos indivíduos em que o médico prescritor foi pediatra é cerca de 2 vezes maior, quando comparados com o odds dos utilizadores a quem o prescritor foi dermatologista.

TABELA 8: REGRESSÃO UNIVARIADA PARA A VARIÁVEL ESPECIALIDADE MÉDICA

Especialidade	Beta	OR	Intervalo de confiança do OR (95%)	p-value
Dermatologista (Ref.)	-3,39			<0,0001
Pediatra	0,66	1,93	[0,98 ; 3,8]	0,0575

Outra especialidade

Para os utilizadores da concorrência, outra especialidade corresponde a 5% das respostas. A tabela seguinte ilustra essas especialidades da qual imuno-alergologista lidera com 50% das respostas. Dermato-venereologia foi referido por 6 respondentes (15%) e é a segunda especialidade mais referida na categoria das outras especialidades.

TABELA 9: OUTRAS ESPECIALIDADES MÉDICAS

Outra Especialidade Médica	n	%
Imunoalergologista	20	50,0%
Dermatovenereologia	6	15,0%
Reumatologia	3	7,5%
Pneumologista	2	5,0%
Urologia	2	5,0%
Cardiologista	1	2,5%
Cirurgia Geral	1	2,5%
Cirurgia Vascular	1	2,5%
Gastroenterologista	1	2,5%
Ginecologista	1	2,5%
Medicina Interna	1	2,5%
Nutricionista	1	2,5%
Total	40	100%

Distribuição da origem da prescrição, segundo a marca

Se o motivo da compra foi prescrição médica, para além da especialidade médica pretendia-se saber ainda a origem da prescrição. Essa variável é composta por 3 categorias, hospital

público, centro de saúde e consultório privado. Obteve-se um total de 652 respostas, (624 para a concorrência e 28 para a marca).

Analisando a distribuição das prescrições médicas, por origem da prescrição verificamos que a maioria teve origem em consultas privadas, no entanto, essa proporção foi maior para a marca (93% versus 71%). Para os produtos da marca não houve prescrições originárias do centro de saúde.

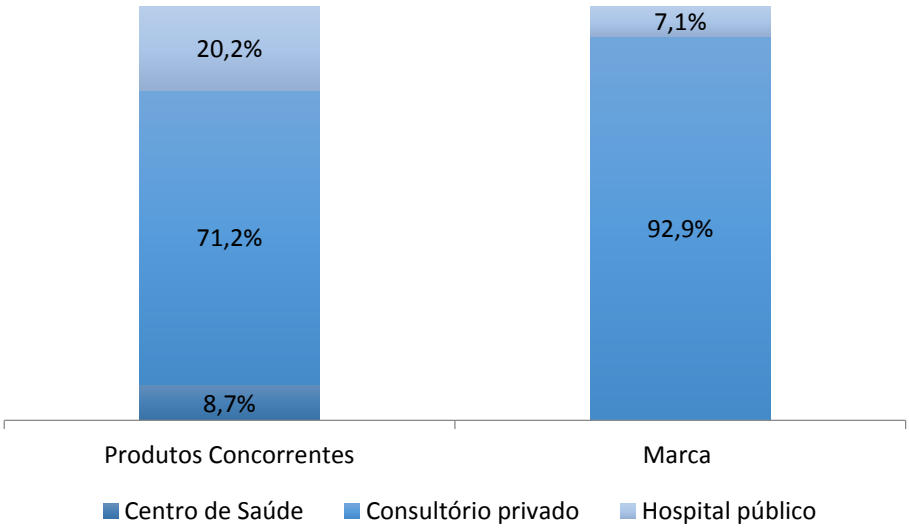

GRÁFICO 10: DISTRIBUIÇÃO DOS RESPONDENTES, POR MARCA, SEGUNDO A ORIGEM DA PRESCRIÇÃO

Para o modelo logístico, agrupou-se as categorias anteriores, hospital público e centro de saúde numa única categoria – Público. Teremos então apenas duas categorias para a análise: Consultório privado e Público (classe de referência)

Obteve-se um $\hat{\beta}_1$= 1,65, o que corresponde a um OR_1= 5,21 o que significa que o odds dos indivíduos cuja prescrição teve origem num consultório privado é 5 vezes maior, quando comparados com o odds dos indivíduos a quem a origem da prescrição foi pública.

TABELA 10: REGRESSÃO UNIVARIADA PARA A VARIÁVEL ORIGEM

Origem	$\hat{\beta}$	OR	Intervalo de confiança do OR (95%)	p-value
Pública(Ref.)	4,51			<0,0001
Consultório privado	1,65	5,21	[1,22 ; 22,15]	0,0256

Motivo da compra anterior

Para os utilizadores prevalentes, ou seja para os utilizadores que já tinham adquirido o produto anteriormente, pretende-se também saber qual o motivo da primeira compra e qual o médico prescritor dessa compra.

Analisando a variável motivo de compra anterior, por marca, verificamos que prescrição médica continua sendo o principal motivo de compra, tanto para a marca como para os produtos da concorrência.

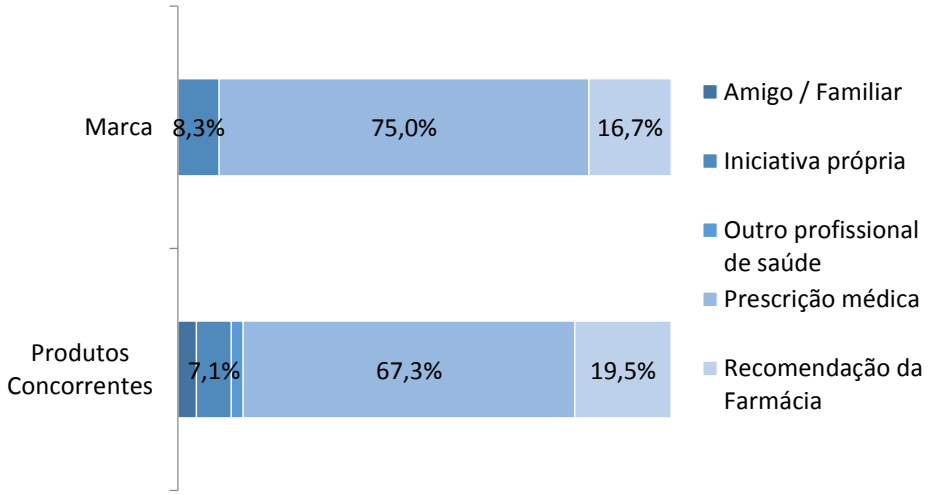

GRÁFICO 11: DISTRIBUIÇÃO DOS RESPONDENTES, POR MARCA, SEGUNDO O MOTIVO ANTERIOR

Especialidade médica anterior

Considerando apenas as especialidades dermatologista e pediatra verificou-se que não existem diferenças estatisticamente significativas entre a proporção de utilizadores da marca e a proporção de novos utilizadores de produtos da concorrência, segundo a especialidade médica anterior (χ_1^2 = 0, 76 ; p_value= 0,3840).

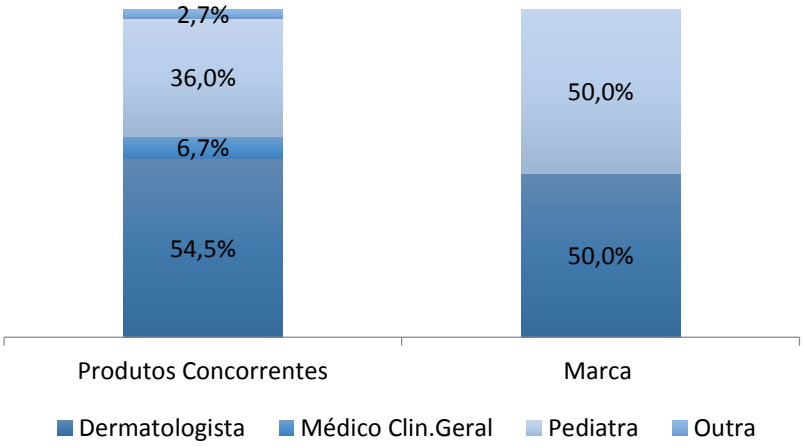

GRÁFICO 12: DISTRIBUIÇÃO DOS RESPONDENTES, POR MARCA, SEGUNDO A ESPECIALIDADE ANTERIOR

Outra especialidade médica

TABELA 11: OUTRA ESPECIALIDADE MÉDICA ANTERIOR

Outra Especialidade Médica	n	%
Imunoalergologista	9	60,0%
Reumatologia	2	13,3%
Cardiologista	1	6,7%
Cirurgia Vascular	1	6,7%
Dermatovenereologia	1	6,7%
Ginecologista	1	6,7%
Total	15	100%

Substituição da marca por produto da concorrência

Pretendia-se ainda saber se houve substituição do produto da marca por outro da concorrência. Registaram apenas seis substituições do produto da marca pelo produto da concorrência (0,6%). Esses produtos foram o Atoderm (n=2), Daveia (n=1), Leti At 4 (n=1), Lipikar (n=1) e Physiogel A.I (n=1).

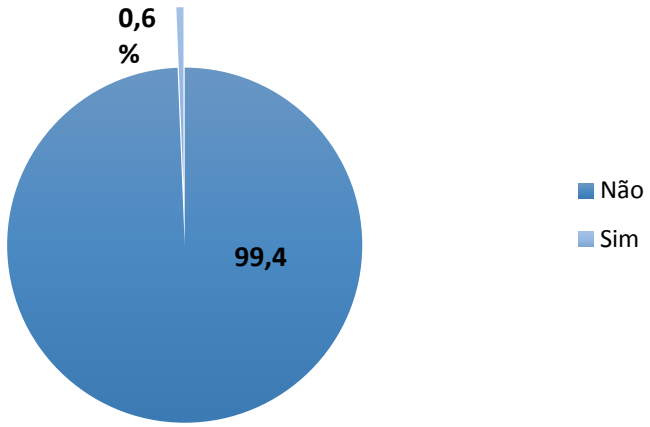

GRÁFICO 13: DISTRIBUIÇÃO DOS RESPONDENTES DA CONCORRÊNCIA, POR SUBSTITUIÇÃO

5. SELEÇÃO DE VARIÁVEIS PARA O MODELO MULTIVARIADO

Após terminar a análise univariada, inicia-se o processo de inclusão das variáveis no modelo multivariado.

Considerou-se um *p-value* de 0,20 para a entrada da variável no modelo e um *p-value* de 0,25 para a saída da variável do modelo.

Tal como foi descrito no método stepwise, incluir-se-ão no modelo todas as variáveis com **p-value** inferior a 0.20 avaliando seguidamente os valores dos *p-values* das variáveis já incluídas no modelo e excluindo-as caso o seu p-value seja superior a 0,25.

5.1. VARIÁVEIS A CONSIDERAR NO MODELO MULTIVARIADO

Algumas das variáveis com *p-values* significativos não foram consideradas nesta seleção por terem muito poucos valores conhecidos.

Por exemplo, se considerássemos as variáveis *especialidade médica* e *origem da prescrição* perderíamos informação relativamente ao motivo da compra. Estas duas variáveis só são consideradas para os utilizadores aos quais o motivo da compra foi prescrição médica.

Contudo era importante discriminar a especialidade médica no motivo da prescrição. Para isso, criou-se uma nova variável *motivo* com quatro categorias. A primeira categoria, *não prescrição médica*, corresponde à junção te todos os motivos exceptuando a prescrição médica.

As outras categorias, referem-se à prescrição médica e dividem-se nas especialidades: *pediatra, dermatologista e médico clinico geral*.

Para essa nova variável, foi também ajustado um modelo de regressão logística a fim de saber se esta entra ou não no modelo multivariado.

Sabe-se que um dos pressupostos para o modelo múltiplo é a independência das variáveis explicativas. Este pressuposto não se verifica com as variáveis diagnóstico eczema, psoríase ou outro diagnóstico porque poderia ser diagnosticado ao utente mais do que uma dessas doenças em simultâneo. Foi então considerado apenas o diagnóstico de eczema para o modelo multivariado. A tabela 12 apresenta um resumo das variáveis do modelo simples candidatas a entrar no modelo múltiplo. Analisando a tabela verifica-se que o menor p-value corresponde á variável *diagnóstico de eczema*, portanto essa será a primeira variável a entrar no modelo.

TABELA 12: VARIÁVEIS CANDIDATAS AO MODELO MULTIVARIADO

Sexo	$\widehat{\boldsymbol{\beta}}$	OR	Intervalo de confiança do OR (95%)	p-value
Feminino (Ref.)	-3,28			<0,0001
Masculino	0,01	1,01	[0,62 ; 1,67]	0,9529
Idade	$\widehat{\boldsymbol{\beta}}$	OR	Intervalo de confiança do OR (95%)	p-value
≥ 18 anos (Ref).	-3,08			<0,0001
≤ 5 anos	-0,32	0,7	[0,43 ; 1,25]	0,2464
6-17 anos	-0,62	0,5	[0,25 ;1,18]	0,1226
Primeira Vez	$\widehat{\boldsymbol{\beta}}$	OR	Intervalo de confiança do OR (95%)	p-value
Não (Ref.)	-3,61			<0,0001
Sim	0,61	1,834	[1,1 ; 3,06]	0,0201
Diagnóstico Eczema	$\widehat{\boldsymbol{\beta}}$	OR	Intervalo de confiança do OR (95%)	p-value
Não(ref)	3,03			<0,0001
sim	1,00	2,723	[1,41 ; 5,23]	0,0028
Motivo	$\widehat{\boldsymbol{\beta}}$	OR	Intervalo de confiança do OR (95%)	p-value
Não Prescrição (Ref.)	-3,42			<0,0001
Pediatra	0,68	1,979	[1,08 ; 3,13]	0,0262
Dermatologista	0,03	1,03	[0,55 ; 1,93]	0,9262
MCG	-0.91	0,4	[0,05 ; 2,99]	0,3736

5.2. SELEÇÃO STEPWISE

Passo (1)

Passemos então á seleção;

TABELA 13: PRIMEIRO PASSO PARA O PROCEDIMENTO DE SELEÇÃO DE VARIÁVEIS

Variável	$\widehat{\boldsymbol{\beta}}$	OR	E.T.W*ald*	p_value	-2LOG verosimilhança
Termo constante	3,03		471,53	<0,0001	548,29
Diagnóstico eczema	1,00	2,72	8,96	0,028	

Passo (2)

No passo 2 selecciona-se, das n-1 variáveis, a que tiver o menor p-value. Neste caso a variável com um menor p-value é a variável *primeira vez*.

TABELA 14: SEGUNDO PASSO PARA O PROCEDIMENTO DE SELEÇÃO DE VARIÁVEIS

Variáveis	$\widehat{\boldsymbol{\beta}}$	OR	E.T.W*ald*	p_value	-2LOG verosimilhança
Termo constante	-4,31		161,31	<0,0001	535,52
Diagnóstico eczema	0,97	2,63	8,29	0,004	
Primeira vez	0,58	1,78	4,88	0,0271	

A estatística G = 548,29-535,52 = 12,77 com 2 graus de liberdade e com um *p-value* 0.0017. Concluímos então que o segundo modelo é melhor que o primeiro, dando-nos mais informação da variável resposta. Analisando agora os valores dos *p-values* da estatística de *Wald*, conclui-se que são todos significativos, assim ambas variáveis permanecem no modelo. Prosseguindo para o passo (3) a próxima variável a ser incluída no estudo será a variável *motivo*.

Passo (3)

TABELA 15: TERCEIRO PASSO PARA O PROCEDIMENTO DE SELEÇÃO DE VARIÁVEIS

Variável	$\widehat{\boldsymbol{\beta}}$	OR	E.T.W*ald*	p_value	-2LOG verosimilhança
Termo constante	-4,38		141,66	<0,0001	512,141
Diagnóstico eczema	0,98	2,67	7,74	0,0054	
Primeira vez	0,50	1,66	3,65	0,056	
Dermatologista	0,55	1,73	3,03	0,082	
Médico Clinico Geral	-0,02	0,98	0,01	0,9436	
Pediatra	-1,00	0,37	0,96	0,3274	

Com a introdução da variável *motivo* a estatística G passa a ser G = 535,52 – 512.14 = 23.38 com 3 graus de liberdade dando um *p_value* <0.0001.

Concluímos então que o terceiro modelo é melhor que o primeiro, dando-nos mais informação da variável em estudo. Analisando agora os valores dos *p-values* da estatística de *Wald*, conclui-se que são todos significativos, assim todas variáveis permanecem no modelo.

Prosseguindo para o passo (3) a próxima variável a ser incluída no estudo será a variável classe etária; contudo, antes de introduzir a variável no modelo, é necessário verificar se a variável idade é ou não linear no *logit* e, caso seja, então está terá de ser introduzida no modelo como uma variável contínua.

5.3. LINEARIDADE NO LOGIT

Utilizaremos o método dos quartis para verificar a linearidade no *logit*. Em primeiro lugar determinam-se os quartis da variável em estudo, em seguida caracterizá-la-emos com base nos quartis obtidos. Esta variável será introduzida no modelo final obtendo assim os coeficientes estimados para as novas variáveis. Após obter os coeficientes estimados, construir-se-á então um gráfico de quartis *vs* $\hat{\beta}_i$.

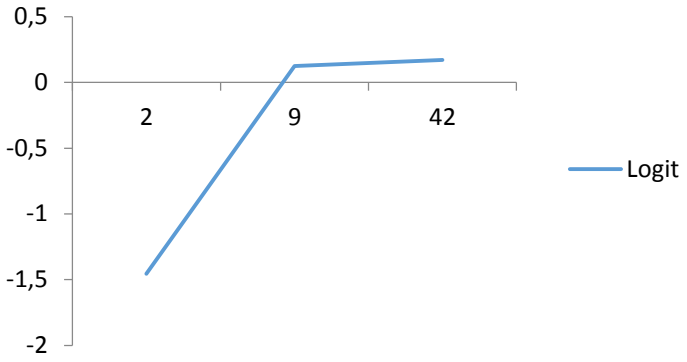

GRÁFICO 14: LINEARIDADE NO LOGIT

A variável idade é considerada línear no *logit*, desde que a sua representação gráfica seja considerada aproximadamente linear o que não se verifica neste caso.

Sendo assim, já que a variável idade não é linear no *logit*, esta será introduzida no modelo como uma variável categórica com as 3 classes já definidas anteriormente.

Passo (4)

Continuando a construção do modelo, precederemos á introdução da variável classe etária no modelo.

TABELA 16: QUARTO PASSO PARA O PROCEDIMENTO DE SELEÇÃO DE VARIÁVEIS

Variável	$\widehat{\beta}$	OR	E.T.W*ald*	p_value	-2LOG verosimilhança
Termo constante	4,04		114,78	<0,0001	501,766
Diagnóstico eczema	1,15	3,17	10,41	0,0013	
Primeira vez	0,54	1,71	4,02	0,0449	
Dermatologista	0,94	2,57	6,93	0,0085	
Médico Clinico Geral	0,21	0,81	0,40	0,5277	
Pediatra	-0,99	0,37	0,92	0,3377	
≤ 5 anos	0,96	0,39	7,59	0,0059	
6-17 anos	-1,01	0,37	5,20	0,0191	

Com a introdução da variável classe etária, obtém-se uma estatística G=512,14-501,77= 10,37 com 4 graus de liberdade, que corresponde a um *p_value* associado igual 0.0346.

Concluímos então que o modelo é melhor que o anterior, dando-nos mais informação da variável em estudo. Analisando agora os valores dos p-values da estatística de *Wald*, concluímos que são todos significativos, assim todas as variáveis permanecem no modelo.

As restantes variáveis que faltam incluir no modelo não são significativos, e nem tão pouco têm *p_value* inferior ao valor de inclusão 0.20. Desta forma, não avançaremos mais com a inclusão de variáveis no modelo.

Conclui-se então que o último modelo será o nosso modelo final construído pelas variáveis, *diagnóstico de eczema*, *primeira vez*, *motivo* e *classe etária*.

6. INTERPRETAÇÃO DOS COEFICIENTES DO MODELO FINAL

Após concluir o modelo multivariado, proceder-se-á à interpretação dos coeficientes estimados, em termos dos OR.

Variável	$\widehat{\beta}_i$	OR	Conclusão
Diagnóstico de eczema	1,15	3,17	O odds dos indivíduos a quem foi diagnosticado eczema/eczema atópico é 3 vezes superior, quando comparados com o odds dos indivíduos a quem não foi diagnosticado eczema/eczema atópico.
Primeira vez	0,54	1,71	Quando comparamos o odds dos utilizadores que compram o produto pela primeira vez com o odds utilizadores prevalentes, este aumenta em 70%.
Classe etária			O que significada que o odds dos utilizadores com menos de 5 anos e com idade compreendida entre os 6 e os 17 anos, diminui em 60% quando comparado com o odds dos utilizadores com idade superior a 18 anos.
≤ 5 anos	-0,96	0,39	
6 a 17 anos	-1,01	0,37	

Cont.

Variável	$\widehat{\beta}_i$	OR	Conclusão
Motivo de compra			
Dermatologista	0,94	2,57	O odds dos utilizadores a quem a compra do produto foi prescrito por um dermatologista aumenta em duas vezes quando comparados com o odds dos utilizadores a quem o motivo da compra não foi prescrição médica.
Médico clinico geral	-0,21	0,81	O odds diminui em 20% quando comparamos utilizadores cujo médico prescritor foi um médico clinico geral com o odds dos utilizadores a quem o motivo da compra não foi prescrição médica.
Pediatra	-0,99	0,37	O odds dos indivíduos cujo médico prescritor foi pediatra diminui em 60% quando comparado com o odds dos utilizadores cujo motivo da compra não foi prescrição médica.

7. DIAGNÓSTICO DO MODELO

7.1. TABELA DE CONTINGÊNCIA

Com a tabela de classificação abaixo indicada, (cut-point = 0,04) verifica-se que 66.5% dos indivíduos estão bem classificados, sendo a sensibilidade e a especificidade 58.1% e 66.8% respetivamente.

TABELA 17: TABELA DE CLASSIFICAÇÃO

Classificação	Marca		
	Sim (1)	Não (0)	Total
Sim (1)	36	543	579
Não (0)	26	1093	1119
Total	62	1636	1698

7.2. CURVA ROC

De acordo com os resultados obtidos, podemos dizer que a área abaixo da curva ROC, é de 70%. Conclui-se então que a discriminação é aceitável.

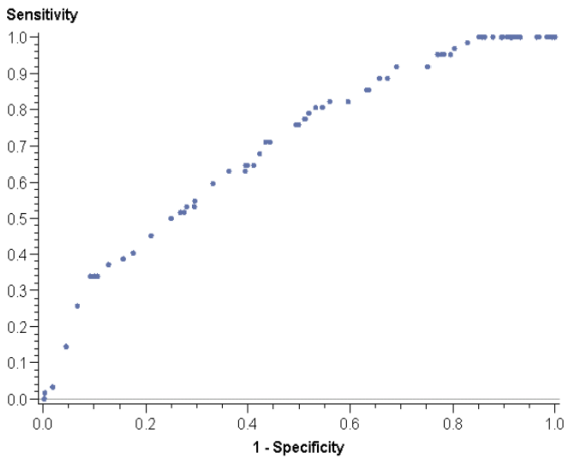

GRÁFICO 15: CURVA DE ROC

7.3. TESTE DE HOSMER-LEMESHOW, PEARSON E DEVIANCE

A tabela seguinte ilustra os resultados obtidos no teste de Hosmer-Lemeshow. O valor da estatística de teste obtido foi $\chi^2_7 = 4{,}89$ a que corresponde um p_value $= 0{,}6732$. Conclui-se que não existe evidência estatistica para rejeitar a hipótese de o modelo se encontrar bem ajustado.

TABELA 18: PARTIÇÃO PARA O TESTE DE HOSMER- LEMESHOW

	Marca		Concorrência		
Grupo	Observado	Esperado	Observado	Esperado	Total
1	0	1,3	171	169,7	171
2	3	2,5	180	180,5	183
3	4	3,3	185	185,7	189
4	4	3,9	184	184,1	188
5	5	4,2	149	149,8	154
6	7	4,1	122	125,0	129
7	5	7,5	184	181,5	189
8	10	11,2	202	200,8	212
9	24	24,0	259	259,0	283

Sob a mesma hipótese, as estatísticas de teste obtidas para os resíduos de Pearson e Deviance foram $\chi^2_{37} = 28{,}8$ e $\chi^2_{37} = 29{,}5$ a que correspondem os p-values 0,7789 0,8053,respectivamente

Analisando estes resultados, e os anteriores, podemos concluir que o modelo se encontra, na generalidade, bem ajustado.

CAPÍTULO IV
CONCLUSÕES

CAPITULO IV – CONCLUSÕES

A amostra em estudo é composta maioritariamente por indivíduos do sexo feminino (marca: 53,8% *vs* concorrência: 54,0%). Analisando a distribuição dos respondentes por faixa etária, verifica-se que não existem diferenças estatisticamente significativas entre os adquirentes da Marca e os de produtos da concorrência, (χ_3^2 =2,99; *p_value* = 0,2241).

Para 60,9% dos utilizadores de Marca esta é a 1ª utilização, enquanto para os aquirentes dos produtos concorrentes a maioria são prevalentes (53,7%).

O diagnóstico de Eczema / Eczema atópico foi o mais prevalente nos utilizadores de Marca (81,5%). Foi ainda registado o diagnóstico de Psoríase, a 10 indivíduos (15,4%).

Para o produto adquirido, a percentagem de prescrição médica, no momento da compra, situou-se nos 57,1% para os produtos Marca e 51,9% para os produtos da concorrência.

As prescrições de todos os produtos em estudo, tiveram maioritariamente origem no consultório privado (92,9% para a Marca e 71,2% para os produtos da concorrência).

A especialidade médica mais referida foi Pediatria para os produtos da Marca e Dermatologista para a concorrência, tanto para os produtos adquiridos no momento do estudo como também para os produtos adquiridos anteriormente.

O modelo logístico multivariado é composto pelas variáveis diagnóstico de eczema, primeira vez, motivo de compra e classe etária.

Quanto á variável *diagnóstico de eczema/eczema atópico* concluímos que o odds dos indivíduos a quem foi diagnosticado eczema/eczema atópico é 3 vezes superior, quando comparados com o odds dos indivíduos a quem não foi diagnosticado eczema/eczema atópico.

No que diz respeito á primeira vez, verificamos que quando comparamos o odds dos utilizadores que compram o produto pela primeira vez com o odds utilizadores prevalentes, este aumenta em 70%.

Para a variável *motivo de compra*, onde a classe de referência foi "não prescrição médica", concluímos que o odds dos utilizadores diminui em 20% e 60% respectivamente para as especialidades médico clinico geral e pediatra. Para a dermatologia o odds é 3 vezes superior.

Relativamente á idade, onde a classe de referência foi " ≥18 anos", verificamos que odds é inferior em 60% para qualquer uma das classes, ≤ 5 anos e de 6 a 17 anos.

Analisando os resultados do diagnóstico do modelo, nomeadamente curva ROC e testes estatísticos aos resíduos, concluímos que o modelo se encontra bem ajustado.

Considerando um cut-point de 0,04 obtivemos 66.5% dos indivíduos bem classificados.

REFERÊNCIAS BIBLIOGRÁFICAS

Ferreira, A. P. (2002). *Modelo de Regressão Logística Multinomial [Dissertação]*.

Gutiérrez, E. Q., & Collantes, D. S. (2011). *Dermatologia Básica em Medicina Familiar*. Lidel- edições técnicas, lda.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression.* Wiley.

Kleinbaum, D., Kupper, L., & Muller, K. (1988). *Applied Regression Analysis and Other Multivariate Methods.* USA: Duxbury Press.

Lopes, J. L. (2007). *Fundamental dos Estudos de Mercado- Teoria e Prática.* Edições Sílado.

Spiegel, M. R. (1993). *Estatística.* São Paulo: Makron books.

Stokes, M. E., Davis, C. S., & Koch, G. G. (2000). *Categorical Data Analysis using The SAS System* (Second Edition ed.). Cary,NC: SAS institute Inc.

Turkman, M. A., & Silva, G. L. (2000). *Modelos Lineares Generalizados - da teoria à prática.* Lisboa: Edições SPE.

http://www.consulteodermatologista.com/Psoriase.aspx

http://www.consulteodermatologista.com/Eczema-atopico-ou-Dermatite-atopica.aspx

Anabela Costa da Silva, *"Análise Estatística de inquéritos Online",* Disponível em:

http://repositorium.sdum.uminho.pt/bitstream/1822/19262/1/Anabela%20Costa%20da%20Silva.pdf

Raquel Maria Jacinto Escola *"Obesidade e PHDA Infantil: Modelo de Regressão Logística",*

Disponível em:

https://dspace.ist.utl.pt/bitstream/2295/575331/1/teseRaquel.pdf