UNIVERSIDADE DE ÉVORA

MODELAÇÃO DO CANCRO DA MAMA NA REGIÃO DO ALENTEJO

DISSERTAÇÃO DE MESTRADO EM MATEMÁTICA E APLICAÇÕES
ESPECIALIZAÇÃO EM ESTATÍSTICA

Helena Isabel Martins de Oliveira

Orientação Científica

Orientadora: Prof.ª Dr.ª M. Manuela Melo Oliveira
Co-Orientadora: Prof.ª Dr.ª Isabel Natário

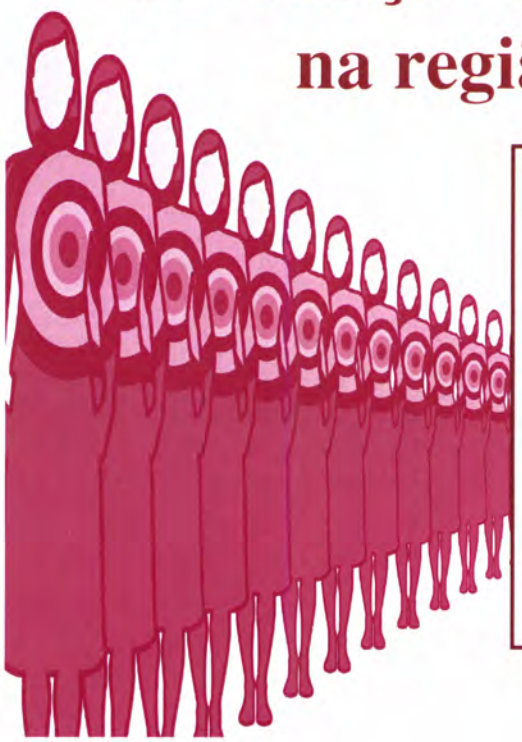
ÉvorA
2010


UNIVERSIDADE DE EVORA

UNIVERSIDADE DE ÉVORA

MESTRADO EM MATEMÁTICA E APLICAÇÕES
Especialização em Estatística

Modelação do cancro da Mama na região do Alentejo



Dissertação apresentada ao Departamento de Matemática da Universidade de Évora para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações (Especialização em Estatística) realizada sob a orientação de Prof.ª Dr.ª M. Manuela Oliveira e co-orientação de Prof.ª Dr.ª Isabel Natário.

Helena Isabel Martins de Oliveira

Dedicatória

Concluo o Mestrado em Matemática e Aplicações e continuo a frequentar o curso da minha vida profissional, não encontrarei mãos que me puxem para cima nesta vida profissional, apenas levo a certeza de encontrar degraus, os quais galgarei passo a passo, ritmados e reforçados por aquilo que aprendi!

À Minha Família, que nos momentos da minha ausência dedicados ao estudo superior, sempre me fizeram entender que o futuro é feito a partir da constante dedicação no presente!

Aos meus amigos e amigas, minha segunda família, companheiros de luta, que sempre acreditaram em mim e me incentivaram a ir mais além, jamais vos esquecerei!

Por fim, aos que me permitiram tudo isto e ao longo de toda a minha vida sempre me acompanharam, os meus pais, o meu muito obrigado, reconheço cada vez mais em todos os meus momentos, que são vocês os meus maiores mestres.

A todas as guerreiras na luta contra o cancro da mama, vós sois vitoriosas!

Agradecimentos

Há tantos a agradecer, por tanto se dedicarem a mim, não somente por me terem ensinado, mas por me terem feito aprender!

A todos os professores – a palavra mestre, nunca fará justiça à vossa dedicação, aos quais, sem nomear nomes, terão o meu eterno agradecimento!

À Universidade de Évora – por me ter dado esta oportunidade de vislumbrar um horizonte superior, o meu Muito Obrigado é mísero por tamanha competência!

Ao Hospital de Évora, sua Direcção e Adminstração – por terem facultado os dados dos vossos pacientes para a realização deste estudo.

Ao Serviço de Anatomia Patológica do Hospital de Évora – pela disponibilidade, carinho e à vontade com que sempre me receberam aquando das visitas para pesquisa dos dados.

Ao Dr. Luis Gonçalves (Médico Patologista) e à Dr. Ruth Sardinha (Técnica de Anatomia Patológica e Responsável pela Unidade de Patologia Molecular do SAP) – pela ajuda no retirar dos dados e na compreensão dos conceitos relacionados com os mesmos, ao Dr. Luís Gonçalves os meus sinceros agradecimentos pelos esclarecimentos e a disponibilidade para me receber no seu gabinete, bem como a disponibilização do seu computador de trabalho.

Às Prof.(as) Dr.(as) Orientadora Maria Manuela Oliveira e Co-Orientadora Isabel Natário, por tanto que me ensinaram, por terem estado do meu lado e terem acreditado sempre no meu trabalho, incentivando-me a não desistir, orientando-me sempre e principalmente quando estava mais perdida entre os números, obrigado pelas ajudas nas análises estatísticas e pelas palavras motivadoras que sempre me animaram!

"Antes de sentirmos que somos bons mestres,
estejamos seguros de que somos bons estudantes!"
Pitágoras

MODELAÇÃO DO CANCRO DA MAMA NA REGIÃO DO ALENTEJO

Resumo

Estudos epidemiológicos são estudos estatísticos onde se procura relacionar ocorrências de eventos de saúde com uma ou várias causas específicas. A importância que os modelos epidemiológicos assumem hoje no estudo de doenças de foro oncológico, em particular no estabelecimento das suas etiologias, é incontornável. Segundo *Ogden, J. (1999)* o cancro é "um crescimento incontrolável de células anormais que produzem tumores chamados neoplasias". Estes tumores podem ter origem benigna (não se espalham pelo corpo) ou maligna (apresentam metastização de outros órgãos). Sendo uma doença actual, com uma elevada taxa de incidência em Portugal quando comparada com outras doenças *(Instituto Nacional de Estatística – INE, 2009)*, aumentando esta taxa com a idade tal como refere *Marques, L. (2003)*, podendo ocorrer o diagnóstico desta doença em qualquer idade. De acordo com *INE (2000)* pode dizer-se que o cancro está entre as três principais causas de morte em Portugal, registando-se um aumento progressivo do seu peso proporcional, sendo o cancro da mama o tipo de cancro mais comum entre as mulheres e uma das doenças com maior impacto na nossa sociedade.

O objectivo principal deste trabalho é a estimação e modelação do risco de contrair uma doença de natureza não contagiosa e rara (neste caso, cancro da mama), usando dados da região do Alentejo. Pretende-se fazer um apanhado das metodologias mais empregues nesta área e aplicá-las na prática, com ênfase nos estudos caso-controlo e nos modelos lineares generalizados (GLM) – mais concretamente regressão logística. Os estudos caso-controlo são usados para identificar os factores que podem contribuir para uma condição médica, comparando indivíduos que têm essa condição ***(casos)*** com pacientes que não têm a condição, mas que de resto são semelhantes ***(controlos)***. Neste trabalho utilizou-se essa metodologia para estudar a associação entre o viver em ambiente rural/urbano e o cancro da mama. Tendo em conta que o objectivo principal deste estudo se prende com o estudo da relação entre variáveis, mais propriamente, análise de influência que uma ou mais variáveis (explicativas) têm sobre uma variável de interesse (resposta), para esse efeito são estudados os modelos lineares generalizados – GLM – unificados na mesma moldura teórica pela primeira vez por *Nelder & Wedderburn (1972)* - e, posteriormente aplicados ao conjunto de dados sobre cancro da mama na Região do Alentejo.

O presente trabalho pretende assim, ser um contributo na identificação de factores de risco do cancro da mama na região do Alentejo.

Palavras-Chave: cancro da mama, risco de doença, epidemiologia, caso-controlo, razão de chances, modelos lineares generalizados.

MODELING BREAST CANCER IN THE ALENTEJO REGION

Abstract

Epidemiological studies are statistical studies where attempts to relate occurrences of health events with one or more specific causes. The importance of epidemiological models that are far in the study of diseases of cancer forum, particularly in establishing their etiology, is inescapable.

According to *Ogden, J. (1999)* cancer is "an incontrollable growth of abnormal cells that produce tumors called cancer". These tumors may be benign (not spread throughout the body) or malignant (show metastasis to other organs). Being a current illness with a high incidence rate in Portugal compared with the same respect to other diseases *(National Statistics Institute – INE, 2009)* having an increasing rate with age as mentioned *Marques, L. (2003)*, and can possibly be diagnosed at any age. According to *INE (2000)* the cancer is among the top three causes of death in Portugal and there is a progressive increase of its proportional weight. Breast cancer is the most common form of cancer among women and the diseases with major impact in our society.

The main objective of this work is to model and estimate the risk of contracting a non-contagious and rare disease (in this case, breast cancer), using data from the Alentejo region. It is intended to summarize some of the methodologies employed in this area and apply them in practice, with emphasis on case-control studies and generalized linear models (GLM) – more specifically the logistic regression. The case-control studies are used to identify factors that may contribute to a medical condition, comparing individuals who have this condition ***(cases)*** with patients who have not the condition but that are otherwise similar ***(controls)***. In this work we used this methodology to study the association between living in a rural/urban and breast cancer. Given that the main objective of this study rather relates to the study of the relationship between variables to analyze the influence that one or more variables *(explanatory)* have on a variable *(response)*, for this purpose we study the generalized linear models – GLM – first unified in the same theoretical framework by *Nelder and Wedderburn (1972)* and subsequently applied to the data set on breast cancer in the Alentejo region.

This work intends to be a contribution in identifying risk factors for breast cancer in the Alentejo region.

Keywords: breast cancer, risk of illness, epidemiology, case-control, odds ratio, generalized linear models.

Índice Geral

Índice de Ilustrações

Figuras

Tabelas

Capítulo 1

Introdução

O termo genérico Cancro é aplicado a várias doenças distintas mas relacionadas, todas elas caracterizadas pelo crescimento e desenvolvimento incontrolado de células anómalas.

O número de pessoas que apresentam doenças relacionadas com cancro continua a aumentar em todo o mundo. Segundo um estudo da *American Cancer Society (2007),* uma em cada oito mortes no mundo deve-se ao cancro, sendo o cancro a segunda causa de morte nos países desenvolvidos (depois das doenças de coração) e a terceira causa de morte nos países em desenvolvimento (depois das doenças de coração e doenças diarréicas). Segundo o mesmo estudo estimou-se que, em 2007, surgissem, diariamente, mais de 33 mil novos casos de cancro e morressem mais de 20 mil pessoas vítimas de cancro no Mundo. Nos países em vias de desenvolvimento, tendo em conta o número total de habitantes, os números são ainda mais elevados (6.7 milhões de casos e 4.7 milhões de mortes) que nos países desenvolvidos (5.4 milhões de casos e 2.9 milhões de mortes). A *Organização Mundial de Saúde (2005)* realizou previsões para 2020 estimando que nesse ano surgirão 10.9 milhões de novos casos de cancro, 24.6 milhões de pessoas a viver com a doença e, não agindo, 10.3 milhões de mortes por ano no Mundo inteiro, verificando-se assim um aumento de quase 50% do número de casos de 2002 para 2020.

Na Europa, e segundo *European Cancer Observatory (2006),* de entre os tipos de cancros mais comuns, nos homens destacam-se: próstata (307013 novos casos e 70381 mortes), pulmão (206161 novos casos e 181854 mortes) e cólon e recto (169896 novos casos e 78353 mortes) e nas mulheres o destaque vai para: mama (331392 novos casos e 89674 mortes), cólon e recto (139654 novos casos e 68114 mortes) e pulmão (73972 novos casos e 66302 mortes).

Em Portugal, segundos dados do *INE (2009)* verificou-se um aumento do número de casos de cancro no período compreendido entre 2002 e 2008, sendo neste período a segunda causa de morte, após as doenças cardiovasculares. Só em 2004, segundo dados da *Direcção Geral de Saúde (2004),* ocorreram 22319 óbitos por cancro em Portugal Continental e Regiões Autónomas, sendo 1217 casos relativos à região do Alentejo. Os distritos que apresentam maior taxa de mortalidade por cancro são: Beja, Setúbal, Lisboa, Porto e Viana do Castelo.

De entre os vários tipos de cancro, *INE (2008)* refere que os tumores da laringe, da traqueia, dos brônquios e dos pulmões ocupam o primeiro lugar como causa de morte nos homens (20.8%) seguido do tumor da próstata (12.3%). Idêntica hierarquia é reservada, nas mulheres, ao tumor maligno da mama (16.4%) e ao cólon, recto e ânus (11.8%). O tumor do estômago verifica um peso

superior nos homens (11.1%) ao das mulheres (9.9%) e detém, em ambos os casos, a terceira posição.

Na década de 90 observou-se uma tendência crescente da mortalidade por cancro da mama, actualmente em declínio em diversos países (em Portugal houve uma diminuição de 2%/ano entre 1992 e 2002) – *Bastos, Barros e Lunet (2007).* Segundo *European Cancer Observatory (2006),* em Portugal, verificaram-se 6335 novos casos de cancro da mama e 1588 mortes devido a essa doença. No período compreendido entre 2002 e 2008, *INE (2009)* refere que não se registaram alterações significativas no número de mortes por cancro da mama em Portugal.

Neste trabalho pretende-se retratar a realidade que se vive na região do Alentejo no que respeita ao cancro da mama. Para tal, explorou-se a base de dados da *Unidade de Anatomia Patológica do Hospital do Espírito Santo em Évora* (*HESE*), relativa a indivíduos a quem foi diagnosticado um tipo de neoplasia mamária no período compreendido entre Agosto de 2003 e Agosto de 2004. Com o intuito de identificar factores de risco foram aplicadas aos dados em estudos algumas metodologias epidemiológicas que se consideraram adequadas.

Durante algum tempo prevaleceu a ideia de que a epidemiologia se restringia ao estudo de epidemias de doenças transmissíveis, contudo, hoje é uma área reconhecida, que trata de qualquer evento relacionado com a saúde (ou doença) da população. Resumidamente, segundo *Last, J. (2001)* a epidemiologia estuda a distribuição e as causas dos estados de saúde ou eventos em populações específicas para posterior aplicação desse estudo na prevenção e controlo de problemas de saúde.

Neste trabalho destacam-se os estudos caso-controlo que tentam identificar os factores de risco para uma doença, comparando indivíduos em que esta está presente (*casos*) ou ausente (*controlos*), de forma retrospectiva (pois a doença e a exposição já aconteceram no momento do delineamento do estudo) na tentativa de encontrar uma possível associação. A medida estatística de associação analisada nos estudos caso-controlo é a razão de chances (*odds ratio*) semelhante ao risco relativo, no caso da doença ser rara. Definimos risco como a probabilidade que um indivíduo ou grupo de indivíduos tem de apresentar, no futuro, uma determinada doença. O risco relativo (*RR*) é uma medida de associação entre a exposição e o evento (resultado), sendo definido como a razão entre a incidência nos indivíduos expostos/não expostos. A *razão de chances* ou *odds ratio* (*OR*) é uma aproximação do risco relativo, representando a comparação da chance de doença sob exposição ao factor de risco com a chance de doença sem exposição ao factor de risco.

Como referimos um dos objectivos do trabalho é estudar a relação entre covariáveis, ou mais particularmente, analisar a influência que uma ou mais covariáveis (explicativas ou independentes) medidas em indivíduos têm sobre uma variável de interesse (variável explicada, endógena ou resposta). A forma de abordar este problema, em termos estatísticos, é através do estudo de um modelo de regressão que relacione essa variável de interesse com as variáveis explicativas.

Dada a natureza dos dados em mãos, tal é conseguido recorrendo a modelos da classe dos modelos lineares generalizados – GLM – *Nelder & Wedderburn (1972)* - que unificam, tanto do ponto de vista teórico, como conceptual um leque de vários modelos estatísticos.

O presente trabalho encontra-se dividido em três partes:

- Uma primeira parte (capítulo 2) em que se descreve o enquadramento do problema em estudo, abordando-se conceitos e termos relacionados com o cancro da mama e, em que, seguidamente, se analisa um conjunto de casos de pacientes residentes na região do Alentejo, a quem foi diagnosticado um tipo de neoplasia mamária na *Unidade de Anatomia Patológica do HESE* (base de dados **CASOS**). Descrevem-se aqui as variáveis em estudo (*Descrição dos dados e Descrição das Variáveis*) e apresenta-se uma análise descritiva preliminar e exaustiva dos dados.

- Uma segunda parte (capítulo 3) em que se aplicam modelos epidemiológicos de análise de doenças crónicas, não contagiosas e raras (em particular, doenças do foro oncológico) aos dados em estudo. Inicia-se uma breve descrição teórica sobre epidemiologia e estudos associados à mesma, com particular ênfase nos estudos *caso-controlo* (o mais adequado à situação em estudo e aos dados em questão). Aplica-se esta metodologia, posteriormente, à comparação da base de dados **CASOS** (anteriormente referida) com uma nova base de dados - **CONTROLOS** (construída tendo em conta pacientes que passaram pela *Unidade de Anatomia Patológica do HESE* a quem não lhes foi diagnosticado qualquer tipo de neoplasia mamária) – realizando-se assim um estudo do tipo *caso-controlo*.

- Uma terceira parte (capítulo 4) em que se descrevem teoricamente os modelos lineares generalizados (**GLM**) e se apresentam três aplicações práticas dos mesmos, para modelar os dados da base de dados **CASOS**, com o objectivo de detectar factores de risco e factores de prognóstico para o cancro da mama, na região em estudo, com recurso ao ***software*** estatístico ***R***.

- Por fim apresentam-se as considerações finais (capítulo 5).

Capítulo 2

Enquadramento do Problema

2.1 A Incidência do Cancro da Mama

Segundo a *Organização Mundial de Saúde – OMS (2009)* o cancro é uma das principais causas de morte no mundo, sendo em 2004, responsável por 7.4 milhões de mortes (cerca de 13% do total de mortes), levando o cancro da mama a 519 000 mortes no mundo.

Segundo dados da *Sociedade Portuguesa de Senologia – SPS (2008)* a incidência de cancro da mama está a aumentar a cada ano, em todo o mundo, sendo, actualmente responsável por, aproximadamente 23% dos novos casos de cancro diagnosticados por ano, ocupando o segundo lugar geral, mas ainda é a causa mais frequente de morte por cancro em mulheres. São diagnosticadas, por ano, 1 100 000 mulheres com cancro da mama, 360 mil só na Europa. O risco de desenvolver cancro da mama, tendo em conta a população mundial, é de uma em cada dez mulheres e a incidência está a aumentar, sendo responsável por mais de um quarto dos novos casos de cancro.

Segundo dados da *Globocan (2008)* as taxas de incidência variam de 19.3 por 100 000 mulheres na África Oriental a 89.9 por 100 000 mulheres na Europa Ocidental e são elevadas (superiores a 80 por 100 000) em regiões desenvolvidas do mundo (com excepção do Japão) e baixas (inferior a 40 por 100 000) na maioria das regiões em desenvolvimento.

No que respeita às taxas de incidência e de mortalidade por cancro da mama tendo em conta o continente Europeu, segundo dados do *European Cancer Observatory (2006)* destacam-se com maior incidência de cancro da mama em 2006: a Bélgica (137.8 novos casos por cada 100 000 e 2609 mortes), a Holanda (128 novos casos por 100 000 e 3267 mortes), a França (127.4 novos casos por 100 000 e 11299 mortes), a Suiça (126.5 novos casos por 100 000 e 1225 mortes), a Suécia (125.8 novos casos por 100 000 e 1434 mortes) e, por fim, a Dinamarca (122.6 novos casos por 100 000 e 1350 mortes).

Em Portugal, e de acordo com a *Liga Portuguesa Contra o Cancro*, todos os dias morrem quatro a cinco mulheres devido ao cancro da mama, anualmente, são diagnosticados mais de 3800 novos casos. Apesar dos avanços no diagnóstico e tratamento deste tipo de cancro, é, actualmente, o mais comum nas mulheres, um dos mais temíveis e também a principal causa de morte nas mulheres entre os 35 e os 55 anos.

Segundo dados recentes do *Plano Nacional de Saúde - PNS (2004/2010, vol.1, pág.58),* a taxa de mortalidade portuguesa padronizada por cancro da mama feminina (/100 000 mulheres), abaixo dos 65 anos tem vindo a aproximar-se da meta estabelecida para 2010 (10,0). Entre 2004 e 2008, a taxa

atrás mencionada diminuiu em Portugal Continental de 12.4 para 11,8 por 100 000 mulheres. Contudo, durante este período, registaram-se variações regionais acentuadas. No Norte, Alentejo e Algarve os valores aumentaram e no Centro e Lisboa e Vale do Tejo verificou-se uma evolução positiva. Tendo em conta o que se passou na região do Alentejo a taxa de mortalidade padronizada por cancro da mama feminina por cada 100 000 mulheres aumentou de 10.9 para 12.8. Contudo, não se pense que o cancro da mama é uma doença que só afecta as mulheres, embora com uma taxa de incidência bem menor, o tumor da mama também afecta os homens podendo até ser mais mortífero, uma vez que o diagnóstico desta doença no sexo masculino é muitas vezes tardio, o que obriga a tratamentos mais agressivos e consequentemente, com piores resultados. Em Portugal, morrem, por ano, cerca de 20 homens com cancro da mama, tal como se pode verificar no *Portal de Oncologia Português*. Segundo dados da *Liga Portuguesa contra o Cancro*, prevê-se, nas próximas duas décadas, um aumento exponencial para o cancro da mama a nível mundial. Segundo o mesmo relatório, esta evolução poderá dever-se ao aumento da esperança de vida. Posto isto, é necessário apostar na prevenção, sensibilização, educação e no rastreio da doença. Segundo o *PNS (2009)*, a cobertura geográfica do rastreio do cancro da mama era já bastante vasta, abrangendo 74.5% dos concelhos de Portugal Continental, o que corresponde à totalidade das regiões do Centro, Alentejo e Algarve, bem como aos distritos de Santarém e alguns distritos do Norte. Prevê-se que no ano de 2010 o número de concelhos abrangidos pelo programa de rastreio do cancro da mama ascenda aos 84.5% e em 2011 aos 100%.

2.2 Morfologia Mamária e o Cancro da Mama

Composição da Mama

As mamas ou seios normais são glândulas cuja principal função é a produção de leite. Cada mama assenta nos músculos do peito (peitorais) que cobrem as costelas. Elas são compostas por lobos (cada mama dividida em 15 a 20 lobos), que se dividem em lóbulos os quais contêm grupos de pequenas glândulas que produzem leite. O leite flui dos lóbulos através de uns tubos finos, os ductos, que permitem a passagem do leite para fora, por meio do mamilo (centro de uma área escura de pele - chamada auréola). Todo o espaço compreendido entre os lóbulos e os ductos é preenchido com gordura. Na mama, tal como nos outros órgãos do corpo humano, também se encontram vasos sanguíneos que permitem irrigar a mama de sangue e ainda, os vasos linfáticos que facultam a circulação da linfa (líquido cuja função é o transporte de nutrientes para as diversas partes do corpo e a recolha das substâncias indesejáveis). Segundo *Morris, E. e Liberman, L. (2005)* os vasos linfáticos terminam nuns órgãos pequenos e arredondados, os gânglios linfáticos. Podem encontrar-se grupos de gânglios linfáticos em muitas partes do corpo humano, no que se refere à zona da mama, podem apresentar-se: perto da mama, nas axilas (debaixo do braço), acima da clavícula ou no peito (atrás do esterno), partindo a drenagem linfática para todas as partes da

mama. Quanto à evolução da mama ao longo da vida da mulher, *Kopans, D. (2007)* refere que o tecido mamário sofre uma constante evolução desde o nosso nascimento, contudo há fases em que essas transformações são mais relevantes, tais como:

- Durante a puberdade (devido ao aumento de estrogénios e progesterona);
- Durante a gravidez e após a menopausa (cessação da produção hormonal).

Assim sendo, na mulher jovem, a mama é maioritariamente constituída por tecido conjuntivo fibroso, mais radiodenso, o que acaba por dificultar a visualização de lesões na mamografia. Já durante a gravidez, o número de lóbulos aumenta tal como o seu tamanho e a mama torna-se quase exclusivamente constituída por lóbulos. Por fim, na mulher idosa, os lóbulos desaparecem quase completamente, e o estroma rico em tecido conjuntivo fibroso é substituído por tecido adiposo, mais radiotransparente, o que segundo *Kumar, V. et al (2005)* facilita a detecção de lesões na mamografia. A anatomia da mama pode observar-se na Figura 1.

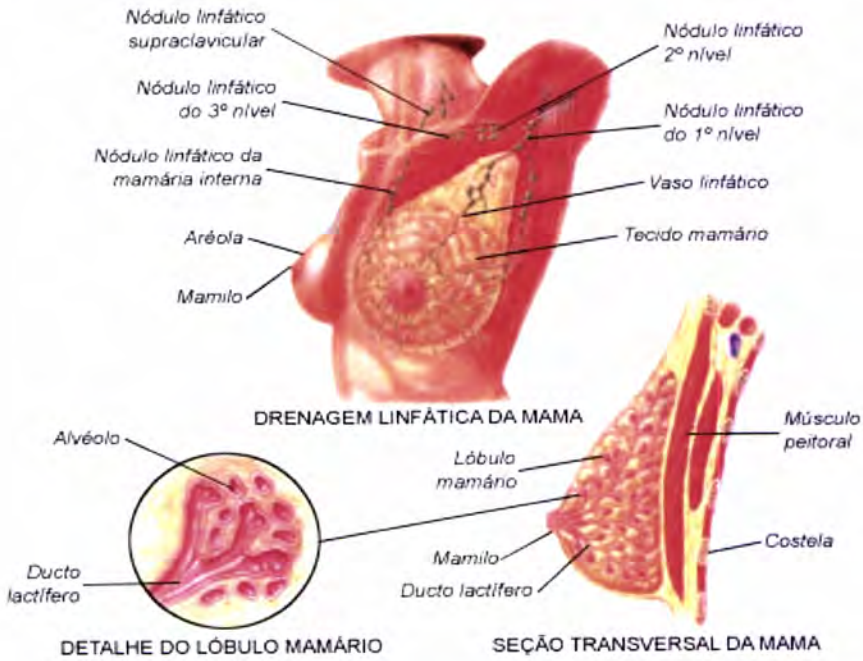


Figura 1: Diagrama esquemático da constituição da mama humana
Fonte: http://www.belezain.com.br/medestetica/mamoplastia.asp

Aparecimento do cancro da mama

O organismo humano é composto por triliões de células que se multiplicam pelo processo de divisão celular. No geral, este processo é ordenado e controlado, responsável pela formação, crescimento e regeneração de tecidos saudáveis do corpo.

Contudo, por vezes, as células perdem a capacidade de limitar e comandar o seu próprio crescimento, passando a dividir-se e a multiplicar-se muito rapidamente e de forma aleatória.

Como consequência dessa disfunção celular, ocorre um desequilíbrio na formação dos tecidos do corpo, no referido local, formando o que se conhece como cancro ou tumor.

Segundo o *Portal da Saúde (2005)* o processo de transformação de uma célula normal numa célula cancerosa designa-se por **carcinogénese**, e este processo apresenta diversas fases. As substâncias responsáveis por esta transformação designam-se agentes carcinogéneos e estes podem ser diversos (radiações ultravioletas, agentes químicos, entre outros).

O termo cancro refere-se a muitas condições malignas – mais de duzentos distúrbios distintos, cada um com o nome derivado do orgão ou tecido no qual é originado.

O Cancro da Mama é, segundo *Ricks, D. (2005)* um processo oncológico que se desenvolve nas células do tecido mamário e que consiste na transformação das células normais em células cancerosas ou malignas, tendo estas últimas a capacidade de se multiplicarem e invadirem os tecidos e outros órgãos.

Como foi referido na secção anterior, a mama é composta por gânglios linfáticos, que funcionam como filtros, cuja principal função é "prender" e "reter" "substâncias estranhas" (bactérias, células cancerígenas, ou outras substâncias malignas) que se podem encontrar no sistema linfático. As células do cancro da mama ao entrarem no sistema linfático podem ser facilmente encontradas nos gânglios linfáticos próximos da mama (os chamados gânglios regionais) - Figura 2.

Na grande maioria das vezes, o cancro da mama, afigura-se como uma massa dura e irregular que, quando palpada, se distingue do resto da mama pela sua consistência

Saliente-se o facto de que, entre as duas mamas existem vias de disseminação tumoral.

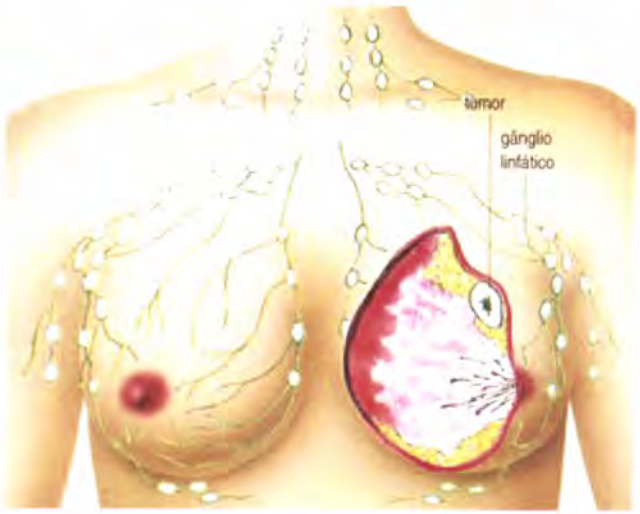


Figura 2: O tumor da mama
Fonte: http://ultimodosprofetas.blogspot.com/2008/05/cancro-da-mama.html

Tipos de Neoplasias da Mama

Neoplasia é também uma designação frequente para tumor, existindo dois tipos de tumores: os ***benignos*** e os ***malignos***.

A neoplasia da mama trata-se do crescimento descontrolado das células mamárias que tem, normalmente, origem em anomalias ou erros genéticos. Primeiramente, a neoplasia da mama pode caracterizar-se segundo *Brown, Z. e Boatman, K. (2008)* em ***invasiva*** ou ***não invasiva***, definindo-se por tumor não invasivo ou "*in situ*" o cancro da mama precoce sem invasão dos tecidos mamários vizinhos ou de outros órgãos. Dentro dos tumores não invasivos os mais frequentes são o ***(CDIS) carcinoma ductal in situ*** (limitado aos ductos) sendo a mamografia a melhor técnica para diagnosticar este tipo de cancro, e o ***(CLIS) carcinoma lobular in situ*** (limitado aos lóbulos), que embora não seja considerado um verdadeira cancro, muitos especialistas pensam que este tipo de carcinoma não se transforme num carcinoma invasivo, ainda que as mulheres que apresentam este tipo de neoplasia tenham maior risco de desenvolver cancro da mama invasivo.

Relativamente aos tumores invasivos, o mais frequente é o ***carcinoma ductal invasivo (CDI),*** seguido do ***lobular (CLI)*** do ***metastásico*** e do ***inflamatório.*** O carcinoma ductal invasivo tem origem nos ductos e invade os tecidos vizinhos, podendo disseminar-se através dos vasos linfáticos ou do sangue, atingindo outros órgãos. Quanto ao carcinoma lobular invasivo, tem origem nos lóbulos, e tal como o anterior, também pode metastizar (disseminar-se) para outras partes do corpo.

O carcinoma metastásico da mama pode reaparecer após o tratamento de um tumor primitivo, ou, pode também, invadir outros órgãos quando os gânglios linfáticos são atingidos; podendo vir a surgir na própria mama, na outra mama, na grelha costal ou num outro orgão (mais frequentemente no pulmão, fígado, osso e cérebro). Este tipo de carcinoma tanto pode surgir um ano após a descoberta do tumor primitivo, como 10 a 15 anos depois, não se conhecendo o período específico da sua manifestação. Por fim, o carcinoma inflamatório da mama é um tipo de tumor muito agressivo mas pouco frequente (corresponde a cerca de 1 a 3% de todos os cancros da mama), que se caracteriza pelo surgimento de sinais inflamatórios, que se podem confundir com mastite aguda da mama.

De entre os carcinomas da mama menos comuns destacam-se: o medular, o mucinoso, a doença de *Paget* da mama, o tubular, o filóide, o metaplásico, o sarcoma e o carcinoma micropapilar. Pode ainda falar-se dos fibroadenomas que por muitos não chega a definir-se como um tipo de neoplasia tratando-se apenas de lesões benignas da mama, resultado de uma alteração do processo de desenvolvimento do lóbulo mamário. São bastante comuns e usualmente presentes numa única mama em pacientes jovens, sendo que o fibroadenoma incide em 9 a 10% das mulheres durante toda a vida e cerca de 7 a 13% das mulheres submetidas a exames periódicos apresentam este tipo de lesão.

A Figura 3 mostra o conjunto de tipologias mais frequentes de neoplasias mamárias. Para uma descrição mais pormenorizada dos vários tipos de neoplasias mamárias ver *Ricks, D. (2005)* e *Brown, Z. e Boatman, K. (2008).*

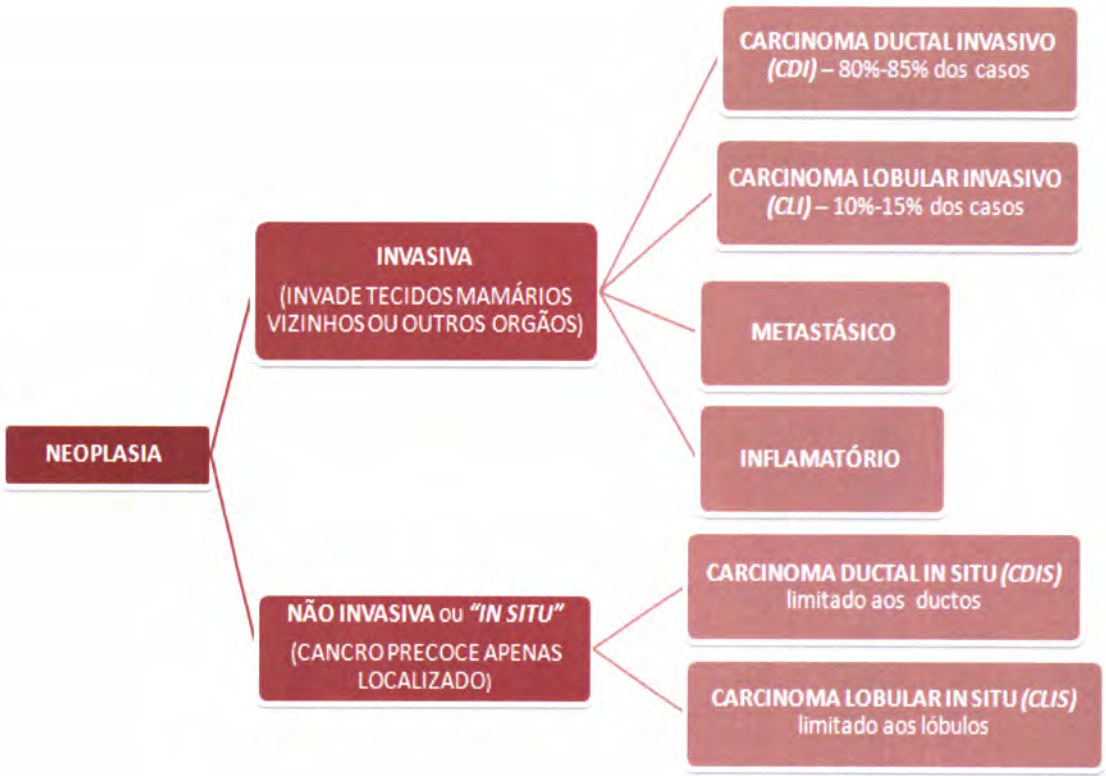


Figura 3: Tipologias mais frequentes de neoplasias mamárias

Factores que contribuem para o cancro da mama e Anamnese

Não é conhecida uma causa específica para o cancro da mama, mas como ponto de partida para o diagnóstico de uma doença, é necessário que um profissional de saúde realize uma espécie de entrevista ao paciente em causa com o objectivo de reunir todos os factos que se relacionem com a doença e com a pessoa em questão. A essa espécie de entrevista é costume chamar-se anamnese, sendo esta fulcral para orientar o médico para o melhor diagnóstico e averiguar quais os factores de risco que apontam para a doença. Segundo *Kopans, D. (2007)* os dois principais factores que determinam o aumento do risco de cancro da mama são: o sexo (ser mulher) e o envelhecimento. Por meio da investigação tem-se demonstrado que há mulheres que apresentam um risco acrescentado para cancro da mama, associado a outros factores.

Factores de risco para o cancro da mama já identificados:

- **Sexo (ser mulher)**
- **Idade**
- **Etnia**
- **História pessoal do cancro da mama**
- **História familiar**

- **Algumas alterações da mama**
- **Alterações genéticas**
- **Primeira gravidez depois dos 30 anos**
- **Não amamentar**
- **História menstrual longa**
- **Terapêutica hormonal de substituição**
- **Radioterapia no peito**
- **Densidade da mama**
- **Obesidade**
- **Inactividade física**
- **Bebidas alcoólicas**

Tal como mencionado acima, *Kopans, D. (2007)* refere que a incidência de cancro da mama aumenta exponencialmente com a ***idade***, sendo cerca de 70% dos casos de cancro da mama descobertos em mulheres com idades superiores aos 50 talvez por se tratar de uma doença mais comum após a menopausa, sendo estas mulheres mais seguidas pelos seus médicos. Segundo este mesmo autor, ***alterações da mama*** bem como ***alterações genéticas*** podem aumentar o risco de cancro da mama, já o facto de a ***mulher não amamentar*** tal como a ***obesidade*** são considerados pelo mesmo como factores de fraca associação com o risco em questão, apesar deste referir que uma ***dieta*** rica em fruta e vegetais e pobre em gordura, está associada a um menor risco. Este refere ainda que, o facto de uma mulher ter feito ***radioterapia*** no peito (incluindo as mamas) antes dos 30 anos aumenta o risco de cancro da mama.

Em contrapartida, o *National Cancer Institute (2004)* refere que alguns estudos demonstram que o ***aumento de peso***, após a menopausa, aumenta o risco de cancro da mama.

Brown, Z. e Boatman, K. (2008) apontam como factores de risco para o cancro da mama: a ***etnia*** (apesar da etnia branca ser mais afectada é na etnia negra que o tumor tem pior prognóstico, talvez devido à escassez dos meios económicos e culturais inerentes a esta população), a ***história pessoal de cancro da mama*** (uma mulher que tenha tido cancro tem maior risco de voltar a ter a doença), a ***primeira gravidez depois dos 30***, a ***história menstrual longa*** (1ªmenstruação em idade precoce e/ou menopausa tardia) e a ***nuliparidade***.

Infocancro aponta ainda como factor de risco para o cancro da mama a ***história familiar*** (cancro da mama na família mais próxima (mãe, tia ou irmã) ou outros familiares do lado paterno ou materno) e caso haja mais de uma parente de 1ºgrau afectada o risco torna-se cinco vezes superior.

Ricks, D. (2005) refere que a ***actividade física*** pode ajudar a diminuir o risco de cancro da mama.

Segundo alguns estudos presentes em *J Natl Cancer Inst (2009)* parece haver relação entre a maior ingestão de ***bebidas alcoólicas*** e o risco de cancro da mama.

Outros estudos referem que mulheres que tomam terapêutica hormonal para a menopausa (apenas com estrogénios ou estrogénios e progesterona) durante 5 ou mais anos após a menopausa parecem, também, apresentar maior possibilidade de desenvolver cancro da mama, nestas mulheres o risco também aumenta caso apresentem na mama, tecido denso (não gordo).
Numa edição recente da *American Cancer Society (2009)*, foi referido ainda que o raloxifeno (droga usada na prevenção e tratamento da osteoporose pós menopausa) reduz o cancro da mama. Para além disso, este revelou-se tão eficaz como o tamoxifeno na prevenção do cancro da mama neste grupo de mulheres em risco, tendo este último, menos efeitos adversos. Para saber mais sobre o tamoxifeno ver *Ricks, D. (2005).*

O presente trabalho procura contribuir para identificar outros possíveis factores de risco para este tipo de cancro bem como reconhecer de que forma factores já identificados podem ser controlados.

2.3 Caso de Estudo: Cancro da Mama na Região do Alentejo

Com o objectivo de indicar os mais adequados instrumentos de prognóstico de cancro da mama realizou-se um estudo estatístico com dados que foram, gentilmente, cedidos pela *Unidade de Anatomia Patológica do Hospital Espírito Santo de Évora (HESE)*, e a partir dos quais se construíram duas bases de dados. Uma primeira base, referente a pacientes residentes na região do Alentejo com diagnóstico de um tipo de neoplasia mamária (*os casos*) e uma segunda base referente a pacientes que não padeciam de qualquer tipo de neoplasia mamária e que foram observados no respectivo Hospital. (*os controlos*).
De referir que, o prognóstico do carcinoma da mama é realizado com base não só, nos exames complementares de diagnóstico (*tamanho tumoral, tipo e grau histológico, margens, invasão vascular ou linfática, estadiamento ganglionar e metástases à distância*), mas também com base em biomarcadores de prognóstico (*receptores hormonais, índice proliferativo – gene ki67, expressão do gene oncossupressor p53, expressão do oncogene Her2 ou da proteína que codifica, c-erB-2*), entre outros.

Inicialmente, o estudo em causa pretendia abarcar vários anos (de Agosto de 2003 a Agosto de 2007), contudo, tal não foi possível, uma vez que:

- Existiam dados em falta para a maior parte das variáveis;
- Os dados encontravam-se registados num programa informático específico do *HESE* e nem sempre se procedeu ao registo do paciente e das características importantes para o estudo, da mesma forma;
- Não existia uniformidade nas designações e nomenclaturas consideradas, de ano para ano;

- A maior parte dos registos referidos, encontrava-se ainda, em fichas médicas (em papel) e às quais não foi possível ter acesso.

Não tendo sido possível a consulta das fichas clínicas pessoais dos pacientes, o que com certeza enriqueceria o estudo já que mais informação poderia ser considerada, a análise estatística, incidiu apenas no período de Agosto de 2003 a Agosto de 2004 e considerou as variáveis que a seguir se descrevem.

2.3.1 Descrição dos Dados

Com o objectivo de determinar os factores de risco em doentes residentes na região do Alentejo, com cancro da mama, procedeu-se à realização de um estudo do tipo descritivo, a partir do levantamento de dados relativos a pacientes (mulheres) em que foram diagnosticadas neoplasias mamárias num dos hospitais da região (Beja, Portalegre, Elvas, Évora e ainda Hospitais do Litoral do Alentejo) e, posteriormente, referenciadas para o *HESE*, mais especificamente, para a *Unidade de Anatomia Patológica* do mesmo, no período de Agosto de 2003 a Agosto de 2004, com o objectivo de se estudar os tecidos, observar-se as peças cirúrgicas, as biópsias efectuadas aos doentes ou as células para se proceder a um diagnóstico. No caso de estarem perante tumores são procurados critérios que indicam qual é o prognóstico da doença, qual o seu estadiamento e fazem o diagnóstico que é depois enviado ao clínico que pediu o exame. Este último, baseado no diagnóstico, aplica ou não o tratamento, consoante as circunstâncias. Os clínicos associados à *Anatomia Patológica* tentam sempre, não dar apenas o diagnóstico, mas também, apresentar factores de prognóstico e de tratamento.

De seguida, os resultados obtidos no diagnóstico são cuidadosamente registados num *software* adequado para esse efeito.

O método de recolha dos dados foi moroso e exigiu várias reuniões no *HESE*, com o Dr. Luís Gonçalves (Médico Patologista) e com outros elementos deste departamento.
Registados os dados, procedeu-se ao seu tratamento estatístico, construindo as bases de dados no ***programa estatístico R***. Conforme referido acima, seleccionou-se apenas um ano de estudo, ficando a nossa amostra com 212 indivíduos.

2.3.2 Descrição das Variáveis

Na Tabela 1 apresentam-se as variáveis que foram consideradas no estudo. Para cada variável é indicado se se trata de uma variável numérica ou categórica e neste último caso, as classes consideradas.

NOME DA VARIÁVEL	DEFINIÇÃO	OPERACIONALIZAÇÃO
Idade	Idade de cada indivíduo em estudo	Variável numérica em anos completos de vida
Idadecat	Idade categorizada tendo em conta as seguintes classes: 1-[15,25[2-[25,35[3-[35,45[4-[45,55[5-[55,65[6-[65,75[7-[75,85[8-[85,100[	Variável categórica
Sexo	Sexo de cada indivíduo em estudo	Variável Categórica
Localidade	Localidade em que reside cada indivíduo em estudo (região do Alentejo)	Variável Categórica
Concelho	Concelho a que pertence cada indivíduo em estudo (região do Alentejo)	Variável Categórica
Distrito	Distrito a que pertence cada indivíduo em estudo (região do Alentejo)	Variável Categórica
Ruralidade	Ruralidade ou não do meio em que reside cada indivíduo em estudo, tomando os valores: 1-Urbano 2-Rural	Variável categórica
Tipo de Amostra	Forma como foi obtida a amostra de tecido: 0-Sem informação 1-Biópsia Histológica 2-Peça Cirúrgica	Variável Categórica
Tipo Histológico	Tipo histológico da neoplasia diagnosticada: 1-Lesões proliferativas intraductais 3-Neoplasias papilares intraductais 4-Proliferações epiteliais benignas 6-Tumores mesenquimatosos benignos 7-Tumores fibroepiteliais (fibroadenoma) 9-Carcinoma ductal invasivo 10-Carcinoma intraductal ou intralobular (Carcinoma in situ) 11-Carcinoma lobular invasivo 17-Carcinoma Papilar	Variável Categórica
Tipo Histológico Binário	Transformação da variável Tipo Histológico numa variável binária: Tipo Histológico entre 1 e 8: Benigno – 0 Tipo Histológico entre 9 e 17: Maligno – 1	Variável Categórica
LateralidadeED	Localização da neoplasia diagnosticada, tendo em conta: 0- Sem Informação 1- Mama esquerda 2- Mama direita	Variável Categórica
Tamanho Tumoral	Medido em centímetros de forma contínua, para fins de análise, agrupada em: 0-Sem Informação 1-até 2cm 2-de 2,1 cm a 5cm	Variável Categórica
Margens Cirurgicas	Referente à invasão ou não do tumor.	Variável Categórica
Grau Histológico	Referente ao grau de diferenciação do tumor: 0-Sem Informação 1-Bem diferenciado/Baixo Grau 2-Moderadamente diferenciado 3-Pouco diferenciado/Alto grau	Variável Categórica

pT	Avaliação da extensão do tumor primário: 0-Não se aplica 1-Sem Informação 2-<2cm 3-Entre 2 e 5 cm	Variável Categórica
pN	Presença ou ausência de metástases regionais: 0-Não se aplica 1-Sem Informação 2-Ausência de metástases 3-Metástases em 1 ou mais gânglios axilares homolaterais móveis 4-Metástases em 1 ou mais gânglios axilares homolaterais fixos uns aos outros ou a outras estruturas	Variável Categórica
pM	Presença ou ausência de metástases à distância: 0-Não se aplica 1-Ausência 2-Presença de metástases à distância	Variável Categórica
IVN	Índice Prognóstico de Van Nuys –referente à avaliação do risco de recidiva apenas para o carcinoma ductal in situ 0-Não se aplica 1-Sem Informação 2-Sem Risco 3-Risco Baixo 4-Risco Médio 5-Risco Alto	Variável Categórica
IPN	Indice Prognóstico de Nottingham – referente à avaliação do risco de recidiva para carcinomas invasivos 0-Não se aplica 1-Sem Informação 2-Inferior a 3,4 =bom prognóstico 3-Entre 3,4 e 5,4 =Prognóstico Intermediário 4-Superior a 5,4 = Mau Prognóstico	Variável Categórica
RE e RP	Receptores hormonais de estrogénio e de progesterona, classificados da seguinte forma: 0-Sem Informação 1-<1% - não favorável ao cancro da mama 2->1% - favorável ao cancro da mama	Variável Categórica
c-Erb-2	Referente à presença e amplificação ou não do gene c-erb-2 0-Sem Informação 1- Não se observou coloração na membrana 2-Coloração da membrana em mais de 10% das células cancerosas com raras ou ausentes manchas 3-Fraca coloração da membrana em mais de 10% das células cancerosas, anel fino de manchas 4-intensa coloração circunferencial da membrana em mais de 10% das células cancerosas e anel de coloração da membrana espesso	Variável Categórica
Ki67	Avaliação da relação entre o gene ki67 e o comportamento tumoral Variável em % e categorizada: 0-Sem Informação 1->10% - não favorável 2-<10% - favorável	Variável Categórica
p53	Avaliação da relação entre o gene p53 e o comportamento tumoral, categorizada em: 0-Sem Informação 1 - >5% - não favorável 2 - <5% - favorável	Variável Categórica
Fase S	Fase de síntese de DNA que permite avaliar o comportamento tumoral em % categorizada em:	Variável Categórica

	0-Sem Informação 1 - >5% - não favorável 2 - <5% - favorável	
IDNA	Índice de DNA/DNA Ploidia Classifica-se da seguinte forma: 0-Sem Informação 1-Diplóide 2-Aneuplóide 3-Tetraplóide 4-Multiplóide e Aneuplóide	Variável Categórica

Tabela 1: Quadro Resumo das variáveis em estudo
Fonte: Unidade de Anatomia Patológica, Hospital Espírito Santo de Évora (HESE)

Descrevem-se, seguidamente, as variáveis da Tabela 1.

Idade e Idadecat

Indica a *idade* de cada indivíduo em estudo. Tendo em conta os dados da *idade*, para sintetizar esta informação tornou-se indispensável a criação da variável *idadecat*, em que se categorizou a idade segundo as classes etárias presentes na Tabela 1.
A *idade* dos indivíduos em estudo varia entre os 18 e os 88 anos.

Sexo

Indica o *sexo* de cada indivíduo. Neste estudo apenas existem dados para o *sexo* feminino, uma vez que restringindo um estudo deste tipo a uma região, como é o caso, e tendo em conta que a neoplasia da mama no *sexo* masculino é bastante rara, o número de homens deste distrito diagnosticados na região do Alentejo, no período em estudo, é bastante reduzido, não sendo de todo relevante para o estudo em causa.

Localidade

Indica a *localidade* em que reside cada um dos indivíduos em estudo. Estão presentes neste estudo 76 localidades da região do Alentejo.

Concelho e Distrito

Indica o *concelho* e o *distrito* a que pertence o indivíduo em estudo, não esquecendo que a região em estudo é apenas a região Alentejana (o que contabiliza 4 distritos e 46 concelhos, como mostra a Figura 4).

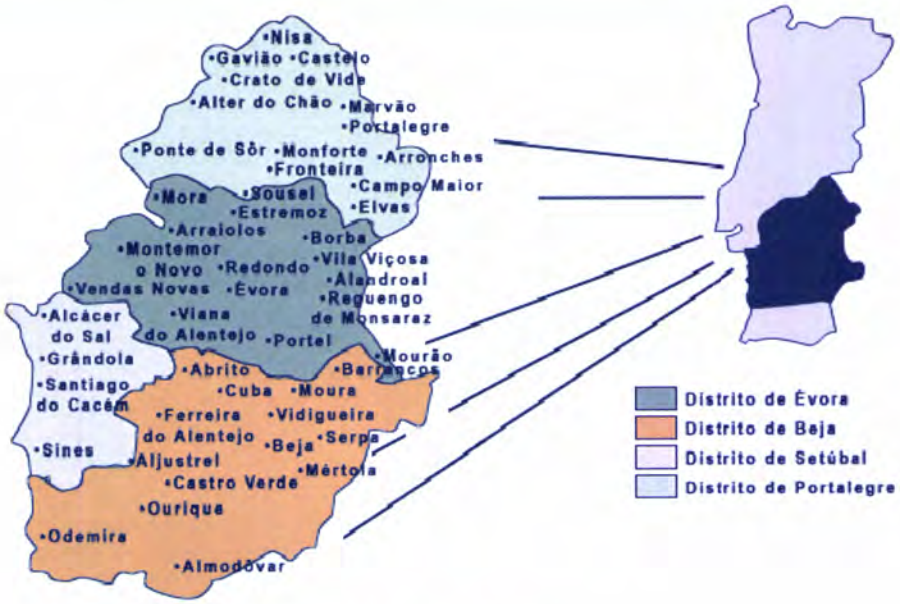


Figura 4: Região do Alentejo

Ruralidade

Esta variável refere-se ao tipo de meio ambiente em que cada indivíduo em estudo reside, tendo em conta o facto de residir numa zona rural (aldeias e lugares) ou numa zona urbana (vilas ou cidades). Para a criação desta variável, considerando que a região em estudo é a região do Alentejo, e que uma das variáveis em estudo é a *localidade* de residência do indivíduo, procedeu-se à classificação dessas localidades como *rurais/urbanas* tendo em conta a *NUTS II* e várias pesquisas realizadas acerca das *localidades* a que pertencem os indivíduos em estudo.

Tipo de Amostra (Biópsia Histológica ou Peça Cirúrgica)

Para os indivíduos em estudo foi necessária a amostra de tecido para posteriormente se realizarem exames, no sentido de detectar o tipo de neoplasia presente bem com as características da mesma. Para tal, pode proceder-se à extracção dessa amostra por um dos dois métodos:

Biópsia Histológica – Procedimento cirúrgico, no qual se colhe uma amostra de tecido ou de um órgão num individuo vivo, destinado a exame microscópico ou análise bioquímica.

Peça Cirúrgica – Procedimento cirúrgico, no qual são retirados órgãos, conjuntos de orgãos e tecidos, na grande maioria das vezes precedidos pelo estudo microscópico através de biópsias cirúrgicas ou punção – biópsia aspirativa por agulha fina e análise citológica de esfregaços de líquidos e secreções humanas (citopatologia), para a cura de uma enfermidade.

Tipo Histológico

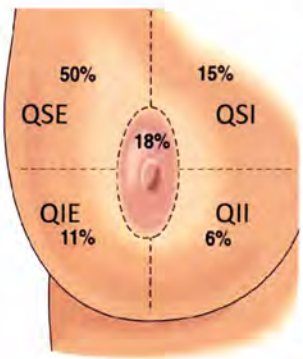
Nome dado ao tipo de neoplasia tendo em conta a sua localização, consistência, mobilidade em relação aos tecidos adjacentes, contornos, dimensões e características. Segundo a *União*

Internacional Contra o Cancro (UICC), o cancro da mama pode ser classificado tendo em conta o tipo histológico, segundo a classificação presente na Tabela 1.

Tipo Histológico Binário

Indica se o tipo de tumor diagnosticado ao indivíduo é *benigno ou maligno*, tendo em conta a classificação para a variável *Tipo Histológico* (Tabela 1) e sabendo que tipos histológicos compreendidos *entre 1 e 8 inclusive* são considerados *benignos*, todos *os restantes* são considerados *malignos*.

Lateralidade



Indica a localização da neoplasia, tendo em conta a divisão da mama em quatro quadrantes, superior externo e interno e inferior interno e externo, tal como se pode observar na Figura 5 e na Tabela 1.

A Figura 5 mostra ainda a distribuição, em percentagem, do cancro da mama pelas diferentes localizações da mesma.

No início do estudo consideraram-se todos os quadrantes, no entanto, devido ao elevado número de valores omissos, considerou-se apenas *mama esquerda* e *mama direita*.

Figura 5: Divisão da mama em quadrantes
Fonte: Portal de Oncologia Português

Tamanho Tumoral

Indica o *tamanho do tumor* em centímetros e as classes em estudo encontram-se na Tabela 1.

Margens Cirúrgicas (invasão do tumor ou não)

Indica a quantidade variável de tecido, supostamente normal, que envolve os tumores, tanto em lateralidade como em profundidade. A margem cirúrgica é verificada em laboratório durante ou após a cirurgia e no presente estudo classificou-se da seguinte forma:

0-Sem Informação

1-Positivo (significa que há invasão do tumor nas margens)

2-Negativo (significa que não há invasão do tumor nas margens)

3-<1mm

Grau Histológico

Refere-se ao grau de malignidade do tumor indicando a sua maior ou menor capacidade de metastização, trata-se de uma avaliação morfológica no que respeita à diferenciação celular de cada tumor. Para definir esta variável é usada a seguinte escala:

- GX - o grau de diferenciação não pode ser avaliado
- G1 - bem diferenciado
- G2 - moderadamente diferenciado
- G3 - pouco diferenciado
- G4 – indiferenciado

No presente estudo alterou-se esta nomenclatura para a que é apresentada na Tabela 1, tendo em conta as categorias existentes para esta variável e a simplificação no tratamento da informação.

pT, pN e pM (Estadiamento TNM)[1]

O cancro da mama é classificado em estádios. O sistema de estadiamento do cancro da mama tem em conta a extensão do tumor (T), o envolvimento de gânglios linfáticos da axila próxima à mama (N) e a presença ou não de metástases à distância (M).

Sendo assim, cada uma das variáveis em causa, pT, pN e pM têm em conta a seguinte classificação, com base em *Atlas TNM Classification of Malignant Tumours (2009)*:

T – Extensão do tumor primário

TX - O tumor primário não pode ser avaliado

T0 – Sem evidência de tumor primário

Tis - Carcinoma in situ: carcinoma intraductal ou carcinoma lobular in situ ou doença de Paget da papila sem tumor

T1 - Tumor ≤ 2 cm

T1a – T ≤ 0,5 cm

T1b - 0,5 cm ≤ T ≤ 1 cm

T1c - 1 cm ≤ T ≤ 2 cm

T2 - 2 cm ≤ T ≤ 5 cm

T3 – T > 5 cm

T4 Tumor de qualquer tamanho, com extensão directa à parede torácica ou à pele.

T4a - extensão para parede torácica

T4b - edema (incluindo peau d'orange), ulceração da pele da mama ou nódulos cutâneos satélites, confinados à mesma mama

T4c - T4a e T4b associado

T4d - carcinoma inflamatório

N – Presença ou ausência de Metástases regionais

[1]**NOTAS:**

1 - O estadiamento TNM apenas faz sentido para carcinomas invasivos (Tipos histológicos 9 e 11, segundo a nomenclatura do tipo histológico).

2 - As variáveis pT, pN e pM, no presente estudo foram categorizadas conforme mostra a Tabela 1, no sentido de facilitar o tratamento da informação.

3 - A variável pM foi eliminada do estudo por se encontrar com um elevado número de valores omissos.

NX - Os gânglios regionais não podem ser avaliados (por ex. foram removidos previamente)

N0 - Ausência de metástases

N1 - Metástase num gânglio(s) auxiliar(es) homolateral (is) móvel (is)

N2 - Metástase nos gânglios axilares homolaterais fixos uns aos outros ou a outras estruturas

N3 - Metástase nos gânglios da cadeia mamária interna homolateral

M – Presença ou ausência de Metástases à distância

MX - A presença de metástases a distância não pode ser avaliada

M0 - Ausência de metástases a distância

M1 - Metástases a distância (incluindo as metástases nos gânglios supraclaviculares)

Estadiamento do carcinoma da mama a partir do sistema TNM:

Estádio 0

Tis + N0 + M0

Estádio I

T1+ N0 + M0

Estádio IIa

T0 + N1+ M0 ou T1 + N1+ M0 ou T2 + N0+ M0

Estádio IIb

T2 + N1 + M0 ou T3 + N0 + M0

Estádio IIIa

T0 + N2 + M0 ou T1 + N2 + M0 ou T2 + N2 + M0 ou T3 + N1, N2 + M0

Estádio IIIb

T4 + qualquer N + M0 ou qualquer T + N3 + M0

Estádio IV

qualquer T + qualquer N + M1

Uma vez identificado o estádio do tumor, é possível ao médico planear o tratamento mais adequado.

Índice de Van Nuys (IVN)

O Índice de Prognóstico de *Van Nuys* (***IVN***) tenta estratificar os doentes portadores de carcinoma mamário em grupos de diferentes riscos de recidiva, com base nas características do tumor e nas margens pós-operatórias.

Este índice foi desenvolvido por *Silverstein, M.J. et al. (1996)* com o objectivo de auxiliar na escolha do tratamento indicado para as doentes com carcinoma **ductal *in situ*** da mama, **fazendo sentido apenas se estivermos perante este tipo de carcinoma. (tipo histológico 10** segundo a Tabela 1). Este índice quantifica quatro factores de prognóstico (*tamanho tumoral, margens, grau*

nuclear e presença ou ausência de necroses), como indicado na Tabela 2. Recentemente, passou a considerar-se também a *idade*.

Pontuação	**1**	**2**	**3**
Tamanho (mm)	≤15	16-40	>40
Margens (mm)	≥ 10	1-9	<1
Histologia	Sem necroses Grau nuclear 1 e 2	Com necroses Grau nuclear 1 e 2	Com necroses Grau nuclear 3
Idade	>60	40-60	<40
Grupos Prognósticos	**Actuação Recomendada**		
Grupo 1: Pontuação de 4 a 6	Tratamento apenas com excisão		
Grupo 2: Pontuação de 7 a 9	Cirurgia conservadora e radioterapia		
Grupo 3:Pontuação de 10 a 12	Mastectomia (com ou sem reconstrução imediata)		

Tabela 2: Índice Prognóstico de Van Nuys
Fonte: Cáncer de mama[2]

No caso em estudo utilizou-se, para a variável *IVN*, a nomenclatura indicada na Tabela 1.

Índice Prognóstico de Nottingham (IPN)

O estudo deste índice, tal como o estadiamento TNM apenas, é realizado no caso de estarmos perante carcinomas invasivos (tipos histológicos 9 e 11)

A fórmula para o cálculo do *IPN* apresentada por *Todd JH et al., (1987)* foi:

$$IPN = (0{,}2 \times tamanho\ do\ tumor\ em\ cm) + (grau\ histológico\ de\ 1\ a\ 3) + (afectação\ ganglionar\ de\ 1\ a\ 3\ (N))$$

Tendo-se classificado esta variável tendo em conta a nomenclatura presente na Tabela 1.

O prognóstico do carcinoma da mama vai depender não só de características ditas clássicas, mas também de **biomarcadores** de prognóstico. São eles:

- Receptores Hormonais (de *estrogénios e progesterona*);
- Expressão do oncogene Her2 ou da proteína que codifica, *c-erB-2*;
- Índice proliferativo (avaliação da expressão do *gene Ki-67*);
- Expressão do gene oncossupressor *p53*.

[2] Cáncer de mama – Atlas de Oncologia clinica, David J. Winchester, 2001

Segundo *Dixon, J.M. (2000)* os chamados biomarcadores ou marcadores biológicos são moléculas que podem ser medidas experimentalmente e indicam a ocorrência de um determinado processo num organismo, pois são substâncias que podem encontrar-se no sangue, na urina ou em tecidos do corpo de alguns pacientes com certos tipos de cancro, em quantidades acima do normal. Estes tipos de marcadores podem ser produzidos pelo próprio tumor ou pelo corpo em resposta à presença do cancro. Os testes para marcadores tumorais não são usados sozinhos para o diagnóstico de um determinado tipo de cancro, uma vez que, a grande maioria dos marcadores pode ser encontrada em níveis elevados em pacientes que não têm qualquer tipo de condições para originar uma doença cancerosa e também, porque nenhum marcador tumoral é específico de um tipo particular de cancro. Para além disso, também é verdade que nem todos os pacientes doentes de cancro apresentam os marcadores tumorais elevados, uma vez que, nas primeiras fases do cancro, os níveis dos marcadores tumorais se encontram numa faixa dita normal.

Embora o uso de marcadores tumorais para diagnosticar o cancro esteja limitado, os investigadores estão a realizar estudos com o objectivo de procurar marcadores que sejam específicos para um determinado tipo de cancro, de forma a puderem utilizá-los para detectar a presença do cancro antes que os sintomas surjam. A mudança dos níveis dos marcadores tumorais pode ser útil para seguir o curso da doença, para medir o efeito do tratamento ou para verificar a reincidência.

Nalguns casos, o nível do marcador tumoral reflecte a extensão da doença ou indica o quão rápido a doença parece estar a progredir.

Receptores Hormonais (RE e RP)

No início da década de 90, a determinação dos níveis do *receptor estrogénio (RE)* e do *receptor progesterona (RP)* começou a ser realizada através da técnica de imuno-histoquímica e é uma das características patológicas que tem valor prognóstico no cancro da mama, sendo cruciais no tratamento hormonal, tal como referem *Hunt K. et al (2001)*.

Um dos efeitos do estrogénio é induzir a expressão da progesterona. A maioria dos carcinomas da mama são *RE* e *RP* positivos, apenas 5% dos tumores são *RP* positivos e *RE* negativos e tumores *RE* e *RP* negativos indicam pior prognóstico. Segundo *Hunt K. et al (2001)*, o *RE*, quando negativo, encontra-se correlacionado com baixa diferenciação tumoral, alta taxa de proliferação celular e outras características desfavoráveis ao prognóstico das pacientes com cancro da mama, sendo a idade também um factor de correlação. Em relação à sobrevivência, os pacientes com tumores *RE* positivos tendem a ter uma sobrevivência maior que os pacientes com tumores *RE* negativos. O *RP* tem-se apresentado, na grande maioria dos estudos como portador de um papel secundário no prognóstico do cancro da mama.

A proteína c-erB-2

Segundo a última classificação da *OMS* existem pelo menos 17 tipos de cancro da mama, havendo dois genes potencialmente relacionados com os diversos tipos tumorais: o gene *HER2* e o gene *FGFR1*.

Antunes A. et al (2004) definem o gene *HER2* (também conhecido como *c-erB-2*) como um oncogene humano que codifica um receptor proteíco que desempenha um importante papel na divisão e crescimento da célula normal. Quando o gene *c-erB-2* é amplificado por mecanismos ainda não conhecidos, ocorre uma produção excessiva de receptores na superfície celular, a divisão celular é estimulada, resultando um crescimento celular acelerado, o qual contribui para o desenvolvimento e progressão do cancro da mama. O *c-erB-2* tem sido extensamente estudado em carcinomas da mama, desde que *Slamon, D. et al. (1987)* demonstraram uma associação entre a sua amplificação e um mau prognóstico.

Como o *c-erB-2* não se trata de uma proteína que se expresse na maioria dos tecidos humanos normais, a sua amplificação pode ser uma estratégia na terapia do cancro da mama no sentido de inibir o crescimento aberrante do tumor, tal como referem *Walther, W. e Stein, U. (2000)*. Um grande número de estudos, de referir *Gasparini, G. et al (1992), Gullick, W. et al (1991), Ioachim, E. et al (1996) e Keshgegian, A. (1995)* demonstram que um valor aumentado deste oncogene leva a um alto risco de recidiva precoce, demonstrando ainda a existência de uma correlação positiva entre esta variável e outras, tais como: *linfonodos axilares* e *grau histológico*. No que diz respeito às metástases, sabe-se que os indivíduos que apresentam expressão aumentada de *c-erB-2* estão mais propensos a desenvolvê-las, observando-se após a sua detecção, um curto período de sobrevivência.

No presente estudo, esta variável foi categorizada conforme descrito na Tabela 1, tendo em conta a nomenclatura[3].

Gene Ki67

Hunt, K. et al (2001) definem o *Ki67* como um antígeno nuclear associado com a proliferação celular, encontrado em todo o ciclo celular (formado pelas fases G1,S,G2 e M, em que G significa intervalo, S vem de síntese e M de mitose) e ausente em G0 (fase em que as células não se dividem mais). Trata-se de uma proteína do ciclo celular correlacionada com maiores taxas de recidiva, menor sobrevivência e período livre de doença. Através da análise imuno-histoquímica tem-se tentado relacionar esta proteína com o comportamento tumoral, pelo que foi descoberta uma

[3] Scoring Guide for the Interpretation of Ventana PATHWAY® HER-2/*neu* (4B5) Rabbit Monoclonal Primary Antibody Staining of Breast Carcinoma, Innovations in Science and Medicine, Ventana

correlação positiva desta variável com o tumor da mama. Pode dizer-se também que existe uma relação directa entre as proteínas *p53* (posteriormente estudada) e *ki67*. Têm-se realizado estudos, no sentido de procurar uma resposta completa para o tratamento do cancro da mama, utilizando os valores desta proteína, a sua importância na progressão tumoral bem como a sua relação com os factores de prognóstico mais clássicos

Gene p53

Varela, S. (2002) define *p53* como uma proteína codificada por um gene (gene *p53*), cuja principal função está relacionada com a preservação do código genético em cada célula. Durante o ciclo da divisão celular, a proteína *p53* faz uma verificação quanto à eventual ocorrência de uma mutação na sequência do código genético em consequência de uma duplicação defeituosa do DNA. Caso seja verificada a existência de uma mutação, é função da proteína *p53*, impedir que esta célula entre em processo de mitose e complete a divisão celular. Para isto podem ser tomados dois caminhos: a correcção da mutação através da activação de proteínas de reparo ou a indução da morte celular através da apoteose (processo de morte celular sem reacção inflamatória). Por exercer esta função de detecção de alterações no DNA e consequente correcção ou morte celular, a proteína *p53* é considerada uma guardiã do genoma, e é um impotante elemento na prevenção do desenvolvimento de tumores, sendo o seu gene codificador classificado como gene supressor de tumor. Se houver uma mutação da proteína *p53* esta não pode cumprir com a sua função e torna-se maior a probabilidade de aparecimento de um tumor maligno, tal como cita *Varela, S. (2002).*

Fenollera, A. (2000) refere que a frequência de mutações no gene *p53* nos cancros da mama oscila entre os 25% e os 50% e aumenta à medida que o estado do tumor progride. A presença de mutações *p53* em tumores, particularmente os da mama e do cólon, indica um cancro mais agressivo com menores perspectivas de sobrevivência.

As experiências laboratoriais mostram que a inserção de um gene *p53* normal em células tumorais resulta numa significativa diminuição na tumorigénese.

Principal conclusão da relação existente entre os biomarcadores e o cancro da mama segundo *Travassoli, F. (1999)*:

Tumores *Ki-67, p53* ou *c.erB-2* positivos são geralmente tumores com mau prognóstico.

Fase S e Índice de DNA (IDNA)

Nos seres vivos pluricelulares, a divisão celular permite a regeneração das células ou parte de órgãos que foram danificados ou, ainda, a renovação das que envelheceram ou morreram. A divisão pode ocorrer a diferentes velocidades e por consequência, em tempos diferentes, conforme, claro, os tecidos em questão.

O ciclo celular é constituído por duas fases principais:

- ✓ Interfase (constituída pelas fases G1, S, G2) – consiste no fim de uma divisão celular e no início da seguinte;
- ✓ Mitose ou período de divisão celular – consiste no período em que ocorre a divisão celular.

De onde se salienta a *Fase S* (**fase de síntese de DNA**), a qual requer um sinal citoplasmático para que se inicie. Nesta fase cada cromossoma é duplicado longitudinalmente, passando a ser formado por dois cromatídeos (replicação semi-conservativa), bem como, numerosas proteínas são também sintetizadas.

Muitos dos defeitos moleculares responsáveis pela transformação das células normais em malignas consistem em mutações em genes codificantes para proteínas que regulam o ciclo celular. Para mais detalhe sobre a *Fase S* ver *Alberts, B.et al (1998) e Jameson, J. (1998).*

Quanto à *ploidia do DNA*, pode dizer-se que cada espécie possui um número característico de cromossomas, que nos seres humanos são dois (daí a designação diplóide), no entanto devido às diversas mutações que o *DNA* pode sofrer, o conjunto básico de cromossomas pode estar em múltiplas cópias (*aneuplóide, tetraplóide e multiplóide aneuplóide*).

Diferentes factores, tais como a *ploidia do DNA* e a fracção da *fase S* do ciclo celular têm sido alvo de estudos no sentido de detectar alguma relação entre estas e o cancro da mama. Aparentemente, *Mareel, M. et al (1991)* dizem haver uma relação entre os *índices de DNA* e as actividades proliferativas, sendo que se verifica que os tumores com conteúdo anormal de *DNA (aneuplóide)* apresentam a *fase S* e o indíce proliferativo elevados, enquanto que tumores com conteúdo de *DNA* normal (*diplóide*) estão associados a níveis baixos da *fase S* e índice proliferativo. Contudo, vários estudos como *Keyhani-Rofagha, S. et al (1990) e Cufer, T. et al (1997)* indicam que a *ploidia do DNA* não é significativa para o prognóstico de cancros da mama malignos nem numa melhor diferenciação entre curta ou longa sobrevivência de pacientes que sofrem deste tipo de cancro.

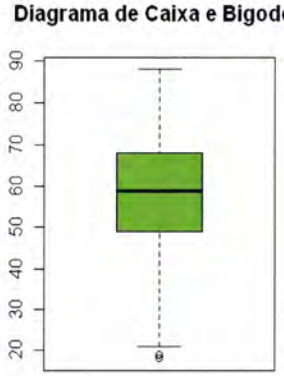
2.3.3 Análise Preliminar

A base de dados em estudo (*Caso*) é constituída por 212 pacientes a quem foi diagnosticado um tipo de neoplasia mamária num dos hospitais da região do Alentejo e foram, posteriormente, referenciadas por estes para a *Unidade de Anatomia Patológica do HESE.*

Para este estudo foram consideradas as 26 variáveis descritas na secção *2.3.2 – Descrição das variáveis*, Tabela 1.

A variável ***pM*** foi eliminada da base de dados por existir um número elevado de valores omissos no que respeita à mesma.

Foram ainda criadas, a partir das variáveis *idade* e *tipo histologico* as variáveis *idadecat* (idade categorizada em classes), a variável *tipo histológico binário* (que permite classificar os diversos tipos histológicos em benignos e malignos, conforme apresentado) e, posteriormente, a variável *tipo carcinoma* (que permite classificar os tipos histológicos 9,10 e 11 como invasivos e *in situ*). Posto isto, passou-se então à realização de uma análise descritiva dos dados em estudo.

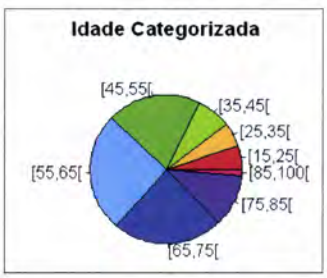


A representação do diagrama de caixa de bigodes da variável *idade* (Figura 6) permite concluir que:

- No que respeita à variável *idade*, esta toma valores compreendidos entre 18 e 88 anos, sendo a média de idades das pacientes em estudo de aproximadamente 57 anos e a mediana de 59 anos de idade.

A análise da Figura 7 permite observar a distribuição dos indivíduos em estudo pelas diferentes classes etárias (*idadecat*). Verifica-se que as classes etárias que apresentam maior número de pacientes são as compreendidas entre os 45 e os 85 anos de idade.

Figura 6: Diagrama de caixa e bigodes da variável idade



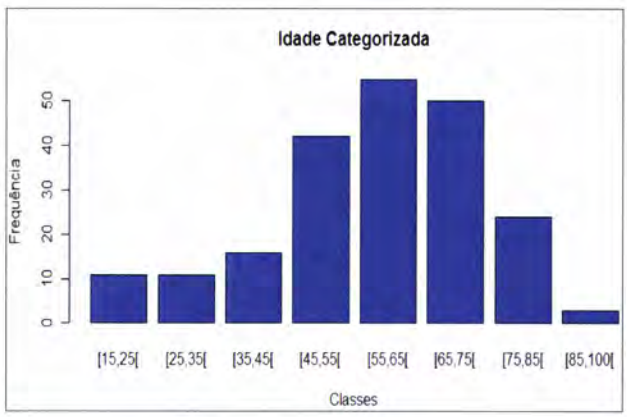


Figura 7: Histograma e diagrama de queijo da variável idadecat

As Figuras 8(A), (B) e (C) mostram que no que respeita às variáveis *localidade*, *concelho* e *distrito*, verifica-se que:

- A maioria dos indíviduos em estudo reside nas *localidades* de Beja (18.87%), Évora (19.34%), Estremoz, Montemor-o-Novo e Vidigueira (com 2.36% cada).
- Estão presentes indíviduos de 33 *concelhos* da região do Alentejo diferentes distribuídos da seguinte forma: Beja (22.17%), Évora (21.7%), Montemor-o-Novo (7.08%), Moura (6.13%), Serpa (4.72%), outros (38.2%).

- Quanto ao *distrito* de residência dos individuos em estudo, a grande maioria dos indivíduos pertence aos *distritos* de Évora (46.7%) e Beja (43.87%).

Figura 8: (A) Localidade, (B) concelho e (C) distrito dos indivíduos em estudo

No que diz respeito ao ambiente em que os casos em estudo residem, tendo em conta se se trata de um *ambiente rural* (lugares ou aldeias) ou de um *ambiente urbano* (vilas ou cidades), por meio da Figura 9 conclui-se que a grande maioria dos indivíduos em estudo vive em zonas *urbanas* (78.3%).

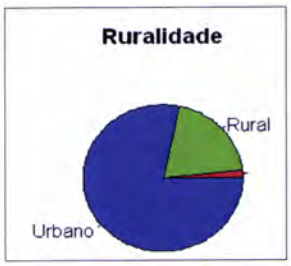


Figura 9: Ruralidade

Figura 10: Tipo de Amostra

A Figura 10 mostra que:

- Para a realização do diagnóstico realizado pela Unidade de Anatomia Patológica a amostra foi maioritariamente obtida como peça cirúrgica (69.81%).

- Tendo em conta o *tipo histológico* das neoplasias diagnosticadas nos indivíduos em estudo, manifestaram-se com maior frequência as seguintes: 7- Fibroadenoma (14.15%), 9-Carcinoma Ductal Invasivo – *CDI* (56.6%), 10 - Carcinoma *In Situ* (8.96%) e o 11 - Carcinoma Lobular Invasivo – *CLI* (6.6%), tal como se observa na Figura 11 (A).

Considerando a *malignidade* ou não da neoplasia diagnosticada, nos 212 casos em estudo estão presentes 162 casos malignos (76.42%) e 50 casos casos benignos (23.58%), tal como se verifica na Figura 11 (B).

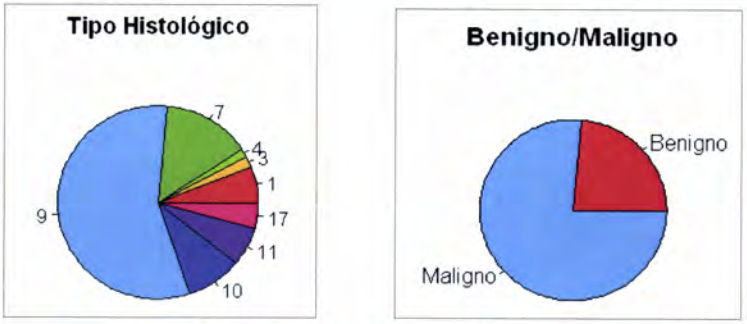


Figura 11: (A) Tipo Histológico e (B) Tipo Histológico Binário

- De entre as neoplasias diagnosticadas, observa-se, nos 212 casos em estudo, 19 *carcinomas in situ* e 134 *carcinomas invasivos*, como se pode verificar por meio da Figura 12.

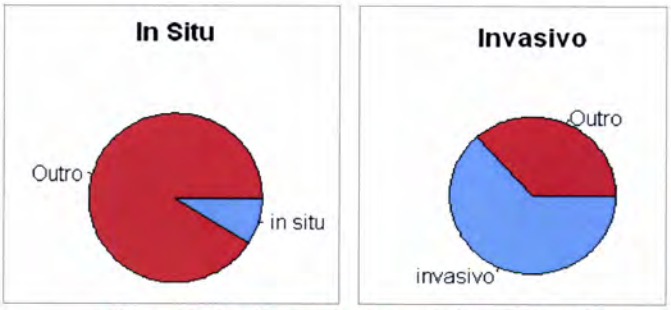


Figura 12: Carcinomas *in situ* e carcinomas invasivos

- No que respeita à variável *lateralidade*, no estudo em questão, observa-se um elevado número de valores omissos para esta variável (95 indíviduos), podendo observar-se 63 neoplasias diagnosticadas na mama direita e 54 neoplasias diagnosticadas na mama esquerda, tal como se observa na Figura 13.

Figura 13: Lateralidade

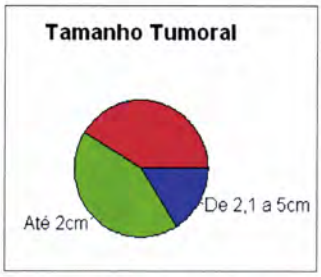


Figura 14: Tamanho Tumoral

Segundo a Figura 14, quanto ao *tamanho tumoral* existe um elevado número de valores omissos (87 dos 212 casos em estudo), sendo que, tendo em conta os restantes casos, a grande maioria das neoplasias apresenta tamanhos até 2 cm (42.92%) e apenas 16.04% apresenta tamanho tumoral entre 2,1 e 5cm.

Por observação da Figura 15, verifica-se que nos 212 casos de estudo, apenas se verificou invasão do tumor nas margens em 32 dos casos (15.1%), saliente-se o facto de haver um número elevado de valores omissos para esta variável (72 dos casos).

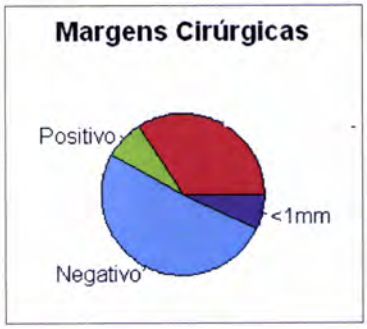
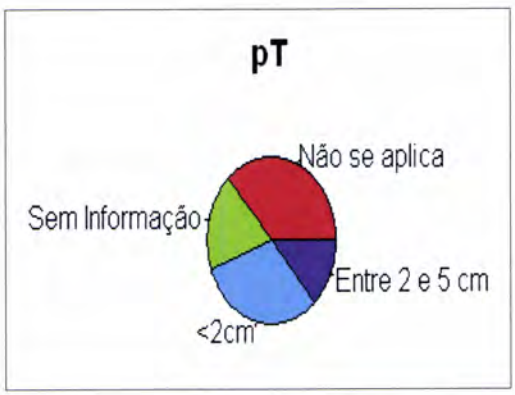


Figura 15: Margens Cirúrgicas

Figura 16: Grau Histológico

Em relação ao *grau histológico* os indivíduos em estudo encontram-se distribuídos de uma forma uniforme, sendo que existem 61 valores omissos, 56 casos de alto grau, 54 de médio grau e 41 de baixo grau, tal como se observa na Figura 16.

- No estudo em causa, tendo em conta a Figura 17, quer para a variável *pT*, quer para a variável *pN* encontramo-nos na presença de um número considerável de valores omissos (40 e 60 casos, respectivamente). Uma vez que, a análise destas variáveis apenas faz sentido para os carcinomas invasivos (134 dos casos), no que se refere à extensão do tumor primário 66 dos casos apresentam extensão inferior a 2cm e 28 casos entre 2 e 5 cm. No que se refere à análise da variável *pN*, verifica-se ausência de metástases regionais em 47 dos casos em estudo e metástases em gânglios móveis ou fixos em 27 dos casos.



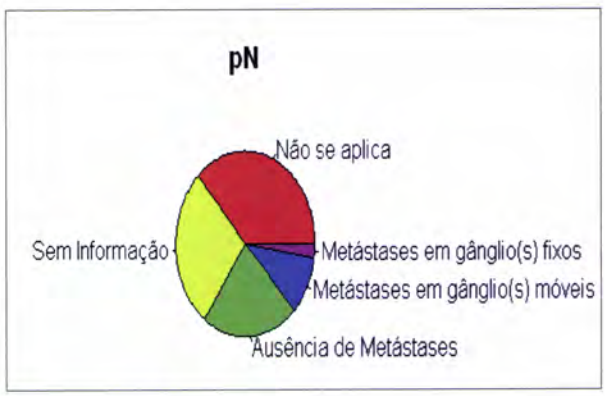


Figura 17: pT e pN dos carcinomas invasivos

Uma vez que o *índice de Van Nuys (IVN)* apenas faz sentido para os carcinomas *in situ* (no estudo, 19 casos), 11 deles apresentam risco baixo e 3 risco médio, tal como se pode observar por meio da Figura 18.

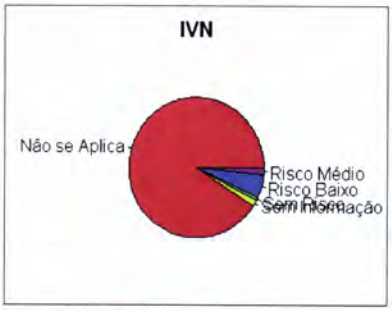


Figura 18: IVN

Figura 19: IPN

- O estudo da variável *IPN* apenas faz sentido para 134 dos casos em estudo (carcinomas *invasivos*), dos quais 59 não possuem informação para a mesma, apresentando 26 dos casos bom prognóstico, 36 dos casos prognóstico intermédio e 13 dos casos mau prognóstico, como mostra a Figura 19.

- No presente estudo, as variáveis *RE e RP* foram categorizadas da forma apresentada na Tabela 1. Saliente-se o facto de estarmos perante 90 casos de valores omissos quer para *RE*, quer para *RP*, sendo que 102 dos casos são favoráveis para *RE* e 87 são favoráveis para *RP*, como mostra a Figura 20.

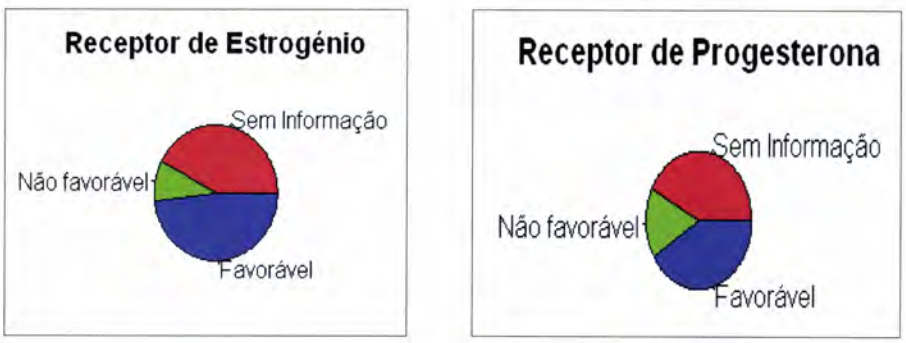


Figura 20: Receptores de Estrogénio e de Progesterona (RE e RP)

- A Figura 21 mostra que existem 90 casos de valores omissos para a variável *c.erB-2*, 90 casos em que não se observou coloração (não existe amplificação do gene c-erB-2) e apenas 13 casos em que se observou intensa coloração (amplificação do gene em causa).

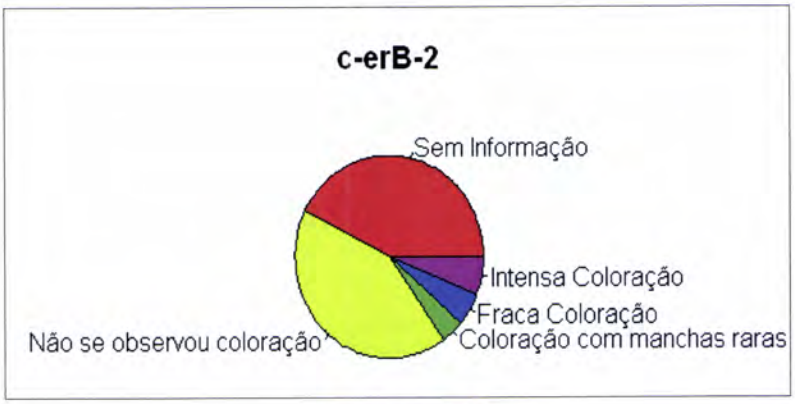


Figura 21: c.erB-2

- No presente estudo verificam-se 90 valores omissos para as variáveis *ki67 e p53*, sendo que, dos restantes casos, 90 apresentam valores de *ki67* favoráveis ao cancro da mama e 32 não favoráveis; 103 apresentam valores de *p53* favoráveis ao cancro da mama e 19 não favoráveis, tal como se pode observar por meio da Figura 22.

Figura 22: ki67 e p53

- Tendo em conta o caso de estudo, as variáveis *fase S* e *Índice de DNA* (*IDNA*) que foram classificadas segundo a Tabela 1, por meio da Figura 23 conclui-se que existem 112 valores omissos para a *fase S* e 101 para o *IDNA*. Quanto à *fase S*, 74 dos casos em estudo apresentaram valores favoráveis ao cancro da mama, no que se refere ao *IDNA*, 68 dos indivíduos apresentam *IDNA* diplóide e 33 indivíduos aneuplóide.

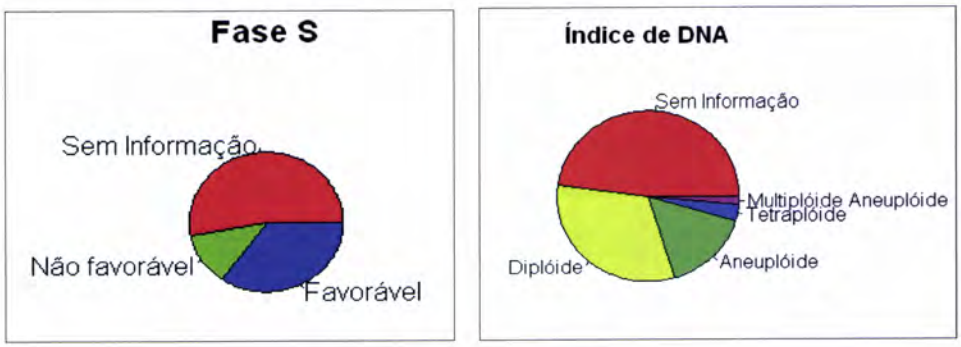


Figura 23: Fase S e IDNA

Procedeu-se, em seguida, à construção de tabelas de contingência e testes de hipóteses entre as várias variáveis consideradas mais relevantes para o estudo.

Tabelas de Contingência e Testes de Independência

As tabelas de contingência, também designadas, tabelas de dupla entrada, são utilizadas para estudar a relação entre duas variáveis categóricas descrevendo a frequência das categorias de uma das variáveis relativamente às categorias da outra.

A relação de independência entre variáveis pode ser estudada por meio do **teste do Qui-Quadrado**, segundo o qual são testadas as hipóteses:

H_0: As variáveis são independentes
H_1: Existe uma relação de dependência entre as variáveis

Tendo em conta os seguintes pressupostos:

1) *As frequências esperadas em cada categoria não devem ser inferiores a 5 unidades sempre que o número total de observações é ≤ 20;*
2) *Se $n > 20$ não deverá existir mais de 20% das células com frequências esperadas inferiores a 5 nem deverá existir nenhuma célula com frequência esperada inferior a 1.*

Neste contexto, duas variáveis são independentes se a probabilidade de cada observação pertencer a uma dada célula for o produto das suas correspondentes probabilidades marginais. Estimam-se as probabilidades marginais como o total de cada linha ou coluna dividido pela dimensão da amostra. Nos testes do Qui-Quadrado, comparam-se os números esperados de observações em cada célula com os respectivos valores observados para se concluir sobre a independência entre as variáveis. Se as diferenças entre os valores observados e esperados não se consideram significativas (o valor observado da estatística de teste está na região de aceitação do teste) conclui-se que as variáveis são independentes. Caso contrário, rejeita-se a hipótese da independência.
Os números esperados em cada célula estimam-se como o produto dos seus totais observados em linha e em coluna divididos por n (tamanho da amostra).

A distribuição da estatística do teste de Qui-Quadrado de *Pearson* segue uma distribuição aproximadamente X^2 com 1 grau de liberdade em tabelas de 2×2 e a sua expressão é a seguinte:

$$X^2 = \frac{n\,(a \times d - b \times c)^2}{(a+b)(c+d)(a+c)(b+d)}$$

em que $X^2 \sim X^2_{(1)}$

			Variável A		
			A1	**A2**	Total
Variável B		**B1**	a	b	a+b
		B2	c	d	c+d
		Total	a+c	b+d	

Tabela 3: Tabela ilustrativa do cálculo dos testes de Qui-Quadrado e de Fisher

O teste do Qui-Quadrado apenas informa sobre a independência entre as variáveis mas nada diz sobre o grau de associação existente.

Nas tabelas de dimensão superior a 2×2, o teste de independência do Qui-Quadrado de *Pearson* utiliza a seguinte estatística de teste:

$$X^2_{rc} = \sum_i \sum_j \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim X^2_{(r-1)(c-1)}$$

em que:
r = número de linhas;
c = número de colunas;
i são as observações em linha e j são as observações em coluna
O_{ij} são as frequências observadas da célula i, j
E_{ij} são as frequências esperadas da célula i, j

Sob a hipótese de independência esta estatística segue uma distribuição aproximada do X^2 com $(r-1) \times (c-1)$ graus de liberdade.

Se os pressupostos do teste do Qui-Quadrado não forem observados, o resultado do teste pode ser enganador, dependendo respectivamente da pequena ou elevada contribuição das células com frequência esperada inferior a cinco, para o valor do teste do Qui-Quadrado. Alguns autores, entre eles *Fisher R. (1938),* recomendam que se observe a seguinte restrição na utilização deste teste:

O teste do Qui-Quadrado pode ser usado se o número de observações em cada célula da tabela for maior ou igual a 5 e a menor frequência esperada for maior ou igual a 5.

A **Correcção de Continuidade de Yates**, sugerida por *Yates, F. (1934),* aplica-se a tabelas de $r \times c$ e procura melhorar o teste do Qui-Quadrado de *Pearson*, contudo, é vista por muitos investigadores como muito conservadora. Este teste obtém-se subtraindo 0,5 aos resíduos positivos e somando 0,5 aos resíduos negativos antes de os elevar ao quadrado. Os resíduos são as diferenças

entre os valores observados e os valores esperados, isto é,

$$X^2 = \sum_i \sum_j \left[\frac{\left(|O_{ij} - E_{ij}| - 0{,}5 \right)^2}{E_{ij}} \right]$$

A distribuição desta estatística, sob a hipótese de independência é Qui-Quadrado com $(r-1) \times (c-1)$ graus de liberdade. O valor obtido desta estatística é sempre inferior ao valor da estatística de *Pearson* e para amostras de grande dimensão os dois testes dão resultados aproximados, levando às mesmas conclusões.

Em amostras pequenas a aproximação da distribuição da estatística do teste de Qui-Quadrado de *Pearson* não é válida, e portanto, o teste não é recomendável, pois o valor do p-value poderá conter um erro apreciável. Foi então apresentado por *Fisher, R. (1938)* outro teste para a independência baseado numa estatística cuja distribuição se obtém de forma exacta – Teste Exacto de *Fisher* - não tem restrições de utilização em tabelas 2×2 e fornece valores exactos para os *p-values* do teste. Tendo em conta que na presença de tabelas 2×2 apenas temos 1 grau de liberdade e caso se verifiquem as seguintes condições:

- O valor de $n < 20$;
- $20 < n < 40$ e a frequência esperada for menor que 5;
- $n \geq 40$ e mais de uma frequência esperada igual a 1;

em que n é o tamanho da amostra, então a alternativa será a utilização do Teste Exacto de *Fisher*.

Segundo a Tabela 3 apresentada acima, pode dizer-se que a probabilidade exacta de obtermos uma dada configuração X de tabela 2×2, no teste de *Fisher*, é dada por:

$$P_a = P(X = a) = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!} \qquad (1)$$

O objectivo principal é determinar quais e quantas distribuições são possíveis na tabela, mantendo fixos os totais marginais.

No caso de haver uma célula na tabela com o valor zero o valor de p-value é o valor obtido em (1), caso não haja nenhuma célula com o valor zero, deve seguir-se o seguinte processo:

1. Calcular a probabilidade tal como se apresenta acima;

2. Construir outra tabela 2×2, subtraindo-se uma unidade aos valores da diagonal que contiver o menor número de acasos e adicionando essa unidade aos valores da outra diagonal;
3. Calcular novamente a probabilidade;
4. Esse processo continuará até que se atinja o valor zero;
5. Somar todas as probabilidades calculadas ficando $p = (P_a + P_{a-1} + \cdots + P_0)$

No caso do valor de p encontrado ser superior aos usuais níveis de significância, a hipótese das características serem independentes (H_0) é aceite, caso contrário, rejeita-se H_0.

Teste de Fisher ou Teste do Qui-Quadrado?

Ao se analisar tabelas de contingência de duas linhas e duas colunas, pode usar-se o teste do Qui-Quadrado ou o teste de *Fisher*.

- O teste de *Fisher* é o mais adequado porque é um teste exacto.
- O teste do Qui-Quadrado é mais fácil de calcular mas a distribuição por amostragem da sua estatística de teste é aproximada.
- O teste do Qui-Quadrado deve ser evitado quando os números presentes na tabela de contingência forem muito pequenos (abaixo de 5). Se os números forem maiores, os resultados dos dois testes são idênticos.

Para mais detalhe sobre tabelas de contingência e testes de independência ver, por exemplo, Everitt, *B. (1992).*

Actualmente, muitos investigadores preferem analisar a relação entre duas variáveis de escala nominal, **para tabelas 2×2,** através da **Razão de chances (Odds Ratio),** por ser mais facilmente interpretável do que o teste do Qui-Quadrado.

A **razão de chances (OR)** mede a associação entre duas variáveis nominais, em que uma das variáveis designadas por factor, é de ocorrência anterior à outra, designada por acontecimento. Basicamente, a razão de chances é definida como a razão entre a chance de um evento ocorrer num grupo e a chance de ocorrer num outro grupo.
Esses grupos, podem ser amostras de pessoas com ou sem uma doença, na qual se quer medir a chance dessa pessoa ter sido exposta a um determinado agente ambiental; ou grupos/amostras, para análises estatísticas, como homens e mulheres, tratados e não tratados, etc.

NOTA: Chance é a probabilidade de ocorrência de um evento dividida pela probabilidade de não ocorrência desse mesmo evento, isto é, $\frac{p}{p-1}$.

Posto isto, uma razão de chances de 1 indica que a condição em estudo é igualmente provável de ocorrer nos dois grupos, tal como refere, por exemplo, *McNeil, D. (1996)*. Uma razão de chances maior que 1 indica que a condição ou evento tem maior probabilidade de ocorrer no primeiro grupo. E, uma razão de chances inferior a 1 indica que a probabilidade é menor no primeiro grupo do que no segundo.
A razão de chances tem de ser igual ou superior a zero. Se a chance do primeiro grupo estiver próxima de zero, OR fica também próximo de zero; pelo contrário, se a chance do segundo grupo se aproximar de zero, OR tende a aumentar atingindo o infinito.

O cálculo da razão de chances é uma boa estimativa para a exposição a um factor de risco, contudo a precisão desta estimativa é, em parte, determinada pelo tamanho da sua amostra, e em geral, quanto maior o tamanho da amostra melhor, pelo que é convencional calcular-se o intervalo de confiança (IC) a 95% para a OR.

NOTA: A fórmula de cálculo desse intervalo de confiança é apresentada na secção *3.2 – Conceitos Básicos utilizados em Epidemiologia,* tal como uma descrição sumária da razão de chances e do seu cálculo.

Contudo, após os resultados da razão de chances e do seu intervalo de confiança, pode testar-se a dependência/independência à exposição/não exposição a determinado factor.

As hipóteses de análise são:

H_0: As variáveis são independentes, isto é, os rácios são iguais a 1
H_1: Existe uma relação de dependência entre as variáveis, isto é, os rácios são diferentes de 1

Não se rejeita H_0 se o valor 1 (igualdade entre o numerador e o denominador dos rácios) pertencer ao intervalo de confiança a 95%, por exemplo, determinado para a razão de chances.

Realizando agora algumas análises de dependência/independência entre as variáveis em estudo como se observa na Tabela 4 e Figura 24.

Idadecat versus *Tipo histologico binário*

	Tipo histológico binário	
Idadecat	Benigno	Maligno
[15,25[	11	0
[25,35[	7	4
[35,45[	6	10
[45,55[	11	31
[55,65[	9	46
[65,75[	4	46
[75,85[	2	22
[85,100[	0	3

Tabela 4: Idadecat *versus* tipo histológico binário

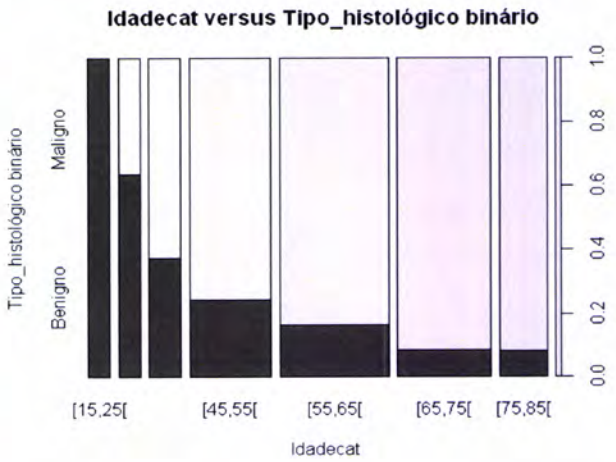


Figura 24: Relação entre as variáveis idadecat e tipo histológico binário

Posto isto, verifiquemos se existe uma relação de dependência entre as variáveis *idadecat* e *tipo histológico binário*, isto é:

H_0: As variáveis tipo histológico binário e idadecat são independentes
H_1: Existe relação entre as variáveis tipo histológico binário e idadecat

O *p-value* do correspondente teste do Qui-Quadrado foi $p=1.764\times10^{-10}$, pelo que se rejeita H_0 para um nível de significância de 5%, concluindo-se então que existe uma relação de dependência entre as variáveis *tipo histológico binário* e *idadecat*.

Tipo histológico binário versus *Ruralidade*

	Ruralidade	
Tipo histológico binário	Rural	Urbano
Benigno	8	41
Maligno	34	125

Tabela 5: Tipo histológico binário *versus* Ruralidade

H_0: As variáveis tipo histológico binário e ruralidade são independentes
H_1: Existe relação entre as variáveis tipo histológico binário e ruralidade

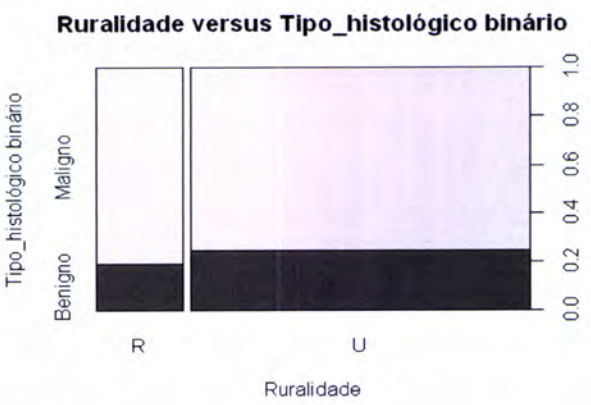


Figura 25: Relação entre as variáveis tipo histológico binário e Ruralidade

O *p-value* do correspondente teste do Qui-Quadrado foi $p=0.5704> 0.05$; pelo que não se rejeita H_0 para um nível de significância de 5% e, conclui-se então, que as variáveis *tipo histologico binário* e *ruralidade* são independentes.

Obteve-se ainda o intervalo de confiança para a razão de chances - $I.C._{95\%O.R.} =]\ 0.27;1.75\ [$-, o que permite reforçar o facto das variáveis serem independentes, uma vez que o valor 1 pertence ao intervalo de confiança.
O valor da razão de chances permite concluir que o **risco de um doente que reside em ambiente rural ter cancro de tipologia benigna excede, significativamente, 0.7 vezes (0.7184461) o dos doentes que residem em ambiente urbano.**

Na Tabela 27 presente no **Anexo 2**, apresentam-se os valores dos *p-values* dos testes do Qui-Quadrado e do teste de *Fisher* para as restantes variáveis consideradas no estudo. De uma forma abreviada podemos dizer que:

- A variável *idade* mostra relação de dependência com as variáveis: *tamanho tumoral, margens cirúrgicas, grau histológico, tipo histológico, tipo histológico binário e carcinoma invasivo;*
- A variável *ruralidade* mostra relação de dependência com as variáveis: *tamanho tumoral, grau histológico, ki67 e tipo histológico;*
- A variável *distrito* mostra-se dependente da variável *tipo de amostra*;
- A variável *tipo amostra* mostra-se dependente das variáveis: *tamanho tumoral, margens cirúrgicas, pT, pN, IPN, Re, Rp, c.erb.2, ki67,p53, fase s, idna, tipo histológico e tipo histológico binário;*
- A variável *lateralidade* mostra-se dependente das variáveis: *tamanho tumoral, margens cirúrgicas, grau histológico, pT, pN, IPN, Re, Rp, c.erb.2, ki67, p53,idna, tipo histológico, tipo histológico binário e carcinoma invasivo;*
- As variáveis *tamanho tumoral, margens cirúrgicas, grau histológico, pT, pN, IPN, c.erb.2,* mostram-se dependentes de todas as variáveis em estudo;
- A variável *IVN* mostra-se dependente das variáveis: *c.erb.2, tipo histológico e carcinoma in situ;*
- As variáveis *Re, Rp, c.erb.2, ki67, p53, fase s e idna* mostram-se dependentes de todas as variáveis em estudo com excepção da variável *carcinoma in situ;*
- As variáveis *tipo histológico, tipo histológico binário, carcinoma in situ e carcinoma invasivo* mostram ser independentes umas das outras.

Capítulo 3

Metodologia Epidemiológica

3.1 Epidemiologia

Lilienfeld, (1980) define Epidemiologia como a ciência que estuda os padrões da ocorrência de doenças em populações humanas e os factores determinantes destes padrões, isto é, trata-se de uma ciência que estuda em termos quantitativos a distribuição de fenómenos de saúde/doença e os seus factores condicionantes e determinantes nas populações Humanas.

É indiscutivel a evolução e as muitas definições de epidemiologia, sendo que, nos últimos 60 anos, esta ciência tem vindo a alargar-se desde a sua preocupação com as doenças infecto-contagiosas e outras doenças transmissíveis, integrando actualmente, todos os fenómenos relacionados com a saúde das populações Humanas.

Enquanto a medicina clínica aborda a doença a um nível individual, a epidemiologia aborda o processo saúde/doença ao nível de grupos de pessoas, podendo considerar-se pequenos grupos ou populações inteiras.

A epidemiologia é considerada uma pedra angular da metodologia de pesquisa da saúde pública, e é uma área bastante apreciada na medicina baseando-se em evidências para a identificação de factores de risco para as doenças e determinação de abordagens de tratamento optimizado na prática clínica. Desta forma, a epidemiologia estuda doenças transmissíveis e não-transmissíveis, focando-se o trabalho dos epidemiologistas na recolha e análise de dados, incluindo o desenvolvimento e aplicação de modelos estatísticos. Sendo assim, a análise da determinação causal das doenças numa colectividade Humana, dividida em classes sociais e/ou grupos específicos de populações (ou a distribuição desigual das doenças nas sociedades) exige da epidemiologia uma interacção transdisciplinar com outras ciências, tais como: Ciências Sociais (Antropologia, Sociologia, Etnologia); Ciência Política, Estatística, Economia, Demografia, Ecologia, História e Medicina, entre outras.

Quanto à origem desta ciência, sabe-se que esta remonta a bastantes séculos atrás, contudo, a palavra epidemia é bem mais antiga. A sociedade, tentou desde sempre, encontrar respostas para explicar as causas das epidemias, tentando sempre classificá-las num dos dois grupos seguintes: doenças infecto-contagiosas ou doenças crónicas, para que a partir daí se determinassem as causas das mesmas. A epidemiologia, segundo *Streiner, D. e Norman, G. (1998)* tem aplicação em diversas disciplinas relacionadas com a saúde e tem contribuído para definir novas síndromes clínicas e respectivas causas a fim de completar o retrato da história natural e curso clínico da doença.

3.2 Conceitos Básicos Utilizados em Epidemiologia

Ao falar de epidemiologia temos, primeiramente, que definir alguns dos conceitos mais utilizados que lhe estão associados. Tendo em conta várias referências, entre as quais *Mausner & Cramer (1984)* e *Clayton & Hills (1993)*, pôde resumir-se abaixo esses conceitos.

Causalidade em epidemiologia

A teoria da multicausalidade ou multifactorialidade tem hoje um papel bastante importante na génese das doenças, uma vez que a maioria das doenças advém de uma combinação de factores (e não, de apenas uma causa, como se pensava anteriormente) que interagem entre si, tendo um papel bastante relevante na determinação dessas mesmas doenças. Como exemplo, dessas múltiplas causas (chamadas causas contribuintes) pode referir-se o facto do cancro do pulmão ser diagnosticado em não fumadores, o que indica que não é apenas o tabaco que está na origem deste tipo de cancro mas que existem outras causas a contribuir para o aparecimento desta doença. Posto isto, pode dizer-se que existem vários factores que contribuem para o aparecimento de determinada doença, factores estes que segundo a proximidade em relação ao desenvolvimento da doença se podem classificar de uma das três formas: distantes, intermédios ou próximos.

Apenas os estudos experimentais estabelecem a causalidade, contudo a maioria das associações encontradas nos estudos epidemiológicos não é causal. Existem nove critérios de causalidade - os chamados critérios de Hill, definidos em *Hill, B. (1965)*.

Indicadores de Saúde

Em epidemiologia tratam-se muitas vezes as variáveis "saúde/doença" como se de uma combinação binária se tratasse, isto é, presença/ausência, uma forma simplista de tratar algo bastante complexo.

Para que seja possível quantificar estas variáveis, bem como proceder a comparações entre as populações, é necessária a existência de indicadores de saúde, devendo estes reflectir com fiabilidade, o panorama da saúde populacional. Apesar desses indicadores serem denominados de "Indicadores de Saúde", estes medem, na grande maioria das vezes, doenças, mortes, gravidade de doenças, o que mostra ser mais fácil medir doença que medir saúde.

Estes indicadores podem ser expressos em termos de frequência absoluta ou frequência relativa (coeficientes e índices), embora os valores absolutos sejam dados mais facilmente disponíveis.

Em estudos que exijam a comparação de frequência de uma determinada doença em diferentes grupos, deve ter-se em conta o tamanho das populações a serem comparadas com a sua estrutura de idade e sexo, expressando-se os dados em forma de taxas ou coeficientes.

Coeficientes ou taxas

Page, R. et al (1995) definem coeficientes ou taxas como as medidas básicas da ocorrência das doenças de uma determinada população num determinado período de tempo. Para o cálculo destes coeficientes ou taxas é necessário determinar se um indivíduo do grupo de interesse tem ou não desenvolvido uma determinada doença ou evento de interesse e esse conjunto de indivíduos representa o numerador da taxa. O denominador da taxa trata-se do número da população em risco de experimentar o tal evento.

Morbilidade

A morbilidade é um dos mais importantes indicadores de saúde, pois existem doenças que causam importante morbilidade mas baixa mortalidade.

Em epidemiologia morbilidade é considerada a extensão da doença, lesão ou deficiência numa população definida, em determinado local e em determinado momento.

Medir morbilidade não é tarefa fácil, pois a sua obtenção apresenta vários graus de dificuldade, como tal, para que se possa acompanhar a morbilidade de uma população é necessária a existência de medidas-padrão de morbilidade, fala-se então das **medidas de prevalência e das medidas de incidência.** Antes de as descrever de uma forma mais concreta é necessário estabelecer uma distinção entre os dois termos (prevalência e incidência) de forma a tornar mais fácil a comparação de frequências. **Incidência** refere-se aos **casos novos** de determinada doença numa dada população num certo período de tempo, e **prevalência** aos **casos existentes** num determinado momento, sendo a primeira dinâmica e a segunda estática (instantânea, de momento).

A incidência reflecte a dinâmica com que os casos aparecem num determinado grupo, informa quantos, entre os saudáveis se tornam doentes num dado período de tempo, ou ainda, quantos, entre os doentes, apresentam uma dada complicação ou morrem decorrido certo período de tempo. É por isso que se costuma dizer que a incidência reflecte a "força da morbilidade" (ou "força da mortalidade", quando se refere a óbitos).

A prevalência é função da incidência, da mortalidade e da efectividade terapêutica. Quanto maior for a incidência (número de casos novos), menor a letalidade (número de óbitos pela doença/número de casos da doença) e menor a efectividade terapêutica (os indivíduos doentes mantém-se mais tempo em estudo) maior é a prevalência. Como tal, tanto a melhoria de determinada doença com um medicamento, fazendo prolongar a vida, mas sem curar a doença, tal como, não tratar doenças curáveis leva ao aumento do número de casos na população, e consequentemente, eleva a taxa de prevalência.

Apresentamos em seguida alguns factores que influenciam a prevalência bem como algumas medidas de morbilidade descritas em *Bonita, R. et al (2006).*

Factores que influenciam a prevalência

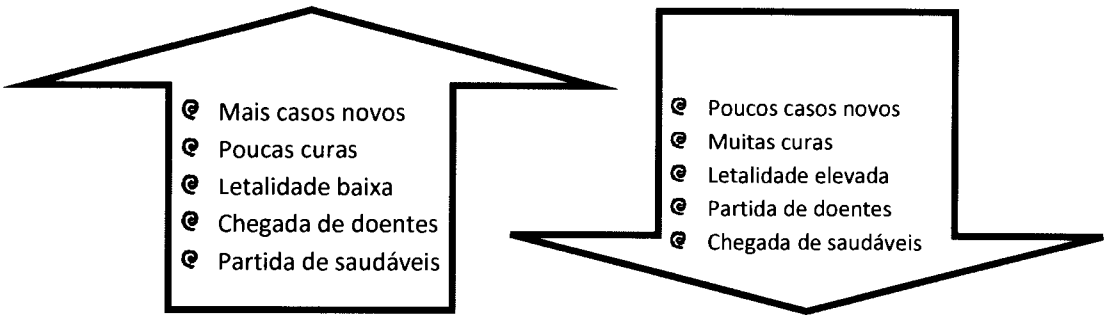


Figura 26: Factores que influenciam a prevalência

A relação entre **incidência e prevalência** segue a fórmula proposta por *Taylor, I. & Knowelden, J., (1957)*:

$$Prevalência \cong Incidência \times Duração\ da\ doença$$

Assim, *Mausner & Cramer (1984)* referem que a variação na prevalência pode ser o resultado de variação na incidência e/ou na duração média. Chama-se condição de equilíbrio ou de estabilidade de uma doença, à situação em que a incidência e a duração média permanecem constantes com o tempo. Evidentemente, uma doença somente atinge uma perfeita condição de estabilidade caso a incidência e a duração se mantenham ambas estáveis durante um período de tempo longo, transformando-se a fórmula acima em

$$Prevalência = Incidência \times Duração\ da\ doença$$

NOTA: Neste caso, conhecidos dois elementos da equação é possível determinar o terceiro.

Sendo assim, as medidas de morbilidade mais utilizadas e que podem ser observadas em *Mausner & Cramer (1984)* são:

Coeficiente de Prevalência (CP): mede a relação entre o número de casos conhecidos de uma doença e a população num determinado período.

$$Coeficiente\ de\ Prevalência\ (CP) = \frac{número\ de\ casos\ conhecidos\ de\ uma\ doença}{número\ de\ pessoas\ na\ população}$$

NOTA: Não devemos esquecer que o denominador deste coeficiente é a chamada população em risco (termo descrito anteriormente).

Coeficiente de Incidência (CI): Mede a relação entre o número de casos novos de uma doença num determinado intervalo de tempo e a população exposta ao risco de adquirir essa doença no mesmo período.

Também existe a incidência cumulativa ou acumulada (IC), que se refere à população fixa, onde não há entrada de novos casos no período de tempo considerado.

$$Incidência\ Comulativa\ (IC) = \frac{número\ de\ casos\ no\ decorrer\ do\ período}{população\ exposta\ no\ início\ do\ período}$$

Densidade de Incidência (DI): é uma medida de velocidade e o seu denominador conta as pessoas multiplicadas pelo seu tempo de seguimento, o denominador diminui à medida que as pessoas, inicialmente em risco, morrem ou adoecem (o que não acontece no caso da incidência cumulativa):

$$Densidade\ de\ Incidência\ (DI) = \frac{número\ de\ casos\ novos}{total\ de\ pessoas\ \times\ tempo\ de\ observação}$$

Mortalidade

O termo mortalidade refere-se ao conjunto dos indivíduos que morrem num dado intervalo de tempo. A taxa de mortalidade ou coeficiente de mortalidade é um dado demográfico do número de óbitos, geralmente para cada mil habitantes num determinado local e num dado período de tempo.

Risco e Grau de Risco

Conceituamos **risco** como probabilidade que um indivíduo ou grupo de indivíduos tem/têm de apresentar no futuro um dano na sua saúde. O conceito de risco é probabilístico e não determinístico, pois a primeira característica do risco é que este é incerto.

O **grau de risco** mede a probabilidade de que o dano ocorra no futuro. Dano aqui refere-se a um efeito não desejado. Enquanto o dano se refere à ocorrência, o risco é a probabilidade de ocorrência do dano, medindo-o como um gradiente que vai de baixo a alto risco. Sendo assim, o risco de que um evento ocorra varia entre zero e um, e quanto mais se aproxima de zero, mais baixo é o risco de ocorrência do dano e quanto mais se aproxima de um, mais alto é o risco para esse dano. Desta forma, é importante medir risco para possibilitar a prioridade ao grupo que apresente maior necessidade, segundo o valor de risco apresentado.

Risco Relativo

Os estudos analíticos (abordados de seguida) têm como objectivo determinar se existe alguma associação entre um factor (ou exposição) e uma doença e, em caso afirmativo, a força dessa associação.

Risco relativo é definido como a relação entre a taxa de incidência nas pessoas expostas a um factor e a taxa de incidência nas pessoas não expostas, o que segundo a Tabela 6, se pode escrever na forma, tal como se pode verificar em *Kahn, H. e Sempos, C. (1989)* e em *Mausner & Cramer (1984)*:

$$Risco\ relativo(RR) = \frac{Taxa\ de\ incidência\ entre\ os\ expostos}{Taxa\ de\ incidência\ entre\ os\ não\ expostos} = \frac{A/(A+B)}{C/(C+D)} \qquad (3.1)$$

Na interpretação do RR, pode dizer-se que caso:

- $RR = 1$, então não há diferença entre os grupos
- $RR > 1$, então a exposição é um factor de risco
- $RR < 1$, então a exposição é um factor de protecção

O risco relativo indica assim quantas vezes a ocorrência do desfecho nos expostos é maior do que aquela entre os não-expostos. A variabilidade amostral desta medida pode ser avaliada através de testes de significância ou via intervalos de confiança. Assim, com o objectivo de saber se um determinado valor de RR representa um efeito presente na população e não apenas na amostra estudada, pode-se calcular um intervalo de confiança (I.C.) para esta quantidade. Para um dado grau de confiança $(1-\alpha)\times 100\%$ este intervalo compreende uma gama de valores mais plausíveis para o RR. Se o valor 1 (referente à nulidade de associação) não estiver contido no intervalo, podemos concluir com uma significância α que na população de onde foi extraída a amostra o RR é diferente de 1, sendo portanto, significativo.

Para a dedução do cálculo do I.C. a 95% do RR, utiliza-se o método de transformação logarítmica. Este método parte do pressuposto que a distribuição amostral de valores de RR possui uma forma assimétrica do tipo log-normal. Assim, por meio de uma transformação logarítmica obtém-se uma curva com forma aproximadamente normal. Desta forma, utilizando fórmulas análogas às utilizadas para o cálculo do intervalo de confiança para variáveis com distribuição normal, pode construir-se o intervalo de confiança para o logaritmo do RR, $ln\ RR$ (no caso específico das medidas de associação utilizam-se os logaritmos naturais). Para expressar o intervalo de confiança na escala original do RR, basta obter a exponencial dos limites encontrados, $exp(\ln RR)$. Posto isto, pode construir-se o intervalo de confiança para o risco relativo usando a seguinte fórmula para o erro padrão (SE) do seu logaritmo$(ln\ RR)$:

$$SE\ (ln\ RR) = \sqrt{\frac{1}{A} - \frac{1}{A+C} + \frac{1}{B} - \frac{1}{B+D}}.$$

A distribuição de amostras de $ln\ RR$ trata-se da distribuição normal, pelo que para que possamos construir, por exemplo, um intervalo de confiança a 90%, para o logaritmo do risco relativo pode usar-se

$$ln\ RR \pm z_{0.95} \times SE(ln\ RR),$$

onde $z_{0.95}$ é o valor apropriado para a distribuição normal. O intervalo de confiança para o risco relativo é obtido por meio da exponencial destes valores. Para mais detalhe ver, por exemplo, *Altman, D. (1991).*

Razão de chances (OR) e o seu Intervalo de Confiança

Não devem considerar-se os termos probabilidade e chance como sendo sinónimos, pois não são, enquanto probabilidade contrapõe o número de casos favoráveis com o número de casos possíveis, a chance contrapõe (ou compara) o número de casos favoráveis com o número de casos desfavoráveis.

Já foi atrás mencionada a razão de chances, podendo concluir-se que OR é, então definida como a probabilidade de que um evento ocorra dividido pela probabilidade de que ele não ocorra e, tendo em conta a Tabela 6, pode ser calculada como:

$$Razão\ de\ Chances\ (OR) = \frac{A/C}{B/D} = \frac{A \times D}{B \times C} \qquad (3.2)$$

		Casos de doença ou de outro tipo de evento		
		Sim	Não	Total
Exposição no início do período em estudo	Sim	A	B	A+B
	Não	C	D	C+D
	Total	A+C	B+D	A+B+C+D

Tabela 6:Casos *versus* exposição/não exposição

Demonstra-se que quando o efeito é raro, é possível aproximar o *RR* num estudo do tipo caso-controlo, pela razão de chances (OR), isto é, quando A é pequeno em relação a B e C é pequeno em relação a D, os resultados entre *RR* e OR são próximos ***(3.1 e 3.2).***

Devido à utilização cada vez maior da razão de chances em epidemiologia, apresentam-se a seguir algumas considerações de *Mausner & Cramer (1984)* sobre esta medida:

- A razão de chances pode ser igual a qualquer valor positivo;
- É igual a zero ou ∞ se algum valor da tabela for nulo (A, B, C ou D) e é indefinido se as duas entradas de uma linha (A e B) ou de uma coluna (C e D) forem nulas;
- É utilizado como medida de associação em estudos caso-controlo, em que os controlos são seleccionados a partir da população em estudo por processo de amostragem. Toda a amostra por melhor que seja feita está sujeita ao acaso, e é por isto que a razão de chances deve ser expressa na forma de intervalo de confiança, calculado a partir de uma margem de erro pré-determinada.

Uma vez calculado o valor da razão de chances, podemos estimar o seu intervalo de confiança para um nível de confiança de 95%, aplicando a seguinte fórmula *(Woolf, B., 1955)*:

$$I.C._{95\%}(OR) \equiv \left]OR \times e^{-z_\alpha}\sqrt{\frac{1}{A}+\frac{1}{C}+\frac{1}{B}+\frac{1}{D}}; OR \times e^{z_\alpha}\sqrt{\frac{1}{A}+\frac{1}{C}+\frac{1}{B}+\frac{1}{D}}\right[$$

Este intervalo de confiança pode ser obtido de forma semelhante ao do risco relativo. Para a demostração pode ver-se, por exemplo, *Ahlbom, A. (1993) ou Hosmer & Lemeshow (2000).*

Interpretando os I.C. para OR segundo *Mausner & Cramer (1984)*, pode dizer-se que:

- Se I.C. inclui o valor 1 – não há diferença, os coeficientes de incidência dos dois grupos são iguais.
- Se I.C. não inclui o valor 1 – há diferença estatisticamente significativa
  - Se o I.C. está todo para a direita de 1 – a exposição é um factor de risco;
  - Se o I.C. está todo para a esquerda de 1 – a exposição é um factor de protecção.

Principais fontes de erro na medição das observações

Deve ter-se sempre presente a possibilidade de erro nas medições das observações a executar em todas as actividades científicas, havendo dois tipos de erros, os que são devido ao acaso (aleatórios) e os sistemáticos (viés). Erros aleatórios referem-se a flutuações à volta de um valor verdadeiro, devido à variabilidade das amostras. O erro sistemático ou viés expressa qualquer diferença entre o valor verdadeiro e o resultado obtido devido a qualquer causa que não seja a variabilidade da amostragem. De entre os dois tipos de erros referidos, o viés é, normalmente, o mais importante, o mais insidioso e o mais difícil de medir, podendo ser entendido como um erro que conduz a uma conclusão tendenciosa. Pode-se classificar em três tipos: **o viés de selecção, o viés de informação e o viés de confundimento**.

Seguindo *Mausner & Cramer (1984)* **o viés de selecção** deve ser considerado quando se está na presença de estudos do tipo caso-controlo. Este ocorre quando casos e controlos diferem entre si sistematicamente, devido à forma de selecção. O recrutamento de casos entre pacientes hospitalizados (ou institucionalizados) é particularmente sujeito ao viés de selecção uma vez que os factores que levam à hospitalização também estão associados a muitos factores de risco. O viés de selecção tem a ver com o facto de seleccionarmos uma amostra através de um método que não garante a sua representatividade.

O viés de informação está relacionado com os erros de classificação da amostra, erros na medição das variáveis, ou na codificação e recolha da informação.

O viés de confundimento relaciona-se com factores que podem ser uma explicação alternativa para associações encontradas uma vez que este pode surgir quando uma determinada variável (denominada confundimento) se intromete alterando ficticiamente a associação entre a variável de exposição e a variável de resposta.

Posto isto, *Mausner & Cramer (1984)* referem que existem várias fontes de erro na medição da morbilidade, entre elas destacam-se: selecção de amostras enviesadas (não randomizadas) da população a estudar; a não participação dos seus elementos; variação entre os observadores; diferentes padrões de resposta; diferenças na percepção da doença (comportamento perante a doença) e diferenças no acesso aos recursos terapêuticos.

Sensibilidade e Especificidade

A sensibilidade mede a capacidade de um teste em identificar correctamente indivíduos que apresentam determinada doença, isto é, o quão sensível é o teste. Já a especificidade mede a capacidade do teste em identificar correctamente aqueles que não possuem a doença, ou seja, o quão específico é o teste.

Resumidamente, tendo em conta a Tabela 7 presente de seguida e a descrição dos seus elementos tem-se:

D^+ - Ter doença
D^- - Não ter doença
T^+ - Teste positivo
T^- - Teste negativo
VPP – Valor preditivo positivo
VPN – Valor preditivo negativo

	D^+	D^-	
T^+	a	b	VPP
T^-	c	d	VPN
	S	E	

Tabela 7: Doente/não doente *versus* teste positivo/negativo

Sensibilidade: Probabilidade de um teste ser positivo, dado que existe a doença

$S = a/(a + c)$

Especificidade: Probabilidade de um testes ser negativo, dado que não existe a doença

$E = d/(b + d)$

Valor Preditivo Positivo: Probabilidade de existir a doença, dado que o teste foi positivo

$VPP = a/(a + b)$

Valor Preditivo Negativo: Probabilidade de existir a doença, dado que o teste foi negativo

$VPN = d/(c + d)$

3.3 Tipologia dos Estudos Epidemiológicos

Os estudos epidemiológicos são um excelente método para colher informações que não são disponibilizadas através dos sistemas de rotina de informação de saúde ou de vigilância.

Este tipo de estudos varia conforme os objectivos estabelecidos, *Hennekens, C. et al (1987)* mencionam que se trata de estudos que pretendem descrever a distribuição da doença ou esclarecer os seus determinantes procurando relações causa-efeito. Também podem ser um modo de avaliar os procedimentos terapêuticos e preventivos alternativos.

Bonita, R. et al (2006) referem que escolher o desenho do estudo adequado é uma etapa crucial na investigação epidemiológica.

Na Figura 27 são apresentados os principais tipos de estudos epidemiológicos, os quais serão brevemente descritos de seguida. São também referidas as principais vantagens e desvantagens associadas a cada um dos tipos de estudos epidemiológicos abordados.

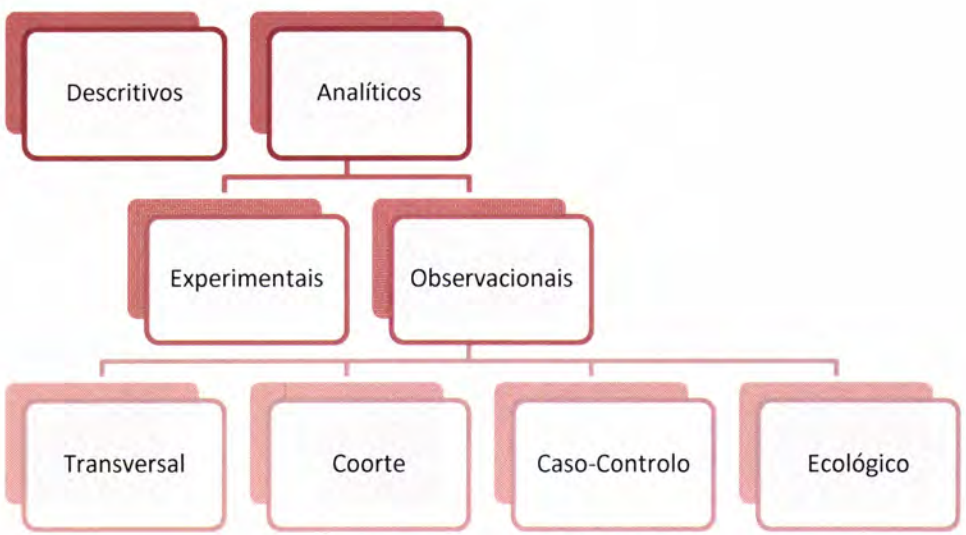


Figura 27: Tipos de estudos epidemiológicos

Os **estudos descritivos** objectivam determinar e informar a distribuição de doenças ou condições relacionadas com a saúde, segundo o tempo, o lugar e/ou as características dos indivíduos, ou seja, responder às questões: quando, onde e quem adoece, tal como referem *Hennekens, C. et al (1987). Bonita, R. et al (2006)* referem tratar-se da primeira etapa da aplicação do método epidemiológico em que o objectivo se prende em descrever as características de uma determinada amostra, não sendo de grande utilidade para estudar a etiologia de uma doença ou a eficácia de um tratamento, devido ao facto de não existir um grupo de controlo para permitir inferências causais. Estes tipos de estudos não testam hipóteses, mas são a base para a formulação destas.

Vantagens – rápidos, baixo custo, são na grande maioria das vezes o ponto de partida para um outro tipo de estudo epidemiológico.
Desvantagem – não existe um grupo de controlo para futuras comparações.

Os **estudos analíticos** estão, normalmente, sujeitos a uma ou mais questões científicas, as ditas "hipóteses", que relacionam eventos: uma suposta "causa" e um dado "efeito" ou "exposição" e "doença" respectivamente. *Bonita, R. et al (2006)* referem que os estudos analíticos vão mais longe, analisando as relações entre o estado de saúde e outras variáveis, uma exposição em particular e um efeito específico, ou a avaliação de procedimentos terapêuticos ou preventivos, testando hipóteses e pressupondo a existência de um grupo de referência, o que permite estabelecer comparações. Os estudos analíticos, no entanto, de acordo com o papel do pesquisador podem ser:

- Observacionais
- Experimentais

Nos **estudos experimentais** o pesquisador controla as variáveis e os indivíduos, estes últimos são submetidos a uma exposição controlada, registando-se o efeito desta.

Vantagens – alta credibilidade, existe um grupo de controlo, permitem uma cronologia de eventos e a interpretação de resultados é simples.
Desvantagens – há situações que não podem ser estudadas por este método, existe a possibilidade de perdas e recusas por parte das pessoas pesquisadas, necessidade em manter uma estrutura administrativa e técnica bem preparada associadas a elevados custos.

Ao contrário, os **estudos observacionais** são aqueles em que o investigador não exerce controlo sobre as variáveis, limitando-se à observação e registo, tal como refere *Altman, D. (1991).*

Os estudos observacionais compreendem:

- Estudo Transversal

- Estudo de Coorte
- Estudo Ecológico
- Estudo Caso-Controlo

As principais diferenças entre os diferentes estudos existem na forma como os indivíduos são seleccionados e na capacidade de medir a exposição no passado.
De seguida, cada um desses métodos será abordado no que respeita aos seus principais pontos de interesse.

Os estudos epidemiológicos podem também ser classificados quanto à sequência temporal do levantamento da exposição e do efeito desta. Posto isto, denominam-se **transversais** quando se faz o registo simultâneo da exposição e do efeito e **longitudinais** quando a exposição e efeito se registam em tempos diferentes. Os estudos **transversais** são também conhecidos como **seccionais ou de prevalência**. Nos estudos **longitudinais** existem dois tipos de delineamento: **caso-controlo e coorte.**

Estudo Transversal

Um estudo transversal ou de prevalência segundo *Dever, G. e Champagne, F. (1984)* examina um determinado grupo de pessoas num determinado período de tempo.
Aplica-se essencialmente, a doenças comuns e de duração longa. O pesquisador delimita uma amostra da população e avalia todas as variáveis dentro dessa amostra, este tipo de estudos envolve um grupo de expostos e um grupo de não expostos a determinados factores de risco. A ideia central do estudo transversal é que a prevalência da doença deverá ser maior entre os expostos que nos não expostos, se for verdade que aquele factor de risco causa a doença. Nestes estudos, "causa" e "efeito" são detectados simultaneamente.
Os estudos transversais têm como objectivo encontrar rapidamente associações comuns entre factores, essa associação é feita como no risco relativo, contudo, denomina-se razão de prevalência (relação entre a prevalência entre expostos e não expostos); pode também ser feito o cálculo pela razão de chances.

Vantagens - rapidez, baixo custo, facilidade na obtenção de uma amostra representativa da população, simplicidade analítica e alto potencial descritivo.
Desvantagens - vulnerabilidade a viés (especialmente de selecção), baixo poder analítico, os pacientes curados ou falecidos não aparecem na amostra, provocando um quadro incompleto da doença (viés da prevalência), a relação cronológica dos eventos pode não ser facilmente detectável, não determina o risco absoluto (incidência).

Estudo de Coorte

Tendo em conta a forma como *Breslow, N. & Day, N. (1980) e Altman, D. (1991)* descrevem este tipo de estudos pode concluir-se que nos estudos coorte, também conhecidos como estudos de incidência, longitudinais ou de seguimento (*follow-up*), existe um grupo de indivíduos com algo em comum, que são acompanhados ao longo do tempo e que periodicamente são investigados por pesquisadores que vão agrupando dados sobre estas pessoas, com o objectivo de se observar a ocorrência de um desfecho. Este delineamento é utilizado para problemas comuns, sendo o seu princípio lógico a identificação de pessoas sadias, a classificação das mesmas em expostas e não expostas ao factor de risco e o acompanhamento destes dois grupos por um período de tempo suficientemente longo para que haja o aparecimento da doença tal como refere *Bonita, R. et al (2006)*. Estes estudos são excelentes para avaliar várias exposições e doenças ao mesmo tempo, estando indicados para doenças frequentes. O aspecto mais importante dos estudos coorte é poder estabelecer a incidência e investigar as potenciais causas que levam a uma determinada condição. A medida de associação nos estudos coorte é a análise do risco relativo, ou seja, quantas vezes os indivíduos expostos desenvolvem a doença quando comparados aos não expostos. Quanto mais forte for a associação, maior será o risco relativo ($RR > 1$) e quando igual a 1, indica que não existe associação tal como referem *Kahn, H. e Sempos, C. (1989)*. Trata-se de estudos que partem da "causa" em direcção ao "efeito", os grupos são formados por "observação" das situações na vida real e existe uma comparação da incidência de casos nos grupos de expostos e não expostos.

Vantagens – é possível calcular o risco relativo, alto poder analítico, simplicidade do desenho, facilidade de análise, não há problemas éticos quanto a decisões de expor as pessoas a factores de risco, a cronologia dos eventos é facilmente determinada, muitos desfechos clínicos podem ser investigados.

Desvantagens – trata-se de um estudo caro, inadequado para doenças de baixas frequências, longo tempo de acompanhamento, vulnerável a perdas, pode ser afectado por mudanças de critérios diagnósticos, por mudanças administrativas e por mudança nos grupos (indivíduos que mudam de hábitos) devido ao longo tempo de seguimento, presença de variáveis confundidoras.

Estudo Ecológico

Segundo *Mausner & Cramer (1984)* nos estudos ecológicos a unidade de observação é um grupo de pessoas e não o indivíduo, trata-se de estudos mais do tipo "gerador de teses" do que do tipo "verificador de hipóteses". O princípio do estudo é o de que, nas populações onde a exposição é mais frequente, a incidência das doenças ou a mortalidade serão maiores. As medidas mais usadas para quantificar a ocorrência de doenças neste tipo de estudos são a incidência e a mortalidade, sendo a análise de correlação a mostrar a associação ente o factor de risco e a doença, não se tratando, no entanto, de uma relação causa-efeito. A limitação deste estudo está no facto de se

atribuir a um indivíduo o que se observou através das estatísticas, a observação da associação de eventos a nível da população não significa, necessariamente, que essa mesma associação se verifique a nível de um indivíduo.

Vantagens – facilidade e rapidez na sua execução, baixo custo, simplicidade analítica, capacidade de gerar hipóteses.

Desvantagens – baixo poder analítico, pouco desenvolvimento das técnicas de análise de dados, não há acesso aos dados individuais, a colheita de dados é feita por diversas fontes, significando pouco controlo sobre a qualidade da informação, existe dificuldade em controlar os viéses.

Estudo de caso-controlo

Breslow, N. & Day, N., (1980) e Hennekens, C. et al (1987) citam que nos estudos do tipo caso-controlo parte-se do efeito para a causa, exigindo para tal a formação de um grupo de indivíduos com determinada doença (efeito) e de um grupo de controlo constituído por indivíduos semelhantes, mas sem a doença. As características (possíveis causas da doença) dos indivíduos de cada grupo são levantadas e verificadas as frequências das mesmas nos dois grupos, tentando identificar os factores de risco da doença.Parte-se da presença (casos) ou ausência (controlos) de doenças e avalia-se retrospectivamente na tentativa de encontrar possíveis associações. Este desenho é retrospectivo, pois doença e exposição já aconteceram no momento do delineamento do estudo, estes estudos estão principalmente indicados para doenças raras, tal como referem *Kahn, H. e Sempos, C. (1989)* que acrescentam ainda que a medida estatística de associação utilizada no estudo caso-controlo trata-se da razão de chances (definida atrás) que se comporta como o risco relativo, para os estudos coorte (quando> 1 existe associação e quando igual a 1 não há associação).

Vantagens – baixo custo, alto potencial analítico, adequado para estudar doenças raras, os resultados são obtidos rapidamente.

Desvantagens – dificuldade em formar o grupo de controlo, vulnerável a inúmeros viéses, os cálculos das taxas de incidência não podem ser feitos directamente (é o pesquisador que determina o número de casos a estudar) e o risco tem de ser estimado indirectamente (*razão de chances*), complexidade analítica.

3.4 Estudo Caso-Controlo

3.4.1 Cancro da Mama e Estudos Caso-Controlo (Breve Revisão da Literatura)

Com o objectivo de encontrar associação entre o cancro da mama e os diversos factores de risco a ele associados foram já realizados diversos estudos do tipo caso-controlo, alguns deles referidos na *secção 2.2* deste estudo, contudo, referem-se aqui mais alguns estudos interessantes realizados nesse âmbito.

Bap-tiste, M.S. et al (1990) realizaram um estudo caso-controlo em 18 condados de Nova Iorque e concluiram que o risco de cancro da mama aumenta com o aumento do consumo diário de álcool, entre as mulheres que consumiram 15 ou mais gramas de álcool por dia, contudo o risco de cancro da mama não parece estar relacionado com o número total de anos que uma mulher bebeu ou restrito a determinados tipos de bebidas alcoólicas.
Zheng, T. et al (2001) concluem uma associação inversa entre amamentação e o risco de cancro da mama.
Tessaro, S. et al (2001) não encontraram associação entre o uso de contraceptivos orais e o cancro da mama em geral, assim como entre faixas etárias e tempo de uso dos contraceptivos orais.
Friedenreich, C. et al (2001) confirmam pesquisas anteriores que associam a realização de desporto e a redução do risco de cancro da mama, sendo essa redução particularmente notável em mulheres na pós-menopausa, para não fumadores e para os que não ingerem bebidas alcoólicas.
Ebrahimi M., Vahdanini M. e Montazeri, A. (2002) realizaram um estudo a pacientes com cancro da mama no Irão e concluíram que uma história familiar positiva de cancro da mama é um forte factor de risco para o aparecimento do cancro da mama em idades jovens e relacionam esta conclusão com a estrutura etária do país, essencialmente constituída por uma população jovem e uma combinação da elevada idade da menarca e idade baixa aquando da primeira gravidez, que são protectoras mais tarde, contudo não houve associação entre cancro da mama e paridade, embora estes autores referenciem estudos que têm relacionado a nuliparidade com a redução do cancro da mama em idade precoce e aumento em idade adulta.

Como estudos interessantes, tendo em conta a exposição ao ambiente rural/urbano destacam-se:

- *Duell, E. et al (2000)* que associam inversamente o risco de cancro da mama e a duração das campanhas de agricultura, concluindo que o facto de se trabalhar ou residir em zonas rurais podem estar associados com um risco reduzido de cancro da mama, no entanto, os seus

resultados sugerem um possível aumento do risco entre um subgrupo de mulheres que trabalham na agricultura e que podem estar mais susceptíveis à exposição a pesticidas.

- *Robert, S. et al (2004)* que associam um maior risco de cancro da mama às mulheres que têm um nível socioeconómico superior ou que vivem em comunidades urbanas.
- E ainda, um estudo realizado a mulheres indianas por *Mathew, A. et al (2008)* refere que a incidência de cancro da mama na Índia é, aproximadamente o dobro nas mulheres que vivem em zonas urbanas quando comparadas com as mulheres que vivem em zonas rurais e para afirmarem tal facto investigaram o papel dos factores antropométricos (índice de massa corporal, tamanho da cintura, tamanho do quadril, etc.) e o tamanho do corpo (grande tamanho de corpo aos 10 anos), e esses dados confirmam a hipótese de que o aumento de factores antropométricos trata-se de factores de risco de cancro da mama na Índia.

Tendo em conta que não só as mudanças no ambiente mas também os hábitos alimentares podem estar intimamente relacionados com o risco de cancro da mama, *Bessaoud, F. et al (2008)* referem que o alto consumo de gordura dietética, de carne e de produtos lácteos parece ser parcialmente responsável pelo aumento do cancro da mama, por outro lado, outros estudos têm mostrado que o alto consumo de frutas, legumes e peixe pode proteger contra este tipo de cancro. Neste estudo refere-se ainda que aleitamento materno bem como actividade física mostraram-se significativas e inversamente associadas ao risco de cancro da mama e longa duração da actividade ovulatória, índice de massa corporal elevado aumenta significativamente o risco de cancro da mama.

Bessaoud, F. et al (2008) vão um pouco mais além, relacionando alguns alimentos com o risco de cancro da mama e concluem que o consumo de cereais e azeite e cancro da mama são inversamente proporcionais e referem que o risco de cancro da mama aumenta 56% por cada 100g/dia adicionais do consumo de carne.

Garicochea, B. et al (2009) referem que o risco de cancro da mama aumenta com a idade.

3.4.2 Caso de estudo: Cancro da Mama e Ruralidade

3.4.2.1 - Material e Métodos

Para a elaboração de um estudo caso-controlo associado ao cancro da mama achou-se interessante e pertinente constituir uma variável de exposição ao ambiente rural – ruralidade.

Pretende-se assim, testar a associação do **cancro da mama** com a **ruralidade**, realizando-se um estudo do tipo caso-controlo. Sobre a variável ruralidade sabe-se que esta foi definida tendo em conta a localidade de habitação de cada um dos pacientes em estudo, considerando que o ambiente rural respeita aldeias e lugares e o ambiente não rural as vilas e as cidades da região em estudo (Alentejo).

A amostra foi constituída por 303 indivíduos do sexo feminino, na faixa etária dos 18 aos 88 anos, sendo todos eles residentes na região do Alentejo. Dos 303 individuos em estudo, 212 eram portadores de algum tipo de neoplasia mamária (casos) e 91 eram normais (controlos). Os casos seleccionados foram todos os indivíduos em que foi diagnosticado um tipo de neoplasia mamária, atendidos na *Unidade de Anatomia Patológica do HESE*, no período decorrente de Agosto de 2003 a Agosto de 2004. Os casos foram seleccionados utilizando-se os seguintes critérios de inclusão: pacientes do sexo feminino (uma vez que o sexo masculino se apresenta numa minoria pouco significativa) em que foi diagnosticada qualquer espécie de neoplasia mamária, com idade superior ou igual a 18 anos na época do diagnóstico definitivo da doença e com residência na região do Alentejo. A identificação dos casos foi realizada através de pesquisa activa, por meio de visitas periódicas ao serviço onde os casos eram diagnosticados. Casos de cancro da mama resumiram-se a casos de neoplasias com o código C.50 de acordo com a 10ª Revisão de Classificação Internacional de Doenças (CID-10) da *OMS*, incluindo todas as subcategorias.

O grupo controlo foi seleccionado entre pacientes do hospital sem história ou suspeita de neoplasia mamária e emparelhados por idade, sexo e localidade de habitação.

Casos e controlos foram registados tendo em conta as informações presentes na ficha médica. O acesso a essas fichas deu-se a partir da Unidade de Anatomia Patológica, que forneceu as referidas fichas, a partir das quais, o médico patologista e assistentes do serviço, procederam a uma explicação breve e concisa acerca das mesmas, com vista à melhor selecção.

Todos os cuidados foram tomados para que não houvesse diferença na abordagem dos casos e dos controlos de modo a garantir a fidedignidade e evitar vícios da informação que pudessem comprometer a validade dos resultados.

Além do *sexo* e da *idade*, uma outra variável de interesse é a *ruralidade*. A variável *idade* foi codificada em 8 categorias, com intervalos de 5 anos, tal como se apresentou na Tabela 1. A *ruralidade* foi codificada em 3 categorias: "sem informação","rural", "urbano", tendo em conta a *localidade* em que reside cada um dos pacientes em estudo. Uma vez que o presente estudo de caso-controlo visa estabelecer uma relação entre as variáveis *cancro da mama* e *ruralidade*, não faz sentido a categoria "sem informação" para a *ruralidade*, procedendo-se á eliminação das (poucas) observações correspondentes a esta categoria quando for conveniente. Foi ainda considerada a variável *distrito*, com as seguintes categorias: Beja, Évora, Portalegre e Setúbal.

Foram realizadas análises univariadas e multivariadas aos dados e, de forma a estimar o risco de cancro da mama associado com as variáveis de interesse, foram calculados os valores das razões de chance (OR) e os $I.C._{95\%}$, com o auxílio do programa **R**, versão **2.9**.

3.4.2.2 - Resultados e sua discussão

A descrição detalhada das variáveis separada pelos dois grupos em estudo (casos e controlos) encontra-se na Tabela 8.

Variável	Grupo dos Casos	Grupo dos Controlos
Idade	Mínima:18 Máxima:88 Média:57,06 Mediana:59	Mínima:30 Máxima:81 Média:54,91 Mediana:55
Localidade[4]	Desconhecido:4 Évora:41 Beja:40 Estremoz:5 Montemor-o-Novo:5 Vidigueira:5 Outras:112	Évora:23 Montemor-o-Novo:5 Estremoz:5 Viana do Alentejo:5 Borba:4 Arraiolos:3 Outras:46
Concelho[5]	Desconhecido:4 Beja:47 Évora:46 Montemor-o-Novo:15 Moura:13 Serpa:10 Estremoz:7 Outros:74	Évora:25 Montemor-o-Novo:7 Viana do Alentejo:7 Elvas:6 Estremoz:6 Borba:4 Outros:36
Distrito	Desconhecido:4 Beja:93 Évora:99 Portalegre:11 Setúbal:5	Beja:15 Évora:65 Portalegre:11
Ruralidade	Desconhecido:4 Rural:42 Urbano:166	Rural:17 Urbano:74
TOTAL	212	91

Tabela 8: Resumo da distribuição dos casos e controlos segundo as variáveis de interesse para o estudo

Realizou-se uma breve análise estatística, com o intuito de proceder a uma comparação entre os grupos de casos e controlos quando cruzados com as variáveis em estudo, com recurso à folha de cálculo *MS Excel* (para obtenção de gráficos ilustrativos e mais aprazíveis de ambos os grupos em estudo) e ao software *R*, atrás mencionado.

No que respeita à variável *idade*, procedeu-se à construção dos diagramas de caixa e bigodes (Figura 28) para ambos os grupos em estudo.

[4] Considerando apenas as localidades que apresentam maior frequência de residência
[5] Considerando apenas os concelhos que apresentam maior frequência de residência

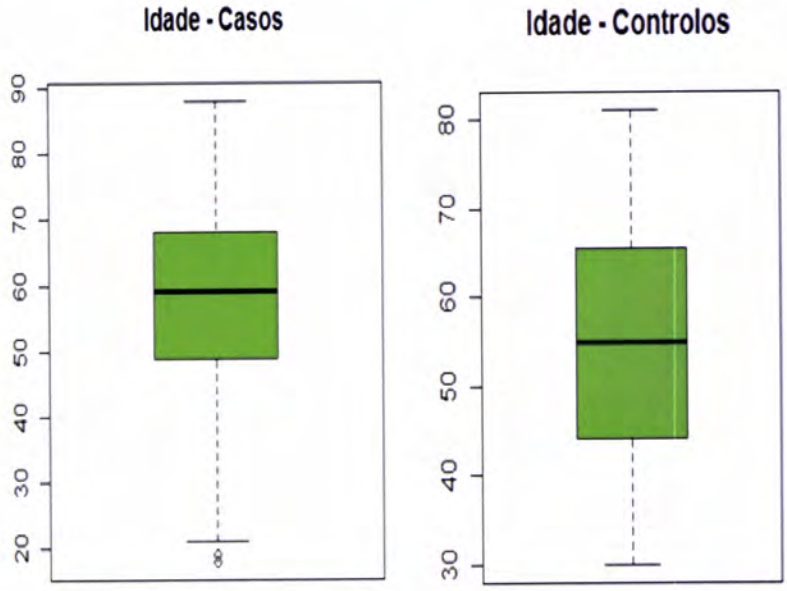


Figura 28:Diagrama de caixa e bigodes da idade em ambos os grupos de estudo

Classe Etária	CASOS Nº (%)	CONTROLOS Nº (%)
[15,25[	11 (5,19)	0 (0)
[25,35[	11 (5,19)	6 (6,59)
[35,45[	16 (7,55)	22 (24,18)
[45,55[	42(19,81)	17 (18,68)
[55,65[	55 (25,94)	20 (21,98)
[65,75[	50 (23,58)	17 (18,68)
[75,85[	24 (11,32)	9 (9,89)
[85,95[	3 (1,42)	0 (0)
TOTAL	212	91

Tabela 9: Distribuição por classe etária

No que respeita à *idade* dos indivíduos em estudo, tendo em conta os grupos dos casos e dos controlos, a distribuição da mesma por faixa etária é a apresentada nas Tabela 9 e Figura 29. É possível verificar que as faixas etárias em que existe maior efectivo de casos são as classes compreendidas entre os 35 anos inclusive e os 85 anos de idade, exclusive, o que também se verifica no caso do grupo dos controlos. De referir que 19.5% dos controlos se encontram na faixa etária entre os 15 e os 25 anos de idade, talvez porque as mulheres jovens se encontram cada vez mais consciencializadas para os problemas de saúde, e façam exames mais frequentemente.

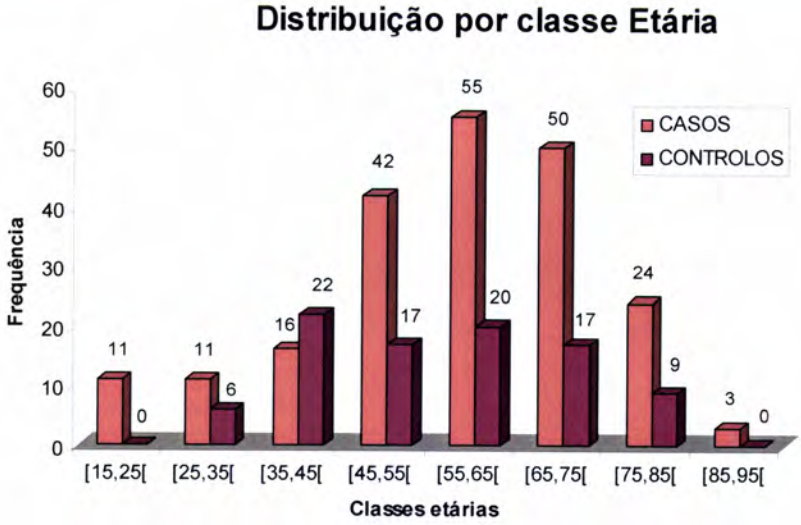


Figura 29: Histograma de distribuição dos grupos por faixa etária

No que respeita ao distrito de residência dos indivíduos em estudo por grupo (casos e controlos) a distribuição encontra-se representada graficamente na Figura 30, onde é possível constatar que a grande maioria, quer dos casos, quer dos controlos, são indivíduos que residem nos distritos de Évora e Beja.

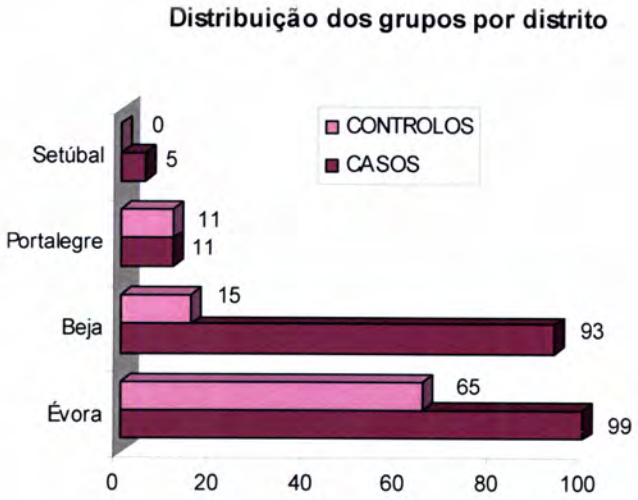


Figura 30: Distribuição dos grupos por distrito de residência

Por fim, quanto à ruralidade a Figura 31 sugere que, tendo em conta os indivíduos em estudo, existe um maior número, quer de casos, quer de controlos, que residem em zonas urbanas da região do Alentejo.

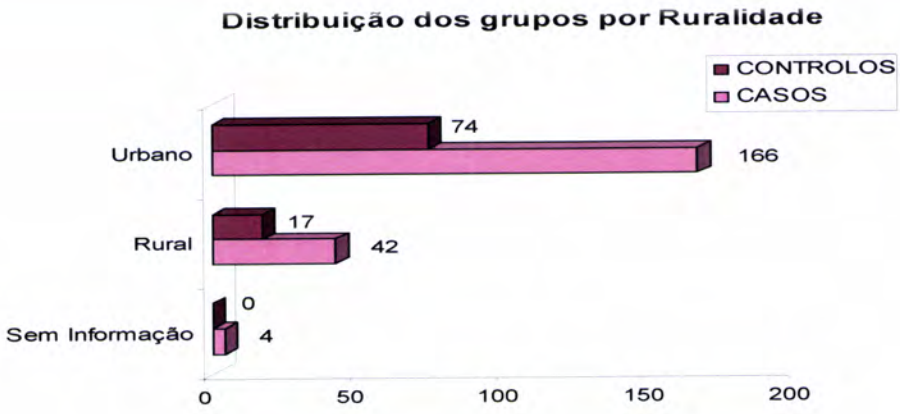


Figura 31:Distribuição dos grupos por Ruralidade

3.4.2.3 - Estudo da associação entre ruralidade e neoplasia da mama

Classe Etária	Grupo Casos (n=203[6])					Grupo Controlos (n=91)				
	Distrito			Ruralidade		Distrito			Ruralidade	
	Beja	Évora	Portalegre	Rural	Urbano	Beja	Évora	Portalegre	Rural	Urbano
[15,25[	7	4	0	1	10	0	0	0	0	0
[25,35[	7	1	3	0	11	0	4	2	0	6
[35,45[	5	10	1	2	14	5	11	6	3	19
[45,55[	24	13	2	8	31	2	15	0	2	15
[55,65[	19	31	5	10	45	4	14	2	5	15
[65,75[	22	24	0	13	33	2	14	1	3	14
[75,85[	8	16	0	6	18	2	7	0	4	5
[85,100[	1	0	0	1	0	0	0	0	0	0
D/I (idade)	P = 0,004287 (D)			P = 0,2213 (I)		P = 0,6342 (I)			P = 0,3018 (I)	

Tabela 10: Tabela de Contingência idadecat *versus* distrito *versus* ruralidade

D/I = Dependentes/Independentes

Quando são estudadas as relações de dependência entre as variáveis em estudo apresentadas na Tabela 10 verifica-se uma relação de dependência entre as variáveis distrito e idade no grupo dos casos, para um nível de significância de 5%.

[6] Não foram contabilizados nesta tabela os 4 casos de que não possuíamos informação acerca do distrito de residência nem os 5 casos residentes no distrito de Setúbal, devido ao reduzido número.

3.4.2.4 - Razão de Chances e Intervalos de Confiança

Após esta breve análise descritiva de ambos os grupos em estudo (casos e controlos) procede-se ao cálculo das *razões de chances* e do I.C. $_{95\%}$ (*ver secção 3.2*) com o objectivo de estimar o risco de cancro da mama associado com as variáveis de interesse.

Na Tabela 11 estão apresentadas os valores das razões de chances e os respectivos intervalos de confiança, para as variáveis de *Ruralidade*, *Distrito* e *Idade*.

Variável	Razão de Chances (OR)	I.C. $_{95\%}$
Ruralidade (Rural)	1.10	0.59 – 2.07
Distrito (Évora)	0.38	0.22 – 0.65
Distrito (Beja)	4.28	2.31 – 7.95
Distrito (Évora e Beja)	2.40	1 – 5.76
Idade (45 aos 75)	1.52	0.91 – 2.54
Idade (35 aos 45)	0.27	0.13 – 0.54

Tabela 11: OR com respectivos intervalos de confiança (95%)

A exposição ao ambiente rural parece elevar um pouco o risco de cancro da mama, uma vez que a chance de cancro da mama é 1.1 vezes maior para as pessoas expostas ao ambiente rural do que para as expostas ao ambiente urbano, ainda que esta diferença não se possa considerar significativa, dado que o valor 1 está incluído no intervalo de confiança a 95%. (OR =1.1; I.C.$_{95\%}$:0.59 – 2.07).

Pode ainda concluir-se que viver no distrito de Beja parece estar associado a maior risco deste tipo de cancro em, cerca de quatro vezes, de maneira estatisticamente significativa (OR=4.28; I.C.$_{95\%}$:2.31 – 7.95).

De salientar de que não se esperavam grandes diferenças entre a exposição ao ambiente rural/urbano uma vez que está aqui a ser considerada a região do Alentejo que de certa forma apesar de ser constituída por vilas e cidades, estas são pouco diferentes das zonas rurais uma vez que se encontram com um reduzido grau de desenvolvimento, indústria, poluição, etc., afinal este estudo trata uma região tipicamente alentejana.

No que se refere à idade, tendo em conta as classes etárias estudadas, verifica-se que o facto de se estar na faixa etária compreendida entre os 45 e os 75 anos de idade corresponde a um risco acrescido de cancro da mama em cerca de uma vez e meia, ainda que este facto não se possa considerar significativo (OR=1.52; I.C.$_{95\%}$: 0.91 – 2.54), facto que vem sustentar o estudo de *Garicochea, B. et al (2009).*

Capítulo 4

Os Modelos Lineares Generalizados

4.1 Introdução

Os modelos lineares generalizados (GLM) foram sistematizados pela primeira vez por *Nelder & Wedderburn (1972)* e constituem uma extensão dos modelos lineares clássicos. O objectivo principal no estudo dos GLM é analisar a influência que uma ou mais variáveis (explicativas ou covariáveis) têm sobre uma variável de interesse, designada por variável resposta (**Y**), através do estudo de um modelo de regressão que as relacione. A variável resposta segue uma distribuição da família exponencial, deixando de ter obrigatoriamente de seguir uma distribuição normal. Vários modelos estatísticos já anteriormente utilizados cabem na classe dos modelos lineares generalizados, entre eles o modelo de regressão logística que será abordado de forma mais pormenorizada posteriormente.

A situação experimental de interesse ao longo deste trabalho é aquela em que há uma variável aleatória dependente ou resposta Y e um vector $\boldsymbol{x} = (x_1, \ldots, x_k)^T$ com k variáveis explicativas, ou covariáveis. A variável resposta Y pode ser contínua, discreta ou categórica/quantitativa. As variáveis explicativas podem ser também de natureza contínua, discreta, qualitativas de natureza ordinal ou não ou categórica/quantitativa.
Então temos os dados na forma

$$(Y_i,\ \boldsymbol{x_i}, i = 1, \ldots n) \qquad \textbf{(4.1)}$$

sendo Y_i as componentes do vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$.

Podemos ainda representar os dados em (***4.1***) na forma matricial

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

4.1.1 Modelo Linear Clássico

Como referimos, os modelos lineares generalizados são uma extensão dos modelos lineares clássicos, pelo que a forma destes últimos é um ponto de partida apropriado para a abordagem dos primeiros. Este modelo é definido na sua forma matricial por

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

onde,

- $\boldsymbol{X}$ é a matriz das variáveis explicativas de dimensão $n \times (k+1)$, cuja primeira coluna é um vector unitário e as restantes correspondem aos valores observados de cada covariável fixa;
- $\boldsymbol{\beta} = (\beta_0,\ldots,\beta_k)$, é o vector de parâmetros cujos valores são geralmente desconhecidos e pretendemos estimá-los a partir dos dados;
- $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)$, é o vector dos erros aleatórios, que capturam a variação da variável resposta ($\mathbf{Y}$) não explicada pelas covariáveis; assume-se para os erros uma distribuição $N_n\ (\mathbf{0},\sigma^2\boldsymbol{I_n})$ (sendo $\boldsymbol{I_n}$ a matriz identidade de ordem n);
- $\mathbf{Y} = (Y_1,\ldots,Y_n)$,vector dos valores observados da variável resposta, com distribuição normal multivariada. Os Y_i´s são variáveis aleatórias normais independentes, de variância constante e igual a σ^2. Temos que $E\,[\boldsymbol{Y} \mid \boldsymbol{X}] = \boldsymbol{\mu}$, com $\boldsymbol{\mu} = (\mu_{1,\ldots,}\mu_n)$ e $\boldsymbol{\mu} = \boldsymbol{X\beta}$, isto é, o valor esperado da variável resposta é uma função linear das covariáveis.

Nos modelos lineares generalizados a distribuição da variável resposta considerada não tem de ser normal, mas sim qualquer distribuição da família exponencial, além disso, a função que relaciona o valor esperado da variável resposta e o vector de covariáveis ($g(\boldsymbol{\mu}) = \boldsymbol{X\beta}$) pode ser qualquer função diferenciável.

4.1.2 A Família Exponencial

Definição 1 (Família exponencial)

Diz-se que uma variável aleatória Y tem distribuição pertencente à família exponencial se a sua função densidade probabilidade (f.d.p.) ou função massa de probabilidade (f.m.p.) se puder escrever na forma

$$f(y|\theta,\phi) = \exp\{\tfrac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\} \qquad \textbf{(4.2)}$$

onde θ e ϕ são parâmetros escalares, a(.), b(.) e c(.) são funções reais conhecidas.

O parâmetro ϕ designa-se por parâmetro de dispersão. A função $a(\phi)$ toma, frequentemente, a forma $a(\phi) = \frac{\phi}{w}$, para um peso conhecido w. θ é a forma canónica do parâmetro de localização e ϕ, muitas vezes é conhecido.

4.1.2.1 Valor médio e variância

Após a definição da família de distribuições possíveis para **Y**, é importante encontrar as expressões gerais para o seu valor médio e variância.

Seja a log-verosimilhança $\ell(\theta, \phi; y) \equiv ln\,(f(y|\theta, \phi))$. Então,

$$\ell(\theta, \phi; y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi).$$

Definindo a função score como sendo $S(\theta) = \frac{\partial \ell(\theta;\phi,Y)}{\partial \theta}$, para famílias regulares sabe-se que:

- $E\,[S(\theta)] = 0$
- $E\,[S^2(\theta)] = E\left[\left(\frac{\partial \ell\,(\theta;\,\phi,Y)}{\partial \theta}\right)^2\right] = -\,E\left[\left(\frac{\partial^2 \ell(\theta;\phi,Y)}{\partial \theta^2}\right)\right]$

Assim, sendo $S(\theta) = \frac{Y - b'(\theta)}{a(\phi)}$, da **primeira relação** sai que:

$$E[S(\theta)] = 0 \Leftrightarrow E\left[\frac{Y - b'(\theta)}{a(\phi)}\right] = 0 \Leftrightarrow b'(\theta) = \mu = E[Y] \qquad \textbf{(4.3)}$$

e da segunda vem $-E\left[\left(\frac{Y - b'(\theta)}{a(\phi)}\right)^2\right] = E\left[\frac{-b''(\theta)}{a\,(\phi)}\right] \Leftrightarrow + \frac{E\,[(Y - E[Y])^2]}{a^2\,(\phi)} = + \frac{b''(\theta)}{a(\phi)}$

$$\Leftrightarrow \frac{Var(Y)}{a^2(\phi)} = \frac{b''(\theta)}{a(\phi)} \Leftrightarrow Var(Y) = b''(\theta)a(\phi) \qquad \textbf{(4.4)}$$

em que, $b'\,(\theta)$ e $b''(\theta)$ são a primeira e a segunda derivadas de b em ordem a θ. Como se pode verificar a variância é o produto de duas funções, uma depende apenas do parâmetro θ e portanto de μ e à qual se chama função de variância e se designa por $V\,(\mu)$ - $b''(\theta) = V(\mu)$ -, enquanto que a outra depende do parâmetro de dispersão.

Nas subsecções seguintes, apresentamos dois exemplos de distribuições da família exponencial que mais nos interessam para análise e tratamento da base de dados em estudo.

4.1.2.2 A distribuição normal

À distribuição normal corresponde a seguinte *f.d.p*

$$f(y|\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}=exp\frac{1}{\sigma^2}\left(y\mu-\frac{\mu^2}{2}\right)-\frac{1}{2}\left[\frac{y^2}{\sigma^2}+\ln(2\pi\sigma^2)\right],$$

com $y \in \mathbb{R}$, $\mu \in \mathbb{R}$ e $\sigma > 0$.

Assim verifica-se que esta distribuição pertence à família exponencial com parâmetro canónico $\theta=\mu$, $a(\phi)=\sigma^2=\frac{\phi}{\omega}$, com $\omega=1, \phi=\sigma^2$, $b(\theta)=\frac{\mu^2}{2}$, $c(y,\phi)=-\frac{1}{2}\left[\frac{y^2}{\sigma^2}+\ln(2\pi\sigma^2)\right]$,

de onde sai que $E[Y]=b'(\theta)=\theta=\mu$ e $Var[Y]=b''(\theta)a(\phi)=\phi=\sigma^2$.

4.1.2.3 A distribuição binomial

À distribuição binomial corresponde a seguinte a *f.m.p.*

$$f(y|\pi)=\binom{m}{ym}\pi^{ym}(1-\pi)^{m-ym}=exp\left(ymln\pi+m(1-y)ln(1-\pi)+ln\binom{m}{ym}\right)$$

$$=exp\left(m\left(y(ln(\pi)-ln(1-\pi))+ln(1-\pi)\right)\right)+ln\binom{m}{ym}=exp\left(m\left(y\theta-ln(1+e^{\theta})\right)+ln\binom{m}{ym}\right),$$ com y $\in \{0,\frac{1}{m},\frac{2}{m},\ldots,1\}$

Verifica-se assim que a distribuição $B(m,\pi)/m$ pertence à família exponencial, com

$\theta=ln\left(\frac{\pi}{1-\pi}\right)$, $a(\phi)=\frac{\phi}{\omega}=\frac{1}{m}$, com $\phi=1$ e $\omega=m$, $b(\theta)=ln(1+e^{\theta})$ e $c(y,\phi)=ln\binom{m}{ym}$,

de onde sai que

$$E[Y]=b'(\theta)=\frac{e^{\theta}}{1+e^{\theta}}=\frac{\frac{\pi}{1-\pi}}{\frac{1}{1-\pi}}=\pi \text{ e } Var(Y)=b''(\theta)a(\phi)=\frac{e^{\theta}}{(1+e^{\theta})^2}\times\frac{1}{m}=\frac{\pi(1-\pi)}{m}.$$

Nota: repare-se que $\theta=ln\left(\frac{\pi}{1-\pi}\right)\Leftrightarrow e^{\theta}=\frac{\pi}{1-\pi}\Leftrightarrow 1+e^{\theta}=\frac{1}{1-\pi}\Leftrightarrow ln(1+e^{\theta})=ln\left(\frac{1}{1-\pi}\right)$

$$\Leftrightarrow -ln(1+e^{\theta})=ln(1-\pi)$$

Existem outras distribuições que pertencem a esta família, como por exemplo a de Bernoulli, a Poisson, a Binomial Negativa, a Exponencial e a Gama, as quais não serão aqui abordadas uma vez que não foram aplicadas na análise e tratamento da base de dados em estudo. Para uma consulta mais detalhada, ver *McCullagh, P. e Nelder, J. (1989)* ou *Turkmann e Silva (2000).*

4.1.3 Descrição dos Modelos Lineares Generalizados

Os modelos lineares generalizados são uma extensão dos modelos de regressão simples, no sentido de abranger situações cuja resposta é não-normal e permitir uma modelação linear nas covariáveis não da média das respostas mas numa função bem comportada da média, pelo que estes modelos são caracterizados através de três componentes:

- *Componente aleatória*, que define a distribuição de probabilidade associada à variável resposta;
- *Componente estrutural ou sistemática*, que especifica a função linear das covariáveis, que é utilizada como predictor;
- *Função de ligação*, descreve a relação funcional entre a componente sistemática e o valor esperado para a componente aleatória.

Componente aleatória

Como referido, a componente aleatória define a distribuição de probabilidade associada à variável resposta. Assumimos que a distribuição de Y pertence à família exponencial tendo a forma da equação ***(4.2)***, com $E[Y|X] = \mu = b'(\theta)$ como já foi referido na secção ***4.1.2***, assumindo-se independência entre os indivíduos.

Componente estrutural ou sistemática

Admitimos que $x_{i1}, x_{i2},\ldots,x_{ik}$, representam os valores de k variáveis explicativas referentes ao i-ésimo indivíduo. A componente sistemática de um GLM fica definida pelo vector coluna $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$, função das k variáveis explicativas através da relação linear

$$\eta_i = \sum_{j=0}^{k} \beta_j x_{ij},\ i = 1, \ldots, n \qquad (4.5)$$

ou escrito na forma matricial

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

sendo $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ preditor linear, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^T$ os parâmetros do modelo e $\mathbf{X}$ a matriz do modelo de ordem $n \times (k+1)$.

Função de ligação

A função de ligação é a terceira componente dos GLM que estabelece a ligação entre a componente aleatória e a componente sistemática, isto é, associa o valor esperado de Y, μ, ao preditor através de:

$$\eta = g(\mu)$$

onde g é uma função monótona e diferenciável. Assim, nos GLM, obtemos a função de ligação entre o valor esperado μ_i e as variáveis explicativas, a partir da seguinte equação:

$$g(\mu_i) = \sum_{j=0}^{k} \beta_j x_{ij}.$$

A escolha da função de ligação depende do tipo de variável resposta. Nos modelos lineares clássicos, a função $g(\mu) = \mu$, conduz à ligação identidade $\mu = \eta$.

Há especial interesse quando o preditor linear coincide com o parâmetro canónico, $\theta = \eta$. A função de ligação correspondente diz-se ligação canónica. Para a ligação canónica $g(\mu_i) = \theta_i$, donde $\theta_i = \sum_{j=0}^{k} \beta_j x_{ij}$.

No caso da distribuição binomial temos $0 < \mu = \pi < 1$ e, daí, é conveniente obter uma função de ligação que transforme o intervalo [0,1] em toda a recta real. A função logit, $\log\frac{\pi}{1-\pi}$ é uma das funções de ligação mais utilizadas nestas situações experimentais. Existem ainda outras funções de ligação possíveis como sendo a probit e a complementar log-log conforme descrito em *McCullagh, P. & Nelder, J. (1989).*

4.1.4 Inferência nos Modelos Lineares Generalizados

O estudo e análise de um modelo linear generalizado compreendem, tal como em qualquer outro estudo estatístico e segundo *Turkman e Silva (2000)* as seguintes fases:

- **Formulação dos modelos** – nesta fase procede-se à escolha da distribuição da variável resposta, com base na análise preliminar dos dados; de seguida escolhem-se as covariáveis e formula-se a matriz de especificação; e por fim, procede-se à escolha e selecção da função de ligação.
- **Ajustamento do modelo (ou modelos)** – engloba a estimação dos parâmetros do modelo.
- **Selecção e validação do(s) modelo(s)** – pretende-se encontrar submodelos com um número adequado de parâmetros, detectar discrepâncias entre os dados e os valores preditos, verificar a existência de outliers e/ou observações influentes. Nesta fase devem ponderar-se 3 factores: adequabilidade, parcimónia e interpretação.

Após a formulação adequada do modelo tendo em conta a escolha acertada da distribuição da variável resposta, das covariáveis e da função de ligação, procede-se ao ajustamento do modelo.
Para ajustar o modelo há que estimar os parâmetros $\beta_j (j = 0, \ldots, k)$, estimação essa feita utilizando-se o método da máxima verosimilhança. No caso de a distribuição ter parâmetro de escala desconhecido, é também necessário estimar o parâmetro ϕ recorrendo ao método dos momentos.

4.1.4.1 Estimação do $\boldsymbol{\beta}$

A estimativa do método da máxima verosimilhança é o valor do parâmetro que maximiza o logaritmo da função de verosimilhança.

Para n observações independentes, a função de verosimilhança de dados com distribuição na família exponencial é:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi) = \prod_{i=1}^{n} e^{\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right)} = e^{\sum_{i=1}^{n}\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right)},$$

vindo o logaritmo da verosimilhança:

$$ln\, L(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right),$$

com $\ell_i = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)$, i = 1,…,n,

e as equações de verosimilhança dadas por:

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = 0, \text{ j=0,…,k.}$$

Estas equações obtêm-se calculando-se

$$\frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_j} = \frac{\partial l_i(\theta_i)}{\partial \theta_i}\frac{\partial \theta_i(\mu_i)}{\partial \mu_i}\frac{\partial \mu_i(\eta_i)}{\partial \eta_i}\frac{\partial \eta_i(\boldsymbol{\beta})}{\partial \beta_j}. \qquad \textbf{(4.6)}$$

Como, $\mu_i = E[Y_i] = b'(\theta_i)$, $Var[Y_i] = b''(\theta_i)a_i(\phi)$ e $\eta_i = \sum_{j=0}^{k}\beta_j x_{ij}$ tem-se que

$$\frac{\partial l_i(\theta_i)}{\partial \theta_i} = \frac{y_i - \mu_i}{a_i(\phi)},$$

$$\frac{\partial \mu_i(\theta_i)}{\partial \theta_i} = b''(\theta_i) = \frac{Var(Y_i)}{a_{i(\phi)}} \Leftrightarrow \frac{\partial \theta_i(\mu_i)}{\partial \mu_i} = \frac{a_i(\phi)}{Var(Y_i)}$$

$$\frac{\partial \eta_i(\boldsymbol{\beta})}{\partial \beta_j} = x_{ij}$$

enquanto que $\frac{\partial \mu_i(\eta_i)}{\partial \eta_i}$ depende da função de ligação $\eta_i = g(\mu_i)$.

Substituindo as expressões acima em ***(4.6)***, obtemos:

$$\frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_j} = \frac{y_i - \mu_i}{a_i(\phi)} \frac{a_i(\phi)}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij}$$

com $z_i = \frac{y_i - \mu_i}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)$, podemos escrever

$$\frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_j} = z_i x_{ij}, j = 0, \ldots, k$$

e as equações de verosimilhança para $\boldsymbol{\beta}$ são dadas por:

$$S_j = \sum_{i=1}^{n} z_i x_{ij} = 0, j = 0, \ldots, k, \qquad \textbf{(4.7)}$$

são em geral funções não lineares de $\boldsymbol{\beta}$. S_j designa-se por *score total* do parâmetro β_j. O vector $S = (S_0, \ldots, S_k)^T$ denomina-se por função score.

Métodos iterativos

As equações de máxima verosimilhança dadas em ***(4.7)*** não têm solução analítica pelo que é necessário recorrer a métodos iterativos. Um método iterativo usado para obter as estimativas de máxima verosimilhança nos modelos lineares generalizados é o método de ***Newton-Raphson***. Aplicando este método, a $(p+1)$-ésima aproximação da estimativa de $\boldsymbol{\beta}$ é dada por

$$\widehat{\boldsymbol{\beta}}^{(p+1)} = \widehat{\boldsymbol{\beta}}^{(p)} - \left[H\left(\widehat{\boldsymbol{\beta}}^{(p)}\right)\right]^{-1} S\left(\widehat{\boldsymbol{\beta}}^{(p)}\right) \qquad \textbf{(4.8)}$$

onde $H\left(\widehat{\boldsymbol{\beta}}^{(p)}\right)$ é a matriz Hessiana, quadrada de ordem $(k+1)$, calculada em $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(p)}$, com elementos:

$$h_{jt} = \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_t}, \quad j, t = 0, 1, \ldots, k,$$

e $S\left(\widehat{\boldsymbol{\beta}}^{(p)}\right)$ é um vector coluna de ordem $(k+1)$ vector dos scores, com elementos $S_j = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j}$, calculado em $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(p)}$.

Um método alternativo e por vezes mais simples do que o método de *Newton-Raphson*, é o ***método dos scores de Fisher***. Este método envolve a substituição da matriz das segundas derivadas da equação ***(4.8)*** pela matriz dos respectivos valores esperados,

$$E\left[\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_t}\right], j, t = 0,1, \ldots, k$$

Estes valores têm uma relação directa com a matriz de informação de *Fisher*, matriz de covariância do vector dos scores,

$$\mathcal{I} = \mathrm{E}[SS^T]$$

com elementos,

$$\mathcal{I}_{jt} = \mathrm{E}[S_j S_t] = \mathrm{E}\left[\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j}\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_t}\right] = -\mathrm{E}\left[\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_t}\right] \qquad \textbf{(4.9)}$$

Aplicando o método dos *scores de Fisher*, a equação ***(4.8)*** é substituída por,

$$\widehat{\boldsymbol{\beta}}^{(p+1)} = \widehat{\boldsymbol{\beta}}^{(p)} + \left[\mathcal{I}\left(\widehat{\boldsymbol{\beta}}^{(p)}\right)\right]^{-1} S\left(\widehat{\boldsymbol{\beta}}^{(p)}\right),$$

onde, $\mathcal{I}\left(\widehat{\boldsymbol{\beta}}^{(p)}\right)$ é a matriz de informação de *Fisher* calculada a partir de $\widehat{\boldsymbol{\beta}}^{(p)}$.
A expressão anterior pode ser re-escrita como:

$$\left[\mathcal{I}\left(\widehat{\boldsymbol{\beta}}^{(p)}\right)\right]\widehat{\boldsymbol{\beta}}^{(p+1)} = \left[\mathcal{I}\left(\widehat{\boldsymbol{\beta}}^{(p)}\right)\right]\widehat{\boldsymbol{\beta}}^{(p)} + S\left(\widehat{\boldsymbol{\beta}}^{(p)}\right) \qquad \textbf{(4.10)}$$

No caso dos modelos lineares generalizados, a contribuição de cada elemento Y_i para $\mathcal{I}_{jt}$ é dada por,

$$\mathrm{E}\left[\frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j}\frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_t}\right] = \mathrm{E}\left[\frac{(Y_i-\mu_i)^2 x_{ij} x_{it}}{[var(Y_i)]^2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right] = \frac{x_{ij}x_{it}}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2,$$

logo, os elementos da matriz de informação são dados por:

$$\mathcal{I}_{jt} = \sum_{i=1}^{n} \mathrm{E}\left[\frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j}\frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_t}\right] = \sum_{i=1}^{n} \frac{x_{ij}x_{it}}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \qquad \textbf{(4.11)}$$

A matriz de informação, $\mathcal{I}$, pode então ser representada matricialmente por:

$$\mathcal{I} = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \qquad \textbf{(4.12)}$$

onde, $\mathbf{W}$ é uma matriz diagonal de ordem n, de elementos não nulos dados por:

$$\overline{W}_i = \frac{1}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \frac{w_i\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2}{\phi V(\mu_i)}. \qquad \textit{(4.13)}$$

Observando a expressão ***(4.7)***, o lado direito da equação ***(4.10)*** é um vector coluna com elemento genérico de ordem t dado por,

$$\Sigma_{j=0}^{k}\left[\Sigma_{i=1}^{n}\frac{x_{ij}x_{it}}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right]\beta_j^{(p)} + \Sigma_{i=1}^{n}\frac{(y_i-\mu_i)x_{it}}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right),$$

na forma matricial fica e dado ***(4.12)***:

$$\left[\mathcal{I}\left(\widehat{\boldsymbol{\beta}}^{(p)}\right)\right]\widehat{\boldsymbol{\beta}}^{(p)} + S\left(\widehat{\boldsymbol{\beta}}^{(p)}\right) = \boldsymbol{X}^T\boldsymbol{W}^{(p)}\boldsymbol{u}^{(p)} \qquad \textit{(4.14)}$$

onde, $\boldsymbol{u}^{(p)}$ é o vector coluna (de ordem n) com elemento genérico,

$$u_i^{(p)} = \Sigma_{j=o}^{k} x_{ij}\beta_j^{(p)} + \left(y_i - \mu_i^{(p)}\right)\left(\frac{\partial \eta_i^{(p)}}{\partial \mu_i^{(p)}}\right) = \eta_i^{(p)} + \left(y_i - \mu_i^{(p)}\right)\left(\frac{\partial \eta_i^{(p)}}{\partial \mu_i^{(p)}}\right) \qquad \textit{(4.15)}$$

com η_i, μ_i e $\frac{\partial \eta_i}{\partial \mu_i}$ calculados no ponto $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(p)}$.

Atendendo às equações ***(4.12)*** e ***(4.14)***, a equação do método dos *scores de Fisher* escreve-se da seguinte forma:

$$\left(\boldsymbol{X}^T\boldsymbol{W}^{(p)}\boldsymbol{X}\right)\widehat{\boldsymbol{\beta}}^{(p+1)} = \boldsymbol{X}^T\boldsymbol{W}^{(p)}\boldsymbol{u}^{(p)} \qquad \textit{(4.16)}$$

Assim, a expressão final da estimativa de $\boldsymbol{\beta}$ na $(p+1)$ −ésima iteração, ou seja, a solução da equação ***(4.16)*** é dada por:

$$\widehat{\boldsymbol{\beta}}^{(p+1)} = \left(\boldsymbol{X}^T\boldsymbol{W}^{(p)}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(p)}\boldsymbol{u}^{(p)} \qquad \textit{(4.17)}$$

O processo iterativo de cálculo das estimativas de máxima verosimilhança de $\boldsymbol{\beta}$ pode ser resumido em duas etapas:

1) Dado $\widehat{\boldsymbol{\beta}}^{(p)}$ (com p a iniciar-se em zero), calcula-se $\boldsymbol{u}^{(p)}$ e $\boldsymbol{W}^{(p)}$ aplicando as equações ***(4.15)*** e ***(4.13)***, respectivamente.
2) A nova iteração, $\widehat{\boldsymbol{\beta}}^{(p+1)}$, é obtido aplicando a expressão ***(4.17)***.

As iterações param quando for atingido o critério definido, por exemplo, se

$$\frac{\|\widehat{\boldsymbol{\beta}}^{(p+1)}-\widehat{\boldsymbol{\beta}}^{(p)}\|}{\|\widehat{\boldsymbol{\beta}}^{(p)}\|} \leq \varepsilon \qquad \textbf{(4.18)}$$

com $\varepsilon>0$, previamente definido.

É importante salientar que, apesar do elemento genérico de **W** conter o parâmetro ϕ, ele por simplificação não entra no cálculo de $\widehat{\boldsymbol{\beta}}^{(p+1)}$. Sem perda de generalidade pode-se fazer $\phi = 1$ o que quer dizer que, o conhecimento ou não do parâmetro ϕ é irrelevante no cálculo de $\widehat{\boldsymbol{\beta}}$.

4.1.4.2 Estimação do parâmetro de dispersão

O parâmetro de dispersão pode ser estimado pelo método de máxima verosimilhança mas, dada a complexidade de cálculo computacional, não compensa o uso deste método. Existe um método mais simples que é baseado na distribuição de amostragem da estatística de *Pearson* generalizada.

No modelo de regressão linear estima-se σ^2 por $\hat{\sigma}^2 = \frac{SSE}{n-p}$, em que $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ representa a soma dos quadrados dos resíduos.

Nos GLM o parâmetro σ^2 estima-se usando a seguinte expressão

$$\hat{\sigma}^2 = \hat{\phi} = \frac{D(y_i,\hat{\mu}_i)}{n-p},$$

em que $D(y_i, \hat{\mu}_i)$ é a função desvio ou *deviance* e é descrita na secção ***4.1.7.1.***
Para uma informação mais detalhada sobre este assunto ver, por exemplo, *McCullagh, P. & Nelder, J. (1989).*

4.1.5 Testes de hipóteses

De uma forma geral, os testes de hipóteses sobre o vector **β** podem ser estabelecidos da seguinte forma:

$$H_0 : C\boldsymbol{\beta} = \boldsymbol{\xi} \quad \textit{versus} \quad H_1 : C\boldsymbol{\beta} \neq \boldsymbol{\xi}, \qquad \textbf{(4.19)}$$

onde C é uma matriz $q \times (k+1)$, $(q \leq k+1)$ com característica q e $\boldsymbol{\xi}$ é um vector de dimensão q previamente definido.
De uma maneira geral interessa testar a hipótese de uma qualquer covariável (x_j) ser irrelevante para o modelo, e daí que ***(4.19)*** possa ser escrita na forma

$$H_0 : \beta_j = 0 \qquad versus \qquad H_0 : \beta_j \neq 0.$$

Assim, a matriz C é dada por C = (0,…,1,0,…,0), ocupando 1 a j-ésima posição e $\boldsymbol{\xi} = \mathbf{0}$.

No caso em que se pretende testar a hipótese de que r vectores de β ($r < k + 1$) são nulos, ou seja,

$$H_0 : \boldsymbol{\beta}_r = \mathbf{0} \qquad \text{vs} \qquad H_0 : \boldsymbol{\beta}_r \neq \mathbf{0},$$

a matriz C = $(I_r, \mathbf{0}_{r\times(k+1-r)})$, onde I_r é a matriz identidade de ordem r, $\mathbf{0}_{r\times(k+1-r)}$ a matriz $r \times (k + 1 - r)$ de zeros e $\boldsymbol{\xi} = \mathbf{0}_r$, onde $\mathbf{0}_r$ é o vector nulo de dimensão r.

Estas hipóteses correspondem a testar submodelos do modelo saturado (modelo que contém n parâmetros linearmente independentes, cuja matriz do modelo é uma matriz identidade $n \times n$), e são úteis na selecção de variáveis explicativas, como veremos posteriormente.
Habitualmente utilizam-se três testes diferentes, baseados em três estatísticas distintas, que construímos á custa das distribuições assintóticas dos estimadores de máxima verosimilhança de $\boldsymbol{\beta}$ ou de funções suas derivadas.

4.1.5.1 Teste de Wald

Seja $\widehat{\boldsymbol{\beta}}$ o estimador de máxima verosimilhança de $\boldsymbol{\beta}$, que tem distribuição assintótica $N_p(\boldsymbol{\beta}, \mathcal{I}^{-1}(\widehat{\boldsymbol{\beta}}))$, sendo $\mathcal{I}^{-1}$a inversa da matriz de Informação de *Fisher*. O estimador de $C\boldsymbol{\beta}$ é $\widehat{C\boldsymbol{\beta}}$, e podemos afirmar que:

$$\widehat{C\boldsymbol{\beta}} \sim N_q(< C\boldsymbol{\beta}, C\mathcal{I}^{-1}(\widehat{\boldsymbol{\beta}})C^T)$$

Se pretendermos testar ***(4.19)***, a estatística de Wald é definida por

$$\mathcal{W} = (C\widehat{\boldsymbol{\beta}} - \boldsymbol{\xi})^T [CI^{-1}(\widehat{\boldsymbol{\beta}})C^T]^{-1} (C\widehat{\boldsymbol{\beta}} - \boldsymbol{\xi}),$$

e tem, sob H_0, a distribuição assintótica de um χ^2 com q graus de liberdade, χ^2_q. Assim H_0 é rejeitado, a um nível de significância α, se o valor observado da estatística de *Wald* for superior ao quantil de probabilidade 1–α de um χ^2_q.

4.1.5.2 Teste da Razão de Verosimilhança

A estatística da razão de verosimilhança ou estatística de *Wilks* é definida por

$$Q = -2\log\frac{max_{H_0}L(\boldsymbol{\beta})}{max_{H_0 \cup H_1}L(\beta)} = -2\{\ell(\widehat{\boldsymbol{\beta}}_0) - \ell(\widehat{\boldsymbol{\beta}}_1)\},$$

onde $\widehat{\boldsymbol{\beta}}_0$ e $\widehat{\boldsymbol{\beta}}_1$ são os estimadores de máxima verosimilhança de $\boldsymbol{\beta}$ sob H_0 e $H_0 \cup H_1$, respectivamente.

O teorema de *Wilks* estabelece que, sob certas condições de regularidade e sob H_0, Q tem distribuição assintótica de um χ^2, em que o número de graus de liberdade é igual à diferença entre o número de parâmetros a estimar sob $H_0 \cup H_1$ $(k+1)$ e o número de parâmetros a estimar sob H_0 $(k+1-q)$.

Assim H_0:$C\boldsymbol{\beta} = \boldsymbol{\xi}$ é rejeitada a favor de H_1 :$C\boldsymbol{\beta} \neq \boldsymbol{\xi}$, a um nível de significância α, se o valor observado da estatística de *Wilks* (Q) for superior ao quantil de probabilidade $1-\alpha$ de um χ^2_q.

4.1.5.3 Estatística *Score de Rao*

A estatística de *Rao* ou estatística *score* é definida por

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

onde $\tilde{\boldsymbol{\beta}}$ é o estimador de máxima verosimilhança de $\boldsymbol{\beta}$ sujeito à restrição imposta pela hipótese nula $C\boldsymbol{\beta} = \boldsymbol{\xi}$. Sob a hipótese (***4.19***) S tem distribuição assintótica de um χ^2 com q graus de liberdade.

A ideia subjacente a este teste é a de que se $\widehat{\boldsymbol{\beta}}$ é o estimador de máxima verosimilhança de $\boldsymbol{\beta}$, $S(\widehat{\boldsymbol{\beta}}) = 0$. Se substituirmos $\widehat{\boldsymbol{\beta}}$ pelo estimador de máxima verosimilhança sob H_0, isto é por $\tilde{\boldsymbol{\beta}}$, $S(\tilde{\boldsymbol{\beta}})$ deverá ser significativamente diferente de zero, se H_0 não for verdadeira. Então concluímos que S mede a diferença entre $S(\tilde{\boldsymbol{\beta}})$ e zero.

Tal como nos outros dois testes, usando a estatística *Score de Rao*, rejeita-se H_0 a favor de H_1, a um nível de significância α, se o valor observado de S for superior ao quantil de probabilidade $1-\alpha$ de um χ^2_q.

Comparação:

Segundo *Turkman e Silva (2000)*, de uma forma abreviada podemos dizer que:

- O teste de *Wald* não requer um esforço computacional tão elevado como o do teste da razão de verosimilhança. Este teste tem grande aceitação na análise de dados binomiais;
- O teste da razão de verosimilhança é considerado como o mais correcto e eficaz. É habitualmente utilizado para comparar modelos encaixados, em que o modelo dado em H_0 é um submodelo do modelo dado em H_1;

- A estatística *Score de Rao* é o menos exigente em termos computacionais, já que não necessita que se calcule $\widehat{\boldsymbol{\beta}}$ mas apenas $\widetilde{\boldsymbol{\beta}}$.

Estas três estatísticas distintas têm uma igual distribuição assimptótica, para realizar testes de hipóteses sobre combinações lineares das componentes de $\boldsymbol{\beta}$. Os resultados obtidos pelas estatísticas nem sempre coincidem, embora para n suficientemente grande tendem a obter resultados semelhantes.

Segundo *Fahrmeir e Tutz (1994)*, é preferível aplicar a estatística da razão de verosimilhança se o número de variáveis explicativas for pequeno e as amostras tiverem uma dimensão moderada. Quando as amostras são de dimensão elevada, aplicam-se os testes de *Wald* e de *Score de Rao*, por serem mais fáceis de calcular.

4.1.6 Selecção de Modelos

Na fase da selecção de um modelo, assume-se que a escolha da distribuição da variável resposta e da função ligação é adequada. Nesta fase, é importante escolher "o melhor modelo", ponderando três factores: bom ajustamento, parcimónia e interpretação. *Turkman e Silva (2000)* referem que: "Um bom modelo é aquele que consegue atingir o equilíbrio entre estes três factores".

No processo de selecção, utilizam-se, frequentemente, dois modelos de referência:

- **Modelo completo ou saturado**

Este modelo contém n parâmetros $\mu_1,\ldots,\mu_n$, linearmente independentes, cuja matriz do modelo é uma matriz identidade $n \times n$. Este modelo atribui toda a variação dos dados à componente sistemática. O modelo saturado serve de referência para medir a discrepância de um modelo intermédio com $k+1$ parâmetros.

- **Modelo nulo**

Este modelo é o mais simples por ter apenas um único parâmetro. Neste modelo assume-se que todas as variáveis Y_i têm o mesmo valor médio μ, neste caso, atribui toda a variação dos dados à componente aleatória. A matriz do modelo corresponde a um vector coluna unitário.

Há então necessidade de estabelecer uma estratégia para a selecção do melhor modelo, ou dos melhores, já que, raramente, se pode falar na existência de um único "melhor modelo".

Seja um submodelo M_1 de um modelo M, correspondendo a um subvector $\boldsymbol{\beta}_1$do vector $\boldsymbol{\beta}$ de dimensão $k+1$ do modelo M.

Então podemos identificar uma partição de $\boldsymbol{\beta}$ em $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^T$. Assim a adequabilidade de um submodelo pode ser testada como:

$$H_0: \boldsymbol{\beta}_2 = \mathbf{0} \; versus \; H_1: \boldsymbol{\beta}_2 \neq \mathbf{0}$$

Esta hipótese pode ser testada aplicando a metodologia da secção ***(4.1.6)***. Designemos $S(\boldsymbol{\beta})$, $\mathcal{I}(\boldsymbol{\beta})$ e $\mathcal{A}(\boldsymbol{\beta})$, a função score, a matriz de informação de *Fisher* e a sua inversa, respectivamente. A partir da partição relativa a $\boldsymbol{\beta}$, temos as seguintes matrizes:

$$S = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \quad \mathcal{I} = \begin{bmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{12}^T & \mathcal{I}_{22} \end{bmatrix} \quad \mathcal{A} = \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A} \\ \mathcal{A}_{12}^T & \mathcal{A}_{22} \end{bmatrix}$$

Sejam ainda $\widehat{\boldsymbol{\beta}} = \left[\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2\right]^T$ e $\widetilde{\boldsymbol{\beta}} = \left[\widetilde{\boldsymbol{\beta}}_1, \mathbf{0}\right]^T$, os estimadores de máxima verosimilhança de $\boldsymbol{\beta}$ sob H_1 e H_0, respectivamente. Se existir um parâmetro ϕ desconhecido, sejam $\hat{\phi}$ *e* $\tilde{\phi}$ estimadores consistentes de ϕ sob H_1 e H_0, respectivamente. De acordo com a secção ***(4.1.5)***, temos as seguintes estatísticas de teste para testar H_0 versus H_1.

- Estatística da razão de verosimilhança

$$Q = -2\{\ell(\widetilde{\boldsymbol{\beta}}_1, \mathbf{0}, \tilde{\phi}) - \ell(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \hat{\phi})\}$$

- Estatística de Wald

$$W = \widehat{\boldsymbol{\beta}}_2^T \hat{\mathcal{A}}_{22}^{-1} \widehat{\boldsymbol{\beta}}_2$$

- Estatística Score de Rao

$$U = \tilde{S}_2^T \tilde{\mathcal{A}}_{22} \tilde{S}_2$$

onde $\tilde{\mathcal{A}}, \tilde{S}, \hat{\mathcal{A}} e \hat{S}$ são os cálculos relativos à matriz $\mathcal{A}$ e à função score em $(\widetilde{\boldsymbol{\beta}}, \tilde{\phi})$ *e* $(\widehat{\boldsymbol{\beta}}, \hat{\phi})$, respectivamente.

Todas estas estatísticas de teste têm uma distribuição assintótica de $\mathcal{X}_r^2$, porque r é a dimensão do vector $\boldsymbol{\beta}_1$.

A estatística de teste a usar pode depender da metodologia aplicada na selecção de modelos. Por exemplo, a estatística de *Wald* é útil quando se começa por formar um modelo com um grande número de parâmetros e se consideram modelos alternativos pela exclusão de covariáveis (selecção *backward*). A estatística *Score de Rao* é útil na selecção de modelos, quando se parte de um modelo nulo, ou de um modelo com poucos parâmetros, e se consideram modelos alternativos pela inclusão de variáveis explicadas (selecção *forward*).

Outro critério de selecção é o *critério de informação de Akaike* (AIC) proposto por *Akaike (1974)*. A estatística correspondente para o modelo em H_0 é

$$AIC = -2\ell(\widetilde{\boldsymbol{\beta}_1}, \mathbf{0}, \tilde{\phi}) + 2r$$

onde r é o número de parâmetros do modelo nulo e $\ell(.)$ é o logaritmo da função de verosimilhança aos dados.

Quanto menor é o valor de AIC menor será a informação perdida, ou seja, é considerado como representativo de um melhor ajustamento. Devemos ter como objectivo minimizar o valor de AIC.

4.1.7. Ajustamento do Modelo

A avaliação da qualidade do ajustamento do modelo escolhido pode ser feita recorrendo a algumas medidas de qualidade de ajustamento e complementada com a análise de resíduos. Nesta secção, vamos apresentar detalhadamente duas medidas de avaliação da qualidade de ajustamento, *deviance* e *estatística de Pearson generalizado*, e alguns tipos de resíduos aplicados na avaliação de modelos lineares generalizados. Estas medidas de avaliação vão medir a discrepância entre as observações Y_i e os valores de $\hat{\mu}_i$, i=1,...,n. Consoante a medida desta discrepância, podemos concluir se o modelo é ou não considerado adequado aos dados em causa.

4.1.7.1 Função desvio ou *deviance*

Segundo *Turkman e Silva (2000)* a função desvio ou *deviance* é a medida de qualidade de ajustamento mais utilizada e é dada pelo logaritmo da razão de verosimilhança. Esta medida compara o modelo saturado com o modelo corrente (de interesse). Consideremos que o modelo saturado é um modelo linear generalizado com a mesma distribuição e a mesma função de ligação do modelo de interesse (corrente).

O logaritmo da função de verosimilhança de um modelo linear generalizado é (sendo $\theta_i = q(\mu_i)$)

$$ln\, L(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{w_i[y_i q(\mu_i) - b(q(\mu_i))]}{\phi} + c(y_i, \phi, w_i).$$

Para o modelo saturado que vamos designar por S tem-se $\hat{\mu}_i = y_i$. O máximo da função log-verosimilhança para este modelo é

$$\ell_S(\widehat{\boldsymbol{\beta}}_S) = \sum_{i=1}^{n} \frac{w_i[y_i q(y_i) - b(q(y_i))]}{\phi} + c(y_i, \phi, w_i)$$

Por outro lado, se designarmos $\hat{\mu}_i$ o estimador de máxima verosimilhança de μ_i, i=1,...,n, o máximo da função log-verosimilhança para o modelo corrente M (de interesse) é

$$\ell_M(\widehat{\boldsymbol{\beta}}_M) = \sum_{i=1}^{n} \frac{w_i[y_i q(\hat{\mu}_i) - b(q(\hat{\mu}_i))]}{\phi} + c(y_i, \phi, w_i).$$

Se comparamos o modelo corrente M com o modelo saturado S através da estatística da razão de verosimilhança, obtemos

$$\begin{aligned} D^*(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}) &= -2\left[\ell_M(\widehat{\boldsymbol{\beta}}_M) - \ell_S(\widehat{\boldsymbol{\beta}}_S)\right] = \\ &= -2\sum_{i=1}^{n} \frac{w_i}{\phi}\left\{\left[y_i q(\hat{\mu}_i) - b(q(\hat{\mu}_i))\right] - \left[y_i q(y_i) - b(q(y_i))\right]\right\} = \\ &= \frac{D(\mathbf{y}, \hat{\boldsymbol{\mu}})}{\phi}. \end{aligned} \qquad \textit{(4.20)}$$

$D^*(\boldsymbol{y}, \widehat{\boldsymbol{\mu}})$ definido em ***(4.20)*** designa-se por desvio reduzido, o numerador $D(\mathbf{y}, \widehat{\boldsymbol{\mu}})$ denomina-se por desvio para o modelo corrente. Note-se que o desvio é só função dos dados. Como se pode observar em ***(4.20)***, o desvio pode ser decomposto

$$\begin{aligned} D(\mathbf{y}, \widehat{\boldsymbol{\mu}}) &= \sum_{i=1}^{n} 2w_i \left\{ y_i\left(q(y_i) - q(\hat{\mu}_i)\right) - b\left(q(y_i) + b(q(\hat{\mu}_i))\right)\right\} \\ &= \sum_{i=1}^{n} d_i \end{aligned}$$

em que d_i mede a diferença dos logaritmos das verosimilhanças observada e ajustada à observação i. Portanto é uma medida de discrepância total entre as duas log-verosimilhanças.
Podemos verificar que o desvio é sempre maior ou igual a zero, e decresce à medida que as variáveis explicativas vão sendo adicionadas ao modelo nulo, tomando valor zero para o modelo saturado ou completo. Uma outra propriedade importante do desvio é a sua aditividade para modelos encaixados.

A estatística do desvio ou deviance tem como objectivo testar a adequabilidade de um modelo. Se o valor observado for superior a $\chi^2_{n-p,\alpha}$, em que p representa a dimensão do vector β, então o modelo é considerado não adequado.

Tendo em conta o modelo binomial, a expressão da função desvio é dada por

$$2\left[\sum_{i=1}^{n} m_i\, y_i ln\frac{y_i}{\hat{\mu}_i} + \sum_{i=1}^{n} m_i(1-y_i)\, ln\frac{1-y_i}{1-\hat{\mu}_i}\right]$$

Trata-se de um teste útil para dados Binomial com grandes contagens, existindo com alternativa o teste do *Hosmer & Lemeshow* para pequenas contagens (descrito na secção ***4.2.5.1***).

4.1.7.2 Estatística de *Pearson* Generalizada

Outra medida importante da adequabilidade de modelos é a estatística de Pearson generalizada, proposta por *McCullagh & Nelder (1989)*, e define-se da seguinte forma:

$$X^2 = \sum_{i=1}^{n} \frac{w_i(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

onde $V(\hat{\mu}_i)$ é a função de variância estimada para a distribuição em causa como referem *Turkman e Silva (2000).*

A estatística X^2 é utilizada para testar a adequabilidade de um modelo, comparando o valor observado desta com o quantil de probabilidade $1-\alpha$ e distribuição X^2 com $n-(k+1)$ graus de liberdade. Em alguns modelos a aproximação pela distribuição de X^2 pode ser má, mesmo para grandes amostras. Por este motivo, é usual agruparem-se os dados o mais possível, de modo que, o número de observações em cada grupo, não seja reduzido.

4.1.7.3 Análise de Resíduos

A análise dos resíduos permite-nos não só uma avaliação local da qualidade do ajustamento no que diz respeito à escolha da distribuição, da função de ligação e das variáveis explicadas a entrar no modelo, como também permite detectar a presença de observações mal ajustadas - *Turkman e Silva (2000)*

4.1.7.3.1 Definição de Resíduos

No modelo linear normal, o vector de resposta **Y** pode ser escrito na forma,

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \text{ com } \boldsymbol{\varepsilon} \sim N_n(0,\sigma^2 \mathrm{I}),$$

o vector de resíduos é dado habitualmente por $\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}}$, onde $\hat{\boldsymbol{y}} = \hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ corresponde ao vector dos valores ajustados.

Outra quantidade de interesse na análise de resíduos, no caso do modelo linear normal, é a matriz de projecção $H = \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}^T$ (designado em inglês, por *hat matrix* pelo facto de $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$). A matriz de projecção é simétrica e idempotente, sendo os seus elementos h_{ij} uma medida de influência exercida por y_j em $\hat{y}_i$. A influência exercida por y_i em $\hat{y}_i$ é reflectida pelo elemento da diagonal, h_{ii},

$$h_{ii} = x_i^T(\boldsymbol{x}^T\boldsymbol{x})^{-1}x_i, \text{ i} = 1,\ldots,\text{n},$$

designado habitualmente por *leverage*. Como se tem $\sum h_{ii} = k+1$ e $0 \leq h_{ii} \leq 1$, *Hoaglin & Welsch (1978)* afirmam que um ponto é considerado influente se $h_{ii} > \frac{2(k+1)}{n}$.

Podemos ainda considerar os resíduos standardizados,

$$r_i^s = \frac{y_i - \hat{\mu}_i}{\sqrt{1-h_{ii}}},$$

obtidos dividindo por um factor $\sqrt{1-h_{ii}}$, o que torna a sua variância constante. Se este resíduo for ainda dividido por s, estimativa de σ, obtemos os resíduos standardizados e studentizados:

$$r_i' = \frac{y_i - \hat{\mu}_i}{s\sqrt{1-h_{ii}}}.$$

Estes resíduos têm aproximadamente uma distribuição normal com valor médio zero e variância um.

McCullagh & Nelder (1989) apresentam três formas de resíduos para os modelos lineares generalizados: os resíduos de *Pearson*, os resíduos de *Anscombe* e os desvios residuais (ou resíduos deviance). Os resíduos de Pearson e os resíduos da *deviance* ou desvios residuais têm sido amplamente usados no processo de validação dos modelos e serão de seguida abordados.

- **Resíduos de Pearson**

À semelhança da definição de resíduo para o modelo linear normal, os resíduos de Pearson são definidos da seguinte forma:

$$R_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{Var(Y_i)}}} = \frac{y_i - \hat{\mu}_i}{\sqrt{a_i(\hat{\phi})V(\hat{\mu}_i)}}.$$

Correspondem aos resíduos simples standardizados pelo desvio padrão de y_i estimado. Tendo em conta que assintóticamente se tem var $(y_i - \hat{\mu}_i) \approx$ var $(y_i(1-h_{ii}))$ (ver, por exemplo, *Turkman e Silva (2000)* o resíduo padronizado é dado por:

$$R_i^{P'} = \frac{y_i - \hat{\mu}_i}{\sqrt{a_i(\hat{\phi})V(\hat{\mu}_i)(1-h_{ii})}}. \qquad \textit{(4.21)}$$

A grande desvantagem da utilização dos resíduos de Pearson é o facto da sua distribuição ser bastante assimétrica para modelos não normais. Mas, a facilidade no cálculo dos mesmos torna a sua utilização bastante vantajosa.

- **Resíduos *deviance* ou desvio residual**

Alternativamente podemos definir novos resíduos usando a contribuição da i-ésima observação para a função desvio definida em ***(4.9)***

$$d_i\left\{y_i(q(y_i)-q(\hat{\mu}_i))-b\left(q(y_i)+b(q(\hat{\mu}_i))\right)\right\}$$

Assim o desvio residual é definido por:

$$R_i^D=\delta_i\sqrt{d_i}$$

onde δ_i = sinal$(y_i-\hat{\mu}_i)$. O desvio residual padronizado é obtido dividindo o desvio residual R_i^D por $\sqrt{\phi(1-h_{ii})}$, isto é

$$R_i^{*D}=\frac{R_i^D}{\sqrt{\phi(1-h_{ii})}}.$$

A Figura 32 mostra a expressão dos resíduos de *Pearson* e os desvios residuais para o modelo Binomial.

Figura 32: Expressões dos resíduos para o modelo Binomial

	R_i^P	R_i^D
Binomial	$\frac{m_i^{1/2}(y_i-\hat{\mu}_i)}{[\hat{\mu}_i(1-\hat{\mu}_i)]^{1/2}}$	$\delta_i\left[2m_i\left(ln\left(\frac{y_i}{\hat{\mu}_i}\right)+(1-y_i)ln\left(\frac{1-y_i}{1-\hat{\mu}_i}\right)\right)\right]^{1/2}$

δ_i = sinal$(y_i-\hat{\mu}_i)$

4.1.8.4 Análise Informal dos Resíduos

A análise informal dos resíduos, baseada essencialmente em representações gráficas, é uma ferramenta de extrema utilidade no processo de detecção de erros, quer na componente sistemática quer na componente aleatória do modelo. Apresentamos a seguir a descrição de alguns gráficos que podem realizar-se na análise de resíduos.

- **Gráfico dos resíduos versus $\hat{\eta}$**

Pela análise deste gráfico, não existem anomalias, quando os resíduos se encontram distribuídos em torno de zero com uma amplitude constante para diferentes valores de $\hat{\mu}$. O aparecimento de uma tendência no gráfico pode resultar de vários factores, bem como, a escolha errada da função de

ligação e da escala de uma das variáveis explicativas ou omissão de um termo quadrático numa variável explicativa.

Para o modelo binomial, *McCullagh & Nelder (1989)* sugerem ainda a transformação seguinte do valor predito $\hat{\mu}$ para a realização deste gráfico

$$-2sin^{-1}\sqrt{\hat{\mu}}\,.$$

- **Gráfico dos resíduos versus variáveis explicativas**

As variáveis explicativas estão presentes no preditor linear. A análise deste gráfico é semelhante à descrita anteriormente, ou seja, a existência de tendências sistemáticas podem indicar a escolha errada da função de ligação ou da escala de uma ou mais covariáveis.

- **Avaliação da omissão de uma variável explicativa**

Para averiguar se uma covariável, digamos u, omitida no modelo deverá ou não ser incluída no preditor linear temos um outro gráfico importante na validação informal dos resíduos, o chamado "*added-variable plot*", que permite cruzar os resíduos aumentados contra u. Para mais detalhe sobre os resíduos aumentados, pode consultar-se, por exemplo, *Cordeiro (1986)*.
McCullagh & Nelder (1989) afirmam que, para este objectivo, não é correcto traçar o gráfico dos resíduos versus u. Em primeiro lugar deve-se obter os resíduos não standardizados do modelo em que u se considera variável resposta, usando, o mesmo preditor linear e as mesmas ponderações usadas para **Y**. O gráfico de diagnóstico é o gráfico dos resíduos de **Y** não standardizados versus os resíduos de u. Se u foi correctamente omitida, então este gráfico não apresentará nenhuma tendência.

- **Validação da função de variância**

A análise informal da função de variância pode ser realizada através do gráfico dos resíduos absolutos versus valores ajustados (ou transformação adequada destes). Com uma função de variância mal escolhida, o gráfico apresentará uma tendência. Se apresentar uma tendência positiva a função de variância cresce lentamente com a média e poderá ser substituída por outra para a qual tal não aconteça.

4.1.7.5. Observações Discordantes

A análise informal dos resíduos descrita na secção anterior, permite-nos averiguar a existência de desvios sistemáticos do modelo. Nesta secção definiremos métodos que nos permitem averiguar se existe uma ou várias observações mal ajustadas pelo modelo, que não seguem o padrão das

restantes observações. Apresentam-se de seguida três medidas sistematizadas em *Turkman e Silva (2000)* que nos permitem avaliar o efeito provocado por este tipo de observações no modelo. Essas medidas são: medida de repercussão, de influência e medida de consistência.

- **Medida de repercussão (*"leverage"*)**

A repercussão mede o efeito que a observação tem nos valores ajustados. A definição geral de repercussão da j-ésima observação no valor ajustado da i-ésima resposta é a amplitude da derivada do i-ésimo valor ajustado $\hat{\mu}_i$ relativamente ao valor observado da j-ésima resposta, y_j. No caso dos modelos lineares generalizados, esta medida é dada pelo (i,j)-ésimo elemento da matriz H (matriz de projecção generalizada que mede a influência que $\boldsymbol{Y}$ tem em $\boldsymbol{\mu}$.)
Assim, uma medida da repercussão da i-ésima observação na determinação de $\hat{\mu}_i$ é dada por h_{ii}. Tendo em conta que

$$tra(H) = \sum_{i=1}^{n} h_{ii} = k + 1 \text{ (número de parâmetros)}$$

e considera-se que um ponto tem elevada repercussão se $h_{ii} > \frac{2(k+1)}{n}$
Os gráficos de h_{ii} contra $\hat{\mu}_i$ ou contra i, são geralmente úteis na identificação de pontos com repercussão elevada.

- **Medida de influência**

Uma observação é considerada influente quando uma pequena mudança no seu valor ou a sua exclusão do modelo provocam uma alteração significativa na estimação dos parâmetros do referido modelo. Saliente-se o facto de serem observações influentes não significa que a elas estejam associados valores de resíduos elevados.

Assim, a influência da i-ésima observação no vector estimado $\widehat{\boldsymbol{\beta}}$ pode ser calculada pela diferença $\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}$, onde $\widehat{\boldsymbol{\beta}}_{(i)}$ representa a estimativa de máxima verosimilhança do vector de parâmetros $\boldsymbol{\beta}$, quando a observação i é omitida, tal como *McCullagh e Nelder (1989)* referem, e $\widehat{\boldsymbol{\beta}}$ representa a estimativa de máxima verosimilhança da amostra com todas as observações.
Cook (1977) sugere como generalização da medida de influência:

$$D_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^T (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{(k+1)\hat{\phi}}. \qquad \textbf{(4.22)}$$

Gráficos de D_i contra $\hat{\mu}_i$ ou contra i são úteis na identificação de pontos influentes.

- **Medida de consistência**

Uma observação é considerada inconsistente se o valor do seu resíduo for elevado. Pode haver observações consistentes com repercussões elevadas. Uma observação inconsistente (*outlier*) não é necessariamente uma observação influente.

Uma das medidas de consistência mais utilizada é o resíduo de eliminação, sendo que, observações com resíduos de eliminação reduzidas são observações consistentes. No caso dos modelos lineares generalizados, pode realizar-se este estudo adaptando o modelo sem uma dada observação (na notação (i)) e calcular os resíduos da observação eliminada em relação ao correspondente valor predito $\hat{\mu}_{(i)} = g^{-1}(x_i^T \hat{\boldsymbol{\beta}}_{(i)})$ por meio de:

$$R_{(i)}^{P'} = \frac{y_i - \hat{\mu}_{(i)}}{\sqrt{a_i(\hat{\phi}) V(\hat{\mu}_{(i)})(1-h_{(ii)})}},$$

onde $h_{ii} = x_i^T (x_{(i)}^T \boldsymbol{W}_{(i)} x_{(i)})^{-1} x_i$

4. 2 Modelos de Regressão logística

No capítulo anterior abordámos de uma forma genérica os modelos lineares generalizados. Este capítulo é dedicado a um dos casos particulares dos modelos lineares generalizados, **o modelo de regressão logística**, em que a variável resposta é binária ou dicotómica. Considerando a variável resposta Y para cada indivíduo ou unidade experimental, esta, assume um de dois valores (0 e 1), sendo o insucesso = 0 e o sucesso = 1. Então define-se a variável aleatória dicotómica, Y da seguinte forma:

$$Y = \begin{cases} 0 & insucesso \\ 1 & sucesso \end{cases},$$

com $P(Y = 1) = \pi$ e $P(Y = 0) = 1 - \pi$, as probabilidades de "sucesso" e "insucesso", respectivamente. O principal objectivo será desenvolver métodos estatísticos que nos permitam investigar a relação entre a probabilidade de sucesso $\pi = \pi(\boldsymbol{x})$ e as variáveis explicativas $\boldsymbol{x}^T = (x_1,\ldots,x_k)$.

4.2.1 Classe de covariáveis

Supondo que para a i-ésima combinação de condições experimentais, definida pelo vector $x_i^T = (x_{i0},\ldots,x_{ik})$, temos observações verificadas em m_i indivíduos, ou seja, para os $n = m_1 + \ldots + m_g$ indivíduos em estudo, os m_i indivíduos do grupo $i, i = 1, \ldots, g$ têm o mesmo vector de covariáveis

$x_i^T=(x_{i0},\ldots,x_{ik})$. Segundo *McCullagh & Nelder (1989)*, estes indivíduos formam uma classe de covariávies.

Quando os dados estão agrupados em classes de covariáveis, as variáveis resposta passam a ter a forma $\frac{Y_1}{m_1},\ldots,\frac{Y_g}{m_g}$, onde $0\leq Y_i\leq m_i$, é o número de sucessos em m_i indivíduos na i-ésima classe de covariáveis. Quando $m_1 = \ldots = m_g = 1$, diz-se que temos dados não agrupados que é um caso particular de dados agrupados. Esta distinção entre dados agrupados e não agrupados é importante, na medida em que:

1. Muitas das análises válidas para dados agrupados, em particular as que envolvem a distribuição normal (gaussiana), não são apropriadas para dados não agrupados;
2. As aproximações assintóticas utilizadas em dados agrupados podem ser baseadas em dois tipos: as m_i-assintóticas ($m_i \to \infty$, $\forall_i$) e as n-assintóticas ($n \to \infty$). Apenas as segundas podem ser utilizadas quando os dados são não agrupados.

Supondo que todos os indivíduos pertencem a uma classe de covariáveis, sendo as observações independentes e a probabilidade de sucesso constante dentro de cada classe, então, Y_i dado m_i tem distribuição binomial com número de provas m_i e parâmetro π_i, ou seja, $Y_i \sim Bin(m_i, \pi_i)$. Neste caso, a variável aleatória Y_i, conta o número de sucessos entre as m_i observações.

4.2.2 Descrição de modelo

A regressão logística assume que a relação entre as variáveis independentes e a variável dependente binária se faz através de:

$$ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

onde π é a probabilidade de sucesso e $x_1, x_2, \ldots, x_k$ são variáveis independentes (preditores). $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ são os coeficientes de regressão estimados. A regressão logística permite estimar a probabilidade de determinado evento ocorrer e, veremos adiante, que também permite estimar razões de chances associadas a diferentes níveis de factores.

Sendo assim, pode dizer-se que o modelo de regressão logística é caracterizado por:

- A componente aleatória é constituída por uma variável aleatória independente, Y_i que tem distribuição binomial, a qual descreve a distribuição dos erros e sobre a qual a análise é baseada;
- A função de ligação é a função logit, $\eta_i = ln\left(\frac{\pi_i}{1-\pi_i}\right)$;
- A componente sistemática do modelo é representada por $\eta_i = x_i^T\beta$.

A função de ligação do modelo de regressão logística pode ser escrita da seguinte forma:

$$\eta_i = ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T \beta$$

ou alternativamente,

$$\pi_i = \frac{e^{\boldsymbol{x}_i^T \beta}}{1+e^{\boldsymbol{x}_i^T \beta}},$$

com $\boldsymbol{\beta} = (\beta_0,\ldots,\beta_k)$, o vector de parâmetros desconhecidos.

Componente aleatória

Usando a função massa de probabilidade da distribuição binomial provou-se que esta distribuição pertence à família exponencial no capítulo anterior (***secção 4.1.2.3***), a partir do qual se podem retirar as seguintes conclusões:

Sabendo que $Y_i \sim Bin(m_i, p_i)$

- ✓ O parâmetro canónico é $\theta_i = ln\left(\frac{\pi_i}{1-\pi_i}\right)$ e consequentemente $\pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}$;
- ✓ A função $a_i(\phi) = \frac{1}{m_i}$ quando $\phi = 1$ (não existe parâmetro perturbador);
- ✓ A função

$$b(\theta_i) = ln\left(\frac{1}{1-\pi_i}\right) = ln\left(\frac{1}{1-\frac{e^{\theta_i}}{1+e^{\theta_i}}}\right) = ln(1+e^{\theta_i});$$

- ✓ A função $c(y_i;\phi) = ln\binom{m_i}{y_i}$.

Função de ligação

Admita-se que a relação existente entre as probabilidades π_i e um vector de $\boldsymbol{x}_i^T = (x_{i0},\ldots,x_{ik})$, covariáveis observadas é da forma linear, então estamos perante um modelo linear generalizado cuja parte determinística é dada por:

$$g(\pi_i) = \eta_i = \boldsymbol{x}_i^T \beta, \qquad \textbf{(4.23)}$$

com π_i a probabilidade de sucesso dado $\boldsymbol{x}_i^T$=$(x_{i0},\ldots,x_{ik})$, ou seja, π_i=P$[Y = 1|\boldsymbol{x}_i^T]$ para $i = 1,\ldots,n$.

Existem vários de tipos de funções de ligação, tal como já foi referido na secção 4.1, de seguida vamos estudar uma das funções de ligação possíveis, a canónica para dados binomiais, que é a função logit.

- **Função logit**

Um dos grandes problemas do modelo apresentado em ***4.23*** prende-se com o facto de a probabilidade π_i, do lado esquerdo puder tomar apenas valores entre zero e um, ao passo que do lado direito o preditor linear pode assumir qualquer valor real, não havendo garantias de que os valores do preditor linear estejam no intervalo correcto, a menos que sejam impostas restrições aos coeficientes. Podemos fazê-lo de duas formas:

- Primeiro, passando as probabilidades π_i para a razão de chances da seguinte forma

$$OR_i = \frac{\pi_i}{1-\pi_i}$$

- De seguida, tomamos os logaritmos e calcula-se o *logit* ou *log-odds* por meio de

$$\eta_i = logit\,(\pi_i) = log\frac{\pi_i}{1-\pi_i}$$

Componente sistemática

O vector das covariáveis $\boldsymbol{x} = (x_0, x_2, \ldots, x_k)$ que constitui a parte sistemática do modelo, podem ser contínuas, qualitativas ou mistas. No grupo das covariáveis contínuas, tem-se como exemplos, a massa, a temperatura, a idade, o peso, etc. Estas variáveis tomam valores numa escala contínua. Como exemplos de covariáveis qualitativas ou categorizadas, temos: sexo, estado civil, classe etária, grupo de tratamentos, etc.

4.2.3 Estimação de máxima verosimilhança do vector β

Admita-se que $Y_1,\ldots,Y_n$, são variáveis aleatórias independentes com distribuição binomial, $Y_i \sim Bin\,(m_i, \pi_i)$, $i = 1, \ldots, n$ e $n = \sum_{i=1}^{n} m_i$. Então, a contribuição da i-ésima observação para o logaritmo da função de verosimilhança é,

$$\ell_i(\pi_i) = ln\binom{m_i}{y_i} + y_i\, ln\left(\frac{\pi_i}{1-\pi_i}\right) + m_i ln(1-\pi_i).$$

Para as n observações independentes, é,

$$\ell(\boldsymbol{\pi}) = \sum_{i=1}^{n} \ell_i(\pi_i) = \sum_{i=1}^{n} \left[ln\binom{m_i}{y_i} + y_i \, ln\left(\frac{\pi_i}{1-\pi_i}\right) + m_i ln(1-\pi_i) \right]. \quad \textbf{(4.24)}$$

A parte sistemática do modelo especifica a relação entre o vector π e as variáveis explicativas incluídas na matriz $\boldsymbol{X}$, de ordem $n \times (k+1)$. Para o modelo de regressão logística, esta relação define-se da seguinte forma,

$$g(\pi_i) = \eta_i = ln\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=0}^{k} x_{ij}\beta_j, \quad \textbf{(4.25)}$$

substituindo a expressão ***(4.25)*** em ***(4.24)***, obtém-se,

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=0}^{k} y_i \, x_{ij}\beta_j - \sum_{i=1}^{n} m_i \, ln\left(1 + e^{\sum_{j=0}^{k} x_{ij}\beta_j}\right), \quad \textbf{(4.26)}$$

o que corresponde ao logaritmo da função verosimilhança expresso em função dos parâmetros desconhecidos $\beta_0,\ldots,\beta_k$. Recorrendo ao procedimento da secção (***4.1.4.1***), podemos deduzir as equações de máxima verosimilhança para o parâmetro $\boldsymbol{\beta}$ da equação (***4.25***). Derivando a equação (***4.24***) em ordem a π_i , obtemos,

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \pi_i} = y_i \frac{\frac{1-\pi_i+\pi_i}{(1-\pi_i)^2}}{\frac{\pi_i}{1-\pi_i}} + m_i \frac{-1}{1-\pi_i} = \frac{y_i}{(1-\pi_i)\pi_i} - \frac{m_i}{1-\pi_i} = \frac{y_i - m_i\pi_i}{\pi_i(1-\pi_i)}.$$

A partir de equação ***(4.26)***, obtém-se a derivada do logaritmo da função de verosimilhança em ordem a β_j,

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - m_i\pi_i}{\pi_i(1-\pi_i)} \frac{\partial \pi_i}{\partial \beta_j}. \quad \textbf{(4.27)}$$

No caso dos modelos lineares generalizados, expressa-se $\frac{\partial \pi_i}{\partial \beta_j}$ como o produto de:

$$\frac{\partial \pi_i}{\partial \beta_j} = \frac{\partial \pi_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \pi_i}{\partial \eta_i} x_{ij}, \quad \textbf{(4.28)}$$

substituíndo a expressão (***4.28***) em (***4.27***), vem:

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - m_i\pi_i}{\pi_i(1-\pi_i)} \frac{\partial \pi_i}{\partial \eta_i} x_{ij}.$$

Para o modelo de regressão logística, obtemos as seguintes equações de máxima verosimilhança para $\boldsymbol{\beta}$:

$$S_j = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}(y_i - m_i\pi_i) = 0$$

ou na forma matricial,

$$\frac{\partial \ell}{\partial \beta_j} = \boldsymbol{X}^T(Y - \mu) = 0.$$

Recorrendo à equação ***(4.11)***, obtemos os elementos da matriz de informação de *Fisher* para **β** do modelo de regressão logística,

$$\mathcal{I}_{jt} = \sum_{i=1}^{n} \frac{m_i x_{ij} x_{it}}{\pi_i(1-\pi_i)} \left(\frac{\partial \pi_i}{\partial \eta_i}\right)^2 = \sum_{i=1}^{n} m_i x_{ij} x_{it}\, \pi_i(1-\pi_i) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X},$$

onde **W** é a matriz diagonal de ponderações com elementos,

$$\overline{W}_{ii} = m_i \pi_i (1 - \pi_i)$$

Recorrendo à equação ***(4.15)***, obtemos o seguinte resultado para o modelo de regressão logística,

$$u_i^{(p)} = \hat{\eta}_i + \frac{y_i - m_i \hat{\pi}_i}{m_i} \frac{\partial \eta_i^{(p)}}{\partial \pi_i^{(p)}},$$

Logo, a expressão final para a estimativa de **β** é dada por (***4.17***) e as iteracções param quando for atingido o critério definido por (***4.18***).

4.2.4 Interpretação das estimativas do modelo

Como interpretar os valores das estimativas de $\boldsymbol{\beta}$ caso as covariáveis sejam qualitativas?

Esta interpretação não é tão simples como no caso do modelo de regressão linear múltipla. No modelo de regressão logística as estimativas dos parâmetros β_k referem-se ao impacto de uma variação unitária de cada uma das variáveis independentes (mantendo fixas as restantes) sobre o logaritmo natural da razão entre a probabilidade de ocorrência do evento e a probabilidade de não ocorrência do evento. Sendo assim, os valores de $\boldsymbol{\beta}$ permitem obter uma indicação do sinal de $\pi(x) = P(Y = 1 | X = x) = 1 - P(Y = 0 | X = x)$ mostrando se esta probabilidade aumenta ou diminui à medida que X aumenta, sendo X uma das covariáveis. Caso o parâmetro estimado apresente um sinal negativo pode dizer-se que a influência do factor no resultado é negativa, caso contrário, o factor influencia positivamente o resultado. A taxa de subida ou descida aumenta à medida que $|\beta|$ aumenta, quando $\beta \to 0$ a curva aproxima-se de uma recta horizontal. Quando o $\beta = 0$ pode significar que Y é independente de X.

- Na presença de uma variável resposta Y binária e uma variável explicativa X, o modelo de regressão logística é dado por

$$\pi(x) = \frac{exp(\alpha + \beta x)}{1 + exp(\alpha + \beta x)}.$$

Equivalente com

$$logit[\pi(x)] = log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x.$$

Aplicando a exponencial a ambos os lados da equação anterior mostra-se que as chances são uma função exponencial de x. Assim, as chances aumentam multiplicativamente com e^{β} para cada aumento de uma unidade de x. Por outras palavras, e^{β} é uma razão de chances, a chance de $X = x + 1$ dividida pela chance de $X = x$.

Isto verifica-se quando a variável independente é codificada como 0 ou 1. Outras codificações podem exigir que se calcule o valor da diferença logística para a codificação específica utilizada, exponenciando de seguida essa diferença para estimar a razão de chances.

4.2.5 Avaliação da qualidade do ajustamento

Depois da selecção do modelo, é necessário avaliar a qualidade do ajustamento do mesmo. Na secção seguinte descrevem-se algumas medidas de avaliação da qualidade do ajustamento de GLM, em particular, para modelos de regressão logística.

4.2.5.1 Medidas globais do ajustamento

Para avaliar globalmente o ajustamento nos modelos de regressão logística empregam-se mais frequentemente os testes X^2 de *Pearson* e o teste baseado na *Deviance*, contudo, podem usar-se ainda, no caso binário, outras alternativas como: teste de *Hosmer-Lemeshow*, erro de predição e curvas de ROC.

Em todos os casos está em causa o teste de hipóteses:
H_0: O modelo é adequado
H_1: O modelo não é adequado

- **O teste χ^2 de Pearson**

O teste χ^2 de *Pearson* utiliza uma estatística que é um caso particular da estatística de *Pearson* generalizada (definida anteriormente).
A estatística do teste χ^2 de *Pearson* aplica-se habitualmente nos modelos de regressão logística, sendo para respostas binomiais calculada da seguinte forma:

$$\chi^2 = \sum_{i=1}^{g} \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)},$$

para dados agrupados (com g classes de covariáveis). No caso dos dados não agrupados, temos:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)}, \text{ já que } m_i = 1.$$

No final compara-se o valor obtido em χ^2 com o quantil de probabilidade $1-\alpha$ e a distribuição χ^2 com $n-(k+1)$ graus de liberdade e se o primeiro for superior ao segundo rejeita-se H_0.

- **A deviance**

A *deviance* define-se como duas vezes a diferença entre o máximo da log-verosimilhança no modelo saturado (atingido em $\tilde{\pi}$) e o máximo da log-verosimilhança no modelo em causa estimado (atingido em $\hat{\pi}$), isto é:

$$D(y;\pi) = 2[\ell(\tilde{\pi};y) - \ell(\hat{\pi};y)] =$$
$$= 2[\sum_{i=1}^{n} ln(\tilde{\pi}_i) + (m_i - y_i)ln(1-\tilde{\pi}_i) - \sum_{i=1}^{n} ln(\hat{\pi}_i) + (m_i - y_i)ln(1-\hat{\pi}_i)].$$

Substituindo $\tilde{\pi} = \frac{y_i}{n_i}$ na equação anterior, obtemos a função *deviance* para o modelo de regressão logística dada por:

$$D(y;\pi) = 2\sum_{i=1}^{g} y_i\, ln\left(\frac{y_i}{m_i\hat{\pi}_i}\right) + (m_i - y_i)ln\left(\frac{m_i - y_i}{m_i - m_i\hat{\pi}_i}\right),$$

para dados agrupados (com g classes de covariáveis). Para dados não agrupados, com $m_i = 1$, com n indivíduos, a função *deviance* passa a ser escrita da seguinte forma:

$$D(y;\pi) = 2\sum_{i=1}^{n} y_i\, ln\left(\frac{y_i}{\hat{\pi}_i}\right) + (1-y_i)ln\left(\frac{1-y_i}{1-\hat{\pi}_i}\right).$$

No que se refere às outras alternativas aplicadas no caso binário, anteriormente abordadas (Teste de *Hosmer-Lemeshow*, Erro de Predição e Curvas de ROC), estas serão detalhadas de seguida:

- **Teste de Hosmer-Lemeshow**

Hosmer & Lemeshow (2000) introduziram um outro teste para avaliar a qualidade do ajustamento no modelo de regressão logística que não é baseado nos resíduos do modelo, como convém a um modelo para dados binários. Este método envolve o agrupamento das observações baseado nas probabilidades estimadas do modelo e posterior comparação simultânea entre os acontecimentos observados e esperados em cada um dos grupos. Assim, o teste requer que as observações sejam ordenadas por ordem crescente da correspondente probabilidade estimada de ocorrência do evento (doença), sendo depois divididas em G grupos. A estatística de *Hosmer-Lemeshow* é então obtida calculando a estatística de qui-quadrado de *Pearson* da tabela de $2 \times G$ das frequências observadas e esperadas para os G grupos.

Sendo assim, supondo $k = n$, pensamos em n colunas que correspondem a n valores de probabilidades estimadas em que a primeira coluna corresponde ao valor mais baixo e a n-ésima coluna ao valor maior. Propõem-se então as duas seguintes estratégias de agrupamento:

- Colapso da tabela com base nos percentis das probabilidades estimadas (resultando em G colunas);
- Colapso da tabela com base em valores fixos das probabilidades estimadas (resultando em G colunas).

Utilizando o primeiro método, o mais usual, usando $G = 10$ grupos resulta que o primeiro grupo contém $n'_1 = \frac{n}{10}$ indivíduos com as menores probabilidades estimadas e o último grupo contém $n'_{10} = \frac{n}{10}$ indivíduos com as maiores probabilidades estimadas. Habitualmente, designam-se estes grupos por decis de risco. Considerando então a tabela de $2 \times G$ das frequências observadas e esperadas nos G grupos, temos que, para qualquer estratégia de agrupamento, o teste de *Hosmer-Lemeshow HL* é obtido através do cálculo da estatística de qui-quadrado de *Pearson* desta tabela, sendo a estatística de teste dada por:

$$HL = \sum_{g=1}^{G} \frac{(o_g - n'_g \bar{\pi}_g)^2}{n'_g \bar{\pi}_g (1 - \bar{\pi}_g)}, k \to g; g \to G$$

onde n'_g é o número de observações existentes no grupo g, $o_g = \sum_{j=1}^{n'_g} y_j$, o número de casos nas n'_g observações e $\bar{\pi}_g = \frac{m_g \bar{\pi}_g}{n'_g}$ é a probabilidade média estimada para o grupo g.

Sob a hipótese nula esta estatística de teste tem uma distribuição aproximada qui-quadrado com $G - 2$ graus de liberdade.

Erro de Predição

O erro de predição avalia o quão bem os valores observados estão próximos dos valores ajustados pelo modelo, através da proporção de casos preditos correctamente. No entanto, um modelo que se ajusta bem aos dados, não faz, necessariamente, uma boa predição. Se a predição é o objectivo da análise então a proporção de casos correctamente classificados é um critério ideal para comparação de modelos. A característica fundamental da validação cruzada é que cada previsão é independente dos dados a que é aplicada. Como consequência, a validação cruzada fornece uma estimativa imparcial do poder preditivo. Sendo assim o erro de predição, que representa a proporção de respostas preditas correctamente pode ser calculada recorrendo à Tabela 12:

		Observados, y_i	
		0	1
Ajustados, $\hat{y}_i$	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

Tabela 12: Valores Observados *versus* Valores Ajustados

donde:

$$EP = \frac{n_{00}+n_{11}}{n}.$$

Este valor mede a aderência dos dados estimados aos dados observados, considerando-se:

$$EP > 0{,}5 \implies sucesso\ no\ ajustamento$$
$$EP \leq 0{,}5 \implies fracasso\ no\ ajustamento$$

Curva ROC

Breve Resumo

A análise ROC (*Receiver Operating Characteristic*) teve a sua origem na teoria de decisão estatística. A aplicabilidade da análise ROC por meio das *curvas ROC* é muito vasta, abrangendo várias áreas, tais como: psicologia, controlo de qualidade, medicina, imagem radiológica, entre outras.

É uma ferramenta muito útil para medir e especificar problemas no desempenho do diagnóstico em medicina. Permite estudar a variação da *sensibilidade* e da *especificidade* de um teste diagnóstico (medidas já definidas na ***secção 3.2***), através de um método gráfico simples e robusto para diferentes valores de corte definidores do próprio teste. Adicionalmente, associa-se a área abaixo da curva ROC ao poder discriminante do teste.

Teoria Estatística

Iremos descrever a análise ROC numa perspectiva da teoria da decisão estatística. Pretendemos encontrar um teste que discrimine bem entre as duas hipóteses abaixo (representadas na Figura 33):

H_0: A população tem média $\mu = \mu_0$

H_1: A população tem média $\mu = \mu_1$

- A curva da esquerda representa a hipótese nula, H_0 (pode ser vista como correspondendo à população sã) e a distribuição da direita a hipótese alternativa, H_1 (que pode ser vista como correspondendo à população doente).

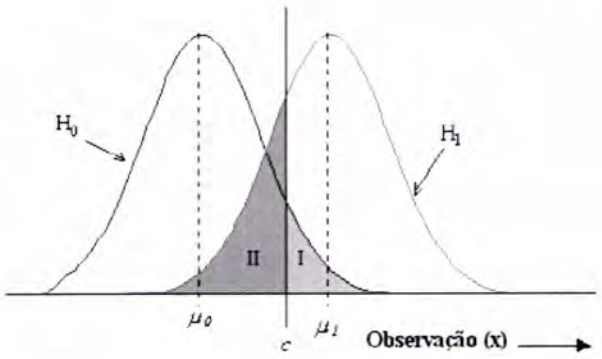


Figura 33: Distribuição de 2 populações (Braga, A., pág.11)

Com base numa observação x uma das hipóteses é aceite, de acordo com o critério de decisão c, a construção do teste estatístico equivale a dividir o eixo x em duas regiões separadas pelo critério de decisão c ($x < c \Rightarrow Aceitar\ H_0$ e $x > c \Rightarrow Aceitar\ H_1$).

- Observações retiradas da Figura 33:

- Conforme o critério de decisão escolhido, poderão determinar-se as probabilidades de cometer um *erro tipo I* ou um *erro do tipo II*;
- A área sombreada à direita do critério de decisão, c, representa um *erro de tipo I* (probabilidade de rejeitar H_0 sendo H_0 verdadeira);
- A área sombreada à esquerda do critério de decisão, c, representa um *erro de tipo II* (probabilidade de não rejeitar H_0 sendo H_1 verdadeira).

Pode calcular-se a significância α e a potência de um teste $1-\beta$, tendo-se:

$$\begin{cases} \alpha = P\,(erro\ Tipo\ I) = 1 - sensibilidade \\ 1-\beta = 1 - P\,(erro\ Tipo\ II) = especificidade \end{cases}$$

Posto isto, é possível construir um gráfico das curvas características da operação, que não são mais do que representações gráficas das probabilidades complementares de α e $1-\beta$, respectivamente, (exemplo na Figura 34).

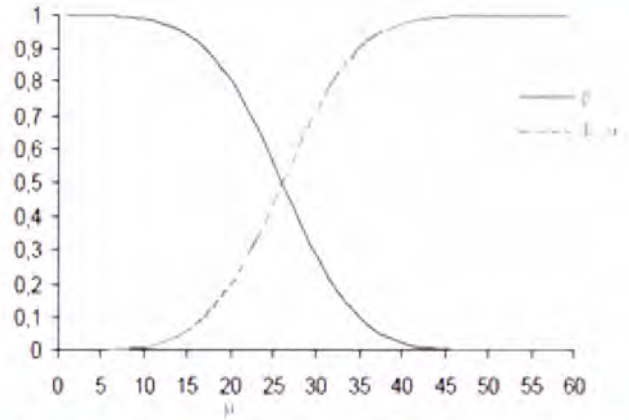


Figura 34: Distribuição das curvas características de operação (Braga, A., pág.13) em que:

$$\alpha = P(erro\ tipo\ I) = P\ (rej.H_0|H_0)$$
$$\beta = P(erro\ tipo\ II) = P(aceitar\ H_0|H_1)$$

A curva ROC mostra como podem variar os dois tipos de erro com a mudança do critério de decisão.

Ponto ou valor de corte

O maior problema da *sensibilidade* e da *especificidade* prende-se com o facto destas medidas dependerem do critério de diagnóstico ou de um *valor de corte*. Como tal, alterando a escolha desse valor de critério pode-se aumentar a *sensibilidade* diminuindo a *especificidade* ou vice-versa. A *sensibilidade* e a *especificidade* dependem de um único ponto de corte para classificar o resultado como positivo.

O *ponto de corte* é assim, o valor que define o limite entre um *teste negativo* e um *teste positivo* (termos descritos de seguida). A definição do *ponto de corte* considera a distribuição das frequências dos resultados observados na população, bem como os valores requeridos para a *especificidade* e a *sensibilidade* do teste em questão. Deverá ainda ter-se em conta que o *ponto de corte* escolhido depende também dos benefícios associados aos resultados correctos e dos custos associados aos incorrectos, tal como refere *Braga, A. (2000),* devendo o critério adoptado para um diagnóstico positivo encontrar-se do lado mais brando.

Se um dos objectivos é escolher um ponto de corte ideal para fins de classificação, *Hosmer & Lemeshow (2000)* sugerem a selecção de um *ponto de corte* que maximize a *sensibilidade* e a *especificidade*. Esta escolha é facilitada através de um gráfico como o que se apresenta na Figura 35 (que mostra os valores da sensibilidade e especificidade versus todos os pontos possíveis), onde vemos que a "escolha" ideal para um ponto de corte pode ser de 0.26 que é, aproximadamente, onde a *sensibilidade* e a *especificidade* cruzam naquele gráfico. Verifica-se ainda, por meio da mesma Figura que quando o valor crítico cresce a especificidade cresce e a sensibilidade decresce, tal como foi já mencionado.

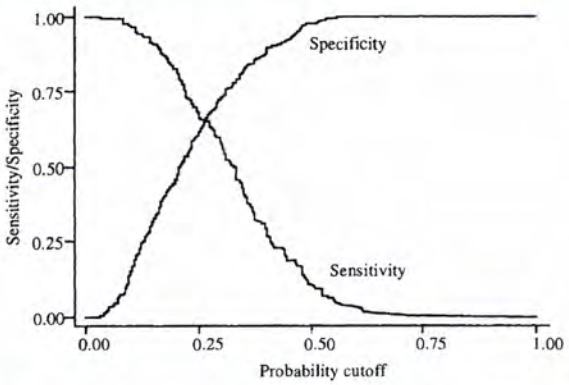


Figura 35: Sensibilidade e especificidade versus todos os pontos possíveis de corte (Hosmer & Lemeshow, pág. 162).

Em suma, valores de corte elevados conduzem a um teste pouco sensível e muito específico, por outro lado, valores de corte baixos conduzem a um teste muito sensível e pouco específico. Na prática interessa-nos ter um teste que seja, ao mesmo tempo, altamente específico e altamente sensível.

Geometricamente, a curva ROC é portanto, um par de parcelas (*1-especificidade, sensibilidade*) num plano denominado ROC unitário, uma vez que as coordenadas deste gráfico representam medidas de probabilidades e, como tal, variam entre zero e um.
A *sensibilidade* diz respeito ao eixo vertical e *(1 – especificidade)* ao eixo horizontal, podendo assim combinar-se estas duas medidas numa curva ROC como se apresenta na Figura 36.

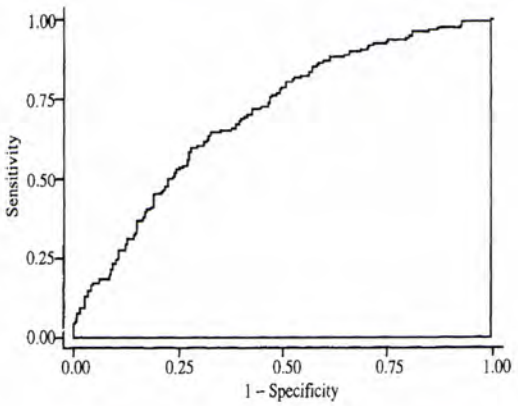


Figura 36: Sensibilidade e 1–especificidade para todos os possíveis pontos de corte, obtém-se assim a curva ROC (Hosmer & Lemeshow, pág. 163).

AUC (*Area Under the Curve*)

A área abaixo da curva ROC *(AUC)* é um dos índices mais utilizados para sintetizar a qualidade da curva, podendo esta área ser calculada por vários métodos, tal como refere *Braga, A. (2000).*

Os valores de *AUC* variam entre 0 e 1 e podem ser usados como uma medida de habilidade do modelo em discriminar quem sofreu o "desfecho" e quem não sofreu. Pode assim interpretar-se o valor de *AUC* como sendo a probabilidade de classificar correctamente um par de sujeitos seleccionados aleatoriamente, um do grupo dos sãos e outro do grupo dos doentes.

Desta forma, de acordo com *Hosmer & Lemeshow (2000)*, o valor de *AUC* pode ser interpretado tendo em conta a Tabela 13 presente abaixo:

AUC	Diagnóstico
AUC = 0,5	Modelo sem poder discriminatório
0,5 ≤ AUC < 0,7	Discriminação fraca
0,7 ≤ AUC < 0,8	Discriminação aceitável
0,8 ≤ AUC < 0,9	Discriminação boa
≥ 0,9	Discriminação excelente

AUC – Area Under the curve

Tabela 13: Valores das áreas da Curva de ROC e correspondente classificação

Análise ROC e Medicina

Num teste de diagnóstico existem dois tipos de erro que podem ocorrer, tal como na medicina:

- A escolha de uma falha (declarar um indivíduo doente como são)
- A escolha de um falso alarme (declarar um indivíduo são como doente)

Nesta situação, um profissional de saúde ao tomar uma decisão, irá preferir um falso alarme a uma falha, optando assim por um teste mais sensível. Contudo, tendo em conta a terapia disponível para a doença em causa que poderá ser cara e deficiente, o teste torna-se pouco específico.

No sentido de resolver este tipo de situações surge a análise ROC associada à curva ROC e área abaixo desta.

Análise ROC e Regressão Logística

A análise ROC pode advir duma tabela de 2×2 como a Tabela 7 apresentada na ***secção 3.2*** aquando da definição das noções de sensibilidade e especificidade. A análise ROC é assim baseada em duas quantidades que contêm toda a informação dessa tabela, uma designada por Valor Preditivo Positivo definida por $a\,/\,(a+b)$ e outra designada por Valor Preditivo Negativo definida por $d\,/\,(c+d)$.

Posto isto, existem dois acontecimentos e duas respostas, no total quatro resultados possíveis, os dois acontecimentos possíveis representam as colunas e as duas respostas permitidas as linhas. Seja X a designação para uma variável aleatória que pode ser discreta ou contínua e tendo em conta que valores baixos de X favorecem a decisão "Não ter doença" (D^-) e valores elevados de X favorecem

a decisão "Ter doença" (D^+) e tendo em conta as distribuições dos valores de X para os valores "Ter doença" e "Não ter doença" verifica-se que essas duas distribuições se sobrepõem, o que significa que alguns dos casos identificados como não doentes poderão ter leituras como doentes e por outro lado, alguns dos casos inicialmente identificados como doentes poderão ter leituras como não doentes.

Estão aqui em análise, as hipóteses:

H_0: O indivíduo é doente (D^+)
H_1: O indivíduo não é doente (D^-)

A representação ROC, em termos de diagnóstico, dá a probabilidade de aceitar H_0, isto é, considerar o indivíduo doente.

A probabilidade estimada da doença é calculada para cada paciente e estas probabilidades são classificadas. Uma curva ROC é obtida pelo cálculo da sensibilidade e especificidade observadas em cada probabilidade estimada da doença.

1.2.5.2 Análise de resíduos

- **Resíduos de Pearson**

No caso dos modelos de regressão logística, o resíduo de Pearson para o i-ésimo indivíduo, define-se por:

$$R_i^P = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}},$$

para dados agrupados. O resíduo de Pearson para a j-ésima classe de covariáveis, define-se por:

$$R_j^P = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{\hat{\pi}_j m_j(1-\hat{\pi}_j)}}.$$

Dado um vector de covariáveis comum a m_i indivíduos, com $y_i = 0$ e $\hat{\pi}_i$ a probabilidade estimada, então, o resíduo de Pearson para i-ésimo indivíduo, com este vector de covariáveis, é dado por:

$$R_i^P = \frac{0 - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}} = -\sqrt{\frac{\hat{\pi}_i}{1-\hat{\pi}_i}}.$$

O resíduo de Pearson calculado para todos os indivíduos na j-ésima covariável é:

$$R_i^P = \frac{0 - m_i\hat{\pi}_i}{\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}} = -\sqrt{m_i}\sqrt{\frac{\hat{\pi}_i}{1-\hat{\pi}_i}}.$$

Verifica-se na expressão anterior que R_i^P aumenta negativamente quando m_i aumenta.

➢ **Resíduos da *deviance* ou desvio residual**

Em (***4.1.7.3.1***) definiu-se para os modelos lineares generalizados os desvios residuais. No caso dos modelos de regressão logística temos para o caso de dados não agrupados

$$R_i^D = \text{sinal}(y_i - \hat{\pi}_i)\sqrt{2\left[y_i ln\left(\frac{y_i}{\hat{\pi}_i}\right) + (1-y_i) ln\left(\frac{1-y_i}{1-\hat{\pi}_i}\right)\right]}$$

ou para dados agrupados, temos:

$$R_i^D = \text{sinal}(y_i - m_i\hat{\pi}_i)\sqrt{2\left[y_i ln\left(\frac{y_i}{m_i\hat{\pi}_i}\right) + (m_i-y_i) ln\left(\frac{m_i-y_i}{m_i(1-\hat{\pi}_i)}\right)\right]}.$$

No modelo de regressão logística, o sinal dos resíduos depende do valor de y_i, isto é, quando y_i=1, o resíduo é positivo e quando y_i=0, é negativo.

Os resíduos podem ter variâncias diferentes, mesmo que a variância da variável resposta seja constante, porque a precisão dos resíduos depende do padrão das covariáveis, pelo que é de todo aconselhável, utilizar resíduos padronizados, pelo que se deve ter:

- Resíduos de *Pearson* e desvios residuais aproximadamente Normais (0,1) para dados que seguem uma distribuição Binomial com grandes contagens (95% dos resíduos devem estar entre - 2 e 2).
- Para dados que seguem uma distribuição Binomial com pequenas contagens, não podemos esperar que os resíduos de *Pearson* e desvios residuais possuam uma distribuição normal padronizada, contudo, devem verificar-se os seguintes critérios:
  - a maioria dos resíduos devem estar entre -2 e 2;
  - a sua variância deve ser unitária.

1.2.5.3 Alguns gráficos

Pode efectuar-se ainda o estudo e à análise dos seguintes gráficos:

- Resíduos padronizados versus preditores lineares η ou versus função dos valores ajustados $\hat{\mu}$ ou versus indíce (deve tentar-se encontrar um padrão nulo, resíduos distribuídos em torno de zero com amplitudes constantes para valores diferentes de $\hat{\mu}$).

- Resíduos versus covariáveis incluídas/não incluídas no modelo.

- Gráfico normal de probabilidades para os resíduos (que pode ser realizado mas não pode ser interpretado).

4.2.5.1 Observações Discordantes

Tendo em conta a descrição apresentada na secção (***4.1.7.5***) pode, no caso binomial, proceder-se ao estudo dos seguintes gráficos:

- Representação gráfica do valor *"Leverage"*
- Representação gráfica dos valores da distância de *Cook, "Cooks Distance"*

4.2.6 Sobredispersão

No caso de se pretender considerar a hipótese de sobredispersão dos dados e de se querer estimar o parâmetro de dispersão deverá ter-se em conta o que foi dito na secção (***4.1.4***). Caso se verifique que de facto há sobredispersão a estimação deve ser feita por quasi-verosimilhança, de acordo com *McCullagh & Nelder (1989, pág. 124-8).*

4.3 Modelos de Regressão Logística - Casos de Estudo

Nas secções anteriores foram apresentados modelos lineares generalizados para analisar dados binários, sendo o modelo logístico o mais popular de entre esses. Apresentam-se, de seguida, dois casos de estudo com aplicação da regressão logística bem como uma breve revisão da literatura de estudos do género.

4.3.1 Caso 1: Tipo de Neoplasia Benigno/Maligno

No estudo de caso-controlo já anteriormente considerado, os casos considerados foram todos os pacientes aos quais foi diagnosticado cancro da mama de origem maligna (159 pacientes de uma população de 208). Estes pacientes foram observados na *Unidade de Anatomia Patológica do HESE* entre Agosto de 2003 e Agosto de 2004. Pretende-se avaliar a associação entre algumas variáveis e o tipo, maligno ou benigno, de neoplasia da mama.

Os controlos são formados por uma amostra de 49 pacientes de uma população de 208, os quais foram também diagnosticados na mesma altura e local e tiveram cofirmada a neoplasia de origem benigna.

O registo de cada paciente e respectivos dados, fez-se depois de extraída e analisada a amostra/peça de tecido retirado da mama. O conjunto de covariáveis consideradas importantes na descriminação dos dois tipos de neoplasias foi o seguinte: *idade, tipo de amostra, ruralidade, lateralidade, tamanho tumoral, margens cirúrgicas e grau histológico*; as quais se encontram descritas e codificadas na Tabela 1 e de uma forma abreviada na Tabela 14.

Variável Resposta		Covariáveis	
Tipo Histológico Binário	Benigno	IDADE	CONTÍNUA
	Maligno		
	Benigno	RURALIDADE[7]	Rural – 8 Urbano – 35
	Maligno		Rural – 32 Urbano-117 **OR[8]≅ 0,72**
	Benigno	DISTRITO	Beja – 25 Évora – 18
	Maligno		Beja - 68 Évora-81 **OR.≅1,65**
	Benigno	LATERALIDADE	Sem Informação – 42 Mama Esquerda – 0 Mama Direita - 1
	Maligno		Sem Informação – 43 Mama Esquerda – 48 Mama Direita – 58
	Benigno	TAMANHO TUMORAL	Sem Informação – 42 <= 2 cm – 1 Entre 2,1 e 5cm - 0
	Maligno		Sem Informação – 34 <=2 cm – 86 Entre 2,1 e 5cm – 29
	Benigno	MARGENS CIRÚRGICAS	Sem Informação - 43 Positiva – 0 Negativa – 0 < 1cm – 0
	Maligno		Sem Informação – 20 Positiva – 17 Negativa – 100 < 1cm - 12
	Benigno	GRAU HISTOLÓGICO	Sem Informação - 42 Baixo Grau – 0 Médio Grau – 0 Alto Grau – 1
	Maligno		Sem Informação – 11 Baixo Grau – 36 Médio Grau – 50 Alto Grau – 52

Tabela 14: Tipo de Neoplasia Benigna/Maligna versus covariáveis em estudo e valores de algumas razões de chances

[7] Não esquecendo que foram consideradas zonas rurais os lugares e aldeias e como zonas urbanas as vilas e cidades dos distritos em estudo.

[8] Os valores de OR aqui apresentam ligeiras diferenças, verificando-se isso também na variável *ruralidade* (estudada no estudo caso-controlo, apresentado abaixo) devido ao facto de não estarmos a considerar o mesmo número de indivíduos.

Desta tabela podem fazer-se as seguintes observações:

- Um maior número de neoplasias diagnosticadas (*quer benignas, quer malignas*) em pacientes que residem em zonas urbanas, o que seria previsível, uma vez que reside um maior número de pessoas nas zonas urbanas.
- As covariáveis *tipo de neoplasia* e *distrito* são independentes (Tabela 27 presente no *AnexoIII*), não se verificando diferenças significativas nos distritos em estudo no que se refere à tipologia de neoplasia, contudo, observa-se um maior número de neoplasias benignas no distrito de Beja e, pelo contrário, um maior número de neoplasias malignas no distrito de Évora. O valor de OR = 1.65 significa que a exposição ao distrito de Beja eleva o risco do cancro da mama, uma vez que a chance de cancro da mama é 1,65 vezes maior para as pessoas que residem no distrito de Beja do que para as que residem no distrito de Évora, ainda que esta diferença não se possa considerar significativa, dado que o valor 1 está incluído no intervalo de confiança a 95% ($I.C._{95\%}$: 0.79 – 3.50) (ver Tabela 11).
- As covariáveis *tipo de neoplasia* e *tipo de amostra* são dependentes uma vez que no caso de neoplasias malignas se elege a peça cirúrgica (resultado esperado).
- Não faz sentido comparar as covariáveis, tamanho tumoral, grau histológico e margens cirúrgicas em neoplasias benignas e malignas, uma vez que não há registos para as neoplasias benignas.

Um dos objectivos deste estudo é, identificar as covariáveis que influenciam significativamente o tipo de neoplasia mamária (*benigna/maligna*) (Y), de modo a determinar a relação do tipo de neoplasia com os factores de risco. Tendo em conta o grupo de variáveis em estudo para este modelo, pode dizer-se desde já que as variáveis com significado para esta análise serão: *idade, distrito e ruralidade*, uma vez que as restantes (*lateralidade, tamanho tumoral, margens cirúrgicas e grau histológico*) se tratam de variáveis intrínsecas ao facto de se ter neoplasia da mama, isto é, de variáveis avaliadas no caso dessa neoplasia existir, não se podendo considerar factores que possam vir a influenciar no grau de malignidade da neoplasia em questão. Desde já fica, este breve comentário acerca desta base de dados que não pode ser tão útil quanto o desejado para concretizar o objectivo proposto. Como tal procedeu-se ao estudo e análise da situação com as variáveis atrás referidas e consideradas mais relevantes com a expectativa de, mais tarde, se possível, este trabalho servir de base para uma análise mais completa no que respeita a variáveis que possam sim tratar-se de factores de risco para o cancro da mama, e assim puder retirar daí conclusões mais pertinentes.

Começar-se-á com o estudo da selecção de covariáveis seguindo-se o estudo da adequabilidade dos modelos para os dados considerados. Os resultados para os ajustamentos dos modelos foram obtidos com base no software *R*.

Selecção de covariáveis e de Modelos

Seja $(x_1, \dots x_k)$ o conjunto de possiveis covariáveis candidatas a ser incluídas no predictor linear $\eta = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$. O problema da selecção do melhor conjunto de covariáveis pode ser visto na perspectiva de um problema de selecção do melhor modelo. Para essa selecção foi usado o método de selecção *stepwise* via AIC. Na Tabela 15 encontram-se os valores do critério de informação de *Akaike* – AIC para os modelos estudados, não devemos esquecer que um bom modelo deverá ter um valor baixo de AIC.

MODELO	AIC
M1	168.99
M2	166.99
M3	165.01
M4	163.33
M5	161.62
M6	**159.63**
M7	**158.3**

Tabela 15: Modelos e valores de AIC

M1 – Tipo histológico binário ~ idade + ruralidade + distrito + idade * distrito + ruralidade * distrito + idade * ruralidade + idade * distrito * ruralidade
M2 – Tipo histológico binário ~ idade + ruralidade + distrito + idade * distrito + ruralidade * distrito + idade * ruralidade
M3 – Tipo histológico binário ~ idade + ruralidade + distrito + idade * distrito + ruralidade * distrito
M4 – Tipo histológico binário ~ idade + ruralidade + distrito + ruralidade * distrito
M5 – Tipo histológico binário ~ idade + ruralidade + distrito
M6 – Tipo_histológico binário ~ idade + ruralidade
M7 – Tipo histológico binário~ idade

Considerando as medidas sumárias, verifica-se que o melhor modelo para o caso em estudo é o modelo M7 com valor de AIC igual a 158.3 e a variável idade significativa considerando um nível de significância de 5% (*p-value* igual a $2.74\times 10^{-9} < 0.05$). As estimativas dos parâmetros de regressão, os respectivos erros padrão e p-values para o teste sobre a sua significância, encontram-se na Tabela 16.

Parâmetro	Estimativa	Erro Padrão	Valor - t	p-value
β_0	-3.263	0.750	-4.35	1.35×10^{-5}
β_1 **Idade**	0.085	0.014	5.95	2.74×10^{-9}

Tabela 16: Estimativas, erros padrões e p-values

Posto isto, pode escrever-se o modelo linear generalizado para a variável resposta (*tipo histológico binário*), tendo em conta que a variável resposta segue uma distribuição *bernoulli* com probabilidade de sucesso (caso benigno) igual a p, tendo em conta que

$$Tipo\ histológico\ binário \sim B\ (p) \quad \text{e} \quad E\ [Tipo\ histológico\ binário] = p$$

e a associação entre o valor esperado da variável resposta ser benigna e as covariáveis é feita através da função de ligação *logit*.

$$logit\ (p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 idade$$

No caso em estudo a covariável é contínua e possui uma relação linear com o *logit* da probabilidade de sucesso da variável resposta, o que significa que à medida que a idade aumenta o logaritmo da chance varia β_1 unidades, logo, para cada ano a mais na idade temos $\exp(\beta_1)$ de variação da chance. Assim, o aumento de 1 ano de idade estima-se que aumente a chance de cancro da mama em 1,09.

Por meio dos valores da Tabela 16, também é possível determinar um intervalo de confiança aproximado para esta razão de chances através de $\exp(\hat{\beta}_1 \pm 1.96\, SE(\hat{\beta}_1))$. Esta análise também é possível por meio de um comando no $\boldsymbol{\mathcal{R}}$ que permite obter os valores da Tabela 17 directamente.

	OR	2.5%	97.5%
β_0	0.038	0.0079	0.1538
Idade	0.089	0.0608	1.1227

Tabela 17: OR e respectivos IC do modelo seleccionado

Tendo em conta que o melhor modelo encontrado para os dados em questão foi:

$$Tipo\ histológico\ binário \sim B\ (p)$$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 Idade$$

Pode ainda comparar-se o modelo corrente com o modelo nulo, por meio de um teste de Qui-Quadrado, de onde se obtém um p-value igual a 1.62×10^{-12}, o que permite concluir que a idade é claramente significativa.

Adequação do Modelo

Depois de seleccionado aquele que se julga ser o "melhor modelo", deve-se perguntar se o modelo é adequado, realizando para tal o seguinte teste de hipóteses:

H_0: *O Modelo adequa-se aos dados*
H_1: *O Modelo não se adequa aos dados*

Para tal podem realizar-se algumas análises, das quais se destacam e foram aqui utilizadas: a percentagem de *deviance* explicada, o erro de predição e as curvas de ROC.

Tendo em conta a informação do modelo escolhido e estimado contida na Tabela 18, apresentada abaixo, pode fazer-se um teste à bondade do ajustamento do modelo, o teste da razão de verosimilhança por exemplo, que resulta num *p-value* elevado (0.97) indicando que não há evidência de falta de ajuste. Apesar disso acontecer não significa que este modelo seja o correcto ou que não existam modelos melhores. Também se verifica que o modelo nulo poderia ser adequado para os dados, uma vez que 204.24 é um valor muito próximo dos 191 graus de liberdade.

***Desvio Nulo:* 204.24 com 191 graus de liberdade**
***Desvio Residual*: 154.34 com 190 graus de liberdade**

Tabela 18: Informação dos desvios do modelo seleccionado

Utilizando os valores da Tabela 18, pode ainda dizer-se que o desvio nulo é o desvio para o modelo com apenas um termo constante, enquanto o desvio residual é o desvio do modelo ajustado. Estes podem ser combinados com o objectivo de encontrar uma percentagem explicada da *deviance* como medida para avaliar a adequação do modelo, uma espécie de equivalência com R^2, calculada por meio de

$$\frac{(Desvio\ Nulo - Desvio\ Residual)}{Desvio\ Nulo}$$

Que no caso em estudo permite verificar que a percentagem da *deviance* explicada é de cerca de 24.4%, valor bastante baixo.

Adicionalmente, no sentido de tentar melhorar o modelo anterior, considerou-se, não apenas a idade como covariável, mas um polinómio de grau 2 ou 3 da idade, do género $Idade + (Idade)^2 + (Idade)^3$.

Novamente, medindo a percentagem da *deviance* explicada, verificou-se que a melhoria não foi notória, pois verificou-se exactamente o mesmo valor para a percentagem de *deviance* explicada.

Relativamente às restantes alternativas que permitem avaliar a bondade do ajustamento do modelo que se podem realizar no caso em estudo obteve-se como erro de predição, uma proporção de "acertos" de 84.38%, de onde se conclui que a percentagen de valores ajustados é superior a 50%, o que leva a afirmar que o modelo se adequa bem aos dados; construiu-se a curva de ROC (apresentada na Figura 37) a qual também leva à mesma conclusão, uma vez que AUC= 0.793, o que de acordo com *Hosmer e Lemeshow* mostra uma discriminação aceitável do modelo seleccionado. A estatística de *Hosmer-Lemeshow* deu o valor de 6.01 e um p-value de 0.646, podendo assim concluir-se que não se rejeita a hipótese nula de que o modelo se ajusta bem aos dados.

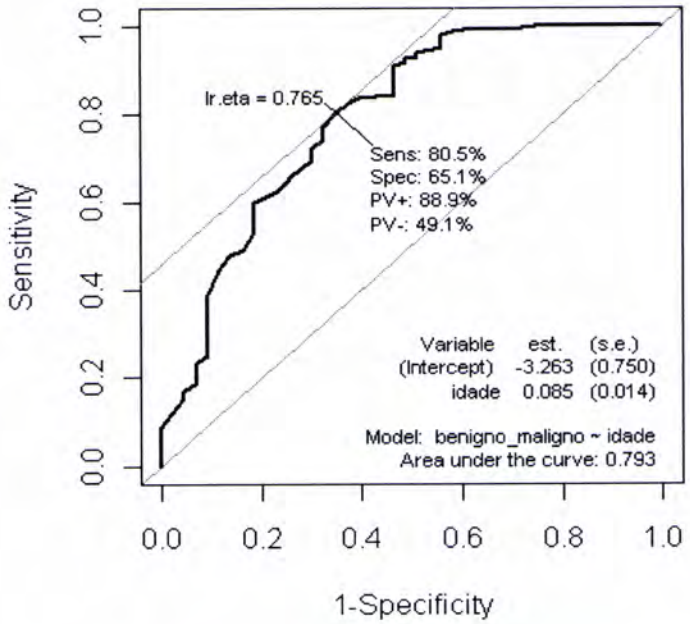


Figura 37: Curva de ROC do modelo seleccionado

Foi ainda estimado o parâmetro de dispersão para o modelo seleccionado, uma vez que, quando se desconfia de sobredispersão a análise por meio do modelo binomial não é a adequada, para o modelo binomial o parâmetro de dispersão é, por defeito, igual a 1. Obteve-se o valor de 1.116, indica que não se verifica sobredispersão do modelo seleccionado.

Análise Informal dos Resíduos

Na adequação do modelo podemos encontrar anomalias, tanto na componente aleatória do modelo, como na componente sistemática, as quais podem ser detectadas através de uma análise informal dos resíduos, usando representações gráficas adequadas.

Na Figura 38 apresentam-se 6 representações gráficas: gráfico aos desvios residuais standardizados, gráfico de probabilidades normal para os resíduos (*Normal Q-Q Plot*), gráficos dos resíduos *versus* valores ajustados, gráfico dos resíduos *versus* valores ajustados transformados, gráfico das distâncias de *Cook* e gráfico dos valores de *Leverage*.

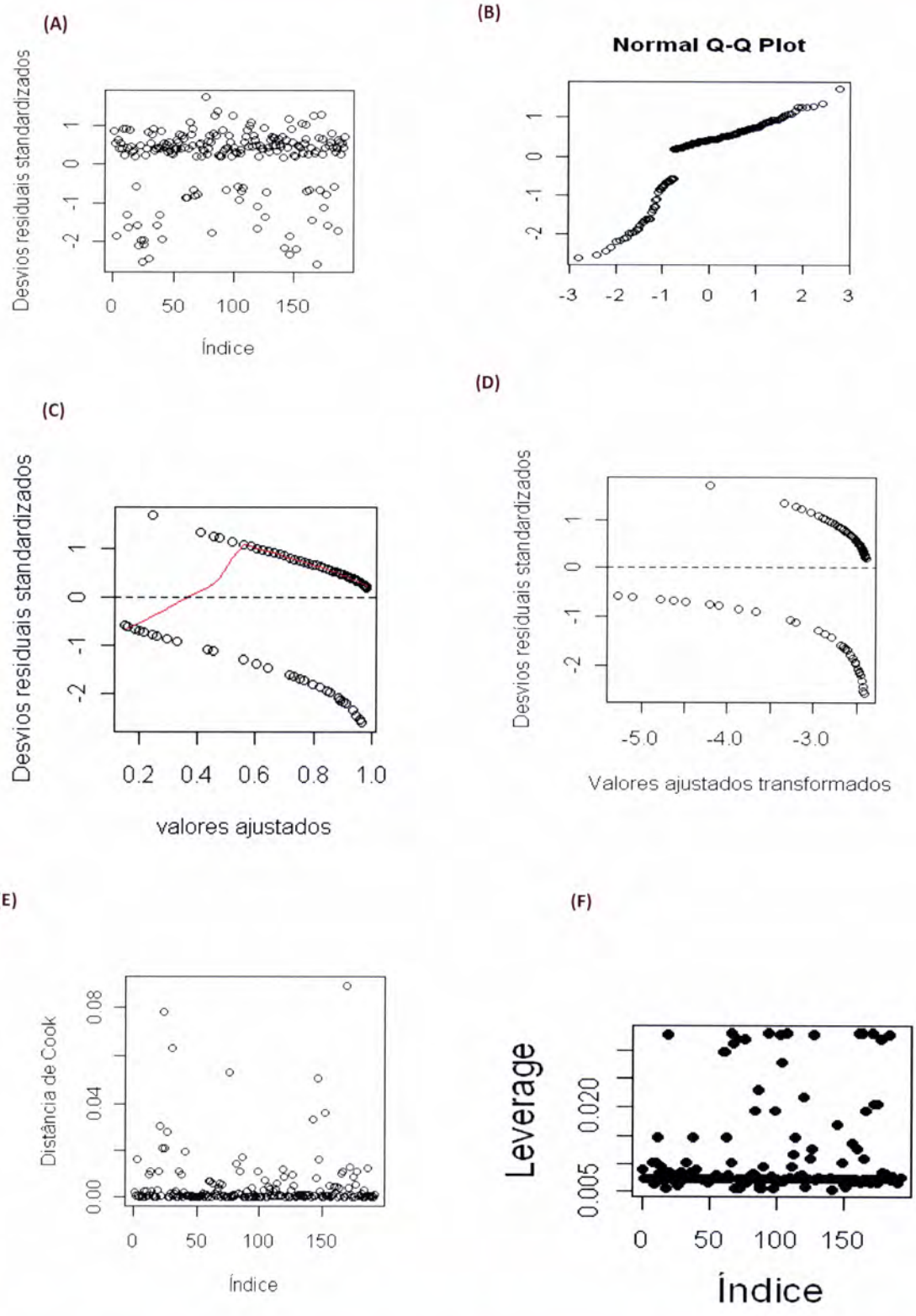


Figura 38:Gráficos (A) Desvios residuais Standardizados, (B) Normal Q-Q Plot dos Resíduos (C) Desvios residuais *versus* valores ajustados, (D) Desvios residuais *versus* valores ajustados transformados (E) Distância de Cook (F) Leverage

Dos gráficos (A) e (C), relativos aos resíduos standardizados e aos resíduos standardizados *versus* valores ajustados, observa-se que 95% dos resíduos se encontram compreendidos entre -2 e 2, o que significa que há validação dos pressupostos do modelo. Nestes gráficos observa-se ainda que existem poucas observações que se encontram abaixo de dois desvios padrões, nada de preocupante. A Figura 38 (B), onde se observa que os resíduos não são normais e nem tinham de o ser, apenas se encontra presente para mostrar, de uma forma grosseira, que para amostras grandes (como é o caso) pode ser de esperar uma espécie de normalidade aproximada dos resíduos. Tentando encontrar uma espécie de curvatura no gráfico que pudesse ser indicativa de vários factores de erro, procedeu-se à transformação dos valores ajustados, sugerida por *McCullagh & Nelder (1989)* e já referida na seccção 4.1.8, de onde surge o gráfico da Figura 38 (D) que nos permite verificar que a transformação não trouxe surpresas. O gráfico da Figura 38 (C) permite ainda realizar uma análise informal da função de variância, de onde se conclui não existiu uma má escolha da função de variância, pois o gráfico não apresenta nenhum tipo de tendência.

É ainda importante analisar a existência de uma ou mais observações mal ajustadas pelo modelo, isto é, observações que não seguem o padrão das restantes observações, as chamadas observações discordantes. No gráfico dos valores de *Cook* (Figura 38 (E)) notam-se algumas possíveis observações discordantes, após a análise no sentido de as detectar, consideraram-se como tal os indivíduos 24, 30, 76, 146 e 169 da base de dados em estudo. Tentando saber um pouco mais acerca destes indivíduos concluiu-se que das 5 observações influentes, um dos indivíduos tem 25 anos de *idade* e constitui um caso *maligno*, o que não é muito habitual, e os restantes indivíduos apresentam *idades* superiores a 68 e constituem casos *malignos*, o que também não é muito comum. Verifica-se ainda que não se dispõe de informação destes indivíduos relativamente às variáveis *tamanho tumoral e margens cirúrgicas*, de apenas um se conhece informação relativamente à *lateralidade* e de apenas dois relativamente ao *grau histológico*. Realizando a mesma análise por meio do gráfico das distâncias de *Leverage* (Figura 38 (F)) *não* se destacam quaisquer indivíduos como sendo possíveis pontos influentes.

Resultados e pesquisa bibliográfica

No que se refere a tentativa de associação entre cancro da mama e diferentes variáveis já vários estudos foram realizados e alguns deles referidos nas secções *2.2* e *3.4.2*, contudo, tendo em conta o caso acima em estudo em que o modelo seleccionado apenas contempla como factor de risco a idade o que vem confirmar, de certa forma, (uma vez que não se tratam de estudos realizados no nosso país), os estudos realizados por *Garicochea, B. et al (2009)*. Também *Mina Bissell (2008)*, uma das maiores especialistas do cancro da mama defendeu que "infelizmente", a idade é o principal factor de risco do cancro, sintetizando "se curarmos o envelhecimento acho que podemos curar o cancro". De todas as outras variáveis em causa, uma vez que a ruralidade não foi

considerada importante pelo modelo seleccionado como factor de risco para o cancro da mama, pode dizer-se que esse facto vem contrapor os estudos de *Robert, S. et al (2004) e Mathew, A. et al (2008).*

4.3.2 Caso 2: Tipo de Carcinoma: In Situ/Invasivo

Neste caso de estudo o objectivo reside na identificação das variáveis que influenciam significativamente o tipo de carcinoma mamário (tendo em conta, apenas os dois principais tipos identificados nos pacientes em estudo – **in situ e invasivo**), de modo a realizar dois estudos:

- Um primeiro, cujo objectivo é determinar a relação entre o *tipo de carcinoma (in situ/invasivo)* e os factores intrínsecos (*idade, ruralidade e distrito*) aos indivíduos em estudo;
-Um segundo em que se pretende determinar factores de prognóstico para o tipo de cancro da mama em causa (*in situ* ou invasivo).

Desta forma, da base de dados inicial seleccionou-se uma amostra de 142 indivíduos em que em 123 destes foi diagnosticado carcinoma invasivo da mama e em 19 carcinoma *in situ* (temos uma amostra pouco equilibrada).
De seguida, apresenta-se uma breve análise descritiva das covariáveis em estudo neste modelo e algumas conclusões da mesma.

Variável Resposta			Covariáveis
Tipo de Carcinoma	In Situ	IDADE	Idade < 50 - 1 Idade >= 50 - 18
	Invasivo		Idade < 50 – 21 Idade >= 50 - 102
	In Situ	TIPO DE AMOSTRA	Biópsia – 5 Peça – 14
	Invasivo		Sem Informação - 2 Biópsia – 28 Peça – 93
	In Situ	RURALIDADE	Rural – 4 Urbano – 15
	Invasivo		Rural – 26 Urbano – 97 O.R.≅1
	In Situ	DISTRITO	Beja – 5 Évora – 14
	Invasivo		Beja – 59 Évora – 64 O.R.≅ 2.565
	In Situ	LATERALIDADE	Sem Informação - 7 Mama Esquerda – 5 Mama Direita - 7

	Invasivo		Sem Informação – 32 Mama Esquerda – 41 Mama Direita - 50
	In Situ	TAMANHO TUMORAL	Sem Informação – 4 <= 2 cm – 15 Entre 2,1 e 5cm - 0
Tipo de Carcinoma	Invasivo		Sem Informação – 28 <=2 cm – 68 Entre 2,1 e 5cm – 27
	In Situ	MARGENS CIRURGICAS	Positiva – 4 Negativa – 12 < 1cm – 3
	Invasivo		Sem Informação – 17 Positiva – 13 Negativa – 85 < 1cm - 8
	In Situ	GRAU HISTOLÓGICO	Sem Informação – 1 Baixo Grau – 8 Médio Grau – 2 Alto Grau – 8
	Invasivo		Sem Informação - 6 Baixo Grau – 27 Médio Grau – 47 Alto Grau – 43
	In Situ	RE	Sem Informação - 7 Re Favorável – 10 Re não Favorável – 2
	Invasivo		Sem Informação - 27 Re Favorável – 79 Re não Favorável – 17
	In Situ	RP	Sem Informação – 7 Rp Favorável – 10 Rp não Favorável - 2
	Invasivo		Sem Informação – 27 Rp Favorável – 65 Rp não Favorável- 31
	In Situ	C.ERB.2	Sem Informação - 7 Não se observou coloração – 6 Coloração com manchas raras – 1 Fraca coloração – 0 Intensa Coloração - 5
	Invasivo		Sem Informação - 27 Não se observou coloração – 72 Coloração com manchas raras – 6 Fraca coloração – 10 Intensa Coloração - 8
	In Situ	P53	Sem Informação - 7 P53 Favorável –11 P53 não favorável – 1

	Invasivo		Sem Informação - 27 P53 Favorável – 79 P53 não Favorável – 17
	In Situ	Ki67	Sem Informação – 7 Ki67 Favorável – 12 Ki67 não Favorável – 0
Tipo de Carcinoma	Invasivo		Sem Informação - 27 Ki67 Favorável – 68 Ki67 não Favorável – 28
	In Situ	IDNA	Sem Informação - 8 Diplóide – 5 Aneuplóide – 4 Tetraplóide – 2 Multiplóide Aneuplóide - 0
	Invasivo		Sem Informação - 37 Diplóide – 52 Aneuplóide – 26 Tetraplóide – 5 Multiplóide Aneuplóide - 3
	In Situ	FASE S	Sem Informação – 8 Fase S Favorável – 7 Fase S não Favorável – 4
	Invasivo		Sem Informação - 46 Fase S Favorável – 60 Fase S não Favorável – 17

Tabela 19: Tipo de carcinoma (In Situ/Invasivo) *versus* covariáveis em estudo

A) Tipo de Carcinoma: In Situ/Invasivo *versus* Factores Intrínsecos

O objectivo deste estudo inicialmente, como já foi referido, era identificar as variáveis que influenciam significativamente o *tipo de carcinoma* mamário (*in situ/invasivo*) (*Y*), de modo a determinar factores de risco para o cancro da mama.

Como a informação relativa a um dos tipos de carcinomas é escassa, havendo apenas na base em estudo 19 casos de *carcinomas in situ* e não havendo informação completa acerca desses casos para as covariáveis em estudo, quando se dividem estes pouquíssimos casos pelas inúmeras categorias dos factores e as suas combinações obtemos uma série de tabelas com várias entradas nulas ou próximas disso o que dificulta a análise pretendida, complicando o processo de estimação. Gostariamos de poder fazer mais no sentido de atingir o objectivo principal mas os dados de que dispomos não se revelaram suficientes. De salientar que esta metodologia poderia ter sido usada com dados mais completos, contudo, fica como um possível trabalho a realizar futuramente a tentativa de completar a base de dados a fim de aplicar esta metodologia e obter resultados fiáveis e mais interessantes.

Contudo, na tentativa de aplicação da metodologia dos modelos lineares generalizados (GLM) a esta base de dados e com o objectivo de realizar um estudo que permita relacionar o tipo de

carcinoma com factores intrínsecos aos indivíduos em estudo somos levados a seleccionar para o modelo inicial covariáveis como *idade, distrito e ruralidade.*

Selecção de covariáveis e de Modelos

Para a selecção do melhor modelo foi usado o método de selecção *stepwise* via AIC. Na Tabela 20 encontram-se os valores do critério de informação de Akaike, AIC (*Akaike Information Criterion*). Um bom modelo deverá ter um valor baixo de AIC.

MODELO	AIC
M1	121.14
M2	117.23
M3	**116.45**
M4	**114.47**
M5	**112.51**

Tabela 20: Modelos e respectivos valores de AIC

M1 – Tipo de carcinoma ~ idade + distrito + rural_urbano + idade*distrito + idade*rural_urbao+distrito*rural_urbano
M2 – Tipo de carcinoma ~ idade + distrito + rural_urbano + idade*rural_urbano
M3 – Tipo de carcinoma ~ idade + distrito + rural_urbano
M4 – Tipo de carcinoma ~ idade + distrito
M5 – Tipo de carcinoma ~ distrito

De salientar que a idade não se mostrou significativa em nenhum dos modelos analisados sendo que quando comparado o modelo M5 com o modelo em que as variáveis consideradas seriam idade e distrito o valor de AIC era superior (114.47) pelo que se excluiu também esse modelo. Em suma, o modelo que deve ser seleccionado deve ser M5 apesar de não se puder considerar um bom modelo, tendo em conta os dados em causa trata-se do melhor modelo e pode escrever-se na forma

$$Tipo\ de\ Carcinoma \sim B(p)$$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 Distrito$$

Procede-se, de seguida, ao ajustamento deste modelo pelo método da máxima verosimilhança.
Foi ainda obtida a estimativa do parâmetro de dispersão para o modelo seleccionado, de valor 1.014 indicando que não se verifica sobredispersão do modelo seleccionado.
As estimativas dos parâmetros de regressão, os respectivos erros padrões (*SE*) e *p-values* associados à sua significância, encontram-se na Tabela 21.

Parâmetro	Estimativa	Erro Padrão	z - value	p-value
β_0	-2.47	0.47	-5.299	1.17×10^{-7}
Distrito – Évora	0.95	0.55	1.72	0.0855

Tabela 21: Estimativas, erros padrões e p-values

Adequação do Modelo

Depois de seleccionado aquele que se julga ser o "melhor modelo", deve-se perguntar se o modelo é adequado, isto é

H_0: *O Modelo adequa-se aos dados*

H_1: *O Modelo não se adequa aos dados*

Para tal podem realizar-se algumas análises, das quais se destacam e foram aqui utilizadas: a percentagem de *deviance* explicada, o erro de predição e as curvas de ROC.

Determinou-se ainda o erro de predição, que resultou numa proporção de "acertos" de 86.62%, de onde se conclui que os valores ajustados são superiores a 0.5, o que leva a afirmar que o modelo se adequa bem aos dados. Quanto à curva de ROC (apresentada na Figura 39) na qual se obtém AUC = 0.417, e, que de acordo com *Fawcett, T. (2006)* se trata de um valor muito pior que um valor de AUC = 0,5 e que indica que esta curva pode ter informação útil mas está a aplicar informação incorrecta, pelo que este não parece ser um teste viável nesta situação.

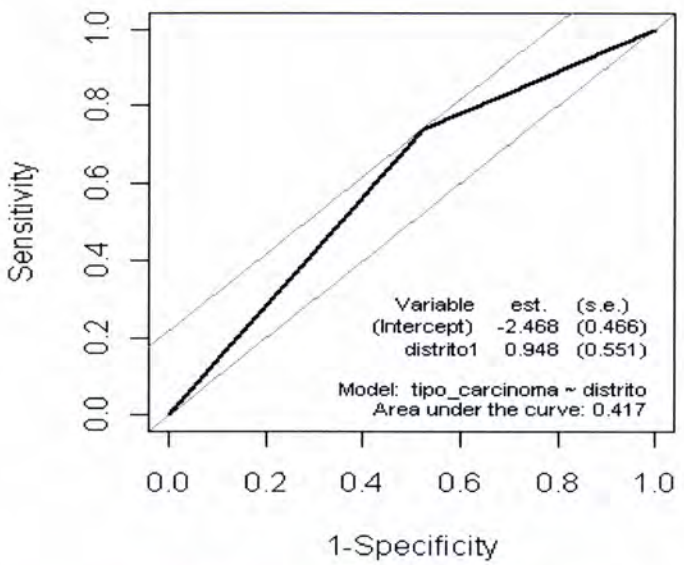


Figura 39: Curva de ROC do modelo seleccionado

Tendo em conta a informação sobre o modelo escolhido e estimado contida na Tabela 21, apresentada abaixo, pode fazer-se um teste da razão de verosimilhança à bondade do ajustamento do modelo, que resulta num *p-value* elevado (0.99) indicando que não há evidência de falta de ajuste. Apesar disso acontecer não significa que este modelo seja o correcto ou que não existam modelos melhores.

Desvio Nulo: **111.77 com 141 graus de liberdade**
***Desvio Residual*: 108.51 com 140 graus de liberdade**

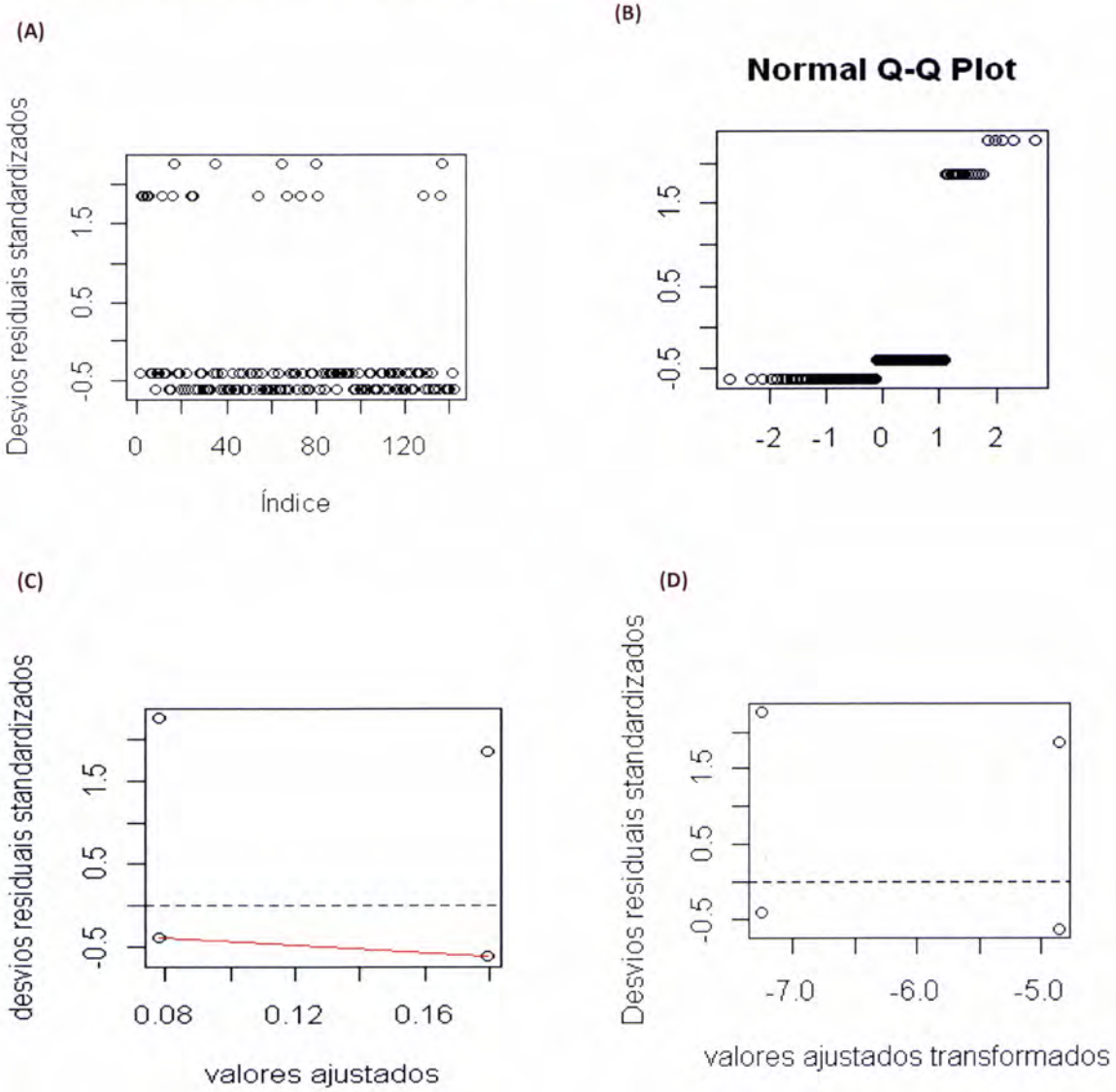
Tabela 22: Informação dos desvios do modelo seleccionado

Utilizando os valores da Tabela 22, pode-se encontrar uma percentagem explicada da *deviance* como medida para avaliar a adequação do modelo, que deu cerca de 2,9%, valor muitíssimo baixo.

Análise Informal dos Resíduos

Na adequação do modelo podemos encontrar anomalias, tanto na componente aleatória do modelo, como na componente sistemática, as quais podem ser detectadas através de uma análise informal dos resíduos, usando representações gráficas adequadas.

Na Figura 40 apresentam-se 6 representações gráficas: gráfico aos desvios residuais standardizados, gráfico de probabilidades normal para os resíduos (*Normal Q-Q Plot*), gráficos dos resíduos *versus* valores ajustados, gráfico dos resíduos *versus* valores ajustados transformados, gráfico das distâncias de *Cook* e gráfico dos valores de *Leverage*.



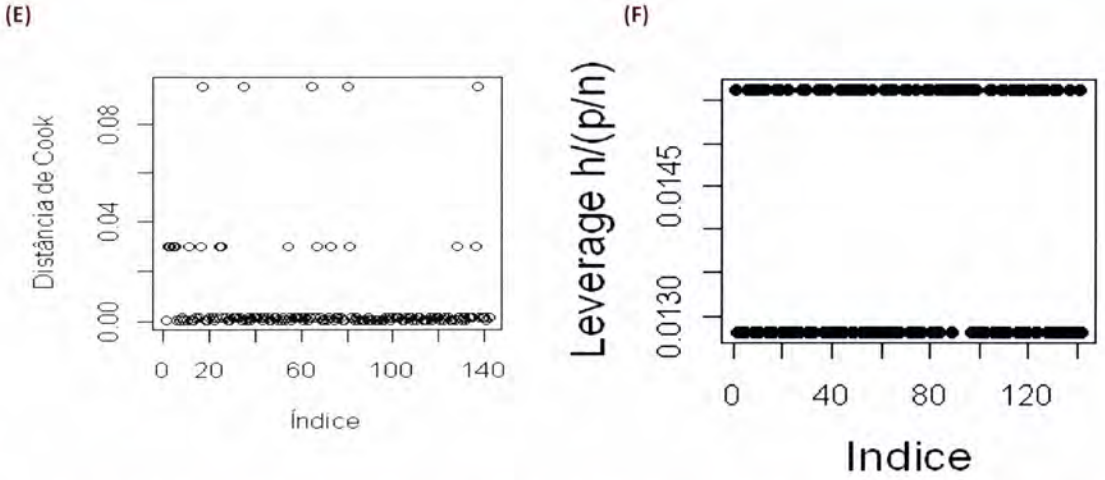


Figura 40:Gráficos (A) Desvios residuais Standardizados, (B) Normal Q-Q Plot dos Resíduos (C) Desvios residuais *versus* valores ajustados, (D) Desvios residuais *versus* valores ajustados transformados (E) Distância de Cook (F) Leverage

Dos gráficos (A) e (C), relativos aos resíduos standardizados e aos resíduos standardizados *versus* valores ajustados, observa-se que 95% dos resíduos se encontram compreendidos entre -2 e 2, o que significa que há validação dos pressupostos do modelo. A Figura 40 (B) onde se observa que os resíduos não são normais e nem tinham de o ser, apenas se encontra presente para mostrar, de uma forma grosseira, que para amostras grandes (como é o caso) pode ser de esperar uma espécie de normalidade aproximada dos resíduos. Tentando encontrar uma espécie de curvatura no gráfico que pudesse ser indicativa de vários factores de erro, procedeu-se à transformação dos valores ajustados, sugerida por *MacCullagh & Nelder (1989)* e já referida na seccção 4.1.8, de onde surge o gráfico da Figura 40 (D) que nos permite verificar que a transformação não trouxe surpresas. O gráfico da Figura 40 (C) permite ainda realizar uma análise informal da função de variância, de onde se conclui que não existiu uma má escolha da função de variância, pois o gráfico não apresenta nenhum tipo de tendência.

É ainda importante analisar a existência de uma ou mais observações mal ajustadas pelo modelo, isto é, observações que não seguem o padrão das restantes observações, as chamadas observações discordantes. No gráfico dos valores de *Cook* (Figura 40 (E)) notam-se algumas possíveis observações discordantes, após a análise no sentido de as detectar, foram bastantes, senão mesmo todas, contudo, realizando a mesma análise por meio do gráfico das distâncias de *Leverage* (Figura 40 (F)) *não* se destacam quaisquer indivíduos como sendo possíveis pontos influentes.

B) Tipo de Carcinoma: In Situ/Invasivo *versus* Factores de Prognóstico

Neste caso de estudo o objectivo reside na identificação de factores de prognóstico para o cancro da mama. Posto isto, pegou-se então nas variáveis da base de estudo, que sejam usadas como **variáveis de prognóstico usualmente associadas a riscos aumentados** (*re, rp, ki67, p53, fase_s,*

idna e *c.erb-2*) de cancro da mama, que se tratou desde sempre, do objectivo principal deste estudo. Como tal, seleccionaram-se como covariáveis para o modelo *re, ki67, p53, idna e fase_s, c.erb-2* e *as suas iteracções*. Desta feita, da base de dados inicial seleccionou-se uma amostra de 142 indivíduos em que 123 destes padeciam de carcinoma invasivo da mama e 19 de carcinoma *in situ*. Continuamos com os problemas já referidos inerentes ao tipo de dados que possuímos o que, de certo modo, dificulta o processo de estimação. Contudo, na tentativa de mais uma aplicação da metodologia dos modelos lineares generalizados (GLM) não poderíamos deixar para trás a tentativa de confirmação de variáveis como preditoras de diagnóstico.

Selecção de covariáveis e de Modelos

Para a selecção do melhor modelo foi usado o método de selecção *stepwise* via AIC. Na Tabela 23 encontram-se os valores do critério de informação de Akaike, AIC (*Akaike Information Criterion*). Um bom modelo deverá ter um valor baixo de AIC.

MODELO	AIC
M1	124.17
M2	120.26
M3	**113.80**
M4	**110.35**

Tabela 23: Modelos e respectivos valores de AIC

M1-Tipo de carcinoma~re+rp+c.erb.2+p53+fase_s+idna+ki67+re*c.erb.2+re*fase_s+re*idna+rp*fase_s+c.erb.2*p53+c.erb.2*fase_s+c.erb.2*idna+p53*fase_s+fase_s*idna
M2 – Tipo de carcinoma ~ re+rp+p53+ki67+c.erb.2+fase_s+idna
M3 – Tipo de carcinoma ~ re+c.erb.2
M4 – Tipo de carcinoma ~ re+c.erb.2+ki67+re*c.erb.2

Em suma, o modelo que deve ser seleccionado deve ser M4 apesar de não se puder considerar um bom modelo, tendo em conta os dados em causa trata-se do melhor modelo e pode escrever-se na forma

$$Tipo\ de\ Carcinoma \sim B(p)$$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 re + \beta 2 c.erb.2 + \beta 3 ki67 + \beta 4 re * c.erb.2$$

Procede-se, de seguida, ao ajustamento deste modelo pelo método da máxima verosimilhança.
Foi ainda obtida a estimativa do parâmetro de dispersão para o modelo seleccionado, de valor 0.9563 indicando que não se verifica sobredispersão do modelo seleccionado.

As estimativas dos parâmetros de regressão, os respectivos erros padrões (*SE*) e *p-values* associados à sua significância, encontram-se na Tabela 24.

Parâmetro	Estimativa	Erro Padrão	z - value	p-value
β_0	-1.4475	0.4335	-3.339	0.00084
$\beta 1 - re$	-2.5974	1.0487	-2.477	0.01326
$\beta 2 - c.erb.2$	-1.0730	0.6780	-1.583	0.11351
$\beta 3 - ki67$	1.7034	0.9744	1.748	0.08043
$\beta 4 - re * c.erb.2$	0.9409	0.4095	2.297	0.02159

Tabela 24: Estimativas, erros padrões e p-values

De salientar que nos modelos M1, M2 e M3 nenhuma das variáveis é significativa e no modelo M4 apresentaram-se significativas as variáveis *re* e *re*c.erb.2* como se pode verificar por meio da Tabela 24.

Adequação do Modelo

Depois de seleccionado aquele que se julga ser o "melhor modelo", deve-se perguntar se o modelo é adequado, isto é

H_0: *O Modelo adequa-se aos dados*
H_1: *O Modelo não se adequa aos dados*

Para tal podem realizar-se algumas análises, das quais se destacam e foram aqui utilizadas: a percentagem de *deviance* explicada, o erro de predição e as curvas de ROC.

Determinou-se o erro de predição, que resultou numa proporção de "acertos" de 86.62%, de onde se conclui que os valores ajustados são superiores a 0.5, o que leva a afirmar que o modelo se adequa bem aos dados. Quanto à curva de ROC (apresentada na Figura 41) na qual se obtém AUC= 0.710, e, que de acordo com Hosmer e Lemeshow indica que o modelo apresenta uma discriminação aceitável.

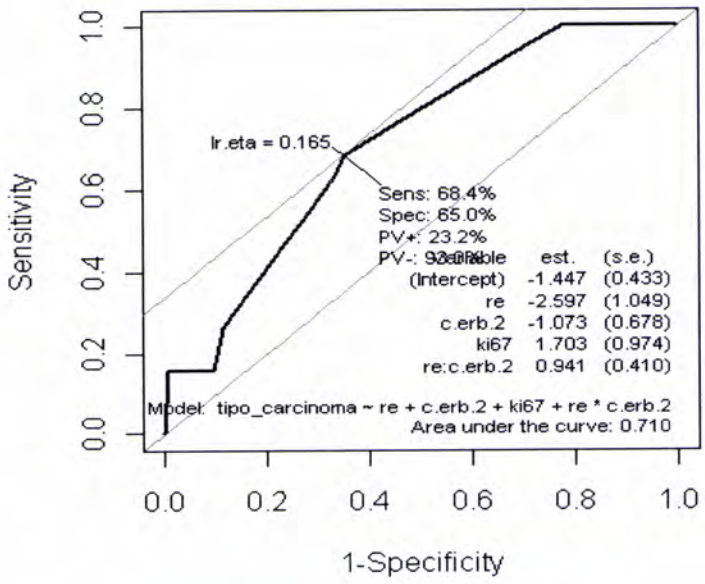


Figura 41: Curva de ROC do modelo seleccionado

Tendo em conta a informação sobre o modelo escolhido e estimado contida na Tabela 25, apresentada abaixo, pode fazer-se um teste da razão de verosimilhança à bondade do ajustamento do modelo, que resulta num *p-value* elevado (0.992) indicando que não há evidência de falta de ajuste. Apesar disso acontecer não significa que este modelo seja o correcto ou que não existam modelos melhores.

Desvio Nulo: **111.77 com 141 graus de liberdade**
***Desvio Residual*: 100.35 com 137 graus de liberdade**

Tabela 25: *Informação dos desvios* do modelo seleccionado

Utilizando os valores da Tabela 25, pode-se encontrar uma percentagem explicada da *deviance* como medida para avaliar a adequação do modelo, que deu cerca de 10,2%, valor baixo.

Análise Informal dos Resíduos

Na adequação do modelo podemos encontrar anomalias, tanto na componente aleatória do modelo, como na componente sistemática, as quais podem ser detectadas através de uma análise informal dos resíduos, usando representações gráficas adequadas.

Na Figura 42 apresentam-se 6 representações gráficas: gráfico aos desvios residuais standardizados, gráfico de probabilidades normal para os resíduos (*Normal Q-Q Plot*), gráficos dos

resíduos *versus* valores ajustados, gráfico dos resíduos *versus* valores ajustados transformados, gráfico das distâncias de *Cook* e gráfico dos valores de *Leverage*.

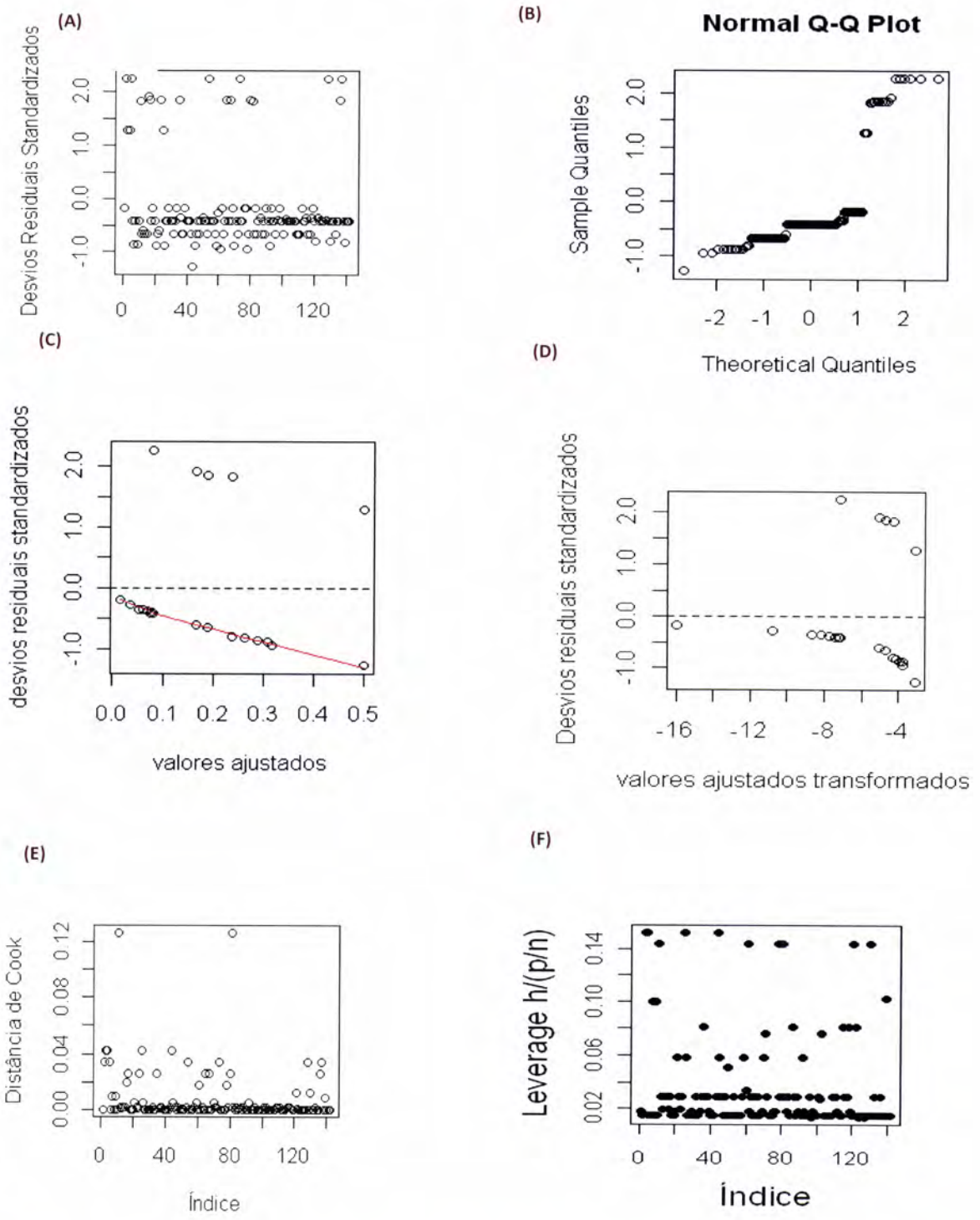


Figura 42:Gráficos (A) Desvios residuais Standardizados, (B) Normal Q-Q Plot dos Resíduos (C) Desvios residuais *versus* valores ajustados, (D) Desvios residuais *versus* valores ajustados transformados (E) Distância de Cook (F) Leverage

Dos gráficos (A) e (C), relativos aos resíduos standardizados e aos resíduos standardizados *versus* valores ajustados, observa-se que 95% dos resíduos se encontram compreendidos entre -2 e 2, o que significa que há validação dos pressupostos do modelo. A Figura 42 (B) onde se observa que os resíduos não são normais e nem tinham de o ser, apenas se encontra presente para mostrar, de uma forma grosseira, que para amostras grandes (como é o caso) pode ser de esperar uma espécie de normalidade aproximada dos resíduos. Este gráfico permite ainda observar a adequação do modelo e a existência de observações atípicas.Tentando encontrar uma espécie de curvatura no gráfico (C) que pudesse ser indicativa de vários factores de erro, procedeu-se à transformação dos valores ajustados, sugerida por *McCullagh & Nelder (1989)* e já referida na seccção *4.1.8*, de onde surge o gráfico da Figura 42 (D) que nos permite verificar que a transformação não trouxe surpresas. O gráfico da Figura 42 (C) permite ainda realizar uma análise informal da função de variância, de onde se conclui que não existiu uma má escolha da função de variância, pois o gráfico não apresenta nenhum tipo de tendência.

É ainda importante analisar a existência de uma ou mais observações mal ajustadas pelo modelo uma vez que grandes resíduos podem prejudicar a precisão da regressão. Pelo que é relevante detectar observações que não seguem o padrão das restantes observações, as chamadas observações discordantes. No gráfico dos valores de *Cook* (Figura 42 (E)) notam-se algumas possíveis observações discordantes, após a análise no sentido de as detectar salientaram-se os indivíduos 3, 4, 11, 25, 44 e 81. Realizando uma breve pesquisa acerca deste indivíduos verifica-se que se tratam de indivíduos que apresentam todos eles intensa coloração circunferencial da membrana em mais de 10% das células cancerosas e anel de coloração da membrana espesso no que respeita à análise de *c.erb.2* (o que acontece apenas para 13 dos casos em estudo) e para além disso, todos os casos apresentam valor de *ki67* favorável ao cancro da mama (80 dos casos em estudo encontram-se nessa situação), tratando-se de 5 casos de *carcinoma invasivo* e apenas 1 caso de *carcinoma in situ.* Constata-se dos dados em mãos que na grande maioria dos *carcinomas invasivos* não se observa coloração aquando da análise de *c.erb.2* (72 casos) e que neste caso o valor de *ki67* é maioritariamente favorável ao cancro da mama (68 casos). No que se refere ao grupo dos *carcinomas in situ,* 5 deles apresentam intensa coloração aquando da análise de *c.erb.2* e 12 deles *ki67* favorável ao cancro da mama.

Realizando a mesma análise por meio do gráfico das distâncias de *Leverage* (Figura 42 (F)) destacam-se estes e outros indivíduos como sendo possíveis pontos influentes, contudo, como os valores dos resíduos não são elevados não há razões para preocupações.

Resultados e pesquisa bibliográfica

Tendo em conta os resultados obtidos neste segundo caso de estudo verificou-se que existe correlação entre muitas das variáveis em causa, muitos valores faltosos em outras e talvez uma má

escolha das variáveis para o alcance do objectivo inicial.

Relativamente às relações de dependências entre determinadas variáveis, as quais se podem verificar por meio da Tabela 27 presente no *Anexo III*, e também aos casos de estudo presentes neste trabalho, estes podem ser de, certa forma, consolidados com alguns estudos já realizados, os quais se passam a citar:

- *Hedley, D. et al (1987)* sugeriram que o significado prognóstico da percentagem da *fase S* estava relacionado com o *grau do tumor,* tal como se verificou por meio do *teste do qui-quadrado* presente na Tabela 27 (p-*value*=$9{,}758 \times 10^{-16}$).

- *Soomro, S. et al (1991)* obtiveram resultados sugestivos de que no *carcinoma da mama invasivo* a imunocoloração para *c.erb.2* é vista principalmente para um subgrupo de tumores – os ductais, e que em quase todos os outros *tipos histológicos*, especialmente aqueles que estão associados a um bom prognóstico, apresentam falta de expressão desta variável. No estudo em causa aquando da análise de *c.erb.2* não se observou coloração na membrana em 72 dos 123 casos de *carcinoma invasivo*. Relativamente aos casos de *carcinomas in situ* (19 casos) de 7 deles não existe informação acerca da expressão da variável *c.erb.2* e em 6 deles não se verificou coloração.

- Segundo *Clark, G.M. et al (1993)* a presença de *receptores de estrogénio(re)* no cancro da mama é hoje aceite como um indicador de sobrevivência prolongada livre da doença. Os dados desse estudo sugerem que a determinação da concentração de *receptores de progesterona (rp)* é de valor igual ou maior do que a determinação de concentração de *receptor de estrogénio* para predizer a sobrevivência livre de doença dos pacientes com cancro da mama.

No último caso de estudo deste trabalho verificou-se que *re* é uma variável significativa de prognóstico na identificação do *tipo de carcinoma*, sendo que 79 dos *carcinomas invasivos* e 10 dos *carcinomas in situ* apresentaram valores de *re* favoráveis ao cancro da mama.Ainda em relação a esta variável destaque-se o facto de esta estar contemplada no modelo seleccionado aquando da tentativa de encontrar factores de prognóstico para o cancro da mama, apresentando-se esta significativa como factor de prognóstico.

Relativamente à determinação de concentração de *rp* esta apresentou-se de valor quase igual ao *re* uma vez que 65 dos *carcinomas invasivos* e 10 dos *carcinomas in situ* apresentaram valores de *rp* favoráveis ao cancro da mama, apesar desta variável não se ter mostrado significativa como possível variável de prognóstico de cancro da mama.

- *Rzymowska, J. (2004)* relata que conteúdo de *DNA* e da *fracção S* parece ser um conjunto de indicadores de grande valor prognóstico no cancro da mama ao qual *Munteanu, D. et al (2004)*

acrescentam que a *ploidia de DNA* pode ser um importante factor importante para estimar a agressividade do tumor.

Relativamente a esta análise e confrontando esta com a análise que se apresenta neste trabalho verifica-se que grande parte dos *carcinomas invasivos* são *diplóides* (52 casos), apresentam *fase_s* favorável ao cancro da mama (60 casos) e apresentam *médio ou alto grau histológico* (100 casos); em relação aos *carcinomas in situ* 5 deles são *diplóides*, 8 apresentam valores de *fase_s* favoráveis ao cancro da mama e 10 apresentam *médio ou alto grau histológico*. Em relação aos dados em estudo nenhuma das variáveis *fase_s, idna e grau_histológico* se apresentaram como factores significativos de prognóstico do cancro da mama.

- *Callahan, R. et al (1993)* indicam que em *carcinomas de mama invasivos* alterações do *p53* são vistas principalmente em tumores medulares e ductal, e que os outros tipos histológicos, especialmente aqueles associados a um elevado nível de diferenciação e de prognóstico favorável, mostram uma muito baixa incidência de mutações do gene *p53* e *Gasco, M. et al (2002)* referem que a análise da expressão de *p53* pode ter valor no diagnóstico, avaliação prognóstica e tratamento do cancro da mama.

No caso em análise 109 dos *carcinomas invasivos* são *ductais* sendo que 68 deles apresentam alterações de *p53* favoráveis ao cancro da mama, não se mostrando esta variável significativa na avaliação prognóstica deste tipo de neoplasia.

- De acordo com os resultados do estudo de *Eisenberg, A. Koifman, S. e Rezende, L. (2001)* as variáveis *idade* da paciente à data do diagnóstico, *grau histológico, rp e p53* foram factores preditivos para a positividade dos *re*, enquanto a *idade das pacientes, idade da menarca e grau histológico* o foram para os *rp*.

Nas análises efectuadas verificou-se 89 carcinomas com alterações de *re* favoráveis ao cancro da mama mostrando-se estas independentes da *idade* dos pacientes e dependente das variáveis *grau histológico, rp e p53*. Já em relação à variável *rp*, 75 dos carcinomas apresentaram valores desta variável favoráveis ao cancro da mama, mostrando-se *rp* independente da *idade* dos pacientes e dependente do *grau histológico*. Relativamente à *idade da menarca* nada se pode referir uma vez que esta variável não se encontra presente neste estudo.

- Tendo em conta o estudo realizado por *Oliveira, A. et al (2004)* sugere-se que a expressão de *c.erb.2* mostrou-se estatisticamente significativa nas lesões proliferativas de risco (*CDIS*) e correlacionou-se com características histopatológicas *(alto grau, presença de necrose, etc.)*.

No presente estudo a variável *c.erb.2* não se mostrou significativa no prognóstico do cancro da mama, contudo, a iteracção desta com *re* já foi significativa, pode ainda acrescentar-se que dos 19

carcinomas *in situ* em 12 deles *c.erb.2* mostrou-se como um possível factor de prognóstico, revelando-se esta correlacionada com *grau histológico.*

- *Karanikas, G. et al (2010)* referiram que a expressão de *ki67* se mostrou correlacionada com o *grau e tamanho tumoral* bem como com *p53 e c.erb.2*, o mesmo se verificando no presente estudo, tal como se pode verificar por meio da Tabela 27. Além disso, *ki67* mostrou-se significativamente relacionado com o *status nodal* e inversamente associado com a *expressão hormonal.* Acrescente-se o facto de que os *carcinomas invasivos* pareciam ter valores maiores quando comparados com *carcinomas de proliferação in situ*, enquanto os *carcinomas ductais* foram correlacionados com maior expressão de *ki67* em relação ao *cancro lobular.*

Relativamente a estes últimos dados na tentativa de os confrontar com a análise em questão apenas se pode dizer que 60 dos *carcinomas ductais invasivos*, 12 dos *carcinomas in situ* e 8 dos *carcinomas lobulares invasivos* apresentaram valores de *ki67* favoráveis ao cancro da mama. Destaque-se ainda o facto de esta variável fazer parte do modelo seleccionado aquando da tentativa de encontrar os factores de prognóstico para o cancro da mama apesar de não ser significativa.

Existem ainda outros estudos que apesar de não nos ser possível confrontar com o estudo presente neste trabalho podem ser aqui referenciados, tais como:

- *Holdaway, I.M. et al (1980)* referiram que pacientes com tumores *re* positivos tinham uma taxa de resposta global significativamente maior de hormonoterapia do que aqueles com tumores *re* negativos o que não se pode verificar por meio dos dados e análises presentes neste estudo.

- *Chavéz-Uribe, E. et al (2002)* citaram que a presença de tecido tumoral com conteúdo de *DNA* anómalo é associado com uma menor sobrevivência entre as mulheres cujo cancro tenha sido tratado por mastectomia.

Capítulo 5

Considerações Finais

O cancro da mama é o tipo de cancro mais comum entre as mulheres e corresponde à sua segunda maior causa de morte, tratando-se de uma das doenças com maior impacto na nossa sociedade. Detectar precocemente o cancro da mama aumenta as hipóteses de cura, pelo que o diagnóstico precoce desta doença é fundamental para que o cancro não se espalhe para outras zonas do corpo, favorecendo o prognóstico, a recuperação e a reabilitação. Era objectivo principal deste estudo detectar factores de risco para o cancro da mama, tendo em conta a região do Alentejo e o período compreendido entre Agosto de 2003 e Agosto de 2004. A primeira tarefa para a realização deste estudo consistiu em construir a base de dados a analisar posteriormente. Para tal, foram realizadas várias reuniões com a *Unidade de Anatomia Patológica do Hospital do Espirito Santo de Évora* e este trabalho de pesquisa de informação tornou-se moroso e, de certo modo, os resultados obtidos ficaram um pouco aquém das expectativas, uma vez que existem bastantes dados faltosos no registo para o efeito e, também, porque a base de dados que serviu de consulta para a criação da base de dados em estudo não contemplava variáveis que, de certeza, seriam mais interessantes para a concretização do objectivo proposto. Posto isto, avançou-se no sentido do objectivo principal, passando-se de seguida a uma análise preliminar dos dados obtidos e a partir dos quais se retiraram algumas conclusões já esperadas. Posteriormente, foram aplicadas várias metodologias estatísticas aos dados em questão, realizando-se um estudo caso-controlo que nos permitiu verificar que a *exposição ao ambiente rural* eleva ligeiramente o *risco de cancro da mama*, talvez porque, segundo o *PNS* o rastreio do cancro da mama ainda não chegou a todo o país, ficando, possivelmente, algumas das zonas rurais alentejanas nessa situação. De salientar, contudo, que apesar desse resultado se ter verificado, o mesmo não se mostrou significativo talvez porque, considerando a região do Alentejo não existe muita diferença entre as zonas rurais e as zonas urbanas. Por falta de informação que poderia ser significativa num estudo caso-controlo não pôde realizar-se outros estudos do género. Tendo em conta que o objectivo principal deste estudo se prende com a detecção de factores de risco para o cancro da mama aplicou-se de seguida a metodologia estatística mais indicada e utilizada nessas situações: *Os modelos lineares generalizados, mais propriamente, a regressão logística.* Para tal foram criadas duas bases de dados a partir da base de dados inicial em que se considerou numa delas a variável resposta o *tipo de neoplasia (benigna/maligna)* e uma outra em que a variável resposta considerada foi o *tipo de carcinoma (in situ/invasivo).* Depois de aplicada a metodologia escolhida nos correspondentes dois conjuntos de dados concluiu-se que:

- Tendo em conta a primeira base de dados o único factor de risco para o cancro da mama que se verificou significativo foi a *idade*. Esta análise ficou aquém das expectativas, uma vez mais porque não existiam informações acerca de variáveis que poderiam ser mais pertinentes e interessantes neste estudo;
- No que se refere à segunda base de dados, foram realizadas duas análises: uma primeira em que se tentaram encontrar factores de risco para o *tipo de carcinoma (invasivo/in situ)* tendo em conta apenas *factores intrínsecos* aos pacientes e chegou-se ao melhor modelo que apenas incluía a variável *distrito* apesar de *não significativa*; uma segunda análise, em que foram tidas em conta apenas *variáveis consideradas no prognóstico de cancro da mama*, no sentido de perceber quais as mais importantes na detecção do *tipo de carcinoma* e aqui as coisas tornaram-se mais complicadas, uma vez que a informação relativa a um dos tipos de carcinomas é escassa, havendo na base apenas *19 casos de carcinomas in situ, faltando ainda alguma informação acerca destes casos*, o que leva a obtenção de uma série de tabelas com várias entradas nulas ou perto disso, quando se dividem esses escassos casos pelas inúmeras categorias dos factores e as suas combinações. Contudo, depois de aplicar a teoria da regressão logística a esta segunda base de dados, chegou-se ao melhor modelo que apenas incluía as variáveis receptor de estrogénio (*re*), os biomarcadores de prognóstico (*c.erb.2, ki67*) e a interacção receptor de estrogénio - *re* com a expressão da proteína - *c.erb.2* como factores de risco, mostrando-se *significativas* apenas o receptor de estrogénio e a interacção referida, isto é, *re e re*c.erb.2.*

Como conclusão pode dizer-se que muito mais gostaríamos de fazer no sentido de alcançar o objectivo principal retirando conclusões pertinentes, mas a verdade é que não possuíamos dados suficientes para o efeito desejado. Apesar de não se ter conseguido alcançar tão bons resultados quanto o desejado investiu-se bastante na aprendizagem das metodologias estatísticas para o fazer e este trabalho fica como base para que *a posteriori* (com esta base de dados aumentada ou outra) se possam obter conclusões mais interessantes.

Como referiu o *Dr.Luís Gonçalves (2010) - "A Medicina não é Matemática, podemos tentar arranjar modelos para explicar doença, mas nem sempre se consegue!"*

Capítulo 6

Referências Bibliográficas

1. Ahlbom, A. (1993). *Biostatistics for epidemiologists*. United States of America: Lewis Publisher.
2. Akaike, H. (1974). *A new look at the statistical model identification*. IEEE Trans. Automatic Control AC-**19**, 716-723.
3. Alberts, B., Bray, D., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P. (1998). *Essential cell biology: an introduction to the molecular biology of the cell*. New York & London: Garland Publishing, Inc.
4. Allen, N., Beral, V., Casabonne, D., Wan Kan, S., Reeves, G., Brown, A., Green, J. (2009). *Moderate Alcohol Intake and Cancer Incidence in Women*. J Natl Cancer Inst.; **101(5)**: 296-305.
5. Altman, D. (1991). *Practical statistics for medical research*. United States of America: Chapman & Hall.
6. Antunes, A., Silva, T., Godinho, I., Amaral, N., Oliveira, C. (2004). *Valor prognóstico da expressão por imuno-histoquímica do C-ERB-2 em doentes sob terapêutica adjuvante em Tamoxifeno por carcinoma primário da mama*. Acta Med Port; **17**: 271-276.
7. Bap-tiste, M.S., Field, N.A., Metzger, B.B., Black, M., Kwon, C.S., Jacobson, H. (1990). *An epidemiological case-control study of breast cancer and alcohol consumption*. International Journal of Epidemiology; **19**:532-538.
8. Bastos, J., Barros, H., Lunet, N.(2007). *Evolução da mortalidade por cancro da mama em Portugal (1955-2002)*. Acta Med Port; **20**: 139-44.
9. Bessaoud, F., Gerber, M. (2008). *Os factores dietéticos e risco de câncer de mama: estudo caso-controle entre uma população do sul de França*, revista Nutrição e Câncer, **60 (2)**, 177-187.
10. Bonita, R., Beaglehole, R., Kjellström, T. (2006). *Basic Epidemiology*, World Health Organization.
11. Braga, A. (2000). *Curvas ROC: Aspectos Funcionais e Aplicações*. Dissertação de Doutoramento no Ramo de Engenharia de Produção e Sistemas, Área de Métodos Numéricos e Estatísticos, Universidade do Minho.
12. Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*, vol.1 – *The analysis of case-control studies*.Lyon: International Agency for Research on Cancer Scientific Publications No.32.

13. Brown, Z., Boatman, K. (2008). *100 Questions & Answers about Breast Cancer.* Third Edition, Jones and Bartlett Publishers, LLC.
14. Callahan, R., Bistocchi, M., Marchetti, A., Buttitta, F., Pellegrini, S., Campani, D., Diella, F., Cecchetti, D. (1993). *p53 mutations and histological type of invasive breast carcinoma I,* Cancer Res. **53(19)**; 4665-9.
15. Chávez-Uribe, E., Viñuela, J., Cameselle-Teijeiro, J., Forteza, J., Puñal, J., Otero, J., Puente-Dominguez, J. (2002). *DNA ploidy and cytonuclear area of peritumoral and paratumoral samples of mastectomy speciemens: a useful prognostic marker?* European Journal of Surgery, **168 (1)**; 37-41.
16. Clark, G.M., McGuine, W.L., Hubay, C.A., Pearson, O.H., Marshall, J.S. (1993). *Progesterone receptors as a prognostic factor in stage II breast cancer.* The New England Journal of Medicine, **309(22)**; 1343-1347.
17. Clayton, D. & Hills, M. (1993). *Statistical Models in Epidemiology.* Oxford Science Publications.
18. Cook, R.D. (1977). *Detection of influential observations in linear regression.* Thecnometrics, **19**,15-18.
19. Cordeiro, G.M. (1986). *Modelos Lineares Generalizados.* VII Simpósio Nacional de Probabilidades e Estatística, Campinas, São Paulo.
20. Cufer, T., Lamovec, J., Bracko, M., Lindtner, J., Us- Krasovec, M. (1997). *Prognostic value of DNA ploidy in breast cancer stage I-II.* Neoplasma; **44**:127-32.
21. Dever, G.E., Champagne, F. (1984). *Epidemiology in health services management.* United States of America: Aspen Publishers, Inc.
22. Dixon, J.M. (2000). *Breast Cancer: Diagnosis and Management.* Elsevier Science B.V.
23. Duell, E., Mililikan, R., Savitz, D., Newman, B., Smith, J., Schell, M., Sandler, D. (2000). *A population based case-control study of farming and breast cancer in North Carolina.* JStor Epidemiology, **11(5)**; 523-531.
24. Ebrahimi, M., Vahdaninia, M., Montazeri, A. (2002). *Risk factors for breast cancer in Iran: a case-control study.* Breast Cancer Res; **4:R10**.
25. Eisenberg, A., Koifman, S., Rezende, L. (2001). *Hormone receptors: association with prognostic factors for breast cancer.* Rev. Bras. Cancerol; **47(1)**:49-58.
26. Everitt, B. (1992). *The analysis of contingency tables.* Second Edition, Chapman & Hall/CRC.
27. Fahrmeir, L. & Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Second Edition, Springer Series in Statistics.
28. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters **27**; 861-874.
29. Fenollera A. (2000). *Mastología, 2.ª edición.* Barcelona: Masson, S.A.
30. Fisher, R., Yates, F. (1938). *Statistical tables for biological, agricultural and medical research.* Edinburgh: Oliver & Boyd.

31. Friedenreich, C., Bryant, H., Courneya, K. (2001). *Case-control study of lifetime physical activity and breast câncer risk.,* American Journal of Epidemiology, **154(4).**
32. Garicochea, B., Morelle, A., Andrighetti, A., Cancella, A., Bós, A., Werutsky, G. (2009). *Age as a prognostic factor in early breast câncer.* Revista de Saúde pública; **43(2)**:311-7.
33. Gasco, M., Shami, S., Crook, T. (2002). *The p53 pathway in breast cancer.* Breast Cancer Research, **4**:70-76.
34. Gasparini, G., Gullick, W.J., Bevilacqua, P.,_Sainsbury J.R.C., Meli Salvatore, Baracchi, P., et al. (1992). *Human breast cancer: prognostic significance_of the c-erbB-2 oncoprotein compared_with epidermal growth factor receptor, DNA_ploidy, and conventioanl pathologic features.* J_Clin Oncol; **10**:686-95.
35. Gullick, W.J., Love, S.B., Wright, C., Barnes, D.M.,_Gusterson, B., Harris, A.L., et al. (1991). *c-erbB-2 protein_overexpression in breast cancer is a risk factor in_patients with involved and uninvolved lymph_nodes.* Br J Cancer; **63**:434-8.
36. Hedley, D., Rugg, C., Gelber, R. (1987). *Association of DNA index and S-phase fraction with prognosis of node positive early breast cancer.* Cancer Research **47**, 4729-4735.
37. Hennekens, C., Buring, J., Mayrent, S. (1987). *Epidemiology in medicine.* First Edition, United States of America: Lippincott Williams & Wilkins.
38. Hill, B. (1965). *The environment and disease: association or causation? Proceedigns of the Royal Society of Medicine*; **58**, 295-300.
39. Hoaglin, D.C., Welsch, R.E. (1978). *The Hat Matrix in Regression and ANOVA.* American Statistician; **32**:17-22.
40. Holdaway I.M., Mountjoy, K.G., Harvey, V.J., Allen, E.P., Stephens, E.J. (1980). *Clinical applications of receptor measurements in breast cancer.* Br J Cancer; **41(1)**:136-9.
41. Hosmer, D. & Lemeshow, S. (2000). *Applied Logistic Regression.* Second Edition, United States of America: John Wiley & Sons, Inc.
42. Hunt, K., Robb, G., Strom, E., Ueno, N. (2001). *Breast Cancer.* New York: Springer-Verlag, Inc.
43. Instituto Nacional de Estatística, *Revista de Estudos Demográficos*, n^{os} 32,34,36,38,40,42,44 e 46, Ano de Edição 2002-2009.
44. Ioachim, E., Kamina, S., Athanassiadou, S., Agnantis,_N.J. (1996). *The prognostic significance of epidermal growth factor receptor (EGFR), c-erbB-2, Ki-_67 and PCNA expression in breast cancer.* Anticancer_Res; **16**:3141-7.
45. Jameson, J. (1998). *Principles of molecular medicine.* New Jersey: Humana Press.
46. Jemal, A.,Siegel, R., Ward, E., Hao, Y., Xu, J., Murray, T.,Thun, M. (2008). *Cancer Statistics, 2008 – American Cancer Society.* CA Cancer J Clin; **58**: 71-96
47. Kahn, H., Sempos, C. (1989). *Statistical methods in epidemiology.* New York: Oxford University Press.

48. Karanikas, G., Koronakis ,N., Lagoudianakis, E.E., Grosomanidis, D., Karavitis, G., Koukoutsis, I., Pappas, A., Kotzadimitriou, K., Papadima, A., Chrysikos, J., Zografos, G., Xepapadakis, G., Manouras, A. (2010). *The value of proliferation indexes in breast cancer.* Eur J Gynaecol Oncol. **31 (2)**: 181-4

49. Keshgegian, A.A. (1995). *c-erbB-2 oncoprotein overexpression in breast carcinoma: inverse correlation with biochemically and immunohistochemically-determined hormone receptors.* Breast Cancer Res Treat; **35**:201-10.

50. Keyhani-Rofagha, S., O'Toole, R.V., Farrar, W.B., Sickle-Santanello, B., DeCenzo, J., Young, D. (1990). *Is DNA ploidy an independent prognostic indicator in infiltrative node-negative breast adenocarcinoma?* Cancer; **65**:1577-82.

51. Kopans, D. (2007). *Breast Imaging*. Third Edition, Philadelphia: Lippincoit Williams & Wilkins.

52. Kumar, V., Abbas, A., Fausto, N., Robbins, S., Cotran, R. (2005). *Robbins and Cotran Pathologic basis of disease*, 7th edition, Philadelphia, PA: Elsevier Saunders.

53. Last, J. (2001). *A dictionary of epidemiology*. 4th edition, Oxford: Oxford University Press

54. Lilienfeld, A.M., Lilienfeld, D.E. (1980). *Foundations of Epidemiology*, 2nd edition, Oxford University Press: Oxford.

55. Mareel, M., Baetselier, P., Van Roy, F. (1991). *Mechanisms of invasion and metastasis*. United States: CRC Press.

56. Marques, L. (2003). *Cancro da Mama*, Rev Port Clin Geral; **19**:463-68.

57. Mathew, A., Gajalakshmi, V., Rajan, B., Kanimozhi, V., Brennan, P., Mathew, B.S., Boffetta, P. (2008). *Anthropometric factors and breast cancer risk among urban and rural women in South India: a multicentric case-control study*, Br J Cancer, 8;99(1):207-13.

58. Mausner & Kramer (1984). *Introdução à Epidemiologia*. 3ªedição – Fundação Calouste Gulbenkian.

59. McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. 2nd edition, London: Chapman and Hall.

60. McNeil, D. (1996). *Epidemiological Research methods*. New York: John Wiley & Sons.

61. Morris, E., Liberman, L. (2005). *Breast MRI: diagnosis and intervation.* United States of America: Springer.

62. Munteanu, D., Zlei, M., Ailiesei, O., Chifu, C., Diaconu, C., Carasevici, E. (2004). *Flowcytometric evidence of dna ploidy in human breast cancer.* The Journal of preventive medicine; **12** (3-4): 59-65.

63. Nelder, J. A. & Wedderburn, R.W.M. (1972). *Generalized Linear Models*. J. Roy. Statist. Soc. Ser. **A 135**; 370-384.

64. Ogden, J. (1999). *Psicologia da Saúde*. Lisboa: Climepsi Editores (Tradução do original em inglês Health Psychology:A textbook. Buckingham: Open University Press,s.d.

65. Oliveira, A., Luca, L., Carvalho, G., Arias, V., Carvalho, L., Assunção, M.(2004). *Imunoexpressão do c-erb-2 nas lesões epiteliais proliferativas intraductais da mama de mulheres*. Rev. Assoc. Med. Bras. **50(3)** São Paulo.

66. Organização Mundial de Saúde (1993). *Classificação Estatística Internacional de doenças e de problemas relacionados à Saúde, 10ª Revisão (CID-10)*, em utilização desde 1999.

67. Page, R., Cole, G., Timmreck, T. (1995). *Basic epidemiological methods and biostatistics: a practical guidebook*. Canada: Jones and Bartlett Publishers.

68. Ricks, D. (2005). *Breast cancer basics and beyond: treatment, resources, sef-help, good news, updates*. First Edition, United States of America: Bang Printing.

69. Robert, S., Strombom, I., Trentham-Dietz, A., Hampton, J., McElroy, J., Newcomb, P., Remington, P. (2004). *Socioeconomic risk factors for breast cancer:distinguishing individual- and community-level effects*. Epidemiology, **15(4)**;442-450.

70. Rzymowska, J. (2004). *DNA index in breast cancers*. Journal Methods in Cell Science, **18 (1)** 1-5.

71. Silverstein, M.J., Lagios, M.D., Craig, P.H., Waisman, J.R., Lewinsky, B.S., Colburn, W.J., Poller, D.N. (1996). *Prognostic Index for Ductal carcinoma in situ of the breast*. A Cancer 1; **77(11)**: 2267-74.

72. Slamon D., Clark G., Wong S., Levin W., Ullrich A., McGuire W. (1987). *Human Breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene*. Science; **235**:177-182.

73. Sobin, L., Rowicz, M., Wittekind, C. (2009). *Atlas TNM Classification of Malignant Tumours*, seventh edition, International Union Against Cancer, Blackwell Publishing Ltd.

74. Soomro, S., Shousha, S., Taylor, P., Shepard, H., Feldman, M. (1991). *C-erb-2 expression in different histological types of invasive breast carcinoma*. J Clin Pathol; **44**:211-214.

75. Streiner, D., Norman, G. (1998). *PQD epidemiology*. United States of America: BC Decker.

76. Taylor, I. & Knowelden, J. (1957) *Principles of epidemiology*. London, Churchill Ltd.

77. Tessaro, S., Béria, J., Tomasi, E., Barros, A. (2001). *Contraceptivos orais e câncer de mama: estudo de casos e controles*, Rev. Saúde Pública, vol.35 no.1 São Paulo.

78. Todd, J.H., Dowle, C., Williams, M.R., Elston, C.W., Ellis, I.O., Hinton, C.P., Blamey, R.W., Haybittle J.L. (1987). *Confirmation of a prognostic index in primary breast cancer*. Br J Cancer 56(4): 489-92.

79. Travassoli, F. (1999). *Pathology of the Breast*. Second Edition, McGraw-Hill.

80. Turkman, A., Silva, G. (2000). *Modelos Lineares Generalizados – da teoria à prática* – DM/IST e CMA, Universidade Técnica de Lisboa.

81. Varela, S. (2002). *El cáncer: Etiologia, Epidemiologia, Diagnóstico y Prevencion*. Elsevier Science.

82. Walther, W., Stein, U. (2000). *Gene therapy of cancer: methods and protocols*. New Jersey: Humana Press.
83. Woolf, B. (1955). *On the estimating the relation between blood group and disease*. Ann. Hum Genet., **19**:251-3.
84. Yates, F. (1934). *Contingency tables involving small numbers and the* χ^2 *test*, Royal Statistical Society Supplement 1 (series B): 217-235.
85. Zheng, T., Holford, T. R., Mayne, S. T., Owens, P .H., Zhang, Y., Zhang, B., Boyle, P., Zham, S.H. (2001). *Lactation and breast cancer risk: a case-control study in Cunnecticut*, British Journal of Cancer, **84**: 1472-147.

Webgrafia

1. Alto Comissariado da Saúde - *Perfil de Saúde na Região do Alentejo – PNS em Foco, Especial Regiões.*Gabinete de Informação e Prospectiva, 2009. Disponível em: URL:http://www.acs.min-saude.pt
2. American Cancer Society, Medicines to Reduce Breast Cancer Risk, 2009 Disponível em: http://www.cancer.org/acs/groups/cid/documents/webcontent/002585-pdf.pdf
3. CiênciaHoje, Entrevista a Mina Bissel, 2008. Disponível em: http://www.cienciahoje.pt/index.php?oid=24870&op=all#cont
4. European Cancer Observatory, Cancer Fact Sheets – Cancer:Breast, 2006. Disponível em: http://eu-cancer.iarc.fr/cancer-13-breast.html,en
5. European Cancer Observatory, Cancer Fact Sheets – Country: European Union, 2006. Disponível em: http://eu-cancer.iarc.fr/country-930-european-union-27.html,en
6. GLOBOCAN 2008, International Agency for Research on Cancer, *Cancer Fact Sheet, Breast Cancer Incidence and Mortality Worldwide in 2008 summary*, 2008
7. InfoCancro. Disponível em: http://www.roche.pt/sites-tematicos/infocancro/
8. Liga Portuguesa Contra o Cancro. Disponível em: http://www.ligacontracancro.pt/
9. Ministério da Saúde, Direcção Geral de Saúde – Plano Nacional de Saúde (2004-2010) – Ganhos em Saúde, Avaliação de Indicadores no período de 2001-2005. 2009. Disponível em: URL:http://www.dgsaude.min-saude.pt
10. Organização Mundial de Saúde,Cancer – Fact Sheets, 2009. Disponível em: http://www.who.int/mediacentre/factsheets/fs297/en/index.html
11. Portal da Saúde, *O que é o cancro? Como preveni-lo?* , 2005. Disponível em: http://www.portaldasaude.pt/portal/conteudos/enciclopedia+da+saude/doencas/cancro/cancro.htm
12. Portal de Oncologia Português, 2009. Disponível em: URL:http://www.pop.eu.com

13. Risco de Morrer em Portugal 2003 (vol II) – Direcção Geral de Saúde. Disponível em: URL:http://www.gds.pt
14. Risco de Morrer em Portugal 2004 (vol II) – Direcção Geral de Saúde. Disponível em: URL:http://www.gds.pt
15. Sociedade Portuguesa de Senologia – SPS – 2008. Disponível em: http://www.spsenologia.pt/

Bibliografia e Webgrafia consultadas mas não referenciadas

1. Agresti, A. (2000). *Categorical Data Analysis* – Second Edition, New Jersey: Wiley & Sons, Inc.
2. Andreozzi, V. (2009). *Modelos Lineares Generalizados*. Apontamentos da disciplina de Métodos Estatísticos em Saúde II.
3. Aragón, T., Enanoria, W. (2007). *Applied Epidemiology Using R*. University of California, Berkeley, School of Public Health.
   Disponível em: https://home.comcast.net/~lthompson221/Splusdiscrete2.pdf
4. Dobson, A. (1990). *An Introduction to Generalized Linear Models*.United States of America: Chapman & Hall/CRC.
5. Knorr-Held, L. (2002). *MATH462: Principles of Epidemiology*. Disponível em: www.maths.lancs.ac.uk/~knorrhel/MATH462.html
6. Rosa, T. (2009). *Factores clínico-patológicos preditivos da metastização ganglionar axilar no cancro da mama.* Dissertação de Mestrado Integrado em Medicina, Universidade da Beira Interior, Faculdade de Ciências da Saúde.
7. Thompson, A.L. (2002). *S- Plus (and R) Manual to Accompany Agresti's Categorical Data Analysis*. Second edition.

Anexos

Anexo I - Comandos do R mais utilizados na análise

a) Instruções do R utilizadas na Análise Estatística Preliminar dos Casos em Estudo

- Criação de uma base de dados a partir de um documento txt e apresentação do nome das variáveis presentes na mesma.

>nome_base_dados <read.table("nome_ficheiro.txt",sep="\t",header=T, fill=T)
>names(nome_base_dados)<-tolower(names(nome_base_dados))

Exemplo:

```
>Caso<-read.table("G_Case.txt",sep="\t",header=T,fill=T)
>names(Caso)<-tolower(names(Caso))
```

- Sumário da base de dados criada

>summary(nome_base_dados)

Exemplo:

```
> summary(Caso)
```

- Dimensão da base de dados

>dim(nome_base_dados)

Exemplo:

```
> dim (Caso)
```

- Categorizando a variável idade – criação da variável idadecat tendo em conta a classificação por classes da variável idade

Exemplo:

```
> Caso$idadecat <-rep(1,dim(Caso)[1])
> Caso$idadecat[Caso$idade>=15&Caso$idade<25]<-1
> Caso$idadecat[Caso$idade>=25&Caso$idade<35]<-2
> Caso$idadecat[Caso$idade>=35&Caso$idade<45]<-3
> Caso$idadecat[Caso$idade>=45&Caso$idade<55]<-4
> Caso$idadecat[Caso$idade>=55&Caso$idade<65]<-5
> Caso$idadecat[Caso$idade>=65&Caso$idade<75]<-6
> Caso$idadecat[Caso$idade>=75&Caso$idade<85]<-7
> Caso$idadecat[Caso$idade>=85&Caso$idade<100]<-8
```

- Criação da Base de dados CASO a partir da base de dados Caso, tendo em conta a criação da variável tipo_histologicoBM conforme descrição desta última.

Exemplo:

```
> CASO<-Caso[!is.na(Caso$tipo_histologico),]
```

```
> CASO$tipo_histologicoBM <-rep(1,length (CASO$tipo_histologico))
> CASO$tipo_histologicoBM[CASO$tipo_histologico<=8] <-0
```

- Indicar ao R quais as variáveis categóricas

>nome_base_dados$nome_variável <factor (nome_base_dados$nome_variável)

Exemplo:

```
> CASOS$margens_cirurgicas<-factor(CASOS$margens_cirurgicas)
```

- Criação dos labels para uma variável

>nome_base_dados <transform(nome_base_dados, nome_variável = factor (nome_variável, label=c("c1",..."cn") levels =1:n))

Sendo:

- o vector c o vector das categorias da variável e portanto os valores de c1,...,cn deverão ser substituídos pelos valores das categorias da variável

- a parte levels=1:n refere-se ao número de níveis ou categorias da variável, sendo esta parte optativa

Exemplos:

```
>CASO<transform(CASO,idadecat=factor(idadecat,label=c("[15,25[","[25,35[","[35,45[","[45,55[","[55,65[","[65,75[" ,"[75,85[","[85,100["),levels=1:8))
>CASO<transform(CASO,ruralidade=factor(ruralidade,label=c("SemInformação","Rural","Urbano")))
```

- Tabela de frequências de uma variável

>tabular(nome_base_dados$nome_variável)

Exemplo:

```
> tabular(CASO$idadecat)
```

- Construção de tabela de Contingência entre duas variáveis e output da mesma

>nome_tabela=(nome_variável1,nome_variável2)
>nome_tabela

Exemplo:

```
> tab=table(idadecat,tipo_histologicoBM)
> tab
```

- Construção de um gráfico que mostre os dados relativos à tabela de contingência

>plot(nome_base_dados$nome_variável1,nome_base_dados$nome_variável2, main="título_gráfico",xlab="nome_variável1",ylab="nome_variável2",pch=19)

Exemplo:

```
>plot(CASO$idadecat,CASO$tipo_histologicoBM,main="idadecat*tipo_histologicoBM",xlab="idadecat",ylab="tipo_ histologicoBM",pch = 19)
```

- Eliminar as observações respeitantes a uma categoria de uma variável (construindo uma nova base de dados)

  >Nome_base_dados_nova <-nome_base_dados [!nome_base_dados$nome_variável =="nome_categoria_variável",]

  Exemplo:
  >Caso.limpo<-CASO[!CASO$ruralidade=="Sem Informação",]

- Eliminar o nível de uma variável (depois do passo anterior)

  >Nome_base_dados_nova$nome_variável<-factor(nome_base_dados$nome_variável)

  Exemplo:
  >Caso.limpo$ruralidade<-factor(Caso.limpo$ruralidade)

- Eliminar as observações respeitantes a uma categoria de uma variável (construindo uma nova base de dados)

  >Nome_base_dados_nova <-nome_base_dados [!nome_base_dados$nome_variável =="nome_categoria_variável",]

  Exemplo:
  >Caso.limpo<-CASO[!CASO$ruralidade=="Sem Informação",]

- "Entrar" numa base de dados

  >attach(nome_base_dados)

  Exemplo:
  >attach(Caso.limpo)

- Teste do Qui-Quadrado

  >chisq.test(nome_tabela)

  Exemplo:
  >chisq.test(tab)

- Teste de Fisher

  >fisher.test(nome_tabela)

  Exemplo:
  >fisher.test(tab)

- Construção de um gráfico do tipo "pie-chart"

>pie(table(nome_variável), main="Título do gráfico", col=rainbow(length (levels(nome_variável))), labels = levels (nome_variável)

Exemplo:

```
> pie(table(rep), main="Receptor de Progesterona", col=rainbow(length (levels(rp))),labels=levels(rp))
```

B) Instruções do R utilizadas no Estudo Caso-Controlo: Tipo de Neoplasia e Ruralidade

Utilizaram-se as mesmas instruções que no estudo dos casos, para o cálculo dos *odds ratio* utilizou-se a instrução do teste de Fisher

C) Instruções do R utilizadas no Estudo dos GLM

Depois de realizada uma análise preliminar à base de dados conforme as instruções já referidas, procedeu-se ao estudo dos GLM utilizando as seguintes instruções:

- Construção de um modelo linear generalizado

>nome_glm <-glm(variável dependente ~ variável_independente_1+variável_ independente_2+....+variável_independente_n, family=binomial, data=nome_base_dados)

Exemplo:

```
> glm1<-glm(benigno_maligno~idade+tipo_amostra,family=binomial, data=Caso_BM)
```

- Sumário de um modelo

>summary(nome_glm)

Exemplo:

```
> summary (glm1)
```

- Estimativa do Parâmetro de Dispersão no Modelo Binomial

>nome_glm <-glm(variável dependente ~ variável_independente_1+variável_ independente_2+....+variável_independente_n, family=quasebinomial, data=nome_base_dados)

- Modelo Saturado

>summary(glm (Variavel_dependente~factor(1:nrow(nome_base_dados)), data=nome_base_dados))

- Modelo Nulo

>summary(glm(variavel_dependente ~1, data=nome_base_dados))

- Realizar o Stepwise a um modelo

>step(nome_glm)

Exemplo:
> step(glm1)

- Comparação de dois modelos

>anova(nome_glm1,nome_glm2, test ="Chisq")

Exemplo:
> anova(glm1,glm2, test ="Chisq")

- Resíduos de Pearson

>res<-residuals(fit,type="pearson")

- Resíduos de Pearson Padronizados

>res<-rstandard (fit, type="pearson")

- Resíduos Deviance

>res<-residuals (fit, type="deviance")

- Resíduos Deviance Padronizados

>res<-restandard (fit, type="deviance")

- Intervalos de Confiança a 95% para os β's

>nome.glm
>nome.sum<-summary(nome.glm)
>nome.glm$coeff
>sqrt(diag(nome.sum$cov.scaled))
>nome.glm$coeff-1.96*sqrt(diag(nome.sum$cov.scaled))
> nome.glm$coeff+1.96*sqrt(diag(nome.sum$cov.scaled))

- Resíduos, gráfico dos resíduos e gráfico dos resíduos versus desvios residuais standardizados

```
>res<- rstandard(nome_glm,type="deviance")
>plot(res)
>plot(nome_glm$fitted.values,res,xlab="valores ajustados",ylab="descios residuais
standardizados ")
lines(lowess(nome_glm$fitted.values,res),col="red")
abline(h=0,lty=2)
```

Exemplo:

```
> res<-rstandard(glm1,type="deviance")
>plot(glm1$fitted.values,res,xlab="valores ajustados",ylab="desvios residuais standardizados")
lines(lowess(glm1$fitted.values,res),col="red")
abline(h=0,lty=2)
```

- Transformação do valor predicto e novo gráfico

```
>res<-rstandard(glm1,type="deviance")
>fit.tran<--2/sin(sqrt(glm1$fitted))
>plot(fit.tran,res,xlab="valores ajustados transformados",ylab="Desvios residuais
standardizados")
abline(h=0,lty=2)
```

- Gráfico dos resíduos vs covariáveis não incluídas no modelo

```
>plot (nome_variável_dependente$nome_variável1, res, xlab="nome_variavel1",
ylab="Resíduos deviance padronizados")
>lines (lowess(variável_dependente$nome_variavel1, res))
>abline (h=0, lty=2)
```

- Gráfico dos resíduos vs covariáveis incluídas no modelo

```
>plot (nome_variável_dependente$nome_variável1, res, xlab="nome_variavel1",
ylab="Resíduos deviance padronizados")
>lines (lowess(variável_dependente$nome_variavel1, res))
>abline (h=0, lty=2)
```

- Gráfico normal de probabilidades para os resíduos

```
>qqnorm (res)
>qqline (res)
```

- Leverage

```
>x<-influence.measures(glm1)
>h<-x$infmat[,"hat"]
```

```
>p<-dim(model.matrix(glm1))
>n<-dim(model.matrix(glm1))
>plot(h/(p/n), ylab='Leverage h/(p/n)',xlab='Indice', cex.lab=1.5,pch=19)
>abline (h=2, lty=2)
```

- Pontos Influentes: Cooks Distance

Exemplo:

```
>library(car)
>plot(cookd (glm1))
```

Ou

```
>library(car)
>plot(cookd(Caso_Tcar.glm1))
>x<-influence.measures(glm1)
>x$infmat[,"cook.d"]
>h<-x$infmat[,"cook.d"]
```

E para detector as observações influentes

```
>h[h>0.04]
```

Adequação do Modelo

- Teste à bondade do ajustamento do modelo

```
>1 – pchisq (deviance(glm1), df.residual(glm1))
```

- Estatística Hosmer e Lemeshow

```
>HL(nome_glm)
```

- Erros de Predição

```
>erropred(nome_glm)
```

- Curvas de ROC (depois de instalada a livraria Epi)

```
>ROC(form = modelo)
>ROC(form = variável_dependente
~variável_independente_1+...+variável_independente_n,data = nome_base_dados)
```

Anexo II - Tabelas de frequências e percentagens para as diferentes categorias das variáveis em estudo

Variável	Categoria	Frequência	%
Idadecat	**[15,25[**	**11**	**5,19%**
	[25,35[	**11**	**5,19%**
	[35,45[	**16**	**7,55%**
	[45,55[	**42**	**19,81%**
	[55,65[	**55**	**25,94%**
	[65,75[	**50**	**23,58%**
	[75,85[	**24**	**11,32%**
	[85,100[	**3**	**1,42%**
Localidade[9]	Beja	40	18,87%
	Estremoz	5	2,36%
	Évora	41	19,34
	Montemor-o-Novo	5	2,36%
	Vidigueira	5	2,36%
Concelho[10]	Beja	47	22,17%
	Évora	46	21,7%
	Montemor-o-Novo	15	7,08%
	Moura	13	6,13%
	Serpa	10	4,72%
Distrito	Sem Informação	4	1,89%
	Beja	93	43,87%
	Évora	99	46,7%
	Portalegre	11	5,19%
	Setúbal	5	2,36%
Ruralidade	Sem informação	4	1,89%
	Rural	42	19,81%

[9] Tendo em conta que neste estudo estão presentes 76 localidades, apenas serão aqui tidas em conta as que apresentam maior frequência

[10] Tendo em conta que neste estudo estão presentes 33 concelhos, apenas serão aqui tidos em conta os que apresentam maior frequência

	Urbano	166	78,3%
Tipo_amostra	Sem informação	3	1,42%
	Biópsia Histológica	61	28,77%
	Peça Cirúrgica	148	69,81%
Tipo_histologico[11]	Fibroadenoma	30	14,15%
	Carcinoma ductal invasivo(CDI)	120	56,6%
	Carcinoma in situ	19	8,96%
	Carcinoma lobular invasivo (CLI)	14	6,6%
Tipo_histologicoBM	Benigno	50	23,58%
	Maligno	162	76,42%
Carcinoma_in.situ	Carcinoma_in.situ	19	8,96%
	Outro	193	91,04%
Carcinoma_invasivo	Carcinoma_invasivo	134	63,21%
	Outro	78	36,79%
Lateralidade	Sem informação	95	44,81%
	Mama Esquerda	33	15,57%
	Mama Direita	44	20,75%
	QSEE	17	8,02%
	QSIE	1	0,47%
	QIEE	2	0,94%
	QIIE	1	0,47%
	QSED	12	5,66%
	QSID	1	0,47%
	QIED	4	1,89%
	QIID	2	0,94%
LateralidadeED[12]	Sem informação	95	44,81%
	Mama Esquerda	54	25,47%
	Mama Direira	63	29,72%
Tamanho_tumoral	Sem informação	87	41,04%

[11] Tendo em conta que no presente estudo se tiveram em conta 8 tipos histológicos de carcinomas mamários (descritos na tabela 12) aqui serão tidas em conta as categorias desta variável que apresentaram maior frequência.

[12] Tendo em conta que as categorias para a lateralidade segundo os quadrantes não se mostraram relevantes nos dados, procedeu-se à criação da variável lateralidadeED (em que apenas se tem em conta se se trata da mama Esquerda ou da mama Direita).

	Até 2 cm	91	42,92%
	De 2,1 a 5cm	34	16,04%
Margens_cirurgicas	Sem informação	72	33,96%
	Positivo	17	8,02%
	Negativo	108	50,94%
	< 1mm	15	7,08%
Grau_histologico	Sem informação	61	28,77%
	Baixo grau	41	19,34%
	Médio grau	54	25,47%
	Alto grau	56	26,42%
pT	Não se aplica	78	36,79%
	Sem informação	40	18,87%
	< 2cm	66	31,13%
	Entre 2 e 5 cm	28	13,21%
pN	Não se aplica	78	36,79%
	Sem informação	60	28,3%
	Ausência de metástases	47	22,17%
	Metástases em gânglio(s) móvel(is)	22	10,38%
	Metástases em gânglio(s) fixo(s)	5	2,36%
IVN	Não se aplica	193	91,04%
	Sem informação	3	1,42%
	Sem risco	2	0,94%
	Risco baixo	11	5,19%
	Risco médio	3	1,42%
IPN	Não se aplica	78	36,79%
	Sem informação	59	27,83%
	Bom prognóstico	26	12,26%
	Prognóstico Intermédio	36	16,98%
	Mau Prognóstico	13	6,13%
RE	Sem informação	90	42,45%
	Não favorável	20	9,43%
	Favorável	102	48,11%
RP	Sem informação	90	42,45%

	Não favorável	35	16,51%
	Favorável	87	41,04%
c-erB-2	Sem informação	90	42,45%
	Não se observou coloração	90	42,45%
	Coloração c/ manchas raras	8	3,77%
	Fraca coloração	11	5,19%
	Intensa coloração	13	6,13%
Ki67	Sem informação	90	42,45%
	Não favorável	32	15,09%
	Favorável	90	42,45%
P53	Sem informação	90	42,45%
	Não favorável	19	8,96%
	Favorável	103	48,58%
Fase_S	Sem informação	112	52,83%
	Não favorável	26	12,26%
	Favorável	74	34,91%
IDNA	Sem informação	101	47,64%
	Diplóide	68	32,08%
	Aneuplóide	33	15,57%
	Tetraplóide	7	3,3%
	Multiplóide Aneuplóide	3	1,42%

Tabela 26: Frequências e percentagens das variáveis em estudo

Anexo III - Tabela 27 -P-Values dos Testes de Independência (Qui-Quadrado e Fisher) entre as variáveis em estudo

X²/F	I	R	D	Ta	L	Tt	Mc	Gh	pT	pN	Ivn	Ipn	Re	Rp	Ce	Ki67	P53	Fs	Idna	Th	Bm	Cis	Ci
I	-	0,390	6,821x10⁻⁶	0,681	0,478	0,025	0,078	0,003	0,051	0,548	0,999	0,195	0,187	0,199	0,548	0,299	0,102	0,103	0,846	0,022	0,001	0,184	0,028
R		—	< 2,2x10⁻¹⁶	0,908	0,1999	0,021	0,4727	0,0157	0,5814	0,5438	0,9051	0,56	0,8318	0,7818	0,2723	0,03	0,64	0,27	0,48	0,013	0,74	0,8142	0,26
D			—	0,023	0,37	0,059	0,2758	0,2235	0,72	0,86	0,92	0,864	0,598	0,52	0,86	0,68	0,88	0,36	0,99	0,002	0,20	0,15	0,38
Ta				—	0,185	8,991x10⁻¹⁶	1,063x10⁻¹⁵	0,19	3,997x10⁻⁷	7,058x10⁻⁸	0,76	5,728x10⁻⁷	9,417x10⁻⁷	9,397x10⁻⁷	1,463x10⁻⁶	1,05x10⁻⁷	8,582x10⁻⁸	5,447x10⁻⁶	8,721x10⁻⁷	0,036	0,002	0,825	0,119
L					—	1,589x10⁻⁸	1,288x10⁻⁸	4,627x10⁻⁸	1,783x10⁻⁸	1,466x10⁻⁸	0,79	1,489x10⁻⁸	0,0003	0,0016	0,0029	0,0005	0,0003	0,061	0,0021	0,013	5,708x10⁻¹¹	0,754	1,403x10⁻⁷
Tt						—	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	0,042	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	5,225x10⁻¹⁶	<2,2x10⁻¹⁶	7,818x10⁻¹⁶	<2,2x10⁻¹⁶	0,0029	1,537x10⁻¹⁶
Mc							—	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	2,584x10⁻¹¹	0,05145	8,97x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	0,002453	3,442x10⁻¹⁶
Gh								—	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	0,002035	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	6,417x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	9,758x10⁻¹⁶	7,169x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	0,0039	<2,2x10⁻¹⁶
pT									—	<2,2x10⁻¹⁶	0,0003419	<2,2x10⁻¹⁶	1,610x10⁻¹⁶	3,989x10⁻¹⁶	1,681x10⁻¹⁶	<2,2x10⁻¹⁶	2,265x10⁻¹⁶	9,92x10⁻¹⁶	9,197x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	S.S.	<2,2x10⁻¹⁶
pN										—	0,003	<2,2x10⁻¹⁶	2,256x10⁻¹⁶	2,065x10⁻¹⁶	7,412x10⁻¹⁶	1,965x10⁻¹⁶	2,225x10⁻¹⁶	1,366x10⁻⁶	4,635x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	S.S	<2,2x10⁻¹⁶
Ivn											—	S.S.	0,06772	0,223	9,329x10⁻⁶	0,36	0,63	0,22	0,38	<2,2x10⁻¹⁶	0,17	<2,2x10⁻¹⁶	S.S.
Ipn												—	2,037x10⁻¹⁶	4,561x10⁻¹⁶	5,865x10⁻¹⁶	<2,2x10⁻¹⁶	9,133x10⁻¹⁶	1,385x10⁻¹⁶	1,457x10⁻¹⁵	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	S.S.	<2,2x10⁻¹⁶
Re													—	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	3,394x10⁻¹⁶	<2,2x10⁻¹⁶	0,874	8,255x10⁻¹⁶
Rp														—	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	4,393x10⁻¹⁶	<2,2x10⁻¹⁶	0,52	3,915x10⁻¹⁶
Ce															—	<2,2x1	<2,2x1	<2,2x10	<2,2x1	7,252x	<2,2x1	0,003	2,837x

Ki67																—	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	2,303x10⁻¹¹	<2,2x10⁻¹⁶	0,067	2,627x10⁻¹⁵
P53																	—	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	2,358x10⁻¹¹	<2,2x10⁻¹⁶	0,66	5,997x10⁻¹⁵
Fs																		—	<2,2x10⁻¹⁶	5,329x10⁻⁵	2,060x10⁻¹⁵	0,41	3,528x10⁻⁷
Idna																			—	1,484x10⁻⁷	8,964x10⁻¹⁵	0,36	1,378x10⁻⁷
Th																				—	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶	<2,2x10⁻¹⁶
Bm																					—	0,024	<2,2x10⁻¹⁶
Cis																						—	9,549x10⁻⁷
Ci																							—

Legenda:

I – Idade
R – Ruralidade
D – Distrito
TA – Tipo Amostra
L – Lateralidade
TT – Tamanho Tumoral
MC – Margens Cirúrgicas
GH – Grau Histológico
pT – Extensão do tumor
pN – Presença ou ausência de Metástases Regionais
IVN – Indice de Van Nuys
IPN – Indice Prognóstico de Nottingham

Re – Receptor de Estrogénio
Rp – Receptor de Progesterona
CE2 – Proteína c.erb-2
Ki67 – Gene ki67
P53 –Gene p53
FS – Fase S
IDNA – Indice de DNA
TH – Tipo Histológico
BM – Benigno /Maligno
CIS – Carcinoma in Situ
CI – Carcinoma Invasivo
S.S. – Sem Significado
Teste do Qui-Quadrado

Anexo IV - Poster presente no congresso da SPE 2009

CANCRO DA MAMA NA REGIÃO DO ALENTEJO

Helena Oliveira[1] **Isabel Natário[2]** **Manuela M. Oliveira[1]**
1.Departamento de Matemática, Universidade de Évora
2.Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
h2oliveira@portugalmail.com

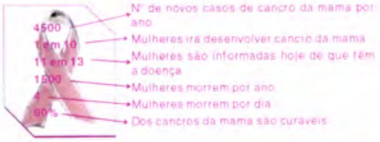
INTRODUÇÃO

O cancro da mama continua a ser uma doença comum e por vezes fatal, trata-se do cancro mais frequentemente diagnosticado em mulheres e a segunda causa de morte por cancro na região do Mediterrâneo Oriental. Mais de 1.2 milhões de mulheres são diagnosticadas com cancro da mama, por ano, no mundo inteiro.[1] No entanto, na década de 90 observou-se uma tendência crescente da mortalidade por cancro da mama, actualmente em declínio em diversos países (em Portugal houve uma diminuição de 2%/ano entre 1992 e 2002) (ver. Bastos, J.,Barros,H. e Lunet, N.(2007)).

O **objectivo** deste trabalho é apresentar, através de uma análise descritiva sucinta a realidade que se vive na região do Alentejo referente ao cancro da mama e também um estudo caso-controle de base hospitalar, no sentido de investigar factores de risco.

OS NÚMEROS DO CANCRO DA MAMA EM PORTUGAL[2]

Segundo dados do PNS (2009), a taxa de mortalidade portuguesa padronizada por cancro da mama feminina (/100 000 mulheres), abaixo dos 65 anos em 2006 decresceu relativamente ao ano de 2001, sendo no Alentejo (14.9%) bastante superior ao Continente (11.1%) e União Europeia (11.0%).
De uma forma resumida verifica-se que:



TIPOLOGIAS DE NEOPLASIAS DA MAMA

***Neoplasia* da mama**: crescimento descontrolado das células mamárias que tem, sempre, origem em anomalias ou erros genéticos. Tendo em conta a sua frequência, pode caracterizar-se por:

Figura 1: Tipologias mais frequentes de neoplasias mamárias

As *Figuras 2 e 3* mostram-nos a diferença entre CLIS (limitado aos lóbulos) e CDIS (limitado aos ductos) e, os locais no corpo humano, mais frequentes de metastização do cancro da mama, respectivamente.

Figura 2: Representação da diferença entre CLIS (limitado aos lóbulos) e CDIS (limitado aos ductos)

Fonte: Portal de Oncologia Português

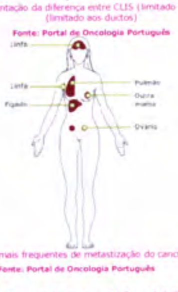


Figura 3: Locais mais frequentes de metastização do cancro da mama

Fonte: Portal de Oncologia Português

CANCRO DA MAMA NO ALENTEJO DE 2003-2007

Os dados relativos ao número de registos de neoplasias mamárias diagnosticadas no período de 2003 a 2007, foram, gentilmente cedidos pelo Serviço de Anatomia Patológica do Hospital do Espírito Santo (HES), em Évora.

@ POR TIPOLOGIA DE NEOPLASIA MAMÁRIA

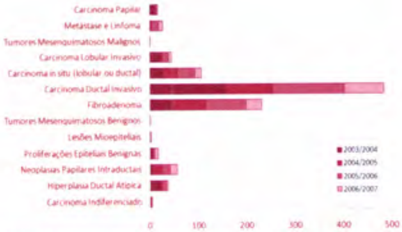


Figura 4: Distribuição dos casos de neoplasias mamárias na Região do Alentejo por tipo e ano de diagnóstico no período de 2003 a 2007

@ Tipologias de neoplasias mamárias mais frequentes na região em estudo: **Carcinomas Invasivos (51.4%)**, **Carcinomas in situ (10.4%)** (ambos malignos) e **Fibroadenomas (22.5%)** (benigno), com destaque para o **Carcinoma Ductal Invasivo (47%)**.

@ Ligeira tendência decrescente do número de cancros de mama diagnosticados na região a partir de 2003, com excepção do período de Agosto de 2005 a Agosto de 2006, em que houve um pico de casos, coincidente com uma alteração de nomenclatura.

@ POR DISTRITO

O número de pacientes diagnosticados que residem no distrito de Beja (45%) é superior aos que residem no distrito de Évora(38%), o que vem contrariar o que seria previsível, uma vez que estamos a falar de diagnósticos realizados no Hospital de Évora.

Figuras 5 e 6: Distribuição dos casos de neoplasias mamárias na Região do Alentejo por distrito e evolução dos casos de cancro da mama nos diferentes distritos no período de 2003 a 2007

Na *Figura 7* apresenta-se a distribuição de cancros benignos/malignos nos diferentes distritos da região do Alentejo. Observa-se que o número de cancros malignos é o dobro do número de cancros benignos, em praticamente todos os distritos da região.

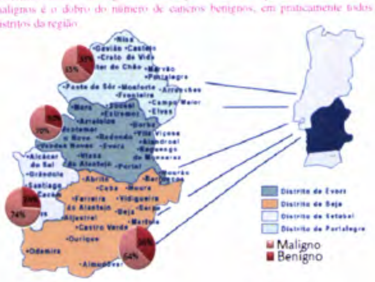


Figura 7: Distribuição dos casos de neoplasias mamárias na Região do Alentejo por distrito e benigno/maligno no período de 2003 a 2007

@ POR GRUPO ETÁRIO

Segundo a OMS, sendo o cancro da mama a forma de cancro mais frequente na mulher, raramente surge antes dos 30 anos de idade, aumentando significativamente a partir dos 45 anos e principalmente depois dos 60 anos. Na *Figura 8* apresenta-se o número de casos de neoplasias mamárias na região do Alentejo por grupo etário e tipo (benigno/maligno) no período de 2003 a 2007.

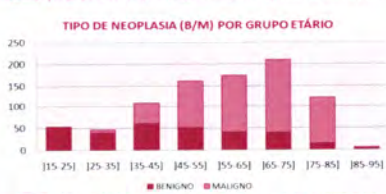


Figura 8: Distribuição dos casos de neoplasias mamárias na região do Alentejo por grupo etário e tipo (benigno/maligno) no período de 2003 a 2007

@ A classe etária]45,75] é aquela em que se registou maior número de casos de cancro da mama.
@ Abaixo dos 45 anos de idade a grande maioria de cancros da mama são benignos.
@ Acima dos 75 anos de idade o número de neoplasias mamárias malignas diagnosticadas ainda foi elevado.

NOTA: Segundo as estimativas do **INE (2009)**, a população do Alentejo envelheceu, verificando-se fortes assimetrias na distribuição da proporção de idosos nos concelhos da região. O índice de envelhecimento foi o mais elevado do país:186 habitantes de 65 ou mais anos para 100 jovens (Perfil de Saúde na Região do Alentejo, (2009)).

ESTUDO CASO-CONTROLE

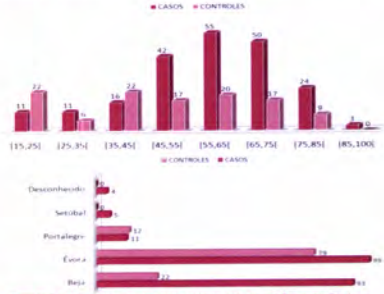
A Epidemiologia é a ciência que estuda os padrões da ocorrência de doenças em populações humanas e os factores determinantes destes padrões (Lilienfeld, 1980), sendo o estudo caso-controle uma das suas ferramentas mais utilizadas.

O estudo caso-controle parte de um grupo de indivíduos doentes – **os casos**, comparando-os com outro grupo de indivíduos em tudo semelhantes aos casos, diferindo somente por não serem doentes – **os controles.**

Os dados de Agosto de 2003 a Agosto de 2004 foram extraídos do conjunto de dados anteriormente analisados, para se efectuar um estudo caso-controle do cancro da mama relativamente à exposição ao ambiente Rural/Urbano dos indivíduos – **os casos.**

A amostra é constituída por 325 indivíduos do sexo feminino, dos quais 212 são casos e 113 são controles.

Os controles foram outros pacientes consultados no hospital no mesmo ano, escolhidos de forma a corresponderem aos casos nas características sexo, idade e distrito de residência.



Figuras 9 e 10: Distribuição dos casos/controles por grupo etário e por distrito no período de Agosto de 2003 a Agosto de 2004

@ A grande maioria dos indivíduos, quer do grupo dos casos (69%), quer do grupo dos controles (48%), situa-se em classes compreendidas entre os 45 e os 75 anos de idade.
@ A grande maioria dos indivíduos em estudo, de ambos os grupos, residem nos distritos de Évora (91%) e Beja (89%).

@RAZÃO DE CHANCES (*ODDS RATIO - OR*) E O SEU INTERVALO DE CONFIANÇA

Identificados os casos e seleccionados os controlos estuda-se a história de ambos os grupos com o objectivo de identificar a presença ou ausência à exposição a determinado factor (ruralidade) que pode ser importante para o desenvolvimento do cancro da mama. Mede-se essa associação através da razão de chances (OR):

		GRUPO	
		CASOS	CONTROLES
EVENTO	EXPOSTOS	A	B
	NÃO EXPOSTOS	C	D

$$ODDS\ RATIO\ (OR) = \frac{A/C}{B/D}$$

Uma vez calculado o OR, é preciso estimar o seu intervalo de confiança de 95%. Esse cálculo é efectuado tendo em conta a expressão

$$I.C_{95\%}(OR) \equiv \left] OR \times e^{-z_{\alpha}\sqrt{\frac{1}{A}+\frac{1}{C}+\frac{1}{B}+\frac{1}{D}}}; OR \times e^{z_{\alpha}\sqrt{\frac{1}{A}+\frac{1}{C}+\frac{1}{B}+\frac{1}{D}}} \right[$$

Com os valores da tabela

		GRUPO	
		CASOS	CONTROLES
RURALIDADE	RURAL	[illegible]	[illegible]
	URBANO	[illegible]	[illegible]

obteve-se um valor de OR = 1.18 e $IC_{95\%}$ (OR) =] 0.65; 2.12[, de onde se conclui que:

@ No que se refere à ruralidade, a grande maioria de ambos os grupos reside em zonas urbanas da região do Alentejo.

@ Estes dados indicam que a chance de cancro da mama é 1.18 vezes maior para as pessoas expostas ao ambiente rural do que as expostas ao ambiente urbano, ainda que esta diferença não se possa considerar significativa - o valor 1 está incluído no intervalo de confiança a 95%.

CONCLUSÃO

Os dados aqui analisados vêm ao encontro da ideia de que a idade é um factor relevante no aparecimento de cancro da mama, pelo que é aconselhável a realização da mamografia anual a partir dos 40 anos de idade. Quanto à ruralidade, este factor não veio acrescentar nada de novo. Foram nítidas as dificuldades em retirar conclusões significativas dos dados, uma vez que estes, apresentavam bastantes omissões em indicadores que poderiam ser bastante relevantes para este estudo.
Apesar do rastreio do cancro da mama já ter chegado a toda a região do Alentejo, a escassez de recursos humanos especializados na prestação de cuidados aos doentes oncológicos é notória, de onde podem advir os resultados obtidos (PNS –(2004-2010).

TRABALHO A DECORRER

Interessará ainda fazer o respectivo mapeamento com vista à detecção de áreas previsivelmente mais afectadas, bem como, identificar variáveis que possam estar relacionadas com o aparecimento deste tipo de cancro. Na modelação recorre-se a modelos lineares generalizados (GLM'S) e ao software estatístico R.

REFERÊNCIAS

1. Bastos, J.,Barros, H. e Lunet, N. (2007) *Evolução da Mortalidade por Cancro da Mama em Portugal (1955-2002)*, Serviço de Higiene e Epidemiologia - Faculdade de Medicina do Porto.
2. Alto Comissariado da Saúde - Ministério da Saúde (2009), *Perfil de Saúde na Região do Alentejo. PNS em Foco – Especial Regiões*, Gabinete de Informação e Prospectiva.
3. Ministério da Saúde (2004-2010). *Plano Nacional de Saúde – Ganhos em Saúde Avaliação de Indicadores (2001-2005)*
4. INE (2009). Estimativas da População Residente
5. Breslow, N.E. e Day, N.E. (1980) *Statistical Methods in Cancer Research: the analysis of case-control studies*,v.1. Lyon: International Agency for Research on Cancer.
6. Clayton, D. e Hills, M. (1993) *Statistical Models in Epidemiology*, Oxford University Press.
7. Agresti, A. (2001) *Categorical Data Analysis* – Second Edition