UNIVERSIDADE DE ÉVORA

DEPARTAMENTO DE MATEMÁTICA

MESTRADO EM MODELAÇÃO ESTATÍSTICA E ANÁLISE DE DADOS

Epidemiologia de HPV na População Feminina da Área de Influência do Hospital Fernando Fonseca

Dissertação de Mestrado sob a orientação do Professor Doutor Paulo Infante

Catarina Duarte Louro da Costa

Évora, 2013

UNIVERSIDADE DE ÉVORA

ESCOLA DE CIÊNCIAS E TECNOLOGIA - DEPARTAMENTO DE MATEMÁTICA

Epidemiologia de HPV na População Feminina da Área de Influência do Hospital Fernando Fonseca

Dissertação apresentada à Universidade de Évora para obtenção do grau de Mestre em Modelação Estatística e Análise de Dados

Sob a Orientação do Professor Doutor Paulo Infante

Catarina Duarte Louro da Costa

Évora, 2013

Ao Fernando e ao Dinis

As variáveis que tornam o meu modelo perfeito!

AGRADECIMENTOS

Ao finalizar esta tese várias pessoas merecem o meu mais sincero agradecimento:

O Professor Paulo Infante, pela forma dedicada como orientou este trabalho, e pelos constantes incentivos nesta recta final;

O Serviço de Anatomia Patológica do Hospital Prof. Doutor Fernando Fonseca, nomeadamente à Dra. Sofia Loureiro dos Santos e à Técnica Nicole Inácio, pela disponibilização dos dados e por toda a ajuda na preparação da base de dados;

O Fernando pelo constante apoio, dedicação e amor ao longo do processo de realização desta tese;

O Paulo pela amizade iniciada com este Mestrado e que em muito contribuiu para o percorrer do caminho que me levou a este momento;

Os colegas do HFF por compreenderem o quão importante este trabalho é para mim, o que me permitiu acabar a tese em tempo útil;

À Rita e ao Ricardo, por estarem sempre tão perto, pela amizade insubstituível que sempre ajudou a superar os momentos mais difíceis e a comemorar as alegrias recebidas;

O Dinis por ser o filho mais querido do mundo e me ter permitido trabalhar de forma continua nesta parte final.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

Epidemiologia de HPV na População Feminina da Área de Influência do Hospital Fernando Fonseca

Resumo:

O vírus do Papiloma Humano é a principal causa no desenvolvimento do cancro colo do útero, mundialmente considerado o segundo cancro mais frequente nas mulheres. Neste trabalho foi analisado o perfil demográfico, hábitos sexuais e resultados clínicos obtidos em 417 mulheres. Como objectivos centrais propõe-se: (1) caracterização da população em estudo e identificação de grupos mais susceptíveis na infecção pelo vírus ou no desenvolvimento de lesões; (2) Identificação dos factores de risco associados à infecção do HPV e (3) Identificação dos factores de risco associados ao desenvolvimento de lesões. A regressão logística binária, permitiu modelar a infecção por HPV, onde se identificou como variáveis significativas a idade, o número de parceiros sexuais, o HIV, o uso de preservativo e utilização de DIU. Relativamente ao desenvolvimento de lesões, os modelos de regressão multinomial e ordinal permitiram identificar como significativas a idade, o HPV de alto risco e de provável alto risco.

HPV Epidemiology among the Women Population within the Fernando Fonseca Hospital Influence area.

Abstract:

The Human Papillomavirus is the main cause for cervical cancer and worldwide considered the second most frequent type of cancer among women. In this work 417 women's demographic profile, sexual habits and clinical outcomes were studied. The key objectives are: (1) characterization of the population in question as well as identification of groups more susceptible of either getting infected by the virus or developing lesions; (2) Identification of risk factors associated with HPV infection and (3) Identification of risk factors associated with the development of lesions. Binary logistic regression allowed us to model HPV infection where variables such as age, number of sexual partners, HIV, use of condom and use of IUD were found significant. Concerning the development of lesions and using multinomial and ordinal regression models we were able to identify as significant age, HPV of high risk and HPV of probable high risk.

1. Introdução

O vírus do Papiloma humano (HPV) pertence à família *Papillomaviridae*, um diverso grupo taxonómico que tem co-evoluído com vários hospedeiros durante milhões de anos. Actualmente já foram identificados mais de 140 HPV (Vermund e Bhatta, 2004), 40 dos quais afectam o tracto genital humano (género *Alphapapillomavirus*) (Wheeler, 2008).

O HPV pode ser identificado no colo do útero, na vagina e vulva das mulheres, na glande, no escroto, no prepúcio, e na pele do pénis dos homens. Em ambos os sexos pode ainda ser encontrado na cavidade anal ou em toda a região peri-anal (Castellsagué, 2008).

A infecção por HPV é reconhecida como sendo a principal causa e um factor necessário no desenvolvimento do cancro cervical, o qual representa a segunda neoplasia mais frequente nas mulheres (9.8%) (Sanjosé *et al.,* 2007) estimando-se ser responsável por mais de 200.000 mortes anuais (Ferlay, 2005 *in* Papachristou *et al.,* 2009).

De acordo com o tipo de lesões que podem induzir, os HPV's classificam-se de alto risco quando estão fortemente associados ao desenvolvimento de neoplasias cervicais intraepiteliais (CIN), que eventualmente progridem para carcinoma (Hameed *et al*., 2001), ou de baixo grau, quando são vírus responsáveis por lesões não oncogénicas, como sejam as verrugas genitais. A categorização de alguns tipos de vírus ainda não é consensual, sendo classificados HPV de risco intermédio (Dunne *et. al*., 2007).

Nos últimos anos, a bibliografia existente sobre o HPV e os factores associados a esta infecção é bastante extensa. O trabalho publicado por Sansojé *et al*. (2007) reúne 78 publicações efectuadas entre 1995 e 2005, onde se analisam resultados provenientes de 157.789 mulheres classificadas com tendo uma citologia normal. De acordo com os resultados apresentados, estimou-se que a prevalência de HPV, em mulheres com citologia normal, é de 10.4%, sendo África a zona mais afectada (22.1%), seguida da América Central e México (20.4%), América do Norte (11.3%), Europa (8.1%) e Ásia (8.0%). Os mesmos autores referem ainda que a elevada prevalência detectada em Africa está relacionada com a falta de programas de rastreio do cancro do colo do útero expondo-as, mais que as restantes, ao desenvolvimento de lesões neoplásicas. Por outro lado, o casamento em idades muito jovens, a união com parceiros mais velhos ou com parceiros que têm simultaneamente mais do que uma mulher e as poucas condições higiénicas são factores que certamente contribuem para a maior prevalência de HPV nesta região. Relativamente às estripes mais frequentes foram

identificados, a nível mundial, o HPV16, HPV18, HPV31, HPV46 e HPV52. Na Europa, destacam-se o HPV16 (1.2%), HPV66 (0.3%), HPV45 (0.3%) HPV31 (0.2%) e HPV42 (0.2%).

Por outro lado, Clifford *et al*. (2003), utilizando dados provenientes de vários países e ajustando os resultados ao tipo histológico, região, espécime de HPV, e *primers* de PCR, estimaram a prevalência de HPV nas mulheres com cancro cervical invasivo (CCI). Este estudo indica que globalmente a prevalência de HPV em CCI se situa entre os 79.3% e os 88.1% (África: 86.5%, Ásia: 79.3% Europa: 86.7%, América do Norte e Austrália: 88.1% e América do Sul e Central: 87.7%). A distribuição dos tipos de HPV pelas várias regiões do globo apresenta também algumas similaridades. O HPV16 é claramente o mais predominante, variando entre 45.9% na Ásia e 62.6% na Europa. O HPV18 fica em segundo lugar na lista dos mais frequentes, variando entre 10% e 14%. Os HPV50, 31 e 33 representam o terceiro, quarto e quinto genótipos mais frequentes, embora não necessariamente com esta ordem.

Na prevenção do cancro cervical têm sido desenvolvidas essencialmente duas abordagens: (1) a vacinação como prevenção primária da infecção por HPV em adolescentes femininas e (2) a utilização de métodos na identificação de HPV de risco oncogénico, representando assim uma via de prevenção secundária e que permite o tratamento de lesões pré-cancerígenas ainda em estádios iniciais (Wheeler *et al*., 2008).

Os programas de rastreio para o colo do útero têm sido implementados pelos sistemas de saúde ao longo do tempo, com maior expressividade em países desenvolvidos, ainda que com protocolos diferenciados e, frequentemente com resultados abaixo do esperado. Os programas para detecção precoce do cancro do colo do útero, como o exame de Papanicolaou, foram propostos há mais de cinco décadas e têm evidenciado um custo-efectividade muito favorável na prevenção do cancro, desde que alcancem a totalidade da população feminina e façam parte das consultas femininas de rotina com adequada indicação do exame, colheita e análise do material, entrega do resultado e conduta terapêutica (Wheeler *et al*., 2008)

Nos Estados Unidos, foi em 2006, recomendada a vacinação profilática contra as estripes 6, 11, 16 e 18. Os estudos clínicos desenvolvidos posteriormente mostraram que a eficácia da vacina quadrivalente estava próxima dos 100% na prevenção da infecção e no desenvolvimento de lesões precursoras de neoplasias em mulheres que antes das vacinas não tinham anticorpos para nenhum destes vírus (Dunne *et. al*., 2007).

A infecção do papilomavírus humano pode ser assintomática ou manifestar-se através de várias lesões benignas ou malignas nas superfícies da mucosa ou cutâneas. A maioria das lesões regride

espontaneamente, ocorrendo apenas em alguns casos progressão para uma condição neoplásica. Estudos longitudinais mostram que a infecção deixa de ser detectável durante o primeiro ou segundo ano. Assim, sendo geralmente uma infecção de curta duração que não exibe sinais patológicos, torna-se praticamente "invisível" para a maioria das mulheres (Vermund e Bhatta, 2004).

A relação entre incidência e prevalência é mediada pela duração média da infecção, um indicador de persistência e que representa a chave no desenvolvimento de lesões de alto grau. Estudos mostram que vírus de alto risco têm uma permanência superior, durando em média, duas vezes mais que vírus de baixo risco (8.1 meses versus 4.9 meses) (Vermund e Bhatta, 2004).

Estima-se que nos Estados Unidos 75% das mulheres activas tenham adquirido pelo menos uma infecção por HPV ao longo da vida e cerca de 24.9 milhões estejam actualmente infectadas (Risser *et al*., 2008; Smith *et al*., 2008).

A transmissão do HPV faz-se essencialmente por contacto sexual, sendo pouco frequentes ou mesmo raras as restantes formas de contágio (e.g. contágio pelo parto) (Vermund e Bhatta, 2004). Desta forma, será espectável que os factores de risco mais relevantes estejam relacionados com o tipo de comportamento sexual, como a idade no primeiro contacto sexual, número de parceiros durante a vida e a existência de parceiros com elevado risco (prostituição, múltiplos parceiros, etc.) (Castellsagué, 2008).

Vários estudos realizados em populações da Europa do Norte e Estados Unidos da América mostram que a maior frequência de HPV genital ocorre antes dos 25 anos, diminuindo progressivamente como aumento da idade. Esses mesmos trabalhos mostram que a incidência cumulativa de infecção por HPV é extremamente elevada em jovens que iniciam a sua actividade sexual (Wheeler, 2008). Tanto a imunidade adquirida do tracto genital como a diminuição do comportamento de risco poderão assim estar na base da relação inversa entre idade e prevalência (Vermund e Bhatta, 2004). Por outro lado, o período correspondente à maturação sexual é de elevada actividade regenerativa e rápida proliferação de células infectadas por HPV. Assim sendo, esta elevada expansão de células infectadas pelo genoma viral de HPV constituirá certamente um dos maiores factores de risco no desenvolvimento de CIN e cancro cervical (Holly, 1996).

Outros estudos evidenciam um segundo pico em idades bastante mais avançadas, sugerindo assim que a distribuição etária pode variar entre diferentes populações. De facto, análises que incluem simultaneamente várias áreas geográficas mostraram diferentes tipologias na prevalência de HPV de acordo com a idade dos indivíduos. Entre estes refere-se o estudo conduzido por Investigadores da Agência Internacional de Pesquisa de Cancro (IARC), onde se verificou que a diminuição da

prevalência de HPV com a idade está associada a países com elevados rendimentos, ao passo que uma tendência praticamente constante foi observada em países mais pobres da Ásia e da Nigéria. Por outro lado, em algumas áreas da América Latina (Colômbia, Chile e México) foi identificada uma relação em forma de "U", corroborando assim algumas das publicações anteriores que apontavam para um possível pico em mulheres mais velhas.

Em Portugal foi, em 2008, incluída no programa nacional de vacinação a vacina quadrivalente para as estripes, 6, 11, 16 e 18. Esta imunização é administrada aos 13 anos em crianças do sexo feminino que nasceram depois de 1995.

Existem alguns estudos realizados em populações portuguesas, nomeadamente Silva *et al.* (2011) e Pista *et al*. (2011;2012).

Em Portugal estima-se uma prevalência de infecção na ordem de 19.4% (n= 451), sendo as estripes mais frequentes o HPV16 (19.7%), HPV31 (11.8%), HPV53 (11.8%), HPV51 (9.8%), HPV66 (8.6%), HPV52 (8.0%), HPV58 (6.9%), HPV59 (6.7%) e HPV18 (4.4%) (Pista, 2011). Assim, estima-se que a vacinação abranja cerca de 32.6% das infecções (Pista *et al.*, 2011). O trabalho de Silva *et al.* (2011) aponta no mesmo sentido, com uma prevalência de 16.7% (n=277) e com valores semelhantes em termos das estripes mais frequentes. Neste trabalho foram ainda analisados indivíduos com e sem vacinação, obtendo-se uma diminuição do risco de 64% em mulheres que lhes foi administrada a vacina quadrivalente.

O trabalho publicado por Pista *et al.* (2012) reúne informação sobre 2.372 mulheres e permitiu identificar factores de riscos associados à infecção, considerando algumas variáveis demográficas, socioeconómicas, estilo de vida e informação médica. Neste estudo foram identificados, recorrendo à regressão logística, alguns factores de risco, nomeadamente, as idades mais jovens, países de nascimento que não Portugal, educação ao nível do ensino secundário, história tabágica há pelo menos 10 anos, elevado número de parceiros sexuais e confirmação de doenças sexualmente transmissíveis nos últimos 12 meses.

Uma vez que os HPV de alto risco são frequentes em mulheres que apresentam uma citologia normal e apenas uma pequena fracção de mulheres infectadas por HPV desenvolvem cancro cervical, a infecção por HPV é um factor importante mas não o único no desenvolvimento de cancro cervical e ano-genital. O risco carcionogénico pode ser potenciado por outros factores como a patogenecidade viral, imunidade e imunosupressão do hospedeiro, comportamento sexual, imunidade adquirida da mucosa, acesso a cuidados de saúde, entre outros.

Por outro lado, mulheres que estejam infectadas por mais do que um tipo de HPV e que apresentem maior carga viral, terão provavelmente um risco superior de desenvolver uma CIN de grau elevado ou mesmo uma neoplasia. Também outras infecções genitais poderão estar relacionadas com a patogenecidade da infecção do HPV (Vermund e Bhatta, 2004). Alguns estudos apontam ainda para que o risco de desenvolver carcinoma das células escamosas será superior em mulheres que tenham tido mais do que uma gravidez. O estudo desenvolvido por Holly (1996) mostrou que mulheres infectadas por HPV que tenham tido mais do que sete filhos têm um risco quatro vezes superior de desenvolver carcinoma do que mulheres infectadas que nunca tenham estado grávidas. Se o número de filhos não for superior a dois, o risco de desenvolver a doença continua a ser mais elevado, cerca de duas vezes mais.

O tabaco é também considerado um factor de risco, uma vez que os componentes tóxicos podem ser identificados, em elevadas concentrações, na mucosa vaginal, favorecendo a mutação de DNA nas células hospedeiras.

Por outro lado, a protecção da integridade estrutural e da imunidade da mucosa poderá ser favorecida através de alguns factores nutricionais, como o consumo de alimentos ricos em retinóides (*e.g.* betacaroteno), anti-oxidantes (vitamina C) e agentes de metilação como seja o caso do ácido fólico (Vermund e Bhatta, 2004). O consumo de químicos e drogas tem sido considerado como um factor de risco destas doenças, nomeadamente pela imunossupressão e favorecimento da persistência do vírus (Vermund e Bhatta, 2004). O uso continuado de contraceptivos orais está também associado com o aumento do risco cervical em mulheres infectadas por HPV. No entanto, estes autores consideram que o risco relacionado com o uso de contraceptivos e com a gravidez poderá não corresponder a factores de risco independentes, uma vez que, a sua associação com as doenças cervicais poderá estar mascarada com a existência de múltiplos parceiros.

Neste contexto, com esta dissertação pretende-se:

- Conhecer a prevalência de infecção genital por HPV numa área populacional com elevada diversidade étnica, servida pelo Hospital Fernando Fonseca (cerca de 750.000 habitantes) e pelos Centros de Saúde dos concelhos da Amadora e Sintra.
- Determinar quais os diversos tipos de HPV prevalentes nessa população.
- Determinar a coexistência de infecção por mais de um tipo de HPV e relacioná-la com as eventuais alterações observadas nas análises citológicas e/ou histológicas.

- Relacionar a infecção pelos diferentes tipos de HPV com as eventuais lesões diagnosticadas nas análises citológicas e/ou histológicas.
- Identificar factores de risco associados à infecção considerando algumas características demográficas, geográficas, hábitos sexuais e informação clínica.
- Identificar factores de risco no desenvolvimento de lesões, considerando as características demográficas, geográficas, hábitos sexuais, informação clínica e infecção do HPV.

Para dar resposta aos objectivos propostos, no capítulo 2 é descrita a amostra e apresentadas as variáveis analisadas neste trabalho. No capítulo 3 estão resumidas as principais metodologias estatísticas aplicadas. No capítulo 4 é apresentada uma descrição sumária das variáveis, relacionando-as com a infecção do HPV e os vários tipos de lesões. Neste capítulo são incluídas algumas abordagens estatísticas que permitiram tecer as primeiras considerações sobre os primeiros quatro objectivos propostos. No capítulo 5 é realizada uma regressão logística binária que permite identificar quais os factores de risco associados à infecção do HPV, respondendo assim ao penúltimo objectivo proposto. Nos capítulos 6 e 7 são identificados os factores que contribuem para o aparecimento de lesões, utilizando a regressão multinomial (capítulo 6) e vários tipos de regressão ordinal (capítulo 7). Estes dois temas permitiram analisar o último objectivo apresentado. Por último, no capítulo 8 serão tecidas as principais conclusões que foram obtidas ao longo deste trabalho.

2. Amostra

O trabalho que se propõe está incluído no Projecto de Investigação designado "Tipos de HPV na População Feminina da Área de Influência do Hospital Fernando Fonseca", liderado pela Dra. Sofia Loureiro dos Santos, Directora do Serviço de Anatomia Patológica do Hospital Prof. Dr. Fernando Fonseca, EPE.

Neste projecto estão incluídas 417 mulheres seguidas em consulta de Ginecologia ou Obstetrícia no hospital, ou acompanhadas em consultas dessas especialidades em Centros de Saúde da área de influência.

São analisados resultados provenientes de:

- Inquérito epidemiológico, com especial ênfase na identificação da etnia, naturalidade e nacionalidade das utentes (ver Anexo 1), o qual foi preenchido pelo(a) enfermeiro(a) que dava apoio à consulta, ou pelo médico(a) que efectuou a consulta, após a assinatura, pela doente, de um consentimento informado. Os dados foram recolhidos durante o ano de 2008 e 2009;
- Genotipagem para HPV de todas as citologias cérvico-vaginais. Utilização do kit CLINICAL ARRAYS ® Papillomavirus Humano (Alfagene), que detecta a presença de 35 tipos virais (6, 11, 16, 18, 26, 31, 33, 35, 39, 40, 42, 43, 44, 45, 51, 52, 53, 54, 56, 58, 59, 61, 62, 66, 68, 70, 71, 72, 73, 81, 82, 83, 84, 85 e 89), em amostras de citologia cérvico-vaginal, processadas em meio líquido (thin prep);
- Inquérito epidemiológico, com especial ênfase na identificação da etnia, naturalidade e nacionalidade das utentes (ver Anexo 1), o qual foi preenchido pelo(a) enfermeiro(a) que dava apoio à consulta, ou pelo médico(a) que efectuou a consulta, após a assinatura, pela doente, de um consentimento informado
- Análise de biópsias, diagnósticos histológicos e citológicos, segundo os critérios definidos por Bethesda 2001: sem lesão, alterações e lesões das células pavimentosas e glandulares (ASC-US, ASC-H, LIBG, LIAG) e neoplasias malignas das células pavimentosas e glandulares (carcinoma, carcinoma pavimentocelular, adenocarcinoma).

3. Métodos

De acordo com os objectivos que se propõe neste trabalho e enumerados no capítulo anterior foi necessário recorrer a vários procedimentos estatísticos. A análise exploratória de dados permitiu de uma forma simples, compreender os indivíduos em estudo e a forma como eles se relacionam de acordo com as variáveis de interesse. Esta abordagem foi construída pela aplicação de testes paramétricos (teste-t, análise de variância) e não paramétricos (teste de Wilcoxon-Mann-Witeney).

Os modelos de regressão logística permitiram identificar as variáveis de interesse na infecção do HPV, enquanto que, os modelos de regressão multinominal e ordinal permitiram identificar as variáveis que contribuem no aparecimento de lesões.

A análise da de dados desta dissertação foi quase toda realizada usando o software StataCorp. 2009. Stata Statistical Software: Release 11. College Station, TX: StataCorp LP. Apenas foi pontualmente necessário recorrer ao SPSS versão 20.

São descritos de seguida, de forma sumária, os vários procedimentos estatísticos utilizados, assim como os comandos Stata utilizados (anexo 3).

3.1 Testes à Normalidade

A distribuição normal é um dos requisitos para muitos dos procedimentos estatísticos, baseados em distribuições teóricas, como sejam o teste-t, a regressão linear, a análise descriminante ou a análise de variância. A violação deste pressuposto coloca sérios constrangimentos na avaliação dos resultados e na sua inferência (Razali e Wah, 2011).

Para avaliar se uma amostra aleatória de observações independentes de tamanho n, provêm de uma população com distribuição normal, podem ser considerados:

(1) Métodos gráficos, como histogramas, diagramas de extremos-quartis, gráfico Q-Q. Este último, é considerado uma das ferramentas mais efectivas na avaliação da normalidade. De referir que estes métodos não devem ser, por si só, considerados suficientes para considerar que uma amostra cumpre o pressuposto da normalidade.

(2) Testes aos coeficientes de achatamento e assimetria, assim como os testes formais à normalidade, nomeadamente, o teste de Shapiro-Wilk , o teste de Kolmogorov-Smirnov com

correção de Lilliefors, o teste de Anderson-Darling, o teste de Cramer Von Mises e o teste de Jarque-Bera.

Os testes de Anderson-Darling, Cramer Von Mises e Kolmogorov-Smirnov comparam a função de distribuição empírica (FDE) dos dados com a função cumulativa da distribuição normal. O teste de Kolmogorov-Smirnov utiliza a máxima distância entre a distribuição empírica e a hipotética, devendo ser utilizado quando são conhecidos os parâmetros da distribuição hipotética. Quando tal não se verifica, ou seja, quando é necessário estimar esses parâmetros, deve ser utilizada uma modificação, o teste de Lilliefors. Por outro lado, os testes de Anderson-Darling e Cramer-von Mises avaliam as diferenças quadráticas entre a distribuição empírica e a hipotética. Por outro lado, o teste de Shapiro-Wilk baseia-se nos valores amostrais ordenados elevados ao quadrado. Este teste é adequado em amostras de pequenas dimensão, menores que 50 observações. Aspectos específicos destes testes podem ver-se, por exemplo, em Razali e Wah (2011).

De salientar que quando as amostras são de grandes dimensões, vários destes testes têm tendência a rejeitar a hipótese da normalidade. Em alternativa, uma abordagem possível é a aplicação dos testes desenvolvidos com base no achatamento e assimetria, como o teste de Jarque-Bera e os testes baseados em D'Agostino, Belanger, e D'Agostino, Jr. (1990) e Royston (1991) (Stata 9, 2005).

Vários autores compararam a eficiência destes testes, utilizando a simulação de Monte Carlo, sob diferentes distribuições e diferentes tamanhos de amostras. De acordo com os resultados obtidos, consideram que o teste de Shapiro-Wilk é o que apresenta maior aderência à normalidade, seguido do teste de Anderson-Darling, Lilierfors e, por último o de Kolmogorov-Smirnov (Razali e Wah, 2011).

3.2 Comparação de médias: Teste t-Student e ANOVA

A comparação de médias obtidas a partir de amostras de duas populações foi avaliada a partir do teste de hipóteses à diferença de médias. Considerando que os dados provêm de uma distribuição normal e que as variâncias das amostras, embora desconhecidas, são idênticas, a estatística de teste (teste t-Student) é dada por:

$$t = \frac{\bar{\mu}_1 - \bar{\mu}_2}{\left[\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right]^{1/2} \left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{1/2}}$$

Onde, $\bar{\mu}_1$ e $\bar{\mu}_2$ correspondem às médias estimadas, s_1^2 e s_1^2 às variâncias amostrais e n_1 e n_2 ao tamanho das amostras das populações.

Esta estatística segue uma distribuição t com $n_1 + n_2 - 2$ graus de liberdade (Sokal e Rohlf, 2009).

Designado por t_0 o valor observado da estatística de teste, num teste bilateral do tipo:

$H_0: \mu_1 - \mu_2 = 0$ *versus* $H_1: \mu_1 - \mu_2 \neq 0$

a hipótese nula é rejeitada se:

$t_0 > t_{\frac{\alpha}{2}, n_1+n_2-2}$ ou $t_0 < -t_{\frac{\alpha}{2}, n_1+n_2-2}$ ou valor $p \leq \alpha$

No caso de um teste unilateral à direita tem-se:

$H_1: \mu_1 - \mu_2 > 0$

e a hipótese nula é rejeitada se:

$t_0 > t_{\frac{\alpha}{2}, n_1+n_2-2}$ ou valor $p \leq \frac{\alpha}{2}$

No caso de um teste unilateral à esquerda tem-se:

$H_1: \mu_1 - \mu_2 < 0$

e a hipótese nula é rejeitada se:

$t_0 > t_{\frac{\alpha}{2}, n_1+n_2-2}$ ou valor $p \leq 1 - \frac{\alpha}{2}$

De acordo com o referido anteriormente, as inferências realizadas serão válidas se cumpridos os pressupostos de independência, normalidade e homogeneidade de variâncias. As violações nestes requisitos podem comprometer de forma mais ou menos severa a análise a realizar (Sokal e Rohlf, 2009).

As violações na igualdade de variâncias comprometem o erro tipo I, o qual pode estar deflacionado ou inflacionado, dependendo da direcção da relação entre o tamanho amostral e a variância da população. Ou seja, como o denominador do teste t é baseado numa média ponderada das estimativas da variância, quando a maior variância se localiza na amostra de maior dimensão, essa maior variância assume maior peso, tornando maior o denominador, e por conseguinte mais pequena a estatística t. Em oposição, quando o tamanho amostral e a variância estão relacionados negativamente, a menor variância assume maior relevo, obtendo-se valores superiores da estatística t, aumentando assim a taxa de rejeição (Myers e Well, 2003).

No caso, da igualdade de variâncias não se verificar, existe uma variação ao teste t , designado de teste t- Welch o qual foi também desenvolvida sob a premissa da normalidade das distribuições. Para mais desenvolvimentos consultar Myers e Well (2003) e Stata (2005).

Apesar de a normalidade ser um dos requisitos do teste t, pequenos desvios não comprometem a aplicação deste procedimento (Montgomery, 2005). No entanto, se as populações são enviesadas ou apresentam *outliers*, a distribuição amostral da diferença de médias tem tendência a mostrar caudas mais pesadas. Nestas situações, as estimativas das diferenças entre populações são menos precisas e o teste t menos potente do que quando o pressuposto da normalidade é cumprido (Myers e Well, 2003).

Em caso de desvios acentuados à normalidade, a abordagem deverá ser via testes não paramétricos, nomeadamente pela aplicação do teste de Wilcoxon-Mann-Witeney. No caso de distribuições normais, a eficiência assimptótica do teste de Wilcoxon-Man-Witeney é 95.5% da eficiência do teste de t-Student (Siegel, 1957).

Este teste permite analisar se as duas populações são iguais em tendência central, ou seja, se as medianas das duas amostras são idênticas entre si. Considerando o caso de um teste bilateral, onde $\widetilde{\mu_1}$ e $\widetilde{\mu_2}$ correspondem às medianas da população:

$$H_0: \widetilde{\mu_1} - \widetilde{\mu_2} = 0 \qquad \textit{versus} \qquad H_0: \widetilde{\mu_1} - \widetilde{\mu_2} \neq 0$$

Considerando W a soma das ordens da amostra de menor dimensão, a estatística de teste é dada pela equação:

$$Z = \frac{W - \mu_w}{\sigma_w}$$

Esta estatística de teste está tabelada para amostra pequenas e segue uma distribuição aproximadamente normal para amostras grandes. De referir que a existência de observações repetidas retira potência ao teste de Wilcoxon-Mann-Witeney.

Para a comparação de médias provenientes de duas ou mais populações foi utilizada a **análise de variância (ANOVA)**. Tal como o nome indica, a ANOVA compara variâncias dentro das amostras com as variâncias entre as amostras. Se a variância dentro dos grupos (amostras) for significativamente inferior à variância entre grupos, ou seja, à que resulta do efeito do factor em estudo, então as médias populacionais estimadas a partir das amostras são significativamente diferentes (Casella e Berger, 2002). No caso do presente trabalho, assume-se que se trata de um modelo de efeitos fixos,

ou seja, considera-se que a análise foi realizada incluindo todos os níveis de interesse ao investigador (Myers e Well, 2003).

Neste trabalho, apenas são realizadas comparações para uma variável independente, tratando-se assim da ANOVA a um factor. Este modelo é escrito:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

onde, Y_{ij} corresponde à observação i da amostra j , α_i ao efeito do tratamento para a amostra i e ε_{ij} ao erro ou resíduo associado a cada observação ij.

A aplicação deste método torna necessário que os resíduos sejam normais e independentemente distribuídos, com média zero e variâncias iguais a σ^2.

O pressuposto da normalidade foi analisado de acordo com o referido no ponto anterior, mediante análise gráfica dos resíduos (nomeadamente o gráfico QQ-plot, considerando-se normal quando os valores descrevem aproximadamente uma recta) e utilizando testes formais à normalidade, nomeadamente pela aplicação do teste de Kolmogorov-Smirnov.

A igualdade de variâncias foi avaliada graficamente, considerando-se válido o pressuposto quando o gráfico dos resíduos *versus* valores previstos tem uma distribuição homogénea, sem ser detectado um padrão específico, ou seja, quando a variabilidade dos resíduos não depende de nenhuma forma dos valores ajustados. Uma distribuição que sugira um determinado padrão é geralmente indicação que os dados necessitam de uma transformação, ou seja, uma análise aos dados utilizando uma métrica diferente. Por exemplo, se a variabilidade dos resíduos aumenta com os valores previstos, Montgomery (2005) sugere uma transformação logarítmica ou pela raiz quadrada.

Dentro dos testes formais à homogeneidade de variâncias destacam-se o teste de Bartllet, Hartley, Cochran e Levene. O teste de Levene foi o seleccionado dado a sua fácil aplicabilidade no Software Stata.

O pressuposto da independência pode ser avaliado confrontado os resíduos com a ordem em que a experiencia for realizada. Mais uma vez, a detecção de um padrão sugere que o pressuposto pode estar comprometido (Montgomery, 2005). Esta avaliação necessita que a experiência tenha sido realizada com uma sequência temporal, característica que não se aplica neste contexto. Assim, considerou-se que a amostra foi bem definida à priori, não comprometendo o pressuposto de independência.

De acordo com o referido anteriormente, a ANOVA permite testar a igualdade de médias entre grupos, $\mu_1, \mu_2, \dots, \mu_k$, ou seja:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k \qquad \textit{versus} \qquad H_1: \exists i,j:\ \mu_i \neq \mu_j\ (i \neq j; i,j = 1, \dots, k)$$

A estatística de teste:

$$F = \frac{QMTrat}{a-1} \Big/ \frac{QME}{N-a} \sim F_{a-1;N-a}$$

onde, $QTrat$ representa a variação explicada pelo tratamento e QME a variação devida ao erro, a o número de grupos e N o tamanho global das amostras.

Ao rejeitar H_0 conclui-se que existe pelo menos um par de médias que é estatisticamente diferente dos restantes, mas não há identificação de qual ou quais os pares que diferem. Para tal, recorre-se a testes de comparação múltipla, efectuados à posteriori, também conhecidos como testes *Post-Hoc*. Como exemplo destes testes, salientam-se o teste de Tuckey, Bonferroni, Scheffé, LSD e Duncan. A aplicação de um determinado teste depende do contexto e do tipo de dados. São tecidas algumas características para estes testes, mas para informação mais detalhada deverá ser consultados os trabalhos de Myers e Well (2003), Casella e Berger (2002) e Oehlert (2010).

As taxas de erro são um aspecto a ter em consideração quando se fazem comparações múltiplas ou simultâneas. Cada teste individual, ou cada intervalo de confiança está associado a um erro tipo I, o qual pode ser controlado pelo investigador Oehlert (2010). Um dos primeiros testes desenvolvidos, o teste LSD (Least Significant Difference), permitia, pela aplicação de múltiplos testes t, comparar dois a dois os níveis de cada factor. Assim, o valor p é considerado pouco correcto, em especial quando existem muitas comparações a realizar Myers e Well (2003).

O teste de Tuckey permite manter os níveis do erro de tipo I, desde que sejam cumpridos os pressupostos de independência, normalidade e homecedasticidade. É ainda considerado mais potente quando as amostras são de igual dimensão, embora exista uma modificação para quando existe um desequilíbrio entre grupos. O teste de Scheffé, apesar de mais conservativo, é provavelmente o teste mais popular e considerado o mais flexível, sendo robusto relativamente aos pressupostos de normalidade e homocedasticidade. Os níveis de significância utilizados nas comparações estão corrigidos e permite ainda comparações mais complexas que envolvem contrastes de mais do que duas médias. O teste de Dunnet deve ser utilizado quando se pretende comparar as médias dos tratamentos com a média de um grupo controlo. O teste de Bonferroni aplica-se para comparar médias duas a duas, sendo sido definido á priori o número de comparações a realizar (Myers e Well , 2003).

É possível que a ANOVA e os testes *Post-Hoc* produzam resultados contraditórios. Ou seja, a ANOVA indicar que existem diferenças em pelos menos um par de médias (rejeitar H_0) e os testes *Post-Hoc* não serem capazes de identificar onde se localizam essas diferenças. Esta discordância acontece porque a ANOVA é um teste mais potente, *i.e.*, mais capaz de rejeitar correctamente H_0, que os testes *Post-Hoc*, que têm associada uma maior probabilidade de erro tipo II.

Muitos autores referem que a ANOVA é relativamente robusta a desvios na normalidade, especialmente quando se tratam de amostras de dimensão elevada, embora Wilcoxon (1998 *in* Ruterford, 2001) refira que pequenos desvios possam originar uma ANOVA menos potente, especialmente no caso de distribuições enviesadas ou leptocúrticas. Isto verifica-se porque este tipo de distribuições têm maior probabilidade de incluir *outliers* do que distribuições normais, o que pode conduzir a um incremento substancial da variância.

A violação de pressuposto da igualdade de variâncias levanta alguns constrangimentos relativamente ao erro de tipo I, em especial se os grupos forem constituídos por tamanhos diferentes. O aumento e diminuição do nível de significância irá depender da relação entre o tamanho da amostra e a variância da população, ou seja, se o grupo de maior dimensão tiver a maior variância, o teste será conservativo, mas se a maior variância estiver localizada no grupo de menor dimensão, o teste poder-se-á tornar demasiado liberal, e por vezes atingir erros de tipo I demasiado elevados (Myers e Well, 2003). Os mesmos autores propõem a utilização de duas estratégias, a transformação dos dados de forma a que as variâncias sejam homodedásticas, ou pela utilização de testes alternativos como o teste de F modificado de Brown–Forsythe, que tem a desvantagem de não ter testes de comparação múltipla e por isso não permitir identificar onde se localizam as diferenças.

Embora se considere que desvios à normalidade não tenham um grande impacto no erro tipo I, o mesmo não se aplica ao poder do teste. Desta forma, existem algumas abordagens, via testes não paramétricos, que não requerem cumprido este requisito. Destes testes salienta-se o teste de Kruskal-Wallis, dada a sua implementação na maioria dos pacotes estatísticos, nomeadamente no *software* STATA. Para uma leitura mais aprofundada sobre este tópico deve consultar-se (Myers e Well, 2003).

Nenhum destes testes exige que as populações tenham uma distribuição normal, no entanto assentam no pressuposto de que as distribuições têm formas idênticas, ou seja, que têm os mesmos valores de variância, assimetria e achatamento. Sob esta restrição, a hipótese alternativa é referida como “hipótese de mudança”, assim designada porque o grupo (factor) alterou a distribuição, mas não a influenciou. No caso de populações que se afastem da normalidade, mas que ainda assim tenham formas idênticas, os testes não paramétricos têm, muitas das vezes, mais poder que o teste

F. O pressuposto de igualdade de variâncias é também aplicado nos testes não paramétricos, sob pena de aumento do erro tipo I (Myers e Well, 2003).

3.3 Regressão Logística Binária

Os modelos de regressão logística fazem parte de um vasto conjunto de modelos designados por modelos lineares generalizados (GLM's).

Os GLM's têm como principal característica o facto da variável resposta seguir uma distribuição dentro de uma família de distribuições com particularidades muito específicas: a família exponencial (Turkman e Silva, 2000). Foram desenvolvidos inicialmente por Nelder e Wedderburn, em 1972, como uma ampliação ao modelo linear clássico. A distribuição considerada não tem que ser normal, desde que seja qualquer distribuição da família exponencial e, embora seja necessário cumprir o pressuposto da normalidade, a função que relaciona o valor esperado e o vector de cavariáveis pode ser qualquer função diferenciável (Turkman e Silva, 2000). Como casos particulares dos GLM, salientam-se os seguintes modelos:

- Modelo de regressão linear clássico;
- Modelos de análise de variância e covariância;
- Modelo de regressão logística;
- Modelo de regressão de Poisson;
- Modelos log-lineares para tabelas de contigência multidimensionais;
- Modelos probit para estudos de proporções.

A regressão logística é uma técnica estatística que permite modelar a ocorrência, em termos probabilísticos, de uma variável dependente de natureza dicotómica. O *design* da regressão logística permite predizer a probabilidade de um evento ocorrer, ou seja, a probabilidade de uma observação ser codificada num determinando grupo.

Considerando Y a variável resposta que assume como valores possíveis 0 e 1, o valor esperado de Y é dado pela seguinte expressão:

$$\hat{\mathrm{Y}} = \mathrm{E}(\mathrm{Y}|\mathrm{X}) = \mathrm{P}[\mathrm{Y} = 1] = \pi$$

Assim, ao modelar π em função das variáveis explicativas, o modelo standard de regressão linear permitiria estimar valores superiores e inferiores a 1, que por se tratarem de probabilidades, são desprovidos de qualquer sentido. Por outro lado, os pressupostos da regressão linear não seriam

cumpridos. A variância de Y, dada por $\pi(X)(1-\pi(X))$, não é constante pelos valores da variável resposta (Agresti, 2007), já que depende de $E(Y)=\pi$. Se $\pi=0{,}5$, a variância é máxima, enquanto que, se $\pi=0$ ou se $\pi=1$, a variância é nula. Como a variável resposta assume apenas dois valores, a distribuição dos erros é binomial e não normal.

Devido aos factores acima enumerados é mais apropriado um modelo que permita uma relação curvo-linear entre $E(Y)$ e cada variável independente X (Agresti, 2007), resultando uma dependência em forma de "S". Existem várias possibilidades que permitem a modelação desta curva. Neste caso, a probabilidade de uma determinada realização j da variável dependente ser o "sucesso" é dada pela expressão abaixo, onde $x_i$representa as variáveis explicativas e β_i os coeficientes do modelo:

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

Ou caso exista mais do que uma variável independente:

$$\pi(x)=\frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_j x_j}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_j x_j}}$$

A função *logit* pode ser linearizada, resultando:

$$Logit\ (\hat{\pi})=g(x)=\mathrm{Ln}\left(\frac{\pi(x)}{1-\pi(x)}\right)=\sum_{k=0}^{k}\beta_i x_{jk}$$

onde a razão $\hat{\pi}/(1-\hat{\pi})$ designada por possibilidades ou *odds*, representa a possibilidade do sucesso (π) relativamente à do insucesso ($1-\pi$) (Powers e Xie, 1999).

Os parâmetros do modelo de regressão logística são estimados pelo método de **Máxima Verosimilhança**, ou seja, serão parâmetros desconhecidos que maximizam a probabilidade de obter os dados observados. A função de máxima verosimilhança expressa a probabilidade dos dados observados em função de parâmetros desconhecidos.

Na equação anterior, $\hat{\beta}$ corresponde à estimativa de máxima verosimilhança, e dependendo do sinal, indica um aumento ou uma diminuição de $\pi(x)$. O exponencial de β, indica a possibilidade do sucesso de um grupo relativamente a um outro de referência. Esta associação entre as duas categorias é quantificada pela razão de possibilidades (*odds ratio*) que no caso de um modelo simples com uma variável explicativa dicotómica se pode escrever como

$$OR=\frac{w_1}{w_2}=\frac{p_1/(1-p_1)}{p_2/(1-p_2)}=\frac{e^{\beta_0+\beta_1 x}}{e^{\beta_0}}=\ e^{\beta_1 x}$$

Após a estimação dos coeficientes, será necessário definir quais as variáveis que devem ser incluídas no modelo. Ou seja, que parâmetros são significativamente diferentes de zero. Turkman e Silva sugerem três abordagens possíveis: (1) Estatística de *Wald*; (2) Estatística de *Wilks* ou Estatística de razão de verosimilhança e (3) Estatística de *Rao* ou Estatística de *score*. Neste trabalho apenas serão abordadas as duas primeiras opções.

A estatística de Wald (W) é conseguida comparando a estimativa por máxima verossimilhança do declive do parâmetro com a estimativa do seu erro padrão.

$$W = (\hat{\beta} - \beta_0)/SE$$

A razão obtida, segundo a hipótese de que $\beta_j = 0$, segue uma distribuição aproximadamente normal (Hosmer e Lemeshow,2000), ou de forma equivalente, W^2segue uma distribuição qui-quadrado, com um grau de liberdade (Agresti, 2007).

Assim, H_0 é rejeitada, para um nível de significância α, se o valor observado da estatística W for superior ao quantil de probabilidade $1 - \alpha$ de χ_q^2, ou equivalentemente, se o valor p do teste não exceder o nível de significância α previamente definido .

A estatística Wé utilizada, em geral, para testar hipóteses sobre variáveis dicotómicas (Turkman e Silva, 2000). Este teste é considerado conservativo, ou seja, a probabilidade de não rejeitar H_0 quando o coeficiente é significativo é maior que a de outros testes (Hosmer e Lemeshow,2000).

O teste de razão de verosimilhança permite, por um lado, testar qual o melhor modelo entre dois modelos encaixados, nomeadamente avaliando se o modelo com uma com uma determinada variável é melhor que o modelo sem essa contribuição, e por outro, julgar sobre a qualidade de ajustamento.

A estatística de Wilks (Λ) ou estatística de razão de verosimilhança é definida por:

$$\Lambda = -2ln\frac{max_{H0}L(\beta)}{max_{H0\cup H1}L(\beta)} = -2\{l(\tilde{\beta}) - l(\tilde{\beta})\}$$

Onde$\hat{\beta}$ é o estimador de máxima verosimilhança e β o valor que maximiza a verosimilhança (Turkman e Silva, 2000).

O teorema de Wilks estabelece que a estatística Λ tem sob H_0, uma distribuição assimptótica de um χ^2, sendo o número de graus de liberdade igual à diferença entre o número de parâmetros que se pretende estimar em H_0. Ou seja,

$$\Lambda = -2\{l(\hat{\beta}) - l(\hat{\beta})\} \sim^{a} \chi^2_q$$

Da comparação entre o modelo saturado e o modelo corrente, através da estatística Λ, obtêm-se a função *deviance*, a qual pode ser decomposta pela soma de parcelas d_i, que medem a diferença entre os logaritmos das verosimilhanças observada e ajustada para cada observação: $D\ (y; \hat{\mu}) = \sum_i^i d_i$. A função *deviance* é sempre positiva, decrescendo à medida que são adicionadas covariáveis no modelo, assumindo o valor zero no modelo saturado (Turkman e Silva, 2000).

Considerando que se pretende analisar dois modelos M_1 e M_2, estando M_2 encaixado em M_1, a estatística da razão de verosimilhança entre os dois modelos pode ser descrita:

$$-2\{l_{M_2}(\hat{\beta}_2) - l_{M_1}(\hat{\beta}_1)\} = \frac{D\ (y; \hat{\mu}_2) - D\ (y; \hat{\mu}_1)}{\phi} \sim^{a} \chi^2_{p1-p2}$$

Ou seja, a comparação de dois modelos encaixados pode ser obtida pela diferença da *deviance* de cada um dos modelos (Turkman e Silva, 2000).

3.3.1 Construção do modelo

O primeiro passo na construção de um modelo passa pela análise individual de cada variável. Esta abordagem pode ser efectuada através de tabelas de contingência, pela estatística de Pearson, ou via modelo logístico univariado, pela razão de verosimilhança.

Qualquer análise univariada ignora a possibilidade de que um conjunto de variáveis, cada uma por si só com uma fraca associação com a variável resposta, se possam tornar em importantes preditores quando associada às restantes, ou de forma oposta, deixar de ter um contributo relevante para o modelo.

Assim, numa segunda fase e de acordo com as recomendações de Hosmer e Lemeshow (2000), deverão ser consideradas como potenciais candidatas todas as variáveis cujo valor p for inferior a 0.25 ou, cuja importância clínica assim o justifique. A significância de cada variável deverá ser avaliada através da estatística de razão de verosimilhança ou pela estatística de Wald, considerando o seguinte teste de hipóteses:

$$H_0: \beta_1 = 0\ versus\ \beta_1 \neq 0$$

Com base nestes critérios, todas as variáveis que não contribuem para o modelo devem ser eliminadas e um novo modelo deve ser ajustado. O modelo resultante deve ser confrontado com o anterior, tanto pela razão de verosimilhança como pela análise dos seus coeficientes. O ciclo de eliminação, ajustamento e verificação dos coeficientes deve ser finalizado quando no modelo

constarem todas as variáveis significativas ou, que revelem alguma importância do ponto de vista clínico. Nesta fase, Hosmer e Lemeshow (2000) sugerem que as variáveis inicialmente excluídas (valor p superior a 0.25 na análise univariada) sejam gradualmente introduzidas, da mais para a menos significativa, de forma a avaliar se conjuntamente com as restantes o seu contributo passa a ser relevante. No caso de variáveis constituídas por mais do que duas categorias, deverá ser avaliado a possibilidade de colapsar alguns dos seus níveis. Esta análise deverá ser efectuada via razão de verosimilhança.

Quando no modelo existem variáveis contínuas, o pressuposto da linearidade das variáveis contínuas com o *logit* deve ser confirmado. Esta análise pode ser efectuada graficamente, obtendo os quartis da variável contínua e criando uma nova variável utilizando os pontos de corte dos quartis. Posteriormente um modelo multivariado deve ser ajustado, substituindo a variável contínua pela categórica (assumindo como classe de referência o primeiro quartil). A análise do gráfico dos coeficientes ajustados *versus* o ponto médio dos quartis permitirá, caso seja linear, aceitar como cumprido o pressuposto da linearidade. Em alternativa, poderá ser utilizado o gráfico de dispersão de alisamento (Hosmer e Lemeshow, 2000). Este tipo de representação, de uma forma muito geral, permite criar uma nova variável que, para cada valor de y_i, contenha o seu valor suavizado. Estes valores são calculados a partir de uma regressão entre x_ie y_i, utilizando uma ponderação, de tal forma que seja atribuído ao ponto mais central (x_i, y_i) a maior ponderação e os pontos mais distantes, baseados na distância $|x_j - x_i|$, o menor peso (StataCorp, 2005).

Esta validação, pode ainda ser conseguida, de modo analítico, utilizando o método dos polinómios fraccionários (Royston e Altman, 1994 *in* Hosmer e Lemeshow 2000), determinando qual o valor de x^p que melhor modela a covariável. O desenvolvimento deste processo implica a criação de vários modelos, os quais são comparados pelo logaritmo da verosimilhança. Assim de acordo com esta metodologia, quando número de covariáveis(J) for igual a um, serão ajustados oito modelos ($p_1 \in \wp$) e seleccionado aquele com maior valor do logaritmo da verosimilhança. Quando J for igual a dois, serão necessários 36 modelos resultantes dos dois pares de potência ($(p_1, p_2) \in \wp * \wp$ e seleccionado aquele que apresentar também o valor mais elevado do logaritmo da verosimilhança.

Royston e Altman (1994) propõem que as potências estejam restringidas ao conjunto:

$$\wp = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}, \text{ onde } p_1 = 0 \text{ corresponde ao logaritmo da variável.}$$

Posteriormente, os modelos serão confrontados com o modelo linear, através do teste de razão de verosimilhança. Ou seja:

- Para $J = 1$ é comparado o logaritmo da verosimilhança do seu melhor modelo ($L(p_1)$) com o do modelo linear inicial $L(1)$:

  $$G(1, p_1) = -2\{L(1) - L(p_1)\}$$

  Este teste segue uma distribuição aproximadamente qui-quadrado com um grau de liberdade, considerando a hipótese nula da linearidade em x.

- Para $J = 2$ é comparado o logaritmo da verosimilhança do seu melhor modelo ($L(p_1, p_2)$) com o obtido em $J = 1$ ($L(p_1)$):

  $$G[p_1, (p_1, p_2)] = -2\{L(p_1) - L(p_1, p_{12})\}$$

  Este teste segue uma distribuição aproximadamente qui-quadrado, com dois graus de liberdade, considerando a hipótese da segunda função de potência ser zero.

- Para $J = 2$ é comparado o logaritmo da verosimilhança do seu melhor modelo ($L(p_1, p_2)$) com o do modelo linear inicial $L(1)$:

  $$G(1, (p_1, p_2) = -2\{L(1) - L(p_{1;}p_2)\}$$

  Este teste segue uma distribuição aproximadamente qui-quadrado com 3 grau de liberdade, considerando a hipótese nula da linearidade em x.

Uma vez definidos todos os efeitos principais e a correcta escala para as variáveis contínuas, deverão ser pesquisadas as interacções entre variáveis, ou seja, se o efeito de uma variável não é constante nos vários níveis de uma segunda variável. A interacção é criada pelo produto aritmético dos efeitos principais e a sua significância pode ser testada pela razão de verosimilhança. Quando é incluída uma interacção não significativa, existe um aumento dos desvios padrão sem no entanto ocorrer uma alteração muito acentuada dos coeficientes. Quando a alteração ocorre tanto na estimativa pontual como nos intervalos de confiança é indicativo que o termo da interacção deverá ser significativo.

Assim, foram criadas interacções de primeiro grau entre as variáveis do modelo, atendendo á premissa de que uma interacção terá que fazer sentido do ponto de vista sentido clínico (Hosmer e Lemeshow, 2000). As interacções não significativas foram retiradas e um novo modelo ajustado, tendo-se comparado os dois modelos via razão de verosimilhança.

Todas as variáveis incluídas no modelo deverão ter um valor p da estatística de Wald inferior a 10% e a 5% no caso das interacções.

3.3.2 Ajustamento do Modelo

Uma vez construído ao modelo final será necessário analisar o modo como este descreve a variável resposta, ou seja, a sua bondade de ajustamento. Deverá ainda ser avaliada a sua capacidade discriminativa, ou seja a sua capacidade de descrever o sucesso e o insucesso, assim como identificar

a existência de observações com impacto nas estimativas dos coeficientes ou qualidade do ajustamento do modelo.

i. Estatística de razão de Verosimilhança e Estatística de Pearson

A medida de discrepância entre os valores ajustados ($\hat{\mu}$) e os valores observados y pode ser efectuada através da estatística da razão de verosimilhanças de Wilks (Λ). Se da comparação do valor observado da *deviance*, para um modelo com k parametros, com o valor crítico de um χ^2_{n-p}, resultar um valor superior a $\chi^2_{n-p,\alpha}$, então o modelo é considerado não adequado (Turkman e Silva, 2000).

A estatística de *Pearson* generalizada permite também averiguar sobre a adequabilidade do modelo. Esta estatística pode ser d escrita da seguinte forma:

$$x^2 = \sum_i \frac{w_i(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

Assim, comparado o valor observado com o quantil de probabilidade $1-\alpha$ de uma distribuiçãp χ^2 com $n-p$ graus de liberdade, conclui-se sobre a adequabilidade do modelo.

No entanto, tanto pela *deviance* como na estatística de *Pearson*, a aproximação pelo χ^2 da distribuição qui-quadrado pode não ser adequada, mesmo quando se consideram grandes amostras. (Turkman e Silva, 2000). A solução consiste em agrupar os dados, de forma a garantir que em cada grupo, o número de elementos não seja pequeno.

ii. Estatística de Hosmer-Lemeshow

Uma das medidas que quantifica a bondade do ajustamento é a estatística de Hosmer-Lemeshow ($\hat{C}$). Este método agrupa as observações de acordo com os percentis das probabilidades estimadas ou através de valores fixos das probabilidades estimadas. Em ambos os casos, a estatística de Hosmer-Lemeshow segue uma distribuição aproximadamente Qui-Quadrado, com o número de grupos (g) -2, graus de liberdade.

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

Onde, n'_k corresponde ao número toral de indivíduos no k-ésimo grupo, c_k o número de padrões de covariáveis no k-ésimo decil, o_k o número de respostas entre o padrão de covariáveis c_k e $\bar{\pi}_k$ a média das probabilidades estimadas.

$$o_k = \sum_{j=1}^{c_k} y_i$$

$$\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n'_k}$$

Estudos posteriores indicam que é preferível o agrupamento baseado nos percentis, na medida que ficam melhor adaptados à distribuição $\chi^2(g-2)$, em especial quando as probabilidades estimadas são baixas*, i.e.*, inferiores a 0.2 (Hosmer e Lemeshow, 2000). Os mesmos autores referem que a maior desvantagem do método é a não identificação de desvios quando os grupos são constituídos por poucos indivíduos.

iii. Tabelas de Classificação e Curva ROC

As Tabelas de Classificação permitem uma análise ao poder preditivo de uma regressão binária. Estas tabelas classificam a variável resposta (y) em função da sua estimativa ($y = 0$ ou $y = 1$). Se a probabilidade estimada é superior a um determinado ponto de corte (c), então $\hat{\mu} = 1$, caso contrário $\hat{\mu} = 0$ (Agresti, 2007).

Desta análise são obtidas três medidas principais:

(1) Sensibilidade = $P(\hat{\mu} = 1|y = 1)$
(2) Especificidade = $P(\hat{\mu} = 0|y = 0)$
(3) Correcta Classificação = $P(\hat{\mu} = 1|y = 1) * P(y = 1) + P(\hat{\mu} = 0|y = 0) * P(y = 0)$

No entanto, estas medidas devem ser interpretadas com precaução e não devem ser aplicadas na comparação entre modelos. Hosmer e Lemeshow (2000) referem que medidas que derivam de tabelas 2X2, como a especificidade e a sensibilidade, dependem fortemente da distribuição de probabilidades da amostra, reflectindo por isso a variabilidade das observações "patient mix" e não a superioridade de um modelo em detrimento de outro. Por outro lado, a própria classificação dos dados é sensível ao tamanho dos dois grupos, favorecendo sempre a classificação dos indivíduos para o grupo de maiores dimensões.

A análise destes parâmetros pode ainda ser efectuada pela curva ROC (*reciver operatin characteristic*), um gráfico que representa a sensibilidade como função de 1-especificidade, para todos os possíveis pontos de corte (π_0). Quando π_0 se aproxima de zero, praticamente todas as estimativas são $\hat{\mu} = 1$, fazendo com que a sensibilidade se aproxime de um e a especificidade de zero e o ponto para as coordenadas (1-especificidade, sensibilidade) fique próximo de (1,1). De modo oposto, quando π_0 se aproxima de um, quase todas as estimativas são $\hat{\mu} = 0$ correspondendo a uma

sensibilidade perto de zero e a especificidade a rondar um, com as coordenadas localizadas em (0,0). Graficamente, os pontos extremos (0,0) e (1,1) são unidos por uma curva concava (Agresti, 2007).

Para uma determinada especificidade, o melhor poder preditivo é aquele que apresentar uma maior sensibilidade. Assim, quanto melhor for o poder preditivo do modelo mais alta será a curva. A área abaixo da curva é usualmente designada por índice de concordância (c) e estima a probabilidade de uma estimativa e observação serem concordantes, ou seja, a observação com o maior valor de y ter também o maior valor de μ (Agresti, 2007). De acordo com Hosmer e Lemeshow (2000), a área abaixo da curva pode ser classificada pelas seguintes categorias:

- $\mathrm{ROC} < 0.5$ Modelo sem poder discriminativo;
- $0.5 \leq \mathrm{ROC} < 0.7$ Modelo com discriminação fraca
- $0.7 \leq \mathrm{ROC} < 0.8$ Modelo com boa discriminação;
- $0.8 \leq \mathrm{ROC} < 0.9$ Modelo com discriminação excelente;
- $\mathrm{ROC} \geq 0.9$ Modelo com poder discriminativo perfeito;

3.3.3 Selecção de Modelos

A análise estatística via modelos lineares generalizados está muitas vezes associada a um número elevado de covariáveis, as quais são essenciais na explicação da variabilidade dos dados. Muitas vezes é necessário a análise da influência de possíveis interacções entre covariáveis. Assim, pode resultar um número elevado de modelos, sendo necessário seleccionar aquele que melhor explica o fenómeno em estudo (Turkman e Silva, 2000).

Um dos métodos aplicados na comparação entre modelos relaciona-se com a função *deviance* através do teste de razão de verosimilhança, abordado anteriormente, como critério de selecção de variáveis. São também descritos o coeficiente de determinação R^2 e os critérios de informação de Akaike e bayesiano.

i. Coeficiente R^2

O poder preditivo do modelo poder ser quantificado pela correlação *R* entre as respostas observadas (y_i) e os valores ajustados ($\hat{y}_i$). Na análise clássica da regressão linear pelo método dos Mínimos Quadrados, R representa a correlação múltipla entre a variável resposta e os respectivos preditores, e R^2, descreve a proporção da variação de *y* que é explicada pelas covariáveis. A vantagem de R^2 sobre *R* é a utilização da escala original, sendo o seu valor aproximadamente proporcional à magnitude do efeito. Na Regressão Logística, R corresponde à correlação entre observações binárias (y_i) e a probabilidade estimada ($\hat{\mu}_i$). A natureza binária da variável resposta irá afectar a amplitude

de variação de R, conduzindo a uma medida de difícil leitura, mas mesmo assim, de bastante utilidade na comparação entre modelos (Agresti, 2007).

Existem uma série de diferentes possibilidades de coeficientes de correlação. No Stata, é possível, para além do índice mais utilizado, o coeficiente de Nagelkerke/Cragg e Uhler's, calcular por exemplo, o coeficiente standart de determinação, o R^2 de máxima verosimilhança, o R^2 de McFadden e o R^2 de Efron . Para mais detalhe consultar Long e Freese (2001).

O R^2 de de Nagelkerke/Cragg e Uhler's corresponde a uma aproximação do R^2 de Cox e Snell, impossibilitando valores superiores a um.

$$R^2 = \frac{1 - (\frac{-2LL_{nulo}}{-2LL_k})^{\frac{2}{n}}}{1 - (2LL_{nulo})^{\frac{2}{n}}}$$

Onde $2LL_{nulo}$ é a verosimilhança do modelo nulo e $-2LL_k$ a verosimilhança do modelo com as variáveis independentes.

ii. Critérios de Informação

Os critérios de informação permitem comparar modelos que utilizam diferentes amostras ou comparar modelos não encaixados. Valores mais baixos são indicativo de um melhor ajustamento e na selecção de modelos deverá ser tida em conta a minimização deste parâmetro (Turkman e Silva, 2000).

O critério de informação de Akaike (AIC) é baseado na função de *log-verosimilhança*, com a introdução de um factor de correcção como modo de penalização da complexidade do modelo (Long e Freese, 2001).

$$AIC = \frac{-2LL_k + 2P}{N}$$

Onde $-2LL_k$ representa a verosimilhança do modelo e P o número de parâmetros.

O critério de informação bayesiano é escrito:

$$BIC = D_k - gl_k \ln N$$

Onde D_k corresponde à Deviance do modelo e gl_k os respectivos graus de liberdade.

Quanto mais negativo for BIC melhor é o ajustamento do modelo. Raftery (1996 *in* Long e Freese, 2001) sugeriu algumas orientações que permitem avaliar, de acordo com a diferença de BIC's, o melhor de dois modelos:

- $0 < BIC_{M2} - BIC_{M1} \leq 2$ Evidência fraca que o modelo 2 seja melhor que o modelo 1
- $2 < BIC_{M2} - BIC_{M1} \leq 6$ Evidência aceitável que o modelo 2 seja melhor que o modelo 1
- $6 < BIC_{M2} - BIC_{M1} \leq 10$ Evidência forte que o modelo 2 seja melhor que o modelo 1
- $BIC_{M2} - BIC_{M1} > 10$ Evidência muito forte que o modelo 2 seja melhor que o modelo 1

3.4 Análise de Resíduos

A análise dos resíduos é útil, não só para a avaliação da qualidade do ajustamento, relativamente à escolha da distribuição, da função de ligação e de termos do preditor linear, mas também na identificação de observações mal ajustadas, que não seguem o padrão das restantes observações (Turkman e Silva, 2000). Genericamente, e seguindo as definições Turkman e Silva (2000), estas observações podem classificar-se em três grandes grupos:

- Repercussão (*leverage*) – A repercussão mede o efeito de uma observação nos valores previstos, indicando quão influente uma observação é;
- Influência – Uma observação é influente, se uma ligeira modificação, ou a sua exclusão do modelo, origina uma alteração significativa nos parâmetros do modelo;
- Consistência – Uma observação com um resíduo elevado é geralmente uma observação inconsistente (*outlier*). Geralmente resultam de um valor extremo da variável resposta e/ou de uma ou mais covariáveis.

Um resíduo deve exprimir a discrepância entre o valor observado y_i e o valor μ_i ajustado pelo modelo.

Relativamente às medidas de consistência, é frequente o uso do resíduo de *Pearson* e do resíduo *Deviance*, para a pesquisa de *outliers*.

O **resíduo de Pearson** corresponde à contribuição de cada observação para o cálculo da estatística de Pearson generalizada, e pode ser definido:

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

Considerando a sua padronização:

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

Agresti (2007) considera que os resíduos standartizados $e_i/\sqrt{1-h_i}$, em valor absoluto, maiores que dois ou três indicam que a observação é um possível *outlier.*

A desvantagem do resíduo de Pearson reside na sua distribuição, que geralmente é bastante assimétrica para modelos não normais. Assim, é frequentemente usado outro tipo de resíduo baseado na função ***Deviance***. Estes resíduos quantificam o desacordo entre uma componente do logaritmo da verosimilhança do modelo ajustado e a componente do logaritmo da verosimilhança, caso cada ponto fosse perfeitamente ajustado. Uma vez que a regressão logística utiliza o princípio da máxima verosimilhança, pretende-se minimizar o somatório dos resíduos *deviance* (Sakar *et. al.*, 2011).

$$R_i^D = \delta[y_i]\,log_e(\hat{\pi}_i) + (1-y_i)\,log_e(1-\hat{\pi}_i)$$

Onde, $\delta = sinal(y_i - \hat{\pi}_i)$

McCullagh e Nelder (1989) referem a vantagem destes resíduos relativamente aos de *Pearson*, uma vez que a sua distribuição é próxima da normal.

Nas medidas de influência considera-se a **distância de Cook**, que avalia a diferença entre as estimativas de máxima verosimilhança do parâmetro $\hat{\beta}$ obtido da amostra sem a observação e a da amostra que inclui todas as suas observações.

$$C_i = = \frac{r_i^2 h_{ii}}{(1-h_{ii})^2}$$

Podem ser também analisados os resíduos **dfbetas**, que representam a diferença standartizada na estimativa do parâmetro quando se elimina uma observação.

$$DFBETA = b_{(-i)j} - b_j$$

Por outro lado, as medidas de repercussão são usualmente avaliadas pela ***leverage***, *i.e.*, pelo ij-ésimo elemento da matriz de projecção generalizada (matriz chapéu - H). Esta matriz, simétrica e idempotente, tem a sua diagonal principal h_{ij} como uma medida da influência que é exercida pelos valores observados nos valores estimados. Os elementos da diagonal principal variam entre zero e

um e os valores elevados de h_{ij} são considerados pontos extremos, em particular se $h_{ij} > \frac{2p}{n}$ (Hoaglin e Welsch, 1978 *in* Turkman e Silva, 2000).

Quando o modelo de regressão é composto por mais do que duas covariáveis, os resíduos descritos anteriormente podem ser pouco informativos na detecção de *outliers* e/ou observações influentes. Neste sentido, existem algumas abordagens que permitem o diagnóstico do modelo mediante a análise do efeito da eliminação de todas as observações com um determinado padrão de covariáveis, tanto na estimação dos coeficientes β como nas medidas do ajustamento, χ^2 e D.

A alteração do valor dos coeficientes estimados é semelhante à proposta por Cook para a regressão linear, sendo obtida pela diferença standartizada de $\hat{\beta}$ e $\hat{\beta}_{-j}$, a qual representa a máxima verosimilhança estimada utilizando todos os padrões de covariáveis j e excluindo todas as observações m_j com padrão x_j, com a standartização via matriz de covariância $\hat{\beta}$.

$$\Delta\hat{\beta}_j = \left(\hat{\beta} - \hat{\beta}_{(-j)}\right)'(X'VX)\left(\hat{\beta} - \hat{\beta}_{(-j)}\right)' = \frac{R_i^P h_j}{(1-h_j)}$$

Uma aproximação semelhante permite avaliar a diminuição do valor do χ^2 de *Pearson* após a eliminação das observações com um padrão de covariáveis x_j:

$$\Delta\chi_j^2 = \frac{r_j^2}{(1-h_j)} = \left(R_i^P\right)^2$$

Assim como a alteração da *Deviance:*

$$\Delta D_j = \frac{d_j^2}{(1-h_j)}$$

Assim, a análise destas estatísticas permite identificar que padrões de covariáveis apresentam um mau ajustamento (valores elevados de $\Delta\chi_j^2$ e ΔD_j) e quais têm uma grande influência na estimação dos parâmetros (valores elevados de $\Delta\hat{\beta}_j$).

3.4 Regressão Multinomial

O modelo de regressão logística, onde a variável resposta é nominal dicotómica, pode ser expandido para o caso em que a variável resposta é nominal policotómica, *i. e.*, com mais de duas classes mutuamente exclusivas.

Estes modelos consistem num conjunto de k modelos logísticos corrigidos, um para cada um das k variáveis dependentes e, da mesma forma que no modelo logístico, existe uma classe de referência com a qual são comparados os *logit* das restantes categorias.

Considerando um modelo cuja variável resposta é constituída por três categorias ($Y = 0,1,2$), o modelo *logit* é escrito da seguinte forma:

$$g_1(X) = ln\left[\frac{P(Y=1|X)}{P(Y=0|X)}\right] = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \cdots + \beta_{1p}x_p = X'\beta_1$$

$$g_2(X) = ln\left[\frac{P(Y=2|X)}{P(Y=0|X)}\right] = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \cdots + \beta_{2p}x_p = X'\beta_2$$

Onde Y corresponde à variável resposta com k categorias e p covariáveis x_1, x_1,..., x_p. A intercepção corresponde a β_0 e a combinação linear das covariáveis é escrita por $X'\beta = \beta_1 x_1 + \cdots + \beta_p x_p$

A expressão geral para a probabilidade condicional num modelo com três categorias é:

$$P(Y=j|X) = \frac{e^{g_j(x)}}{\sum_{j=0}^{2} e^{g_j(x)}}$$

Como existe mais do que uma combinação de coeficientes que conduzem às mesmas probabilidades, é necessário normalizar o sistema relativamente a uma das categorias da variável dependente. As probabilidades são:

$$P(Y=0|X) = \frac{1}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

$$P(Y=1|X) = \frac{e^{g_1(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

$$P(Y=2|X) = \frac{e^{g_2(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

Considerando a mesma nomenclatura que o modelo binomial, $\pi_j(x) = P(Y=j|x)$, para $j = 1, 2\ e\ 3$.

De acordo com as equações anteriores, para uma variável dependente com k classes serão necessários $k-1$ modelos. Os *odds ratio* são calculados para cada uma das $k-1$ classes relativamente à classe de referência. Assim, o *odds ratio* de $Y = j$ relativamente a $Y = 0$ para uma covariável com valor $x = a$ *vesus* $x = b$:

$$OR_j(a,b) = \frac{P(Y = j|x = a)/P(Y = 0|x = a)}{P(Y = j|x = b)/P(Y = 0|x = b)}$$

3.4.1 Construção do modelo

As estratégias de construção de um modelo multivariado para uma variável resposta de natureza multinomial são muito idênticas às aplicadas no modelo de regressão logística. O processo de exclusão/inclusão das variáveis pode também ser efectuado mediante a estatística de Wald ou via teste de razão de verosimilhança. Hosmer e Lemeshow (2000) aconselham a segunda hipótese, dados os múltiplos graus de liberdade das variáveis.

O pressuposto da linearidade do *logit* quando no modelo estão incluídas variáveis de natureza contínua também carece de validação. Esta análise, assim como o estudo futuro de resíduos, apoiam-se nos mesmos métodos que os apresentados para a regressão logística binária.

Dadas as limitações da maioria dos pacotes estatísticos neste tipo de análise, a metodologia utilizada seguiu as recomendações de Begg e Gray (1984 *in* Hosmer e Lemeshow, 2000). Assim, no caso de uma variável resposta constituída por três níveis, deverão ser ajustados dois modelos logísticos binários:

- $Y = 1$ *versus* $Y = 0$, ignorando $Y = 2$;
- $Y = 2$ *versus* $Y = 0$, ignorando $Y = 1$;

Estes autores mostraram que as estimativas obtidas dos coeficientes são consistentes e que a perda de eficiência não é muito acentuada.

A linearidade do *logit* será então avaliada de acordo com o método dos polinómios fraccionários, métodos dos quartis e pela avaliação do gráfico de dispersão de alisamento (Hosmer e Lemeshow, 2000).

Definidos os efeitos principais do modelo deverá ser analisada a interacção entre as várias covariáveis. Esta análise segue os mesmos princípios de inclusão exclusão das restantes variáveis, pela aplicação do teste de Wald e razão de verosimilhança.

Após a identificação de todas as variáveis explicativas, importa avaliar o ajustamento global do modelo bem como examinar a contribuição individual de cada individuo. Existem algumas abordagens possíveis, nomeadamente extensões aos testes de bondade de ajustamento e diagnóstico, como propostas por Lesaffre e Albert (1989 *in* Hosmer e Lemeshow, 2000). Estes métodos continuam a não estar disponíveis na maioria dos programas estatísticos, sendo recomendado por Hosmer e Lemeshow (2000) a mesma abordagem proposta por Begg e Gray em 1984 (Hosmer e Lemeshow, 2000). Ou seja, subdividir o modelo multinomial em várias equações de regressão logística e proceder aos testes usuais de diagnóstico e ajustamento, como os descritos no capítulo anterior.

3.5 Regressão Ordinal

Os modelos de regressão ordinal logística são utilizados na análise de dados cuja resposta é do tipo ordinal. Nos últimos anos têm sido de grande aplicação em análises epidemiológicas, nomeadamente em estudos de qualidade de vida em escalas intervalares, indicadores de saúde e, como no presente caso, na análise da gravidade de doenças (Abreu *et. al.*, 2009).

Existem vários tipos de modelos de regressão logística ordinal, mas consideram-se como principais o modelo da categoria adjacente, o modelo da razão contínua e o modelo de riscos proporcionais.

Assumindo que a variável resposta é ordinal e pode ser constituída $k+1$ valores $(0,1,2,\dots,k)$. A expressão geral para a probabilidade da variável resposta Y é igual ao condicional k no vector x de p covariáveis, ou seja, $P[(Y=k|X)] = \phi_k(x)$.

No contexto dos modelos ordinais, o modelo de regressão multiniomial é considerado o modelo base. Ou seja, os modelos são parametrizados de modo a que os coeficientes sejam o *log-odds* de uma categoria, $Y = K$, com uma categoria base, $Y = 0$.

a. Modelo da Categoria Adjacente

Este modelo pode ser considerado como uma extensão do modelo multinominal e compara cada categoria da variável resposta ($Y = k$) com uma categoria de referência ($Y = 0$), em geral a primeira ou a última. A natureza ordinal dos dados implica uma estrutura linear dos coeficientes, fazendo com que os OR estimados tenham uma tendência de crescimento. São modelos muito utilizados quando a variável resposta é ordinal composta por categorias discretas (Abreu *et. al.*, 2009).

Assumindo que o logaritmo dos *odds* não dependem da resposta e que pode ser escrito linearmente, então os *logits* para as categorias adjacentes podem ser escritos da seguinte forma (Hosmer e Lemeshow,2000):

$$a_k(x) = ln\left[\frac{P(Y=k|X)}{P(Y=k-1|X)}\right] = ln\left[\frac{\phi_k(X)}{\phi_{k-1}(X)}\right] = \alpha_k + x'\beta$$

para j=1,2,…,k. O modelo da categoria adjacente é uma versão constrangida do modelo base, com a intercepção em $\beta_{k0} = (\alpha_1 + \alpha_2 + \cdots \alpha_k)$ e declive $\beta_k = k\beta$.

$$ln\left[\frac{\phi_k(X)}{\phi_0(X)}\right] = ln\left[\frac{\phi_1(X)}{\phi_0(X)}\right] + ln\left[\frac{\phi_2(X)}{\phi_1(X)}\right] + \cdots + ln\left[\frac{\phi_k(X)}{\phi_{k-1}(X)}\right]$$

$$= a_1(X) + a_2(X) + \cdots + a_k(x)$$

$$= (\alpha_1 + x'\beta) + (\alpha_2 + x'\beta) + \cdots + (\alpha_k + x'\beta)$$

$$= (\alpha_1 + \alpha_2 + \cdots \alpha_k) + kX'\beta$$

Considerando um modelo com uma variável explicativa composta por dois níveis, o *odds ratio* é dado por:

$$\mathrm{OR}(k) = ln\left[\frac{P(Y=k|X=1)/P(Y=k-1|X=1)}{P(Y=k|X=0)/P(Y=k-1|X=0)}\right]$$

b. Modelo da Razão Contínua

Este modelo, proposto por Fienberg (1980 *in* Long e Cheng, 2004), foi desenhado para processos ordinais que representam um conjunto de eventos ou estados sucessivos. Permite comparar a probabilidade de uma resposta igual a uma determinada categoria com a probabilidade de uma resposta maior, $Y > y$ (Abreu *et. al.*, 2009).O modelo *logit* é escrito por:

$$r_k(x) = ln\left[\frac{P(Y=k|X)}{P(Y<k|X))}\right] = ln\left[\frac{\phi_k(X)}{\phi_0(X) + \phi_1(X) + \cdots + \phi_{k-1}(X)}\right] = \theta_k + x'\beta_k$$

Onde k=1,2,…,K. De acordo com a parametrização da equação anterior, para cada comparação existem diferentes intercepções e declives resultantes dos vários *logits* que são ajustados (Hosmer e Lemeshow, 2000).

São apropriados quando existe interesse em uma categoria específica da variável resposta (Abreu *et. al.*, 2009).

c. Modelo de Riscos Proporcionais

São modelos que comparam a probabilidade de uma resposta menor ou igual a uma categoria, $\mathrm{Y} \leq \mathrm{k}$, com a probabilidade de uma resposta maior, $\mathrm{Y} > \mathrm{k}$. São considerados $(\mathrm{k} - 1)$ pontos de corte das categorias, sendo que o último é baseado na comparação das probabilidades acumuladas (Abreu *et. al.*, 2009).

Estes modelos são geralmente utilizados na análise de variáveis ordinais, ou provenientes de uma variável contínua que foi agrupada.

O modelo é escrito na seguinte forma:

$$o_k(x) = ln\frac{[P(Y \leq k|X)]}{P(Y > k|X)} = ln\left[\frac{\phi_0(X) + \phi_2(X) \ldots + \phi_k(X)}{\phi_{k+1}(X) + \phi_{k+2}(X) \ldots + \phi_k(X)}\right] = \tau_k + X'\beta$$

para k=1,2,…,K.

Ou seja, a relação existente entre a categoria mais baixa e as restantes é a mesma que entre a segunda categoria mais baixa e as seguintes. Esta característica depende do pressuposto de riscos proporcionais, que deverá ser assumido para cada covariável incorporada no modelo (Abreu *et. al.*, 2009; Hosmer e Lemeshow,2000).

No *software* estatístico Stata, estão disponíveis dois testes que permitem avaliar o pressuposto dos riscos proporcionais:

- O teste de razão de verossimilhança aproximado desenvolvido por Wolfe e Gould (1998) que compara a verosimilhança do modelo ordinal ajustado com o modelo obtido a partir da junção de $j - 1$ modelos logísticos binários, fazendo um ajustamento para a correcção da resposta binária. Para que o pressuposto seja cumprido será necessário não rejeitar a hipótese nula, de que não existem diferenças entre os coeficientes dos dois modelos. Este teste não indica se a violação do pressuposto ocorre em determinada variável (Long e Cheng, 2004).
- O teste Wald de Brant, também designado de teste para o pressuposto da regressão paralela, permite testar simultaneamente e individualmente os coeficientes do modelo, indicando assim quais as covariáveis mais problemáticas (Long e Cheng, 2004).

Caso se verifique que o pressuposto dos riscos proporcionais é violado, será necessário recorrer a outras abordagens, tal como o modelo de regressão multinomial ou mediante os modelos logísticos ordinais generalizados. No Stata, estes modelos são construídos pela utilização do comando gologit,

desenvolvido por Vicent´s Fu (1998 *in* Williams, 2006), ou pelo comando gologit2, desenvolvido a partir do primeiro, por Williams (2006). Este último, considerado menos restritivo, permite calcular outros tipos de modelos:

- Modelos ordinais generalizados, em que nenhuma variável necessita de cumprir o pressuposto da proporcionalidade de riscos:

$$P(Y_i > j) = g(X\beta_j) = \frac{e^{\alpha_j + X_i\beta_j}}{1 + e^{\alpha_j + X_i\beta_j}}, j = 1,2, \dots, M-1$$

  Onde M representa o número de categorias da variável resposta.

  Assim, para $M = 2$, obtêm-se um modelo de regressão logística binário. Quando $M > 2$, surgem vários modelos logísticos binários, com combinações das categorias da variável resposta. Por exemplo, se $M = 3$, obtêm-se três equações de regressão, onde para $J = 1$ se confronta a categoria 1 com a 2, 3 e 4; para $J = 2$ as categorias 1 e 2 com a 3 e 4; e para $J = 3$ as categorias 1, 2 e 3 *versus* 4.

- Modelos de riscos proporcionais parciais, em que apenas algumas variáveis validam o pressuposto da proporcionalidade. Assim, alguns $\beta's$ são idênticos para todos os valores de j, enquanto outros podem variar. Por exemplo, na seguinte equação $X1$ e $X2$ são idênticos, variando apenas $X3$:

$$P(Y_i > j) = g(X\beta_j) = \frac{e^{\alpha_j + X1_i\beta 1 + X2_i\beta 2 + X3_i\beta 3_j}}{1 + e^{\alpha_j + X1_i\beta 1 + X2_i\beta 2 + X3_i\beta 3_j}}, j = 1,2, \dots, M-1$$

  Este tipo de modelo é mais parcimonioso que o anterior e as suas estimativas são de mais fácil de interpretação do que as obtidas nos modelos não ordinais, como no caso da regressão multinomial (Williams, 2006).

Peterson e Harrel (1990 *in* Williams, 2006) propuseram uma parametrização equivalente aos modelos anteriores, designada de modelo de riscos proporcionais parciais irrestrito. De acordo com estes autores, cada variável explicativa é constituída por coeficiente β e por $M - 2\gamma$ coeficientes, onde M representa o número de categorias em Y e γ representa os desvios à proporcionalidade.

3.5.1 Construção do Modelo

As etapas utilizadas na construção de um modelo logístico ordinal são idênticas às utilizadas para o modelo Logístico Binário.

Os modelos ordinais não têm nem as estatísticas de diagnóstico nem os testes de bondade de ajustamento solidamente desenvolvidos, sendo necessário recorrer às técnicas do modelo de regressão logística binário, com a desvantagem de não se estar a avaliar o modelo ajustado actual,

mas sim uma aproximação deste. Assim, todos os pontos identificados com tendo um mau ajustamento ou classificados como influentes, devem ser testados, eliminando estas observações e ajustando um novo modelo.

4. Análise Exploratória de Dados

De acordo com o referido anteriormente, foram incluídos no estudo mulheres seguidas em consulta de Ginecologia ou Obstetrícia no Hospital Prof. Dr. Fernando da Fonseca ou acompanhadas em consulta dessas especialidades em Centros de Saúde da área de Influencia desse Hospital.

As duas subamostras, perfazem um total de 417 indivíduos, das quais 55% são provenientes do HFF e os restantes 45% dos Centros de Saúde (Figura 4.1).

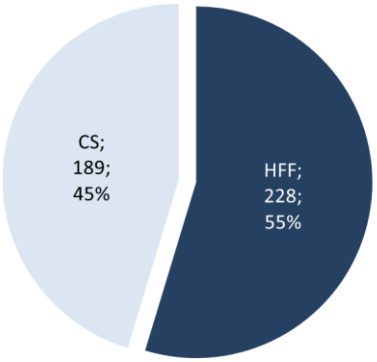

Figura 4.1 Distribuição dos indivíduos por área de recrutamento.

4.1 Características Gerais da Amostra

Este estudo contemplou mulheres cuja idade variou entre os 17 e os 79 anos de idade, obtendo-se uma idade média de 40.9 anos e um desvio padrão de 12.5 anos. Pela observação da Figura 4.2 e da Figura 4.3 podemos analisar a distribuição de frequências da idade, podendo-se verificar que esta é relativamente simétrica com achatamento platicúrtico. O diagrama de extremos e quartis, reforça a simetria da distribuição, sendo ainda visível um *outlier* que corresponde ao indivíduo mais idoso da amostra. Os testes à normalidade de, Shapiro-Francia (valor $\mathrm{p} = 0.00319$) e Kolmogorov-Smirnov (valor $p = 0{,}011711$) levam à rejeição da hipótese de que a distribuição da idade é normal, estando os problemas concentrados no achatamento, tal com o observado na Tabela 4.1. O Gráfico de Qualtil-Quantil (Figura 4.2) mostra que o afastamento à normalidade é mais acentuado nas caudas da distribuição, ou seja, nas classes jovens e mais velhas da amostra. No entanto, e analisando globalmente os resultados, verifica-se que apesar da amostra não ter uma

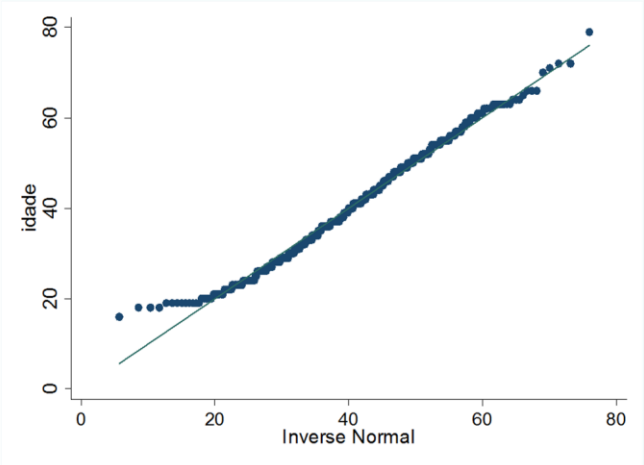

Figura 4.2 Gráfico de Quantil-Quantil para a distribuição da variável idade.

distribuição tipicamente normal, o seu afastamento não é muito acentuado. Posto isto, as comparações que serão efectuadas relativamente à idade serão abordadas mediante testes paramétricos e os seus equivalentes não paramétricos.

Tabela 4.1. Teste à normalidade relativamente à simetria e achatamento da variável idade.

Skewness/Kurtosis tests for Normality
------- joint ------

Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	Prob>chi2
Idade	417	0.1130	0.0005	13.04	0.0015

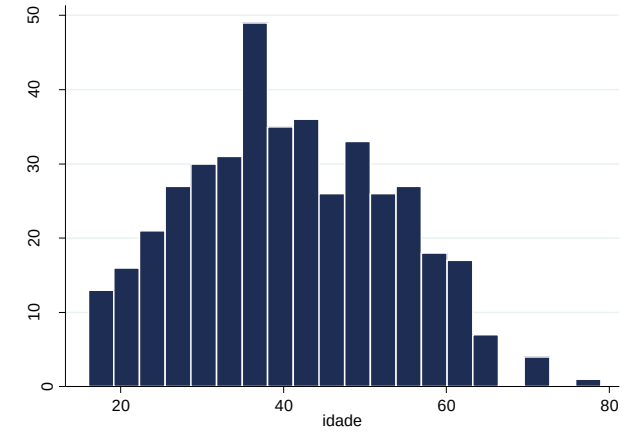

Figura 4.3 Distribuição de frequências de idade dos indivíduos da amostra.

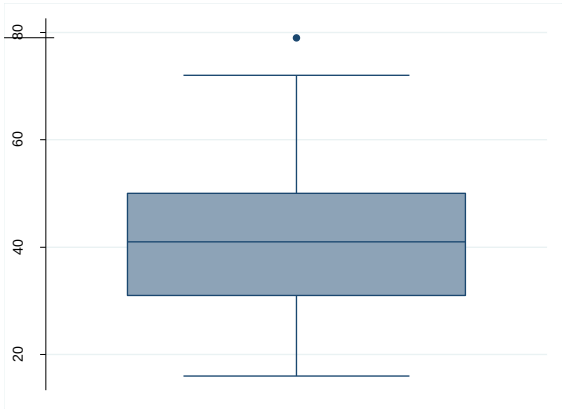

Figura 4.4 Gráfico de extremos e quartis de idade dos indivíduos da amostra.

Quanto ao grupo populacional, os indivíduos foram classificados em 5 classes: negróides (N=87; 20.9%), caucasóides (N=310; 74.3%), ameríndios (N=2; 0.5%), asiáticos (N=1; 0.2%) e outros (N=4; 0.99%). Houve ainda 13 mulheres onde não foi registado o seu grupo populacional, as quais representam 3.1% da amostra.

Outra variável considerada de interesse é a nacionalidade (portuguesa ou outra). A sua análise revelou que maioritariamente as mulheres são de nacionalidade portuguesa (N=329; 78.9%), onde apenas 18.5% (N=77) referiu ter outra nacionalidade e 2.6% (N=11) não respondeu ou não foi perguntado.

Como esta informação também está disponível na base de dados do HFF, foi analisada a concordância entre as duas fontes de informação, na esperança de que, caso a concordância fosse aceitável, se poderia usar a informação do HFF para preencher os dados em falta. No entanto, verificou-se que a base de dados de utentes é bastante incompleta no que respeita aos seus dados demográficos/geográficos. A nível da nacionalidade, apenas quatro utentes tinham registos coincidentes, sendo que os restantes 413 não tinham informação sobre a variável. Quanto ao grupo

populacional, verificou-se que no sistema informático do HFF, todas as utentes estão identificadas como tendo nacionalidade Portuguesa.

Assim, o objectivo inicial, de que amostra fosse o reflexo cultural/biológico das mulheres residentes nestes concelhos, não foi alcançado. Por outro lado, o tamanho reduzido da amostra também impossibilitou a comparação entre os diferentes grupos. Neste sentido, as comparações foram efectuadas, entre os dois maiores grupos, os caucasianos e negróides, que em conjunto representam 98.2% da amostra.

Foram também inquiridos hábitos dobre o comportamento sexual, nomeadamente se a mulher era sexualmente activa e quantos parceiros sexuais tinha tido ao longo da sua vida.

A análise dos questionários mostrou que 87.8% (N=366) das mulheres era, no momento do estudo, sexualmente activa. Em oposição, apenas 8.6% (N= 36) referiu não ter qualquer contacto sexual e 15 não respondeu ou não foi registada resposta. Embora esta variável seja considerada uma das mais importantes no estudo epidemiológico do HPV, a sua leitura no questionário está comprometida. Não é claro que uma resposta negativa signifique que nunca tenha ocorrido contacto sexual, ou porque no momento do inquérito a mulher não é sexualmente activa, e mesmo que a pergunta tenha sido interpretada correctamente, não existe informação sobre durante quanto tempo é que existe abstinência sexual. E esta é uma das questões fundamentais do estudo.

A questão relacionada com o número de parceiros sexuais foi inicialmente desenvolvida como uma variável ordinal composta por 4 níveis: (1) nenhum parceiro; (2) de 1 a 5; (3) de 6 a 10 e (4) mais de 11. De acordo com a Tabela 4.2, verificou-se uma enorme desproporcionalidade entre as classes, com 90.4% dos casos situados entre 1 e 5 parceiros. Apenas uma mulher respondeu não ter tido nenhum parceiro, pelo que se juntou esta classe com a seguinte. Pela mesma razão, as classes três e quatro foram também colapsadas. A nova configuração das classes está também apresentada na Tabela 4.2.

Tabela 4.2 Nº de observações por nível, antes (I) e depois do reagrupamento (II).

Observações por nível	Nº (I)	Nº (II)
Nenhum parceiro	1	378
1-5 Parceiros	377	
6-10 Parceiros	18	26
+ 11 Parceiros	8	
Indeterminado	13	13
Total	417	417

A comparação da média de idades entre as mulheres com e sem actividade sexual (39.7 e 51.4 anos, respectivamente) mostrou-se bastante díspar entre grupos (Figura 4.5). O pressuposto da normalidade é somente cumprido no grupo sem actividade sexual, onde se obtêm um valor p, para o teste de Kolmogorov-Smirnov, superior a 0.20. O grupo com actividade sexual apresenta um valor p para este teste de 0.016. Mais uma vez, os problemas na normalidade encontram-se no achatamento da distribuição, embora o afastamento à normalidade não seja muito pronunciado. Assim, a diferença de médias foi comprovada pelo teste t-Student, para amostras homocedásticas ($p = 0.4843$), onde se obtém um valor p inferior a 0.00001. Em alternativa, a abordagem não paramétrica, o teste de Wilcoxon-Mann-Whitney, indica que existe evidência estatística suficiente para considerar que são diferentes as medianas das duas amostras (valor p inferor a 0.00001).

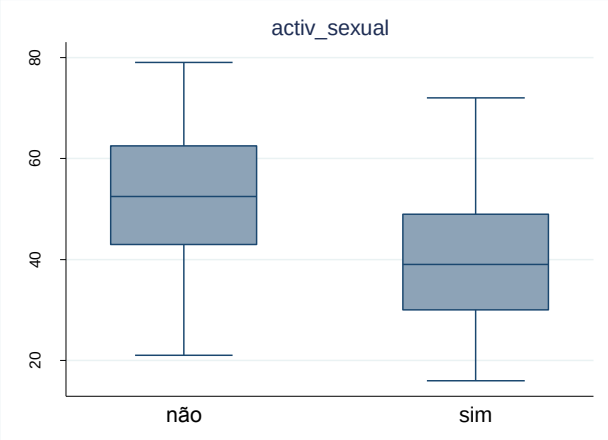

Figura 4.5 Gráfico de extremos e quartis de idade entre indivíduos com e sem actividade sexual.

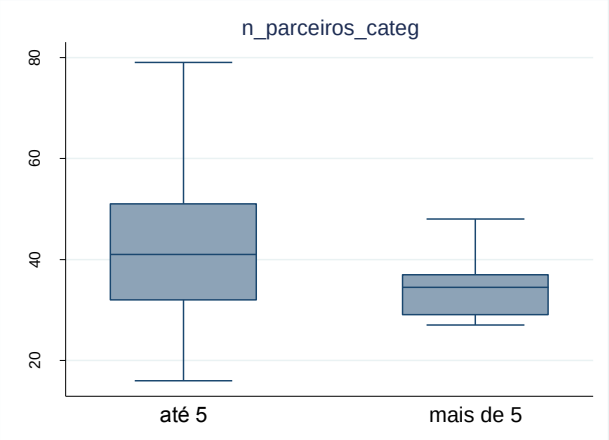

Figura 4.6 Gráfico de extremos e quartis de idade entre indivíduos com menos e mais de 5 parceiros.

A análise da Figura 4.6 permite verificar que são as mulheres mais jovens que têm mais de cinco parceiros. O teste de Kolmogorov-Smirnov mostrou que a idade tem uma distribuição aproximadamente normal nas duas amostras em estudo (valor p de 0.057 e 0.195), embora no grupo com mais de cinco parceiros o seu valor p esteja muito próximo do α considerado. O teste t –Student com a aplicação da fórmula de Welch, para variâncias desiguais (valor p inferior a 0.00001), e considerando um teste unilateral à direita, indica que a média de idades do grupo que têm no máximo até cinco parceiros é estatisticamente superior ao grupo com mais de cinco parceiros. São obtidos resultados semelhantes mediante aplicação do teste de Wilcoxon-Mann-Whitney cujo valor p é 0.00115.

O tipo de contraceptivo usado foi outra variável de interesse e contemplada no estudo. Uma vez que a mesma mulher pode usar diferentes tipos de contraceptivos, foi mantida uma variável para cada tipo de método (Figura 4.7).

No total, 134 (32.1%) mulheres não usam nenhum tipo de contraceptivo. Das que usam (N=239, 65.0%), o método mais frequente é a pílula (62.4%), seguido do preservativo (18.1%). Quinze das inquiridas respondem que utiliza simultaneamente mais do que um método, o preservativo com outro tipo, geralmente a pílula.

Das mulheres que responderam ter actualmente contacto sexual, cerca de 28.2% não usam nenhum tipo de contraceptivo. O teste do χ^2 permite identificar associação entre estas duas variáveis, obtendo-se um valor p inferior a 0.0001, com o teste exacto de Fisher.

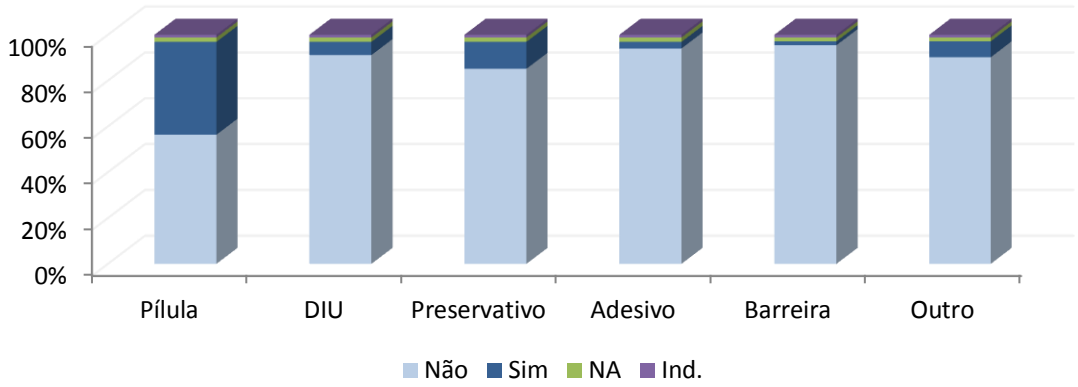

Figura 4.7 Percentagem de indivíduos que utilizam os vários tipos de contraceptivos.

Pela análise do gráfico da Figura 4.8 é possível observar que a média de idades do grupo que usa pelo menos um tipo de contraceptivos é mais jovem que a amostra que não usa nenhum método. O teste de Kolmogorov-Smirnov indica um possível afastamento à normalidade destas distribuições, no entanto, a análise dos coeficientes de achatamento e simetria, mostram que o afastamento não é muito pronunciado, e que resulta essencialmente em problemas na simetria. A comparação de médias entre estas amostras, foi efectuada pelo teste t, que para variâncias idênticas ($\mathrm{p} = 0.0557$), mostra um valor p (para um teste unilateral à esquerda)

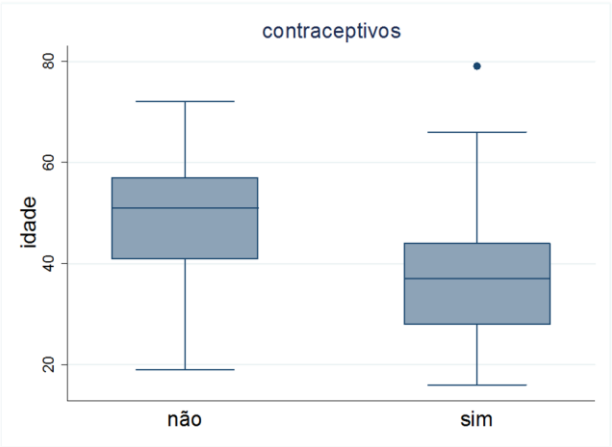

Figura 4.8 Gráfico de extremos e quartis de idade para os indivíduos que usam ou não contraceptivos.

inferior a 0.00001. O mesmo resultados é obtido considerando o teste de Wilcoxon-Mann-Whitney, de onde resulta, para o teste unilateral à esquerda, um valor p inferior a 0.00001.

Por último, foi questionado a existência de doenças sexualmente transmissíveis, onde se considerou as seguintes opções: (1) sífilis, (2) HIV, (3) clamídea e (4) outra. A análise dos resultados mostrou que este tipo de patologia estava presente em 17 indivíduos: uma infecção simultânea por sífilis e HIV, uma de sífilis, duas por HIV, três por clamídea e dez casos que reportaram ter outra DST. De salientar que estes resultados não foram confirmados clinicamente.

Foi ainda perguntado se actualmente faziam terapia local e de qual tipo. Contudo, dado a natureza da variável e os poucos casos a que se aplica (N=4, 1.0%), não foi considerada na análise estatística.

4.2 Tipos de HPV

Como se pode observar na Tabela 4.3, das 417 amostras analisadas, 310 (74.34%) foram negativas para a presença do vírus. As restantes, correspondendo a 25.66%, estavam infectadas com pelo menos um tipo de HPV. As infecções simples foram identificadas em 71 indivíduos (17.03%), enquanto que, infecções múltiplas estavam presentes em 36 indivíduos (8.63%).

Tabela 4.3 Resultados de infecção por HPV (negativos, infecções simples e infecções múltiplas).

Tipo de Infecção	Nº de HPV's	Nº e % de Indivíduos	
(I) Negativo	0	310	74,34%
(II) Positivo com Infecção Simples	1	71	17,03%
(III) Positivo com Infecção Múltipla	2	22	5,28%
	3	9	2,16%
	4	4	0,96%
	6	1	0,24%

Uma das questões fundamentais, e também incluídas nos objectivos iniciais deste trabalho, prende-se com a relação entre o número de infecções e o tipo de lesão associada. No entanto, a pequena dimensão da amostra, associada à desproporcionalidade de indivíduos que representam cada lesão , impossibilita as comparações desejadas. Assim, nas análises futuras, optou-se por incluir os indivíduos em três grandes grupos: negativo, positivo com infecção simples e positivo com infecção múltipla (Tabela 4.3).

A Figura 4.9 permite analisar os vários grupos de HPV tendo em conta a sua frequência. Os vírus de alto risco representam 57.93% das infecções, sendo as estirpes 16, 52, 31 e 58 as frequentes (Figura 4.10). Por outro lado, as infecções de baixo grau, correspondem ao segundo grupo mais comum (17.68%), seguidas dos vírus de risco indeterminado (14.02%) e por último dos vírus de provável alto risco (10.37%).

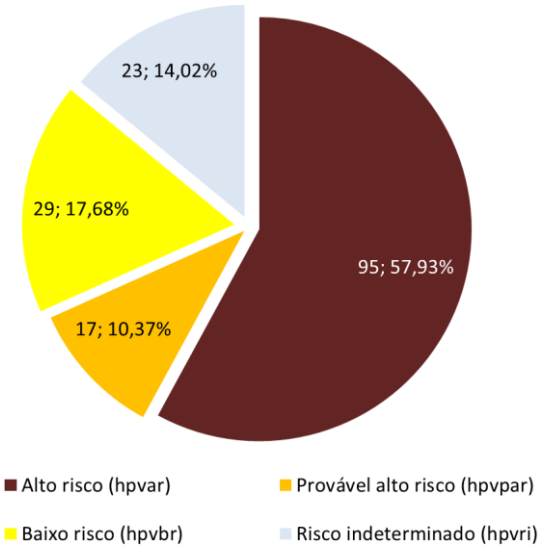

Figura 4.9 Distribuição dos grupos de HPV's.

As infecções múltiplas estão quase sempre associadas a um HPV de alto risco (94.44%), sendo a associação mais frequente a coexistência entre dois ou mais vírus deste grupo. Isto é, das 36 infecções múltiplas, 17 têm dois tipos de HPV de alto risco, duas têm três e num caso foram identificados 4 vírus. Tal como observado na Tabela 4.4, as restantes co-infecções são variáveis, não sendo observado nenhum padrão específico.

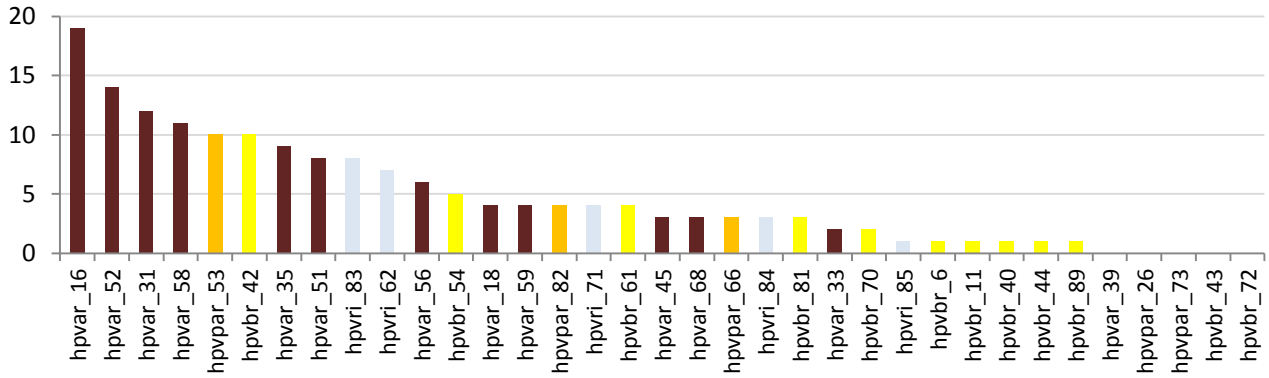

Figura 4.10 Distribuição das estripes de HPV's na população. Legenda: vermelho: alto risco (hpvar); laranja: provável alto risco (hpvpar); amarelo: baixo risco (hpv_br) e azul: risco indeterminado (hpvri).

Considerando que a distribuição de idades é aproximadamente normal (valor p do teste de Kolmogorov-Smirnov 0.093 e superior a 0.200 para o grupo sem HPV e com HPV, respectivamente), e que existe homocedasticidade entre grupos (valor $p = 0.1691$), a aplicação do teste t para amostras independentes, indica que, média de idades do grupo com HPV (34.5 anos) é inferior à média de idades do grupo livre de infecção (43.0 anos). Neste caso, obtém-se um valor p para um teste unilateral à direita inferior a 0.00001.

Por outro lado, quando se observam as médias de idades entre indivíduos sem infecção, com infecção simples e com infecção múltipla, levanta-se de imediato a suspeita que poderão existir diferenças entre grupos (Tabela 4.5).

Tabela 4.4. infecções múltiplas de acordo com a classificação e designação dos HPV's.

Individuo	Nº HPV's	Alto risco	Provável alto risco	Risco indeterminado	Baixo risco
A	2	16	53		
B	2	59		54	
C	2	58		54	
D	3	51 58	82		
E	2	31	53		
F	3	31		62	42
G	2	52 56			
H	3	15 56		84	
I	2			83	42
J	2	35 58			
K	2	51	82		
L	2	31		62	
M	4			82 83	
N	2	16	82		
O	2	16 68			
P	2	16 58			
Q	3	35		83	54
R	2	18			11
S	2	33 58			
T	2	45 52			
U	2	56 58			
V	3	18			42 81
W	3	33 35	53		
X	3	35 52 58			
Y	4	51 52		84	6
Z	6	16 35 51 56			40 81
AA	3	31 59			54
AB	2	51			54
AC	4	52 68		71	70
AD	2	82		85	
AE	2	52 56			
AF	2	31			42
AG	2	58			61
AH	3	31		83	70
AI	2	45	53		
AJ	4		66 82	83	42
Total de HPV´s	**93**	**54**	**9**	**13**	**17**
	50%	**29%**	**5%**	**7%**	**9%**
Nº de indivíduos com pelo menos 1 tipo	36	34	8	12	15
	100,00%	94,44%	22,22%	33,33%	41,67%

Tabela 4.5 Média de desvio padrão da variável n_hpv_cat.

n_hpv_cat	Summary of idade		
	Mean	Std. Dev.	Freq.
sem infecção	43.06129	12.229181	310
inf. simples	35.732394	11.471652	71
inf. múltipla	32.027778	9.3884005	36
Total	40.860911	12.469765	417

Considerando que: (1) existe independência, (2) cumprido o pressuposto da homogeneidade de variâncias pela aplicação do teste de Bartlett ($p = 0.135$) e análise do gráfico dos valores ajustados contra os resíduos standartizados (Figura 4.12), e (3) verificada ainda a normalidade dos resíduos, pelo teste à normalidade de Kolmogorov-Smirnov aos resíduos *standartizados* ($p = 0.0682$) e análise do gráfico Quartil-Quartil (Figura 4.11, estão reunidas as condições necessárias para a realização de uma análise de variância a um factor.

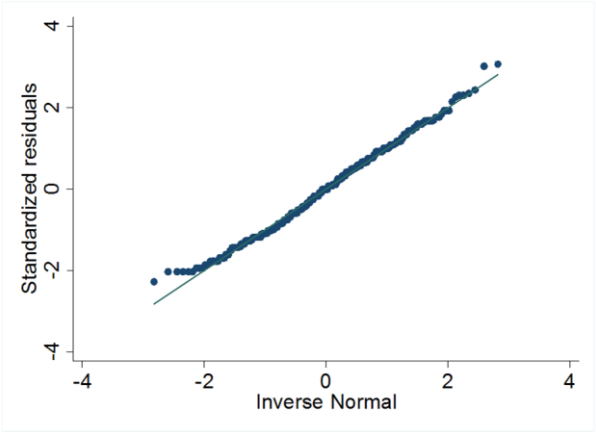

Figura 4.11 Gráfico de Quantil-Quantil para os resíduos *standartizados* do modelo ANOVA.

Embora a análise dos resíduos *standartizados* evidencie dois pontos com valor próximo de três (Figura 4.12), nenhum deles se destaca dos demais, indicando que não deverão haver problemas com *outliers*.

A análise da Tabela 4.6 permite constatar que o valor de F é significativo ($p < 0.00001$), concluindo-se assim que existem diferenças em pelo menos um par de médias. Dado, a desproporcionalidade entre classes, a identificação das diferenças pode ser avaliada pelo teste de Scheffé, método este robusto relativamente aos pressupostos de normalidade e igualdade de variâncias. Os resultados do teste de comparação múltipla encontram-se na Tabela 4.7, na qual se pode observar que a média do grupo sem infecção

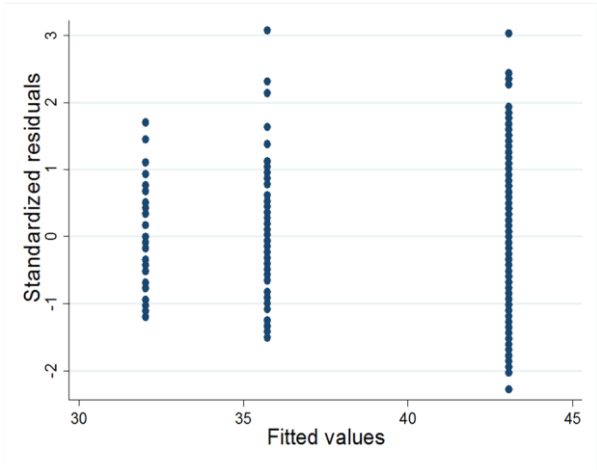

Figura 4.12 Gráfico dos valores ajustados *versus* resíduos standartizados.

difere estatisticamente dos restantes, mas que a os grupos com infecção simples e múltipla têm médias semelhantes.

Tabela 4.6 Resultados da análise de variância relativamente às comparações entre médias de idade e tipo de infecção.

```
                           Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      6177.20965      2   3088.60483     21.85     0.0000
 Within groups      58508.7232    414   141.325418
------------------------------------------------------------------------
    Total           64685.9329    416   155.495031

Bartlett's test for equal variances:  chi2(2) =   4.0017  Prob>chi2 = 0.135
```

Tabela 4.7 Resultados do teste de Scheffé relativamente às comparações entre médias de idade e tipo de infecção.

```
        Comparison of idade by n_hpv_cat
                      (Scheffe)
Row Mean-     |
Col Mean      | sem infecção   inf. simples
--------------+------------------------------
inf. simples  |    -7.3289
              |      0.000
inf. múltipla |   -11.0335      -3.70462
              |      0.000         0.315
```

Embora tenham sido considerados cumpridos os pressupostos requeridos para a ANOVA, a normalidade dos resíduos relevou alguns constrangimentos, optando-se por comparar os resultados com o equivalente não paramétrico, o teste de Kruskal-Wallis. De acordo com este método, existe pelo menos um par de médias que difere dos restantes (valor $p = 0.0004$). A diferença entre grupos, calculada pela aplicação do teste de Man-Whitney, com a correcção de Bonferroni, indicou que, tal como nos testes de comparação múltipla da ANOVA, a média de idades do grupo sem infecção difere da média dos restantes grupos ($p < 0.00001$, em ambos os casos), mas que a média do grupo com infecção simples não difere da média do grupo com infecção múltipla ($p = 0.0834$). A correcção de Holmes aponta no mesmo sentido, indicando que o único par de médias que não difere entre si são os grupos com infecção com o valor p a corresponder a 0.166.

A análise da Figura 4.13 permite identificar que a distribuição das várias estripes de HPV é diferente entre nacionalidades, verificando-se que na amostra de nacionalidade Portuguesa, as estripes mais frequentes são, por ordem decrescente, 16, 52, 42, 31 e 58, enquanto que, na amostra de outra nacionalidade, as estripes predominantes, são a 83, seguida da 53, 16, 31 e 58. A falta de representatividade de indivíduos em algumas classes impediu que as comparações fossem

efectuadas a este nível, optando-se por comparar os agrupamentos efectuados tendo em conta a severidade do vírus (Figura 4.14).

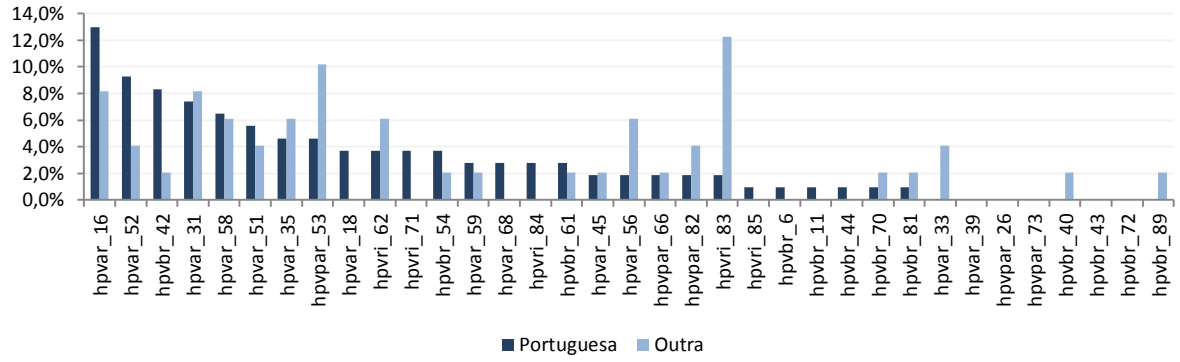

Figura 4.13 Distribuição das estripes de HPV's por Nacionalidade (Portuguesa *vs* outra).

Assim, os resultados obtidos a partir do teste χ^2, mostraram que o tipo de HPV (por grupo de severidade) é independente da nacionalidade ($p = 0.1072$). No entanto, há que ter em consideração a discrepância na representatividade dos grupos, onde 78.9% dos indivíduos refere ter nacionalidade Portuguesa.

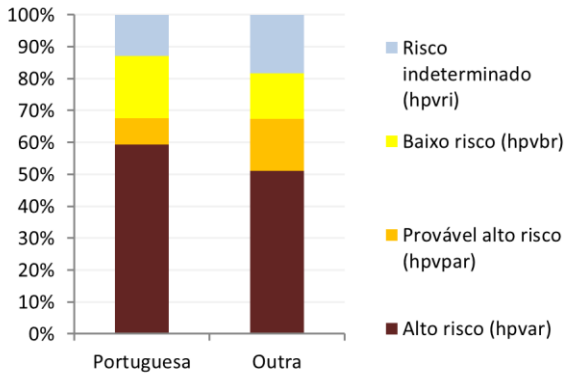

Figura 4.14 Distribuição dos grupos de HPV por Nacionalidade.

Embora as mulheres sexualmente activas tenham uma maior incidência de infecção (89.7% *versus* 95.1%) (Figura 4.15), a análise inferencial não permite concluir que a proporção de infecção seja diferente entre grupos ($p = 0.111$, pelo teste de χ^2).

Por outro lado, a comparação entre mulheres com menos e mais de cinco parceiros (Figura 4.16), mostrou ser altamente significativa, concluindo-se que a infecção por HPV não é independente desta variável ($p = 0.001$) . Na Tabela 4.8 estão apresentados os valores para a frequência observada, para os valores

Tabela 4.8 Valores obtidos para os valores observados, valores esperados e resíduos de Pearson na Tabela (hpv *vs* n_parceiros_categ_2).

```
observed frequency
expected frequency
Pearson residual

-------------------------------
           | n_parceiros_categ_2
       hpv |     até 5  mais de 5
-----------+-------------------
       não |       292        12
           |   284.436    19.564
           |     0.449    -1.710
           |
       sim |        86        14
           |    93.564     6.436
           |    -0.782     2.982
-------------------------------
Pearson chi2(1) =  12.6284         Pr = 0.000
likelihood-ratio chi2(1) = 10.858 Pr = 0.001
```

esperados e para os resíduos de Pearson. Assim, o maior afastamento entre os valores previstos e observados, traduzindo-se também num resíduo mais elevado, permite concluir que as mulheres com mais de cinco parceiros têm tendencialmente mais infecções por HPV.

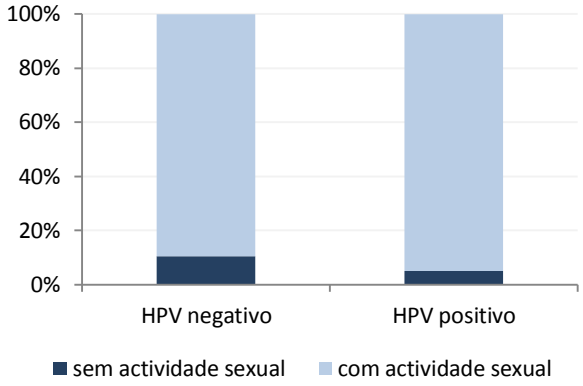

Figura 4.15 Distribuição dos grupos de HPV nas mulheres com e sem actividade sexual.

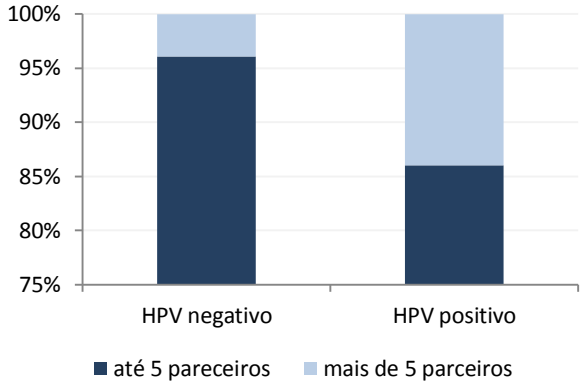

Figura 4.16 Distribuição dos grupos de HPV nas mulheres com menos e mais de 5 parceiros.

A infecção por HPV mostrou estar mais associada aos indivíduos que usam algum tipo de contraceptivo ($\mathrm{p} = 0.011$, pelo teste de χ^2). No entanto, esta relação poderá ser consequência destes indivíduos serem sexualmente activos e por isso estarem mais susceptíveis à infecção (Tabela 4.9). Assim, para testar a independência condicional relativamente ao ter ou não actividade sexual, realizou-se o teste de Cochran-Mantel-Haenszel, o qual com um valor p de 0.0078, leva a rejeitar a hipótese de independência, ou seja, existe associação entre ter HPV e usar contraceptivos condicionalmente a ter actividade sexual.

Tabela 4.9 Comparação entre resultado de HPV e uso de contraceptivos relativamente a existência ou não de actividade sexual.

	Sem contraceptivos		Com contraceptivos	
	HPV Negativo	HPV Positivo	HPV Negativo	HPV Positivo
Sem actividade sexual	22	2	6	2
	91,67%	**8,33%**	75,00%	**25,00%**
Com actividade sexual	84	18	181	79
	82,35	**17,65**	69,62	**30,38**
Total	106	20	187	81
	84,1%	**15,9%**	69,8%	**30,2%**

Relativamente ao tipo de contraceptivo usado, as diferenças significativas, calculadas a partir do teste de χ^2, foram identificadas nas mulheres que usam a pílula ($p = 0.011$), o adesivo ($p = 0.016$), o DIU ($p = 0.030$), ou outro qualquer método ($p = 0.049$). Para o preservativo e para a barreira não foram encontradas diferenças entre grupos.

4.3 Resultados Citológicos

De acordo com o estabelecido na metodologia deste projecto, foram realizadas citologias cervico-vaginais a todas as mulheres que participaram no estudo. Os resultados citológicos permitiram identificar a existência de alterações resultantes de inflamação por microorganismos, bem como, alterações, lesões (ASC-US, ASC-H, LSIL, HSIL) e neoplasias malignas das células pavimentosas e glandulares (carcinoma, carcinoma pavimentocelular, adenocarcinoma).

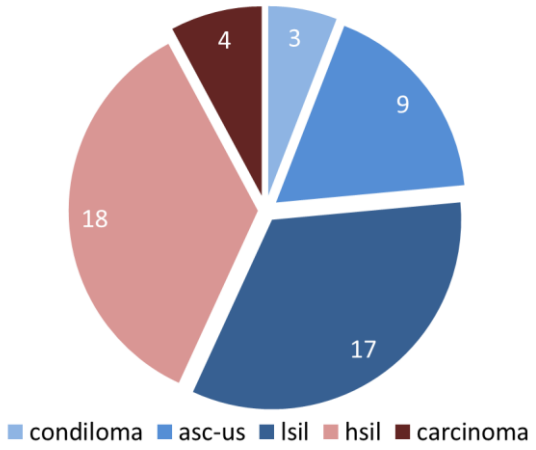

Figura 4.17 Distribuição do tipo de lesões (resultados citológicos).

Das 417 mulheres que compõem a amostra, 12.0% apresentam evidência de alterações patológicas (n=50) (Figura 4.17). As lesões de baixo grau, *i.e.*, os condilomas, ASC-US e LSIL, estão presentes em 6.7% dos casos, enquanto que, condições mais severas, HSIL e carcinomas, foram diagnosticados em 5.3% das mulheres (Tabela 4.10).

Tabela 4.10. Frequência e percentagem dos vários tipos de lesões.

HPV	baixo grau		condiloma	asc-us	lsil	alto grau		hsil	carcinoma
Negativo	**10**	**2,4%**	3	4	4	**3**	**0,7%**	2	1
Positivo	**18**	**4,3%**	0	5	13	**19**	**4,6%**	16	3
Total	**28**	**6,7%**	3	9	17	**22**	**5,3%**	18	4
									417

A comparação entre idades médias relativamente ao tipo de lesão (ligeiro *versus* severo) (Figura 4.18) pode ser efectuada através do Teste t-Student. Considerando cumprido o pressuposto de independência, que a idade tem uma distribuição aproximadamente normal (ambos os grupos com valor p para o teste de Kolmogorov-Smirnov de 0.200) e que existe homecedasticidade entre grupos ($p = 0.3330$), é possível rejeitar a hipótese nula das médias serem idênticas ($p = 0.0044$). Assim, conclui-se que mulheres mais jovens têm predominantemente lesões ligeiras, enquanto que, lesões pré-cancerígenas ou cancerígenas atingem mulheres com idades próximas dos 40 anos (Figura 4.15).

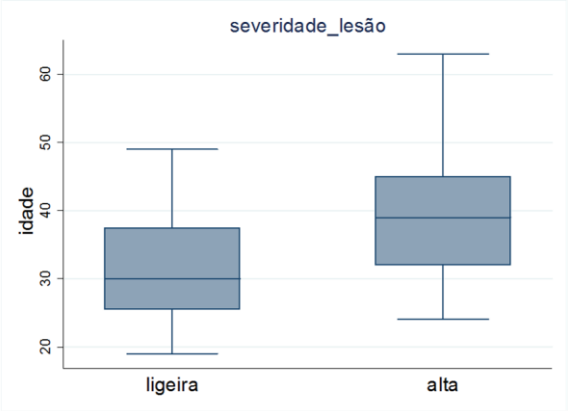

Figura 4.18 Gráfico de extremos e quartis de idade para as mulheres com lesões ligeiras e severas.

A análise da Figura 4.19 aponta também no mesmo sentido, sendo possivel identificar uma concentração das lesões ligeras (a azul) no extremo esquerdo do gráfico, enquanto que as lesões severas (a rosa) se encontram à direita, em faixas etárias superiores.

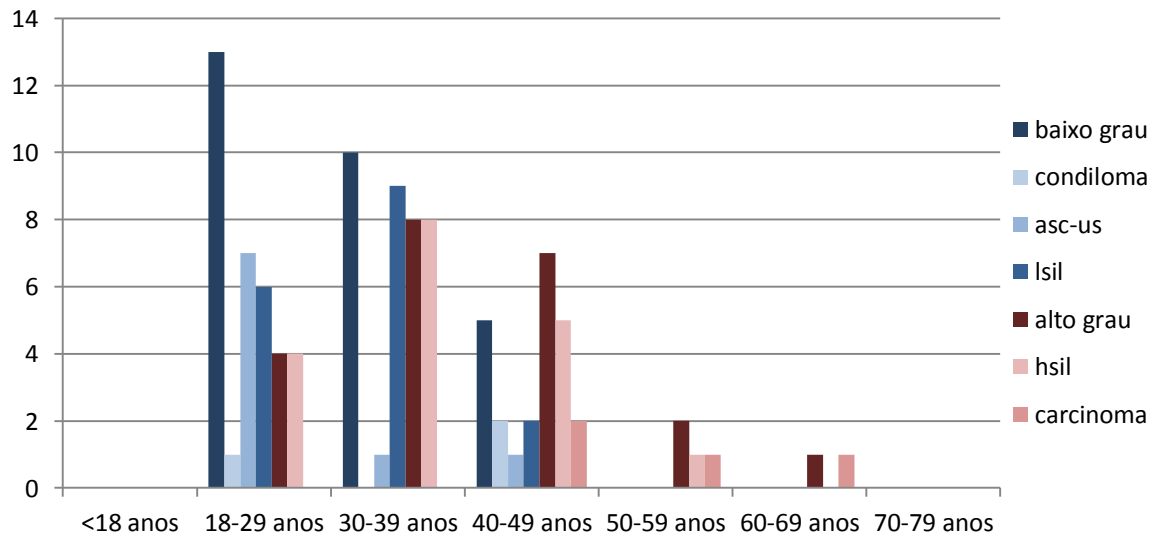

Figura 4.19 Distribuição das lesões por faixa etária. Legenda: lesões ligeiras a azul e lesões severas a rosa.

A presença de alterações patológicas é consistente com a existência de HPV, observando-se que, 64% das lesões ligeiras ocorrem em indivíduos infectados pelo vírus, subindo essa percentagem para os 86% quando as lesões são de grau elevado. Independentemente do tipo de lesão, os HPV de alto risco são sempre os mais frequentes, apesar de representarem 53% das infecções nas lesões ligeiras

e 70% nas mais severas (Tabela 4.11). Como seria de esperar, pelas proporções apresentadas, o tipo de lesão (ligeira *versus* severa) é dependente da existência de HPV de alto risco, obtendo-se um valor p de 0.017, pela utilização do teste exacto de Fisher.

Tabela 4.11 Classes de HPV por tipo de lesão.

HPV	baixo grau		condiloma	asc-us	lsil	alto grau		hsil	carcinoma
hpv_alto_risco	**15**	**53,6%**	0	4	11	**19**	**70,4%**	16	3
hpv_prov_alto_risco	**3**	**10,7%**	0	0	3	**2**	**7,4%**	2	0
hpv_risco_indet	**5**	**17,9%**	0	2	3	**3**	**11,1%**	3	0
hpv_baixo_risco	**5**	**17,9%**	0	1	4	**3**	**11,1%**	3	0
Total	**28**	**100,0%**	0	7	21	**27**	**100,0%**	24	3

Por outro lado, as infecções múltiplas aparentam ser independentes ao tipo de lesão (valor p do teste exacto de Fisher 0.612), correspondendo aproximadamente a 32% casos (Tabela 4.13). O número de HPV's por indivíduo é variável, observando-se conforme já referido anteriormente, uma mulher com infecção simultânea de 6 estirpes (4 de alto e 2 de baixo grau).

Tabela 4.13 Número de infecções múltiplas por tipo de lesão.

HPV	baixo grau	condiloma	asc-us	lsil	alto grau	hsil	carcinoma
Infecções Simples	**19**	3	8	9	**15**	11	4
Infecções Múltiplas	**9**	0	1	8	**7**	7	0
2 HPV's	**3**	0	0	3	**5**	5	0
3 HPV's	**2**	0	0	2	**2**	2	0
4 HPV's	**3**	0	1	2	**0**	0	0
6 HPV's	**1**	0	0	1	**0**	0	0
Total Geral	**32,1%**	0,0%	11,1%	47,1%	**31,8%**	38,9%	0,0%

5. Regressão Logística para a infecção por HPV

5.1 Regressão Logística Univariada e Multivariada

Neste capítulo pretende-se averiguar o efeito das variáveis idade, grupo populacional, nacionalidade, actividade sexual, número de parceiros sexuais, existência de doenças sexualmente transmissíveis (sífilis, HIV, clamídia, outra), uso de contraceptivos (preservativo, pílula, dispositivo intra-uterino, adesivo, barreira, outro) e uso de terapia local na infecção por HPV.

Antes da construção do modelo logístico multivariado serão analisadas individualmente as variáveis, tendo em conta o seu valor p referente à estatística de Wald. Na Tabela 5.1 apresentam-se os valores dos coeficientes do modelo de regressão, bem como o seu exponencial (*odds ratio*: OR), desvio padrão, valor da estatística de Wald e respectivo valor p e intervalos de confiança para OR.

Tabela 5.1 Regressão Logística univariada relativamente à infecção por HPV (valores estimados de OR, desvio padrão, estatística de Wald e IC 95% para OR).

Variáveis	β	OR	OR (Er.Padrão)	Estatística Wald (valor p.)	OR (Int. Conf.)	
					Inferior	Superior
proveniência	-0,3548	0,7013	0,1739	0,1530	0,4313	1,1402
idade	-0,0625	0,9394	0,0100	0,00001	0,9200	0,9591
grupo_pop_2	-0,6465	0,5239	0,1385	0,0140	0,3120	0,8796
grupo_pop_2(1)	0,6419	1,9000	0,9517	0,2000	0,7119	5,0711
nacionalidade	0,5289	1,6971	0,4653	0,0540	0,9916	2,9047
activ_sexual	0,8045	2,2357	1,1095	0,1050	0,8452	5,9134
n_parceiros_categ_2	1,3766	3,9612	1,6324	0,0010	1,7663	8,8838
dst	0,1964	1,2170	0,6628	0,7180	0,4185	3,5390
sífilis	1,0696	2,9143	4,1346	0,4510	0,1807	47,0063
hiv	1,7723	5,8846	7,2380	0,1500	0,5281	65,5677
outra	0,2218	1,2483	0,8731	0,7510	0,3169	4,9168
anticoncept_pílula	0,5937	1,8106	0,4145	0,0100	1,1560	2,8359
anticoncept_diu	0,9456	2,5745	1,0982	0,0270	1,1158	5,9400
anticoncept_preservativo	-0,4992	0,6070	0,2354	0,1980	0,2839	1,2981
anticoncept_adesivo	1,4351	4,2000	2,5076	0,0160	1,3033	13,5349
anticoncept_barreira	0,1360	1,1456	0,9674	0,8720	0,2189	5,9958
anticoncept_outro	-1,2127	0,2974	0,1842	0,0500	0,0883	1,0016
terap_local	1,0759	2,9327	2,9515	0,2850	0,4079	21,0833

Assim, da sua análise podemos concluir que:

- Por cada ano de idade, a possibilidade de ser positivo para o HPV diminui cerca de 6%. Caso se considere um intervalo de 5 anos, o risco diminui cerca de 30% ($e^{(-0.0625*5)} = 0.269$), ou diminui praticamente para metade ao considerar 10 anos ($e^{(-0.0625*10)} = 0.465$);

- Relativamente ao grupo populacional, as mulheres classificadas como outras não diferem estatisticamente das caucasianas, mas as negróides têm apenas metade do risco de estar infectadas comparativamente às caucasianas;
- A possibilidade de estar infectado aumenta quase quatro vezes mais quando se mais de cinco parceiros sexuais;
- As mulheres que usam dispositivo intra-uterino têm o seu risco aumentado em cerca de 80%;
- O uso do adesivo quadruplica a possibilidade de infecção.

A construção do modelo final foi efectuada de acordo com as recomendações de Hosmer e Lemeshow (2000) e descritas no Capítulo 3.

O modelo encontrado para a explicação da variável resposta (presença de HPV) é constituído apenas pelos efeitos principais da variável idade, grupo populacional, número de parceiros sexuais, infecção por HIV, uso de preservativo e uso de dispositivo intra-uterino. A Tabela 5.2 resume para cada variável o seu coeficiente, OR, valor da estatística de Wald e respectivo valor p e intervalos de confiança para OR.

Tabela 5.2 Regressão Logística multivariada relativamente à infecção por HPV (valores estimados de OR, desvio padrão, estatística de Wald, IC 95% para OR e teste de razão de verosimilhança entre o modelo actual e o modelo sem a variável).

Variáveis	β	OR	OR (Er.Padrão)	Estatística Wald (Signif.)	OR (Int. Conf.)		T. Razão Verosi. (valor p
					Inferior	Superior	
Idade	-0,0636	0,9383	0,0109	0,00001	0,9172	0,9600	<0.00001
n_parceiros_categ_2	1,1240	3,0772	1,3242	0,0090	1,3239	7,1524	<0.00001
Hiv	2,5484	12,7868	17,1029	0,0570	0,9295	175,9058	0.0482
anticoncept_diu	1,0352	2,8156	1,2672	0,0210	1,1654	6,8025	0.00244
anticoncept_preservativo	-0,8926	0,4096	0,1828	0,0450	0,1708	0,9821	0.0320

Sendo a idade uma variável contínua, um dos pressupostos que carece de validação é a sua linearidade com o *logit*.

Uma das abordagens possíveis corresponde à análise do gráfico de dispersão com alisamento da escala do *logit*. De acordo com a figura 5.1, observa-se um decréscimo relativamente constante até aos 60 anos de idade, seguindo-se de um aumento progressivo até aos 80 anos, idade máxima das mulheres estudadas. A tendência do alisamento do *logit* parece ser indicativo do não cumprimento do pressuposto da linearidade. No entanto, há que ter em consideração que este tipo de análise é sensível ao tamanho da amostra. Neste caso, verifica-se que a partir dos 60 anos o número de

indivíduos por idade é reduzido, fazendo com que um ponto mais afastado possa produzir um efeito semelhante ao observado no gráfico.

Para confirmação da linearidade, a idade foi subdividida de acordo com a distribuição dos seus quartis, sendo esta nova variável incluída no modelo de regressão. O gráfico da figura 5.2 (ponto médio dos quartis *versus* coeficientes da regressão) mostra uma diminuição sensivelmente constante do logaritmo dos *odds*, indicando assim poderá não haver problemas com linearidade da idade.

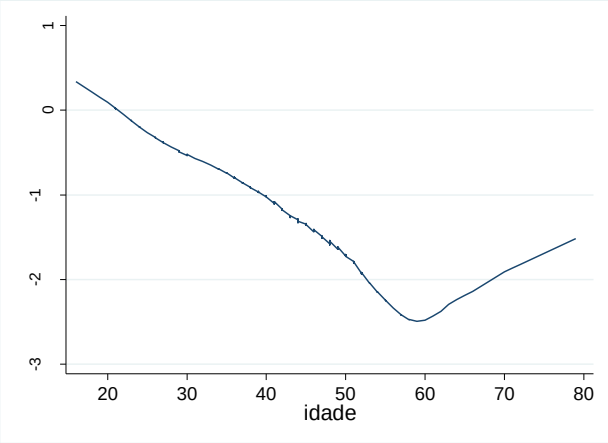

Figura 5.1 Gráfico de dispersão com alisamento da escala do *logit versus* idade.

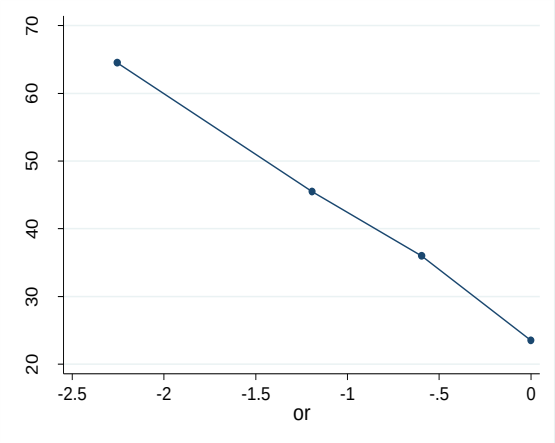

Figura 5.2 Gráfico dos coeficientes de regressão *versus* ponto médio dos quartis da variável idade.

Por último, foi aplicada a técnica dos polinómios fraccionários, com os seus resultados apresentados na seguinte Tabela.

Tabela 5.3 Resumo do método do polinómio fraccionário para a variável idade.

```
Fractional polynomial model comparisons:
------------------------------------------------------------------
idade             df       Deviance      Gain   P(term) Powers
------------------------------------------------------------------
Not in model       0        424.008        --     --
Linear             1        389.244     0.000    0.000  1
m = 1              2        389.044     0.200    0.655  .5
m = 2              4        388.281     0.963    0.683  3 3
------------------------------------------------------------------
```

- O nível de significância $p < 0.0001$ corresponde teste de razão de verosimilhança (para *df=1*) entre modelo com a variável idade *versus* o modelo sem esta variável.
- O nível de significância $p = 0.655$ corresponde ao teste de razão de verosimilhança, para um grau de liberdade, entre o modelo com a variável idade linear e o modelo com a transformação idade$^{0.5}$, *i.e.,* $G = 0.200$ *e* $P[\chi^2(1) \geq 0.200] = 0.655$.
- O nível de significância $p = 0.683$ corresponde ao teste de razão de verosimilhança, para dois graus de liberdade, entre o modelo com a transformação idade$^{0.5}$ e o modelo com a

transformação idade^3 e idade^3, *i.e.*, $G = 0.963 - 0.200 = 0.763$ *e* $P[\chi^2(2) \geq 0.763] = 0.683$.

- O nível de significância $p = 0.773$ corresponde ao teste de razão de verosimilhança, para três graus de liberdade, entre o modelo com a variável idade linear e o modelo com a transformação idade^3 e idade^3, *i.e.*, $G = 389.244 - 388.281 = 0.963$ e $P[\chi^2(3) \geq 0.963] = 0.810$.

Assim, se conclui que as melhores transformações obtidas pelo polinómio fraccionário não são estatisticamente melhores do modelo linear, assumindo-se que a variável idade é linear com o *logit*. Esta conclusão é ainda suportada, de acordo com o referido anteriormente, pela análise da idade através dos seus quartis.

5.2 Bondade do Ajustamento

O agrupamento das observações para o cálculo da bondade do ajustamento mediante a estatística de Hosmer-Lemeshow ($\hat{C}$), foi efectuado com base nos percentis das probabilidades estimadas. A Tabela 5.4, mostra para os 10 grupos calculados (decis de risco) as frequências estimadas e observadas. Tomando como exemplo, o quinto decil de risco, constituído por 31 indivíduos, a frequência observada para o grupo com HPV é nove, sendo 5.9 a sua frequência esperada. De modo análogo, a frequência observada para o grupo sem HPV é 22, sendo 25.1 a sua frequência esperada.

Tabela 5.4 Frequências observadas (Obs) e Estimadas (Exp) em cada decil de risco.

Logistic model for hpv, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0700	3	2.2	38	38.8	41
2	0.0993	0	3.6	40	36.4	40
3	0.1316	5	5.0	37	37.0	42
4	0.1724	7	6.9	37	37.1	44
5	0.2065	9	5.9	22	25.1	31
6	0.2574	13	10.2	30	32.8	43
7	0.3123	8	10.8	29	26.2	37
8	0.3958	13	14.4	27	25.6	40
9	0.4897	20	17.8	20	22.2	40
10	0.8059	21	22.3	17	15.7	38

number of observations = 396
number of groups = 10
Hosmer-Lemeshow chi2(8) = 9.10
Prob > chi2 = 0.3341

Uma vez que a estatística $\hat{C}$ depende de m-assimptótico, é necessário que as frequências estimadas não sejam inferiores a cinco. No presente caso, existem particulares constrangimentos nos três

primeiros decis de risco. Para solucionar este problema, dever-se-á agregar grupos adjacentes, aumentando assim o número das frequências esperadas e diminuindo simultaneamente os graus de liberdade. Condensando os três primeiros grupos, a $P[\chi^2(6) \geq 0.573] = 0.454$, ou seja, a estatística $\hat{C}$ parece indicar que o modelo tem um ajustamento adequado aos dados.

Na Tabela 5.5 está apresentada a Tabela de Classificação considerando um ponto de corte de 0.5. Da sua análise é possível concluir:

- A percentagem correcta de classificação é 76.01% [(21+280)/396];
- A sensibilidade, *i.e.*, a percentagem de indivíduos com HPV correctamente classificados é de 21.21% (21/99) e
- A especificidade, *i.e.*, a percentagem de indivíduos sem HPV correctamente classificados é de 94.28% (280/297).

Tabela 5.5. Tabela de Classificação para c=0.5.

```
Logistic model for hpv

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        21            17  |         38
     -     |        78           280  |        358
-----------+--------------------------+-----------
   Total   |        99           297  |        396

Classified + if predicted Pr(D) >= .5
True D defined as hpv != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   21.21%
Specificity                     Pr( -|~D)   94.28%
Positive predictive value       Pr( D| +)   55.26%
Negative predictive value       Pr(~D| -)   78.21%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    5.72%
False - rate for true D         Pr( -| D)   78.79%
False + rate for classified +   Pr(~D| +)   44.74%
False - rate for classified -   Pr( D| -)   21.79%
--------------------------------------------------
Correctly classified                        76.01%
--------------------------------------------------
```

Da análise anterior é possível concluir que apenas uma minoria de indivíduos com HPV são correctamente classificados. Assim, serão analisados outros pontos de corte de forma a maximizar o par sensibilidade/especificidade. A Tabela 5.6 mostra a evolução destes parâmetros quando se faz variar c entre 0.6 e 0.05.

Graficamente, o ponto que maximiza a sensibilidade e especificidade é obtido na zona onde as curvas se intersectam, ou seja entre 0.25 e 0.3. Seleccionando $c = 0.25$, obtêm-se uma sensibilidade e especificidade de 66.98% e 64.67%, respectivamente.

No entanto, o objectivo deste trabalho não é utilizar os modelos de regressão logística como método de classificação, mas identificar quais as variáveis com contributo relevante para a infecção. Para além do mais, dado a gravidade que é atribuída ao vírus do papiloma humano, seria aconselhável garantir que quando existe infecção, ela é identificada pelo modelo, *i.e.*, um modelo com melhor sensibilidade do que especificidade. Nesse sentido, diminuir-se-ia o valor do ponto de corte para os 0.15 onde a sensibilidade ronda os 89% e a especificidade os 43.7%.

Tabela 5.6. Resumo da Sensibilidade e Especificidade para uma variação do ponto de corte entre 0.6 e 0.5, com incremento de 0.05.

Ponto Corte	Sensibilidade	Especificidade	1-Especificidade
0,6	10,10%	97,98%	2,02%
0,55	12,12%	97,31%	2,69%
0,5	21,21%	94,28%	5,72%
0,45	27,27%	91,58%	8,42%
0,4	40,40%	87,54%	12,46%
0,35	48,48%	80,81%	19,19%
0,3	58,59%	76,09%	23,91%
0,25	**64,65%**	**64,98%**	**35,02%**
0,2	79,80%	55,89%	44,11%
0,15	88,89%	43,77%	56,23%
0,1	96,97%	26,26%	73,74%
0,05	9,00%	4,04%	95,96%

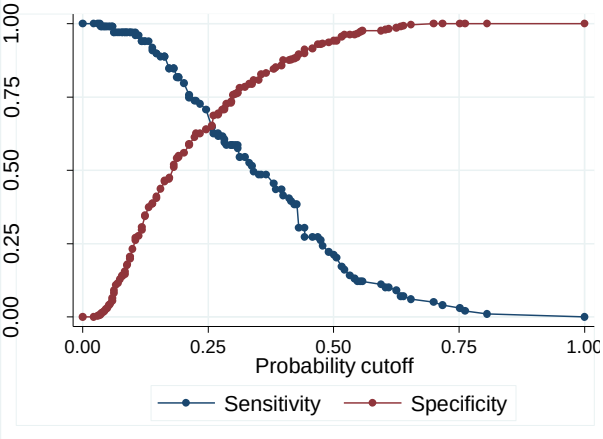

Figura 5.3. Gráfico da sensibilidade *vs* Especificidade para os vários pontos de corte.

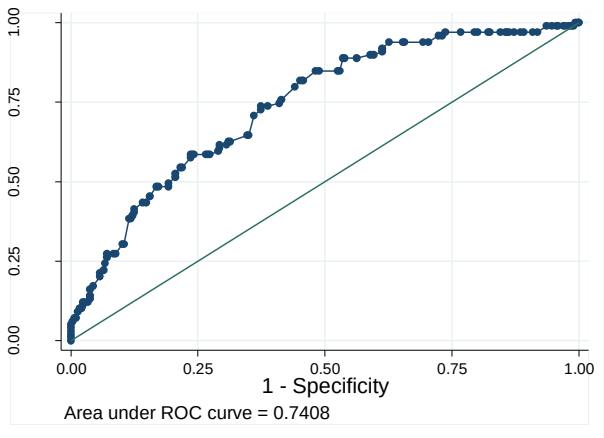

Figura 5.4. Curva ROC para a sensibilidade *vs* 1-Especificidade para os vários pontos de corte.

A figura 5.3 mostra o gráfico da sensibilidade *versus* 1-Sensibilidade, quando se fazem variar todos os pontos de corte possível. A área abaixo da curva é igual a 0.74 (IC 95%: 0.68-0.80) o que pela classificação adaptada por Hosmer e Lemeshow (2000), corresponde a um modelo com ajustamento aceitável.

5.3 Diagnóstico do Modelo

Numa primeira etapa a pesquisa de *outliers* e observações influentes foi efectuada graficamente pela análise dos vários resíduos contra as observações ou contra as probabilidades estimadas.

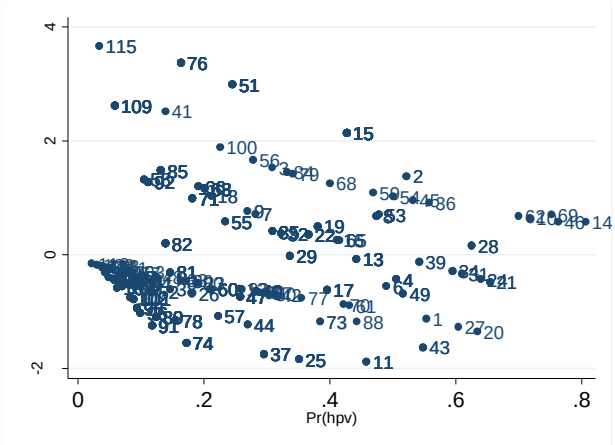

Figura 5.5 Gráfico dos resíduos de Pearson standartizados *versus* valores previstos.

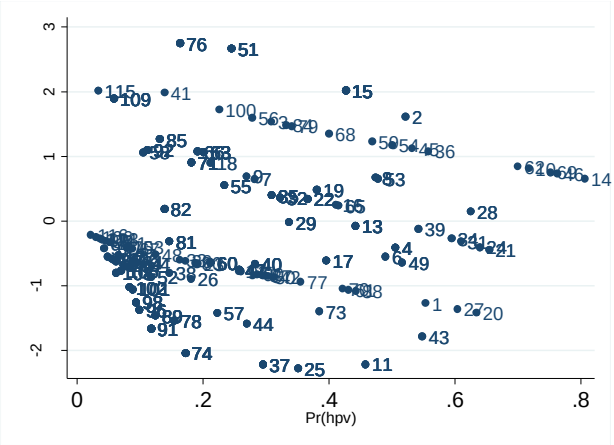

Figura 5.6 Gráfico dos resíduos *Deviance versus* valores previstos.

Os resíduos standartizados (Figura 5.5) mostram que algumas observações ultrapassam o limite 2-3 (padrões 41, 109 e 51). Em particular, um padrão de covariáveis atinge valores muito próximos de quatro (padrão 115).

A mesma informação é obtida através dos resíduos *deviance*, com os mesmos indivíduos a apresentarem os valores mais extremos (Figura 5.6). Apesar destes pontos exibirem resíduos elevados e ser por isso potenciais *outliers*, nenhum deles apresenta valores elevados de *leverage* (Figura 5.7), indicando que, caso estas observações sejam excluídas do modelo, as estimativas do modelo logístico não serão muito diferentes das actuais. De modo oposto, observam-se três padrões de covariáveis com resíduos *leverage* bem mais destacados (padrões 14, 69 e 88), sugerindo que estes pontos deverão corresponder a observações influentes nas estimativas do modelo.

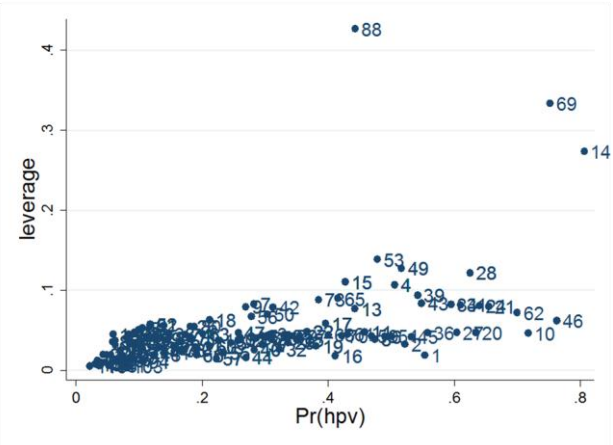

Figura 5.7 Gráfico dos resíduos *Leverage versus* valores previstos.

Foi também quantificado o ajustamento das observações pela eliminação de determinados indivíduos que apresentam padrões de covariáveis semelhantes. Esta abordagem consistiu na análise gráfica do $\Delta\chi^2$, ΔD e $\Delta\hat{\beta}$ contra as probabilidades estimadas pelo modelo e ainda ΔD contra a probabilidades estimadas mas com o tamanho dos pontos a representar a magnitude da influência de $\Delta\hat{\beta}$.

De acordo com Hosmer e Lemeshow (2000), o mau ajustamento pode ser detectado pela existência de pontos no canto superior esquerdo ou direito destes gráficos, ou ainda pela existência de pontos mais destacados dos restantes.

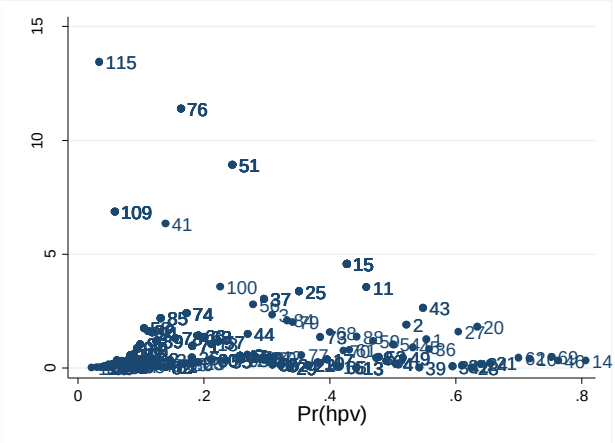

Figura 5.8 Gráfico $\Delta\chi^2_j$ *versus* probabilidades previstas ($\hat{\pi}_j$).

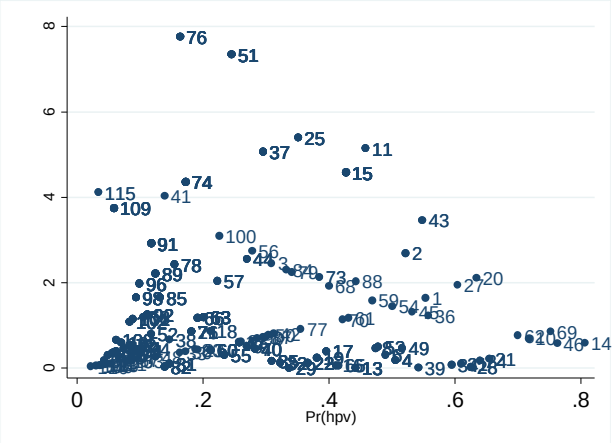

Figura 5.9 Gráfico ΔD_j *versus* probabilidades previstas ($\hat{\pi}_j$).

A análise da figura 5.8 mostra que a exclusão dos pontos já anteriormente identificados nas figuras 5.5 e 5.6, está associada a uma variação elevada do χ^2, que no caso do padrão 76 e 115 ultrapassam as 10 unidades. As propriedades inerentes aos resíduos *Pearson* e *Deviance*, fazem com que a amplitude de $\Delta\chi^2$ seja superior ao de ΔD (Hosmer e Lemeshow, 2000), conforme se pode observar pela comparação dos gráficos das figuras 5.8 e 5.9 .

Apesar de existirem pontos que evidenciam algumas suspeitas, a maioria dos valores de $\Delta\chi^2$ e ΔD são inferiores a 4, ou pelo menos não são muito distantes deste valor. O valor quatro é utilizado como referência, pois corresponde ao percentil 95 da distribuição de $\Delta\chi^2$ e ΔD $(\chi^2_{0.95}(1) = 3.84)$.

Na figura 5.10 está representada a influência do diagnóstico $\Delta\hat{\beta}$ contra a probabilidade estimada. Um dos pontos apresenta um afastamento mais acentuado da nuvem, no entanto, há que ter em consideração a pequena magnitude da escala e, adicionalmente, de acordo com as recomendações de Hosmer e Lemeshow (2000), a influência deve ser superior a uma unidade para que tenha um efeito nos coeficientes estimados.

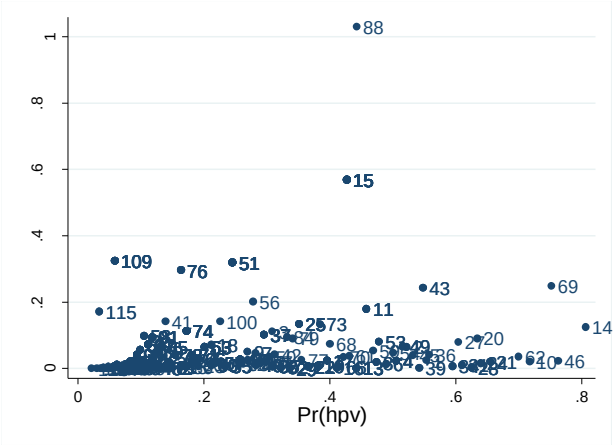

Figura 5.10. Gráfico $\Delta\hat{\beta}$ *versus* probabilidades previstas ($\hat{\pi}_j$).

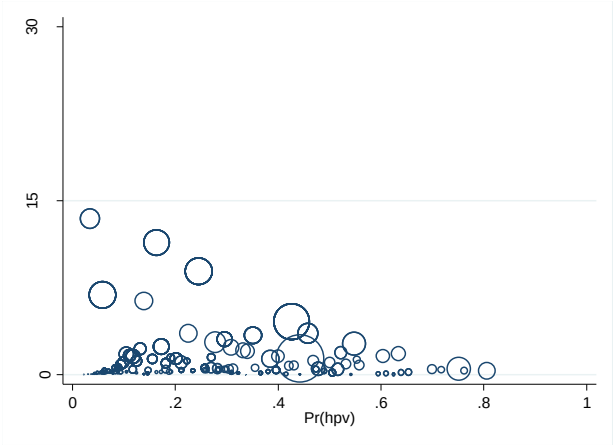

Figura 5.11. Gráfico $\Delta\chi^2_j$ *versus* probabilidades previstas ($\hat{\pi}_j$), sendo o tamanho dos pontos proporcional a $\Delta\hat{\beta}$.

O gráfico 5.11 confronta o $\Delta\chi^2_j$ com a probabilidade estimada, permitindo adicionalmente identificar em cada ponto a contribuição dos resíduos e da *leverage* em $\Delta\hat{\beta}$. Os círculos de maiores dimensões são observados no canto superior esquerdo, a que correspondem os valores mais elevados de $\Delta\chi^2_j$, mas também na região próxima de $\hat{\pi} \approx 0.4$, onde existe um valor baixo de $\Delta\chi^2_j$, mas numa região onde se espera um valor máximo da *leverage* (Hosmer e Lemeshow, 2000).

A análise conjunta dos gráficos indicam a existência de quatro conjuntos de observações que se destacam das demais (três com valores elevados de $\Delta\chi^2$ ou ΔD e outra relativamente ao $\Delta\hat{\beta}$). Estes conjuntos estão resumidos nas Tabela 5.7 e 5.8.

Considerando em primeiro lugar a eliminação das nove observações que constituem o grupo 51, a diminuição de χ^2 e D não parece ser relevante. O mesmo se aplica à variação percentual dos coeficientes, sempre inferior a 10%. Os restantes grupos apresentam comportamentos semelhantes, indicando que deverão permanecer no modelo. Ou seja, em qualquer dos casos, a eliminação das observações não conduz a alterações significativas nos coeficientes ou nas medidas de ajustamento. De facto, as propriedades flexíveis do modelo de regressão logístico, fazem com que raramente se consiga um modelo com melhor ajustamento, exceptuando no caso em que observações tenham probabilidades muito baixas ou muito altas ou que o ajustamento do modelo seja demasiado débil (Hosmer e Lemeshow, 2000).

Tabela 5.7 Valores dos padrões de covariáveis para cada uma das variáveis do modelo, valor observado yi_i, número de elementos m_j, probabilidade estimada $\hat{\pi}$, e valores das respectivas estatísticas de diagnóstico ($\Delta\beta$, $\Delta\chi^2$, ΔD e h).

$p\#$	51	76	115	88
idade	37	45	72	49
n_parceiros_categ_2	até 5	até 5	até 5	até 5
dst_hiv	não	não	não	sim
anticoncept_diu	não	não	não	não
anticoncept_preservativo	não	não	não	não
yi_i	7	3	1	1
m_j	9	6	2	1
$\hat{\pi}$	0.245	0.164	0.034	0.442
$\Delta\beta$	0.319	0.296	0.172	1.030
$\Delta\chi^2$	8.931	11.394	13.450	1.384
ΔD	7.343	7.759	4.119	2.038
h	0.035	0.025	0.013	0.427

Tabela 5.8. Coeficientes estimados considerando todas as observações, percentagem de variação quando são eliminados os conjuntos de covariáveis idênticas e valores da bondade do ajustamento para cada modelo considerado.

Variáveis	Todas Observações	Conjuntos Observações Eliminados				
		51	76	115	88	Todos os padrões
Idade	-0,064	0,70%	5,31%	1,25%	-8,37%	16,28%
n_parceiros_categ_2	1,124	6,25%	-1,79%	-3,49%	-0,54%	12,34%
dst_hiv	2,548	1,47%	0,54%	-0,54%	-	-
anticoncept_diu	1,035	7,35%	-0,46%	-3,50%	-3,07%	19,15%
anticoncept_preservativo	-0,893	-7,72%	3,19%	4,94%	-5,88%	-14,97%
ΔD	-194,622	4,8%	4,0%	1,8%	0,6%	11,5%
$\Delta\chi^2$	127,550	3,6%	3,8%	8,1%	2,1%	21,2%
Observações Excluídas	0	9	6	2	1	18

5.4 Interpretação dos Coeficientes

Uma vez analisado o ajustamento do modelo, poder-se-á prosseguir para a análise dos seus coeficientes.

A interpretação de cada coeficiente é realizada sob a premissa de que as restantes covariáveis permanecem constantes:

- O risco de ter HPV diminui cerca de 6% por cada aumento de um ano de idade. O intervalo de confiança a 95% é bastante restrito indicando que o risco de contrair infecção pode diminuir entre 4% e 8%. Considerando intervalos de idade de 5 anos, o risco de infecção diminui aproximadamente 27% (IC 95%: $[5 * (-0.0636) \pm 1.96 * 5 * 0.0109] = [19.0\%; 34.6\%]$) e ao considerar intervalos de 10 anos, o modelo indica que a possibilidade de adquirir HPV diminui quase para metade (47% (IC 95%: $[10 * (-0.0636) \pm 1.96 * 10 * 0.0109] = [34.5\%; 57.3\%]$.
- A possibilidade de ter HPV triplica em mulheres que já tiveram mais de cinco parceiros sexuais. Com uma confiança de 95% podemos dizer que a possibilidade de infecção pode ser apenas 1.6 vezes mais e no máximo 8.4 vezes mais.
- Mulheres que estejam infectadas pelo HIV apresentam um risco muitíssimo superior (12 vezes mais elevado), do que aquelas que não se encontram infectadas por esta doença. Como se pode verificar o erro padrão é bastante elevado, conduzindo assim a uma estimativa do intervalo de confiança bastante alargado.
- Relativamente ao uso de contraceptivos, duas variáveis mostraram serem significativas a 5%: (1) o uso de dispositivo intra-uterino, que favorece o risco de infecção (2.8 vezes mais) e o uso de preservativo com efeito protector (diminuição de 59%).

Na Figura 5.12 está apresentado o gráfico das probabilidades estimadas de contrair infecção em função da idade comparando as duas categorias de parceiros sexuais (até cinco e mais de cinco). Assim, de acordo com o previsto, observa-se que a probabilidade de infecção vai diminuindo com a idade, sendo ainda superior em mulheres que têm mais de cinco parceiros sexuais.

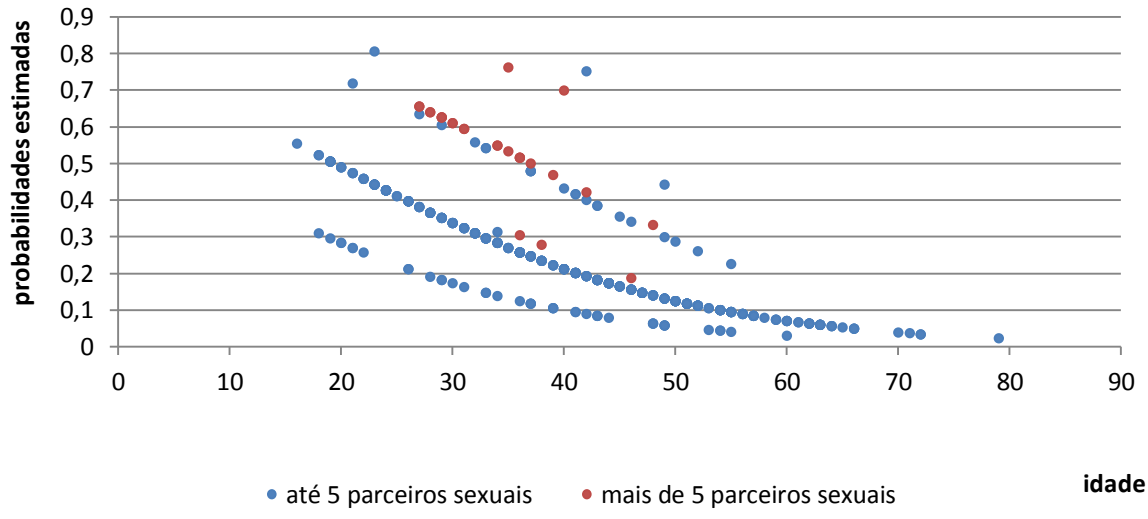

Figura 5.12 Probabilidades estimadas de contrair infecção em função da idade (até 5 parceiros *versus* mais de cinco parceiros sexuais).

Por outo lado, é também visível que o grupo de mulheres que não usa preservativo têm maior probabilidade de adquirir a infecção (Figura 5.13).

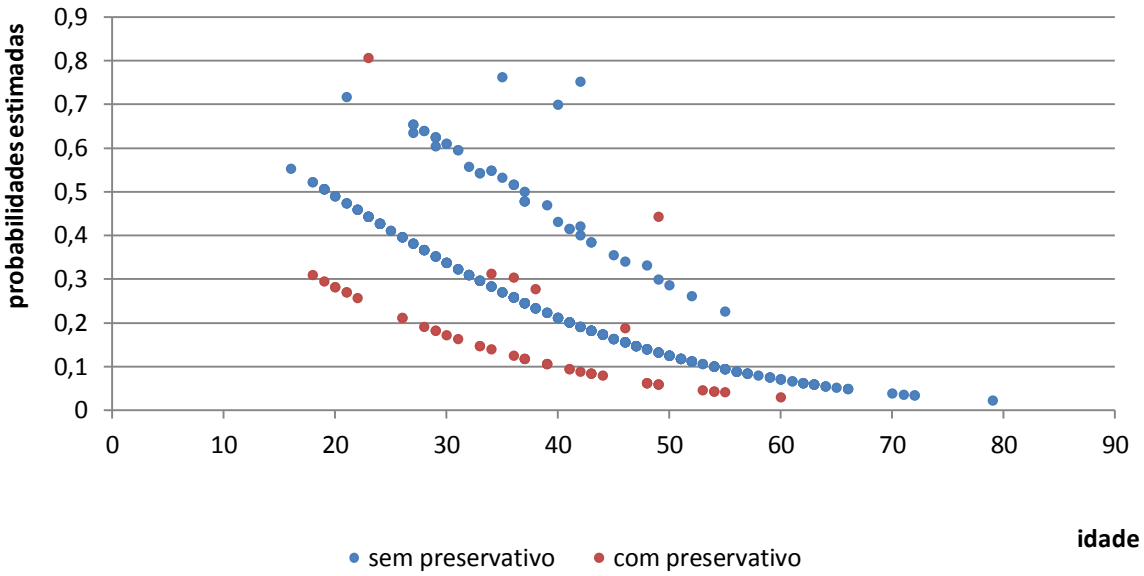

Figura 5.13 Probabilidades estimadas de contrair infecção em função da idade e do uso de preservativo.

6. Regressão Multinomial para o Desenvolvimento de Lesões

6.1 Regressão Multinomial Univariada e Multivariada

Posteriormente ao estudo das variáveis que contribuem para a infecção pelo vírus do papiloma Humano, pretende-se analisar quais os factores associados ao desenvolvimento de lesões. Como o número de lesões identificadas na amostra é reduzido, optou-se por agrupa-las em classes tendo em conta a sua severidade. Assim, a variável resposta é composta por três níveis: sem lesão, com lesão ligeira e com lesão severa. Como variáveis explicativas serão consideradas a idade, os vários tipos de HPV (agrupados em classes de severidade), a actividade sexual, o número de parceiros sexuais (em categorias) e as doenças sexualmente transmissíveis. Os HPV´s (n_hpv_cat e n_hpv_alto_risco_cat) foram ainda agrupados em três categorias (zero, infecções simples e infecções múltiplas) e a variável número de parceiros agrupada em duas classes (até cinco e mais de cinco parceiros).

Sendo uma variável composta por três níveis, do modelo de regressão multinomial resultam duas equações, que comparam separadamente as lesões ligeiras e severas (alta) com o grupo de referência, sem lesão.

A análise foi iniciada pelo estudo individual das variáveis (Tabela 6.1), tendo sido incluídas no modelo multivariado, as variáveis com valor p da estatística de Wald inferior a 0.25 (Tabela 6.2).

Tabela 6.1 Análise univariada para a regressão logística relativamente à infecção por HPV (valores estimados de OR, desvio padrão, estatística de Wald e IC 95% para OR).

Logit	Variáveis	β	OR	OR (Er.Padrão)	Estatística Wald (valor p)	OR (Int. Conf. 95%)	
						Inferior	Inferior
Ligeira	hpv	7,6	3,1784	4,8800	<0,00001	3,3781	17,2658
Alta	hpv	26,9	17,0717	5,1800	<0,00001	7,7360	93,3398
Ligeira	hpv_multiplos	8,2	3,8251	4,5300	<0,00001	3,3008	20,4627
Alta	hpv_multiplos	8,1	4,1476	4,0800	<0,00001	2,9667	22,0973
Ligeira	n_hpv_cat (1)	5,3	2,5887	3,4600	<0,00101	2,0694	13,8104
	n_hpv_cat (2)	13,4	6,8733	5,0400	<0,00001	4,8777	36,6202
Alta	n_hpv_cat (1)	23,8	15,7612	4,7800	<0,00001	6,4744	87,1949
	n_hpv_cat (2)	34,6	25,2150	4,8700	<0,00001	8,3230	144,2542
Ligeira	Idade	0,9	0,0185	-4,0900	<0,00001	0,8859	0,9583
Alta	Idade	1,0	0,0179	-0,9600	0,3350	0,9481	1,0183
Ligeira	hpv_alto_risco	9,2	3,7944	5,3600	<0,00001	4,0789	20,6358
Alta	hpv_alto_risco	50,4	32,3790	6,1000	<0,00001	14,2811	177,5707

Tabela 6.1 Análise univariada para a regressão logística relativamente à infecção por HPV (continuação).

Logit	Variáveis	β	OR	OR (Er.Padrão)	Estatística Wald (valor p)	OR (Int. Conf. 95%)	
						Inferior	Inferior
Ligeira	n_hpv_alto_risco_cat (1)	4,6	2,2778	3,1300	0,0020	1,7752	12,1446
	n_hpv_alto_risco_cat (2)	14,7	9,3862	4,2000	<0,00001	4,1974	51,3944
Alta	n_hpv_alto_risco_cat (1)	9,6	4,5887	4,7700	<0,00001	3,7984	24,5066
	n_hpv_alto_risco_cat (2)	4,3	4,7518	1,3100	0,1920	0,4831	37,7869
Ligeira	hpv_prov_alto_risco	3,9	2,6541	1,9900	0,0470	1,0175	14,8238
Alta	hpv_prov_alto_risco	3,2	2,5966	1,4600	0,1430	0,6716	15,5953
Ligeira	hpv_risco_indet	5,5	3,0898	3,0200	0,0030	1,8159	16,5462
Alta	hpv_risco_indet	4,0	2,7008	2,0400	0,0420	1,0533	15,0477
Ligeira	hpv_baixo_risco	4,0	2,1772	2,5300	0,0120	1,3634	11,6279
Alta	hpv_baixo_risco	2,9	1,9215	1,6000	0,1100	0,7864	10,6355
Ligeira	n_parceiros_categ_2	3,1	1,8465	1,9200	0,0540	0,9783	9,9522
Alta	n_parceiros_categ_2	4,4	2,6700	2,4500	0,0140	1,3430	14,4502
Ligeira	n_hpv_cat (1)	5,3	2,5887	3,4600	0,0010	2,0694	13,8104
	n_hpv_cat (2)	13,4	6,8733	5,0400	<0,00001	4,8777	36,6202
Alta	n_hpv_cat (1)	23,8	15,7612	4,7800	<0,00001	6,4744	87,1949
	n_hpv_cat (2)	34,6	25,2150	4,8700	<0,00001	8,3230	144,2542
Ligeira	activ_sexual	1,3	0,9778	0,3400	0,7350	0,2931	5,6946
Alta	activ_sexual	0,9	0,6870	-0,1500	0,8840	0,1985	4,0300
Ligeira	Dst	0,9	0,9446	-0,1000	0,9180	0,1140	7,0629
Alta	Dst	1,2	1,2344	0,1500	0,8840	0,1467	9,2807
Ligeira	dst_sífilis	0,0	0,0053	-0,0100	0,9920	0,0000	.
Alta	dst_sífilis	0,0	0,0060	-0,0100	0,9930	0,0000	.
Ligeira	dst_hiv	0,0	0,0096	-0,0200	0,9870	0,0000	,
Alta	dst_hiv	9,1	11,2967	1,7700	0,0770	0,7878	104,2814
Ligeira	dst_clamídea	0,0	0,0040	-0,0100	0,9910	0,0000	.
Alta	dst_clamídea	0,0	0,0046	-0,0100	0,9920	0,0000	.
Ligeira	dst_outra	1,5	1,6312	0,3900	0,6970	0,1849	12,4686
Alta	dst_outra	0,0	0,0017	-0,0100	0,9900	0,0000	.

Dado que a variável hpv_baixo_risco não é significativa em nenhum dos *logit*, foi retirada do modelo e calculada a diferença da *deviance* entre estes dois modelos encaixados.

O valor da estatística de teste é:

$$G = -2 * [-129.8480 - (-129.7304)] = 0.2353$$

que para dois graus de liberdade, origina um valor p de 0.8890, concluindo-se que hpv_baixo_risco não origina um melhor modelo, pelo que deverá ser excluída.

Tabela 6.2 Resultados da regressão Logística Multivariada para o modelo 1 (valores estimados de OR, desvio padrão, estatística de Wald, IC 95% para OR e teste de razão de verosimilhança entre o modelo actual e o modelo sem a variável).

Logit	Variáveis	β	OR	OR (Er.Padrão)	Estatística Wald (valor p)	OR (Int. Conf. 95%)		T. Razão Verosi. (valor p)
						Inferior	Superior	
Ligeira	Idade	-0,07	0,94	0,0211	0,0030	0,8956	0,9783	0.0015
	n_hpv_alto_risco_cat (1)	1,49	4,44	2,2736	0,0040	1,6279	12,1132	0.00001
	n_hpv_alto_risco_cat (2)	2,42	11,26	7,5589	<0,00001	3,0198	41,9735	
	hpv_prov_alto_risco	0,62	1,86	1,4225	0,4160	0,4162	8,3244	0.7267
	hpv_risco_indet	0,70	2,02	1,3625	0,2960	0,5398	7,5743	0.4057
	hpv_baixo_risco	0,32	1,38	0,9186	0,6260	0,3758	5,0847	0.2353
	n_parceiros_categ_2	0,66	1,94	1,2851	0,3180	0,5288	7,1076	-
Severa	Idade	0,03	1,04	0,0264	0,1700	0,9851	1,0886	0.0015
	n_hpv_alto_risco_cat (1)	4,28	72,28	51,5283	<0,00001	17,8706	292,3129	0.00001
	n_hpv_alto_risco_cat (2)	2,57	13,06	17,1420	0,0500	0,9957	171,1746	
	hpv_prov_alto_risco	0,31	1,37	1,3650	0,7540	0,1930	9,6781	0.7267
	hpv_risco_indet	0,97	2,63	2,3773	0,2830	0,4495	15,4442	0.4057
	hpv_baixo_risco	0,14	1,15	0,9753	0,8690	0,2181	6,0625	0.2353
	n_parceiros_categ_2	0,86	2,36	1,8012	0,2620	0,5263	10,5429	-

A partir deste novo modelo foram retiradas sucessivamente as variadas não significativas:

- hpv_prov_alto_risco: $G(2) = 2.191, p = 0.7418$;
- n_parceiros_categ_2: por não se tratarem de modelos encaixados (amostras constituídas por diferentes dimensões), não foi aplicado o teste da diferença de *deviance*. No entanto, a variável é excluída pelo valor p associado à estatística de Wald não ser significativo em nenhum dos seus *logit*s (0.3471 e 0.2420);
- hpv_risco_indet: $G(2) = 3.1300, p = 0.2091$.

Este procedimento permitiu chegar a um modelo constituído por duas variáveis: a idade e o número de HPV's de alto risco (categorizada) (Tabela 6.3). Sendo a idade apenas significativa numa das suas equações, ponderou-se assim a sua exclusão:

$$G = -2 * [-143.9751 - (-136.9938)] = 13.9626$$

Neste caso, o valor p obtido (0.0009) permite concluir que o melhor modelo é aquele que a inclui, devendo esta permanecer. A exclusão da idade, não conduz a uma alteração muito acentuada dos coeficientes das restantes covariáveis, obtendo-se a maior variação no primeiro nível do n_hpv_alto_risco_cat (21%).

Tabela 6.3 Resultados da regressão logística multivariada para o modelo 2 (valores estimados de OR, desvio padrão, estatística de Wald, IC 95% para OR e teste de razão de verosimilhança entre o modelo actual e o modelo sem a variável).

Logit	Variáveis	β	OR	OR (Er.Padrão)	Estatística Wald (valor p)	OR (Int. Conf. 95%)		T. Razão Verosi. (valor p)
						Inferior	Inferior	
Ligeira	Idade	-0,07	0,94	0,0202	0,0020	0,8974	0,9767	0.0516
	n_hpv_alto_risco_cat (1)	1,56	4,78	2,3396	0,0010	1,8306	12,4752	0.00001
	n_hpv_alto_risco_cat (2)	2,28	9,82	5,9583	<0,00001	2,9909	32,2524	
Severa	idade	0,03	1,03	0,0253	0,1600	0,9865	1,0858	0.0516
	n_hpv_alto_risco_cat (1)	4,46	86,90	61,7638	<0,00001	21,5778	349,9545	0.00001
	n_hpv_alto_risco_cat (2)	2,77	15,89	19,7112	0,0260	1,3959	180,7904	

Por outro lado, a variável nº de HPV's de alto de alto risco não apresenta a mesma magnitude entre as duas funções *logit*, ou seja, as infecções múltiplas potenciam o risco de desenvolver lesões ligeiras, mais do que as infecções simples, mas nas lesões severas, o risco é mais levado nas infecções simples do que nas múltiplas. Deve ainda salientado a pouca representatividade de indivíduos com mais de duas infecções, os quais representam 4.1% da amostra (17 indivíduos). Estes resultados sugerem que a variável poderá ser dicotomizada em dois níveis, ou seja, reagrupada em: (1) sem HPV ou (2) com pelo menos um HPV de alto risco.

$$G = -2 * [-139.9585 - (-136.9938)] = 5.9294$$

O valor da estatística G, para 2 graus de liberdade, permite obter um valor p de 0.0516, que embora esteja dentro do nível de significância considerado (10%), dado os constrangimentos acima enumerados, optou-se pelo modelo mais parcimonioso (Tabela 6.4).

Tabela 6.4 Resultados da regressão Logística Multivariada para o modelo final (valores estimados de OR, desvio padrão, estatística de Wald, IC 95% para OR e teste de razão de verosimilhança entre o modelo actual e o modelo sem a variável).

Logit	Variáveis	β	OR	OR (Er.Padrão)	Estatística Wald (valor p)	OR (Int. Conf. 95%)		T. Razão Verosi. (valor p)
						Inferior	Superior	
Ligeira	idade	-0,07	0,94	0,0200	0,0020	0,8984	0,9768	0.0066
	hpv_alto_risco	1,80	6,03	2,5891	<0,00001	2,6020	13,9907	<0.00001
Severa	idade	0,04	1,04	0,0253	0,1280	0,9894	1,0885	0.0066
	hpv_alto_risco	4,28	72,14	50,9551	<0,00001	18,0677	288,0161	<0.00001

De referir que a inclusão das variáveis que não tinham sido incluídas inicialmente no modelo multinomial, não originaram um melhor modelo que o apresentado na tabela 6.4.

Estando uma variável de natureza contínua incluída no modelo, será necessário averiguar qual a escala adequada, *i.e.*, testar a hipótese da linearidade do *logit*. Esta análise permitirá também avaliar se o facto de a idade não ser significativa numa das equações, poderá estar relacionado com o incumprimento deste pressuposto.

Após categorização dos indivíduos em grupos de idades de acordo com os respectivos quartis e obtidas as estimativas dos coeficientes do modelo de regressão logística a partir da nova variável categorizada e restantes variadas, representou-se o gráfico de dispersão entre os quartis e os seus coeficientes estimados. De acordo com os gráficos da figura 6.1 e 6.2 observa-se uma tendência ligeiramente quadrática. Por outro lado, os gráficos de dispersão com alisamento da escala do *logit* (Figura 6.3 e 6.4) apontam no mesmo sentido, sugerindo a necessidade de inclusão no modelo da variável $idade^2$.

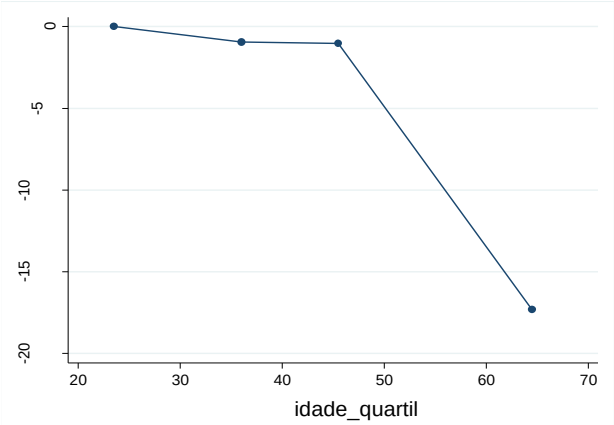

Figura 6.1 Gráfico dos coeficientes de regressão versus ponto médio dos quartis da variável idade referente ao *logit* 1.

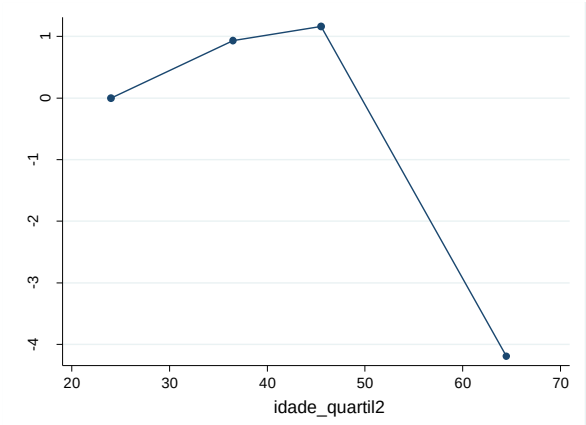

Figura 6.2 Gráfico dos coeficientes de regressão versus ponto médio dos quartis da variável idade referente ao *logit* 2.

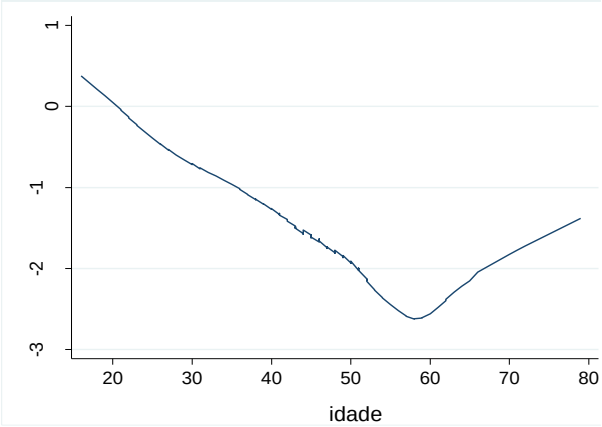

Figura 6.3 Gráfico de dispersão com alisamento da escala do *logit* versus idade (*logit* 1).

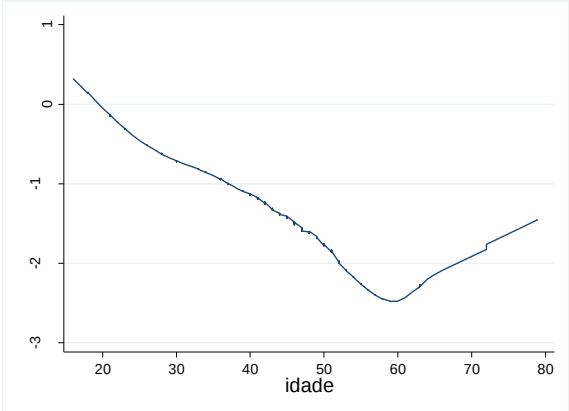

Figura 6.4 Gráfico de dispersão com alisamento da escala do *logit* versus idade (*logit* 2).

A diferença de *deviance* entre os dois modelos não é significativa, concluindo-se que $idade^2$ não têm um contributo significativo ($G(1) = 2{,}1859, p = 0.1393$).

O método dos polinómios fracionários indica que nem os melhores termos de $m = 1$ ou $m = 2$ melhoram o modelo linear (Tabela 6.5 e Tabela 6.6). Assim, a variável idade deverá permanecer sob a sua forma continua e linear.

Tabela 6.5 Resumo do método do polinómio fraccionário para a variável idade relativamente ao primeiro *logit*.

```
Fractional polynomial model comparisons:
-----------------------------------------------------------------
idade            df       Deviance      Gain   P(term) Powers
-----------------------------------------------------------------
Not in model      0        175.367        --     --
Linear            1        164.919     0.000    0.001  1
m = 1             2        162.985     1.934    0.164  3
m = 2             4        162.629     2.290    0.837  3 3
-----------------------------------------------------------------
```

Tabela 6.6 Resumo do método do polinómio fraccionário para a variável idade relativamente ao segundo *logit*.

```
Fractional polynomial model comparisons:
-----------------------------------------------------------------
idade            df       Deviance      Gain   P(term) Powers
-----------------------------------------------------------------
Not in model      0        109.077        --     --
Linear            1        106.196     0.000    0.090  1
m = 1             2        104.372     1.824    0.177  -2
m = 2             4        103.861     2.335    0.774  -2 -2
-----------------------------------------------------------------
```

Na etapa seguinte de construção do modelo, foram avaliadas as interacções dos efeitos principais, neste caso entre a idade e o HPV de alto risco. Tal como observado na Tabela 6.7 a interacção não é significativa e a alteração dos coeficientes é muitíssimo elevada, sempre acima dos 33%, chegando a atingir os 96% no hpv_alto_risco do primeiro *logit*. Assim se conclui que este termo não origina um melhor modelo. Esta informação é ainda corroborada pelo teste de razão de verosimilhança de onde se obtém um valor p de 0.229 ($G(2) = 2.9525, p = 0.2285$).

Tabela 6.7 Resultados da regressão Logística Multivariada para o modelo 2 com interacção entre os efeitos principais (valores estimados de OR, desvio padrão, estatística de Wald e IC 95% para OR).

Logit	Variáveis	β	OR	OR (Er.Padrão)	Estatística Wald (valor p)	OR (Int. Conf. 95%)	
						Inferior	Superior
Ligeira	Idade	-0,09	0,92	0,0260	0,0020	0,8675	0,9694
	hpv_alto_risco	0,06	1,06	1,5572	0,9670	0,0600	18,8014
	idade*hpv_alto_risco	0,06	1,06	0,0468	0,2120	0,9689	1,1527
Severa	Idade	0,07	1,07	0,0536	0,1750	0,9701	1,1807
	hpv_alto_risco	5,99	399,54	1155,8490	0,0380	1,3774	115892,3000
	idade*hpv_alto_risco	-0,03	0,97	0,0556	0,5470	0,8630	1,0813

6.2 Bondade do Ajustamento

Estando o modelo final concluído e antes de se realizar qualquer tipo de inferência, importa realizar um estudo detalhado do ajustamento e diagnóstico das observações. Esta análise foi efectuada, tal como referida no capítulo 3, dividindo o modelo em duas equações de regressão logística:

(1) *Logit* 1: modelo que compara mulheres com lesões ligeiras relativamente à classe de referência (sem lesão);
(2) *Logit* 2: modelo que compara mulheres com lesões severas relativamente à classe de referência (sem lesão);

Na tabela 6.8 estão resumidos os resultados das medidas utilizadas na avaliação da bondade do ajustamento. No modelo 1 e 2 são analisados, respectivamente, 82 e 85 padrões de covariáveis distintos.

Tabela 6.8 Resumo das estatísticas para a bondade do ajustamento e respectivos valores p para os modelos individuais de Regressão Logística.

Logit	Hosmer Lemeshow ($\hat{C}$)	Pearson (χ^2)	Stukel (S)
1	2,19 (p=0,335)	61,13 (p=0,960)	2,94 (p=0,087)
2	0,15 (p=0,699)	54.60 (p=0,984)	4,27 (p=0,120)

Em qualquer uma das equações e independentemente do método, o valor p é não significativo, indicando que globalmente o ajustamento do modelo é adequado. Para garantir que no cálculo da estatística de Hosmer e Lemeshow não houvesse grupos constituídos por menos do que cinco observações, procedeu-se ao agrupamento de categorias adjacentes. Assim, para o modelo 1 e 2, foram considerados dois e um grau de liberdade, respectivamente.

6.3 Diagnóstico do Modelo

A pesquisa de *outliers* e observações influentes foi realizada mediante análise dos resíduos Pearson, *deviance* e *leverage*, bem como, através das estatísticas de diagnóstico de $\Delta\hat{\beta}$, $\Delta\chi^2$e ΔD.

No primeiro *logit* destacam-se dois padrões de covariáveis (4 e 50) com resíduos de Pearson e *Deviance* ligeiramente superiores aos restantes, embora com valor inferior a 3 (Figuras 6.5 e 6.6). O gráfico com os resíduos *Leverage* (Figura 6.7) mostra que apenas o padrão 4 se evidência, indicando que este conjunto de observações poderá originar diferentes estimativas caso sejam excluídos do modelo.

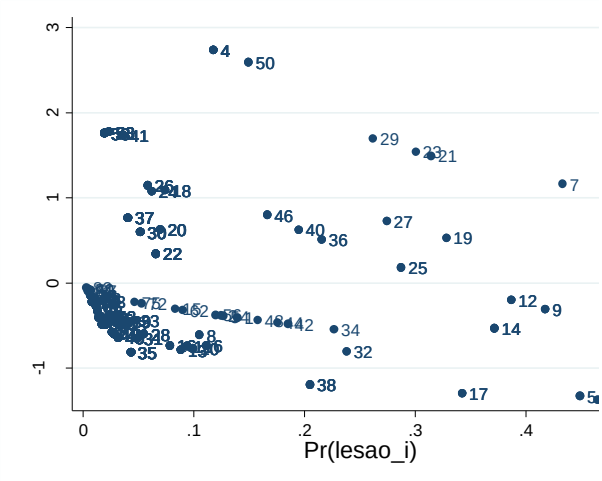

Figura 6.5 Gráfico dos resíduos de Pearson standartizados *versus* valores previstos para a equação do *logit* 1.

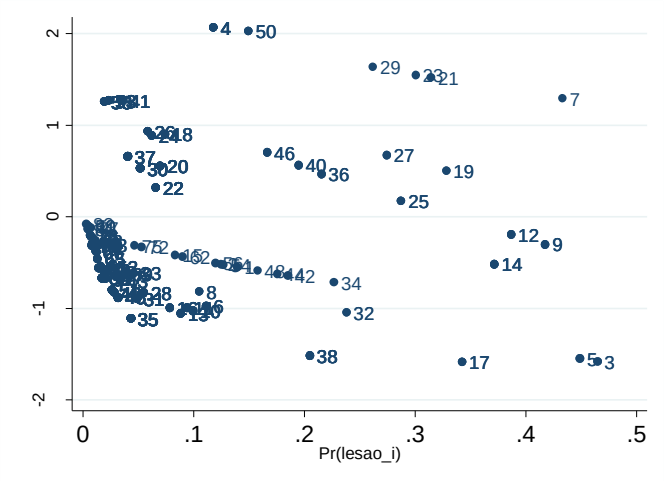

Figura 6.6 Gráfico dos resíduos *Deviance versus* valores previstos para a equação do *logit* 1.

De novo, na figura 6.8 e 6.9 voltam a destacar-se os mesmos padrões de covariáveis com valores elevados de $\Delta\chi^2$ e ΔD.

Caso o padrão 4 fosse excluído da amostra, a estatística χ^2 do ajustamento de Pearson sofreria um decréscimo na ordem dos 7.5 e a *deviance* uma alteração de 4.9 (23.1%) unidades (5.3%). Por outro lado, a eliminação do grupo 50, implicaria uma modificação no ajustamento e na *deviance* de 6.7 (22.5%) e 4.4 (4.0%), respectivamente.

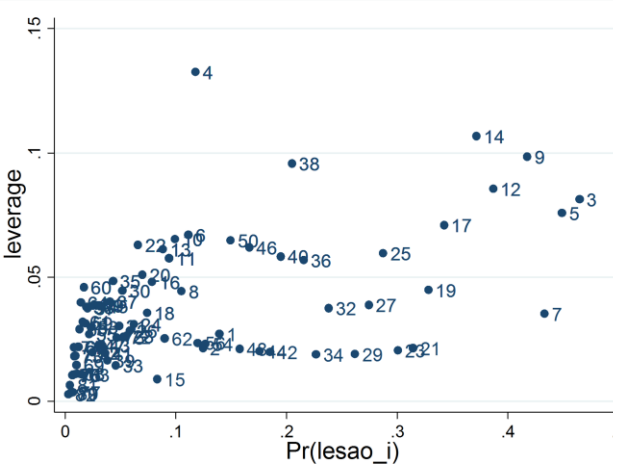

Figura 6.7 Gráfico dos resíduos *Leverage versus* valores previstos para a equação do *logit* 1.

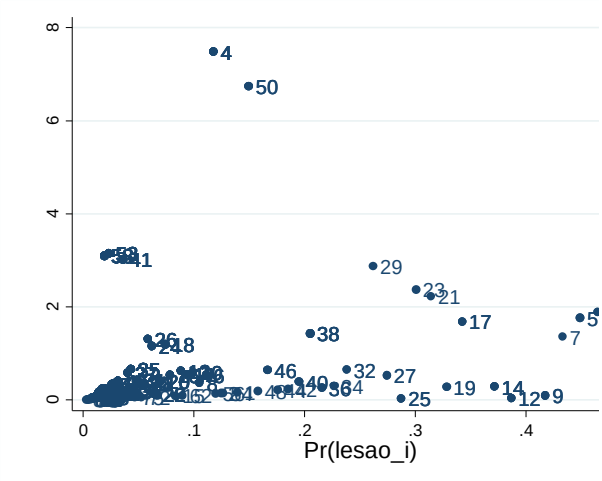

Figura 6.8 Gráfico $\Delta\chi_j^2$ *versus* probabilidades previstas ($\hat{\pi}_j$).

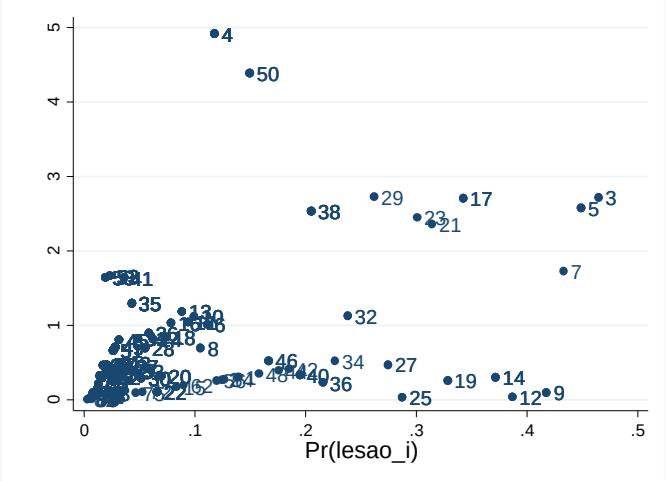

Figura 6.9 Gráfico ΔD_j *versus* probabilidades previstas ($\hat{\pi}_j$).

A eliminação dos indivíduos que constituem o padrão 4 (n=7) e o padrão 50 (n=3) não estão associadas a uma alteração na magnitude dos seus coeficientes. De acordo com o sugerido anteriormente, tendo em conta os valores de *leverage*, as maiores variações nos coeficientes acontecem quando o padrão 4 é excluído do modelo, obtendo-se uma variação de 27.3% na idade e 16.6% no hpv_alto_risco (Tabela 6.8). A exclusão do padrão 50 origina uma variação máxima de 16.4% e quando se eliminam os dois padrões, obtém-se uma variação não superior a 11.09% (Tabela 6.9).

A pequena alteração nas estatísticas de χ^2 e D (Tabela 6.9), a variação reduzida dos coeficientes que constituem o modelo e sem alteração da sua magnitude, sugerem que o modelo conseguido sem os sete indivíduos do padrão 4, ou sem os três indivíduos do padrão 50, não representa um melhor modelo. Assim, não havendo suspeita para que estes indivíduos devam ser retirados da amostra (*e.g.* erros de medição), optou-se por não perder a informação que está incorporada nestes 10 indivíduos, permanecendo estas observações no modelo final.

Tabela 6.9 Coeficientes estimados considerando todas as observações, percentagem de variação quando são eliminados os padrões de covariáveis idênticos e valores da bondade do ajustamento para o modelo do *Logit* 1.

Variáveis	Todas Observações	Conjuntos Observações		
		4	50	Todos os grupos
Idade	-0,0653	-27,34%	16,38%	-11,09%
hpv_alto_risco	1,7973	16,58%	-11,09%	4,89%
Deviance ΔD	-279,91702	5,3%	4,0%	9.0%
$\Delta\chi^2$		23,1%	22,5%	22,5%

Relativamente ao segundo *logit*, verifica-se que quatro padrões de covariáveis têm resíduos de Pearson superiores a duas unidades (Figura 6.10). Em particular oito observações (padrão 42) têm um resíduo próximo de quatro. Este valor mais elevado resulta da diferença entre a probabilidade estimada, 0.072, e a probabilidade observada, y_j/m_j, 0.125. Os valores obtidos para a *deviance* são relativamente homogéneos (Figura 6.11) e não existem observações com valores de *leverage* (Figura 6.12) muito destacados ou acentuados.

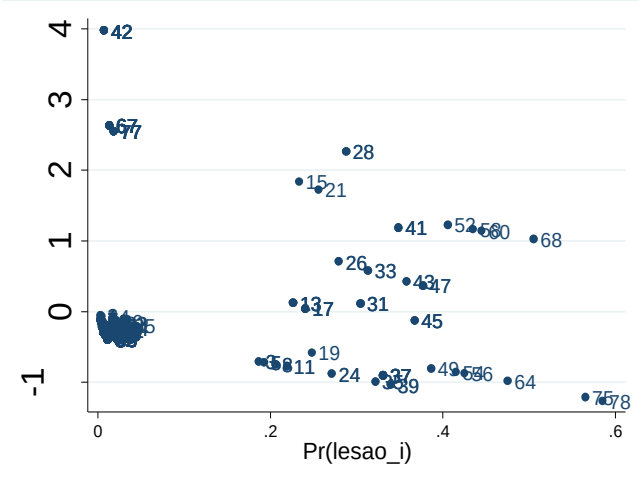

Figura 6.10. Gráfico dos resíduos de Pearson standartizados *versus* valores previstos para a equação do *logit* 2.

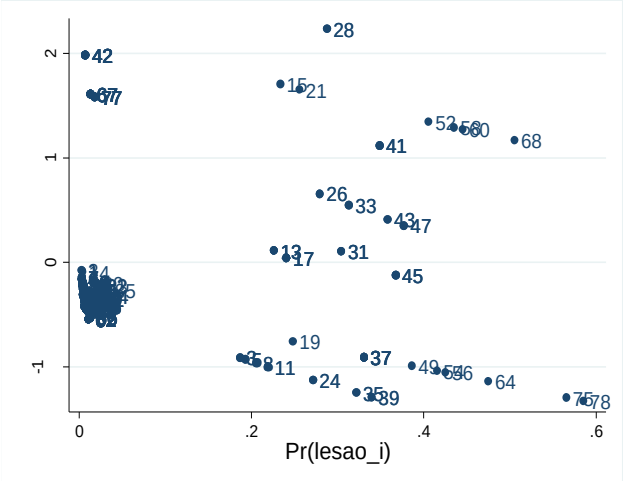

Figura 6.11 Gráfico dos resíduos *Deviance versus* valores previstos para a equação do *logit* 2.

As estatísticas abaixo apresentadas mostram também problemas no ajustamento no padrão 42, com valores de $\Delta\chi^2$ e ΔD relativamente elevados (Tabela 6.10). Por outro lado, as estimativas dos coeficientes praticamente não sofrem alterações quando se eliminam estas alterações (Tabela 6.11).

Assim, de modo semelhante ao primeiro *Logit*, não são retiradas nenhumas observações ao modelo.

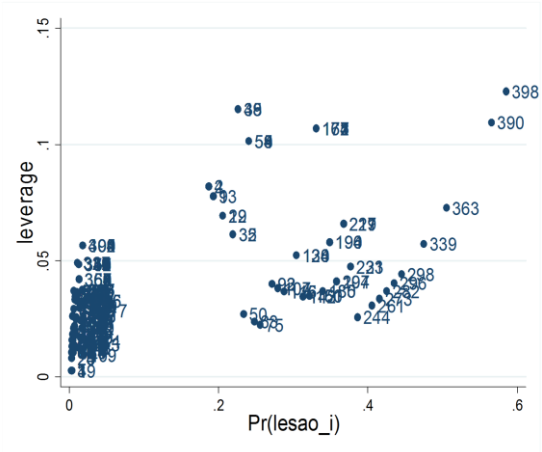

Figura 6.12 Gráfico dos resíduos *Leverage versus* valores previstos para a equação do *logit* 2.

Tabela 6.10 Valores dos padrões de covariáveis para cada uma das variáveis do modelo, valor observado yi_i, número de elementos m_j, probabilidade estimada $\hat{\pi}$, e valores das respectivas estatísticas de diagnóstico ($\Delta\beta$, $\Delta\chi^2$, ΔD e h) para cada um dos *logit* .

	Logit 1		*Logit* 2			
$p\#$	4	50	42	67	77	28
Idade	19	43	40	55	63	32
hpv_alto_risco	não	não	não	não	não	sim
yi_i	3	2	1	1	1	2
m_j	7	3	8	9	7	2
$\hat{\pi}$	0,118	0,150	0.072	0.013	0.018	0.288
$\Delta\beta$	1,143	0,4671	0.355	0.305	0.390	0.195
$\Delta\chi^2$	7,483	6,742	15.824	6.947	6.511	5.141
ΔD	4,920	4,389	4.023	2.7093	2.654	5.173
h	0,132	0,065	0.214	0.042	0.566	0.366

Tabela 6.11 Coeficientes estimados considerando todas as observações, percentagem de variação quando são eliminados os conjuntos de covariáveis idênticas e valores da bondade do ajustamento para o modelo do *Logit* 2.

Variáveis	Todas Observações	Conjuntos Observações				
		42	67	77	28	Todos os grupos
Idade	0,0370	1,18%	1,41%	0,53%	-0,95%	0,76%
hpv_alto_risco	4,2786	0,75%	-0,51%	-0,09%	-0,37%	0,28%
Deviance ΔD	-278,7916	0,4%	0,1%	0,0%	0,33%	0,8%
$\Delta\chi^2$		3,4%	1,5%	0,3%	0,6%	1,0%

6.4 Interpretação dos Coeficientes

Finalizada a análise do ajustamento e diagnóstico do modelo, podem então ser interpretados os seus coeficientes (Tabela 6.4).

- As infecções por HPV de alto risco devem ser consideradas um forte factor de risco no desenvolvimento de lesões. As mulheres com este tipo de vírus têm seis vezes mais possibilidades de terem lesões ligeiras e 72 vezes mais possibilidade de apresentarem lesões severas. De salientar a ampla magnitude do intervalo de confiança no segundo *logit*, onde o risco de vir a desenvolver lesões graves pode ser tao baixo quanto 18 vezes ou tão alto quanto 288 vezes mais.

De acordo com o referido anteriormente, a categorização desta variável em três níveis (sem infecção, até 5 HPV's e mais de 5 HPV's) não origina um melhor modelo, pelo que se optou no agrupamento das duas últimas classes.

- A idade mostrou apenas ser significativa na primeira equação, indicando um efeito protector, ou seja, o aumento de um ano de idade está associado a uma diminuição de cerca de 6% no desenvolvimento de lesões ligeiras. Se considerarmos um intervalo de 5 anos, o risco diminui 28% e 48% em intervalos de 10 anos.

7. Regressão Ordinal para o Desenvolvimento de Lesões

No capítulo anterior foi realizada uma regressão logística multivariada, considerando a variável resposta composta por três níveis: (1) sem lesão; (2) com lesão ligeira e (3) com lesão severa. Neste capítulo, a variável resposta mantem-se, mas considera-se que os seus níveis são ordenados.

Como variáveis explicativas serão utilizadas a idade, os vários tipos de HPV (agrupados em classes de severidade), a actividade sexual, o número de parceiros sexuais (em categorias) e as doenças sexualmente transmissíveis.

A natureza da variável resposta permite uma abordagem via regressão ordinal considerando o modelo de riscos proporcionais. Assim, descrever-se-á os passos na construção do modelo.

Tabela 7.1 Resultados da regressão logística ordinal univariada (valores estimados de OR, desvio padrão e estatística de Wald).

Variáveis	β	OR	OR (Er.Padrão)	Estatística Wald (valor p)	OR (Int. Conf.)	
					Inferior	Superior
hpv	2,5235	12,4716	4,3424	<0,00001	6,3030	24,6772
hpv_multiplos	2,0333	7,6394	2,8115	<0,00001	3,7136	15,7152
n_hpv_alto_risco_cat (1)	2,3221	10,1968	3,9165	<0,00001	4,8031	21,6474
n_hpv_alto_risco_cat (2)	2,8671	17,5860	7,5534	<0,00001	7,5782	40,8100
idade	-0,0472	0,9539	-0,0128	<0,00001	0,9291	0,9793
hpv_alto_risco	2,8978	18,1346	6,2522	<0,00001	9,2266	35,6430
n_hpv_alto_risco_cat (1)	3,0414	20,9339	7,7174	<0,00001	10,1637	43,1171
n_hpv_alto_risco_cat (2)	2,4773	11,9088	6,3171	<0,00001	4,2106	33,6815
hpv_prov_alto_risco	1,2365	3,4437	1,8968	0,0250	1,1700	10,1360
hpv_risco_indet	1,4943	4,4563	2,0440	0,0010	1,8136	10,9498
hpv_baixo_risco	1,2010	3,3236	1,4764	0,0070	1,3915	7,9385
n_parceiros_categ_2	1,3043	3,6852	1,6594	0,0040	1,5247	8,9071
n_hpv_cat (1)	2,3221	10,1968	3,9165	<0,00001	4,8031	21,6474
n_hpv_cat (2)	2,8671	17,5860	7,5534	<0,00001	7,5782	40,8100
activ_sexual	0,0773	1,0804	0,5980	0,8890	0,3651	3,1967
dst	0,0230	1,0233	0,7857	0,9760	0,2272	4,6084
dst_sífilis	-12,3972	0,0000	-0,0040	0,9900	0,0000	,
dst_hiv	1,6200	5,0531	6,3490	0,1970	0,4306	59,3001
dst_clamídea	-12,5208	0,0000	-0,0030	0,9880	0,0000	,
dst_outra	-0,2205	0,8021	-0,8516	0,8350	0,1001	6,4265

A partir da análise univariada (Tabela 7.1) foram seleccionadas para o modelo as variáveis significativas na estatística de Wald, considerando um nível de significância de 25%. As etapas seguintes, aplicando as recomendações de Hosmer e Lemeshow (2000), permitiram chegar a um modelo que inclui apenas os efeitos principais das variáveis n_hpv_alto_risco_cat e hpv_risco_indet.

Tabela 7.2 Resultados da regressão logística ordinal multivariada ((valores estimados de OR, desvio padrão, estatística de Wald, IC 95% para OR e teste de razão de verosimilhança entre o modelo actual e o modelo sem a variável).

Variáveis	β	OR	OR (Er.Padrão)	Estatística Wald (Signif.)	OR (Int. Conf.)		T. Razão Verosi. (valor p)
					Inferior	Inferior	
n_hpv_alto_risco_cat (1)	3,0314	20,7268	7,6773	<0,00001	10,0287	42,8372	<0,00001
n_hpv_alto_risco_cat (2)	2,1279	8,3976	4,7773	<0,00001	2,7537	25,6092	
hpv_risco_indet	1,1163	3,0535	1,6679	0,0410	1,0468	8,9070	0.0014

De acordo com o referido na metodologia, a averiguação do pressuposto dos riscos proporcionais foi efectuada através de três testes estatísticos. O resultado do teste de razão de verosimilhança, cuja estatística de χ^2para três graus de liberdade, permite obter um valor p de 0.04 ($\chi^2 = 8.30, gl = 3$) rejeitando assim, para um α de 5%, a hipótese da proporcionalidade dos riscos. De salientar que se trata de um teste aproximado e que o valor obtido está muito próximo do limiar de significância considerado.

Por outro lado, o valor do teste de Brant para a hipótese de que todos os coeficientes são paralelos ($\chi^2 = 6.31, gl = 3, p = 0.097$), indica que o pressuposto não deve ser rejeitado (Tabela 7.3). Embora os resultados dos dois testes tenham conclusões contraditórias, os seus valores p estão muito próximos da fronteira dos 5%.

O *output* do teste de Brant fornece ainda um conjunto de regressões logísticas binárias. Neste caso existem duas equações, dado a natureza da variável resposta, composta por três categorias. Esta metodologia começa por comparar a primeira categoria com as restantes e depois confronta as duas primeiras com a última categoria. Caso o pressuposto das linhas paralelas não seja violado é esperado que todos os coeficientes (excepto a intercepção) sejam semelhantes ao longo das equações, e que as diferenças resultem apenas da variabilidade da amostra. De acordo com os resultados, verifica-se que os coeficientes são bastante semelhantes, embora as maiores diferenças estejam concentradas na segunda categoria da variável n_hpv_alto_risco_cat, sendo esta variável a que apresenta um valor p mais próximo da significância ($p = 0.064$).

Por último, o teste aproximado de Hosmer que compara da verosimilhança do modelo multinomial com o modelo ordinal, indica uma estatística de teste, $G = -2 * [-146.1741 - (-142.24212)] = 7.8657147$, que para 3 graus de liberdade, permite obter um valor $p = P(\chi^2\ (3) > 7.8657147) = 0{,}0489$ permitindo assim rejeitar o pressuposto dos riscos proporcionais.

Considerando que a rejeição pelos testes de razão de verosimilhança verificam-se apenas para uma significância de 5% e que no teste de Brant, o pressuposto é cumprido, tanto globalmente como individualmente, considerou-se válido o pressuposto dos riscos proporcionais.

Tabela 7.3 Resultado do teste de razão de verosimilhança para a proporcionalidade de *odds*.

```
. brant, detail

Estimated coefficients from j-1 binary regressions

                        y>0         y>1
  _In_hpv_alt_1   2.8478878   3.8959962
  _In_hpv_alt_2   2.3823905   1.7220614
hpv_risco_indet   1.1049946   .73749008
          _cons  -3.0775182  -4.7646181

Brant Test of Parallel Regression Assumption

    Variable |      chi2   p>chi2    df
-------------+--------------------------
         All |      6.31    0.097     3
-------------+--------------------------
_In_hpv_al~1 |      3.44    0.064     1
_In_hpv_al~2 |      0.32    0.573     1
hpv_risco_~t |      0.23    0.629     1
----------------------------------------

A significant test statistic provides evidence that the parallel
regression assumption has been violated.
```

Ainda a partir da Tabela 7.3 é possível analisar o teste de razão de verosimilhança entre o modelo ajustado e modelo nulo, concluindo-se assim que o modelo encontrado é globalmente melhor que modelo sem nenhum preditor.

Na avaliação do ajustamento e no diagnóstico das observações, foram realizadas duas regressões logísticas binárias, onde a variável resposta é constituída:

(1) *Logit* 1: modelo que compara mulheres com lesões ligeiras relativamente à classe de referência (sem lesão);
(2) *Logit* 2: modelo que compara mulheres com lesões severas relativamente à classe de referência (sem lesão);

Esta análise foi realizada anteriormente aquando a avaliação do modelo multinomial (capítulo 6), onde se concluiu que ambos os *logits* tinham um bom ajustamento e que não existem *outliers* ou observações influentes que devessem ser excluídas do modelo final.

O modelo final, ajustado para o número de HPV's de alto risco e HPV's de risco indeterminado, permite tecer as seguintes considerações:

- Quando existe infecção por um vírus de alto risco, a possibilidade de desenvolver lesões severas aumenta 20 vezes relativamente à possibilidade de desenvolver lesão ligeira ou não ter qualquer tipo de alteração, considerando que a restante covariada permanece constante. Da mesma forma, e dada a proporcionalidade de riscos, o risco de ter lesão severa ou ligeira aumenta vinte vezes comparativamente ao risco de não ter lesão.
- Em mulheres com infecções múltiplas de HPV's de alto risco, a possibilidade de terem lesões com maior severidade aumenta pouco mais de oito vezes. De referir que, para esta variável, em ambos os seus níveis, o valor elevado de erro padrão conduz a um intervalo de confiança bastante amplo. Neste último caso, o aumento pode ser tão pequeno quanto 2.7 vezes ou tão grande quanto 25.6 vezes.
- Por último, as infecções com HPV's de risco indeterminado, independentemente do número de vírus, favorecem o desenvolvimento de lesões graves, *i.e.*, as mulheres infectadas por este grupo de vírus têm três vezes mais hipótese de desenvolver lesões graves (considerando que a restante covariada permanece constante).

Anteriormente considerou-se que o pressuposto dos riscos proporcionais era cumprido, apesar de dois dos testes realizados não o indicar e, no teste de Brant, uma das variáveis ter um valor p muito próximo de 0.5. Assim, serão também analisados outros tipos de modelos, propostos por Williams (2007), mediante a utilização do comando go*logit*2.

O primeiro modelo considerado é construído sobre a premissa de que nenhuma variável cumpre o pressuposto da proporcionalidade dos riscos (**modelos ordinais generalizados**). De acordo com a Tabela 7.4, as estimativas são muito semelhantes às das equações de regressão logística binária apresentadas no teste de Brant (Tabela 7.3). A sua interpretação é também idêntica, *i.e.*, a primeira equação confronta a categoria 0 com as categorias 1 e 2 e, a segunda equação as categorias 0 e 1 com a 2. Os coeficientes positivos indicam que valores mais elevados da variável resposta aumentam a possibilidade de se estar num nível mais elevado de Y, ou seja de ter lesão grave, enquanto que os coeficientes negativos indicam que os níveis mais elevados da variável resposta aumentam a possibilidade de se estar no nível actual ou inferior de Y. Uma vez que o pressuposto de riscos proporcionais não é aplicado, existem tantos $\beta's$ quantos os níveis da variável resposta menos um, obtendo-se um modelo com mais parâmetros com uma complexidade idêntica ao modelo de regressão multinomial.

Tratando-se de modelos encaixados, o melhor modelo entre o modelo ordinal generalizado e o modelo ordinal pode ser testado via razão de verosimilhança. Assim, o resultado do valor p permite concluir que o modelo generalizado, embora mais complexo, é melhor que o modelo ordinal ($G(3) = 8.41,\ p = 0.0383$).

Tabela 7.4 Resultados para a regressão do modelo ordinal generalizado (go*logit*2).

```
. xi: gologit2 lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet
i.n_hpv_alto_~t   _In_hpv_alt_0-2     (naturally coded; _In_hpv_alt_0 omitted)

Generalized Ordered Logit Estimates               Number of obs   =        417
                                                  LR chi2(6)      =      90.51
                                                  Prob > chi2     =     0.0000
Log likelihood = -141.96898                       Pseudo R2       =     0.2417

------------------------------------------------------------------------------
     lesao_i |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ausente      |
_In_hpv_al~1 |    2.85475   .3716034     7.68   <0.0001   2.126421     3.58308
_In_hpv_al~2 |   2.347408   .5916144     3.97   <0.0001   1.187865    3.506951
hpv_risco_~t |   1.098566   .6023404     1.82   0.068    -.0819996    2.279131
       _cons |  -3.079272   .2619339   -11.76   <0.0001  -3.592653   -2.565891
-------------+----------------------------------------------------------------
ligeira      |
_In_hpv_al~1 |   3.930667   .6585044     5.97   <0.0001   2.640022    5.221311
_In_hpv_al~2 |    1.48253   1.243278     1.19   0.233    -.9542515    3.919311
hpv_risco_~t |   1.288097   .8632559     1.49   0.136    -.4038538    2.980047
       _cons |  -4.832757   .6028282    -8.02   <0.0001  -6.014278   -3.651235
------------------------------------------------------------------------------
```

Ainda com o objectivo de conseguir um modelo mais parcimonioso, recorreu-se ao **modelo ordinal de riscos proporcionais parciais**, onde se obteve os resultados apresentados na Tabela 7.5.

Esta opção inicia um processo iterativo, que tem início num modelo totalmente irrestrito, igual ao apresentado na Tabela 7.4. Posteriormente, utilizando a estatística de Wald, testa-se se os coeficientes são idênticos entre as várias equações, *i.e.*, se cada variável cumpre o pressuposto das linhas paralelas. Se existir uma ou mais variáveis com um valor p não significativo, a variável menos significativa é restringida de forma que o seu efeito seja idêntico entre as várias equações. Um novo modelo é estimado e o processo é repetido até que não existam variáveis que cumpram o pressuposto. Neste caso, a primeira variável a assumir o mesmo β entre as equações é o hpv_risco_indet, seguindo-se o terceiro nível da variável n_hpv_alto_risco_cat. O segundo nível de n_hpv_alto_risco_cat, não apresenta, pelo teste de Wald, o mesmo coeficiente entre as equações.

É também realizado um teste de Wald global que compara o modelo restrito com o modelo irrestrito. O valor de p não significativo ($p = 0.77$) indica que o modelo final não viola o pressuposto das linhas paralelas. Este modelo, conforme referido anteriormente, tem dois dos seus parâmetros restringidos. O teste de razão de verossimilhança indica também que o modelo actual é melhor que o modelo nulo ($p < 0.0001$).

Embora o modelo resultante pareça tão complexo quanto o anterior, existem dois parâmetros com coeficientes idênticos entre as equações. Assim, neste modelo existem quatro estimativas para análise, e não seis como no modelo generalizado. As variáveis restritas são interpretadas como anteriormente descrito, enquanto que as variáveis irrestritas, que neste caso corresponde à segunda categoria de n_hpv_alto_risco_cat (ter um HPV), apresentam um risco evolutivo:

- Quando existe infecção por um vírus de alto risco, a possibilidade de desenvolver lesões ligeiras ou severas aumenta 17 vezes (IC 95%: [8.47;30.03]), enquanto que a possibilidade de desenvolver lesões severas comparativamente às lesões ligeiras ou sem lesão aumenta quase 63 vezes (IC 95%: [18.85; 208.51]).
- Em mulheres com infecções múltiplas de HPV's de alto risco, a possibilidade de terem lesões com maior severidade aumenta praticamente dez vezes (IC 95%: [3.11;30.07]).
- As infecções por vírus de risco indeterminado triplicam o risco de desenvolvimento de lesões mais graves (IC 95%: [1.04;9.09]).

Tabela 7.5 Resultados para a regressão do modelo de riscos proporcionais parciais (go*logit*2 - autofit).

```
. xi: gologit2 lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, autofit lrforce
i.n_hpv_alto_~t   _In_hpv_alt_0-2     (naturally coded; _In_hpv_alt_0 omitted)

Testing parallel lines assumption using the .05 level of significance...
Step  1:  Constraints for parallel lines imposed for hpv_risco_indet (P Value = 0.8295)
Step  2:  Constraints for parallel lines imposed for _In_hpv_alt_2 (P Value = 0.4904)
Step  3:  Constraints for parallel lines are not imposed for
          _In_hpv_alt_1 (P Value = 0.01488)
Wald test of parallel lines assumption for the final model:
 ( 1)  [ausente]hpv_risco_indet - [ligeira]hpv_risco_indet = 0
 ( 2)  [ausente]_In_hpv_alt_2 - [ligeira]_In_hpv_alt_2 = 0
           chi2(  2) =    0.52
         Prob > chi2 =    0.7707

An insignificant test statistic indicates that the final model
does not violate the proportional odds/ parallel lines assumption
If you re-estimate this exact same model with gologit2, instead
of autofit you can save time by using the parameter
pl(hpv_risco_indet _In_hpv_alt_2)
------------------------------------------------------------------------------
Generalized Ordered Logit Estimates               Number of obs   =        417
                                                  LR chi2(4)      =      89.92
                                                  Prob > chi2     =     0.0000
Log likelihood = -142.26208                       Pseudo R2       =     0.2402

 ( 1)  [ausente]hpv_risco_indet - [ligeira]hpv_risco_indet = 0
 ( 2)  [ausente]_In_hpv_alt_2 - [ligeira]_In_hpv_alt_2 = 0
------------------------------------------------------------------------------
     lesao_i |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ausente      |
_In_hpv_al~1 |   2.860291   .3694912     7.74   0.000     2.136102     3.58448
_In_hpv_al~2 |    2.26865    .578973     3.92   0.000     1.133883    3.403416
hpv_risco_~t |   1.123707   .5528647     2.03   0.042     .0401124    2.207302
       _cons |  -3.079003   .2612538   -11.79   0.000    -3.591051   -2.566955
-------------+----------------------------------------------------------------
ligeira      |
_In_hpv_al~1 |   4.138278   .6131227     6.75   0.000     2.936579    5.339976
_In_hpv_al~2 |    2.26865    .578973     3.92   0.000     1.133883    3.403416
hpv_risco_~t |   1.123707   .5528647     2.03   0.042     .0401124    2.207302
       _cons |  -5.031011   .5461517    -9.21   0.000    -6.101448   -3.960573
------------------------------------------------------------------------------
```

O teste de razão de verosimilhança permitiu comparar este modelo com o modelo generalizado, onde se conclui que o modelo com menos parâmetros (riscos parciais) é preferível ao modelo generalizado ($G(2) = 0.59,\ p = 0.7459$).

Para comparação dos resultados, foi ainda analisada a parametrização proposta por Peterson e Harrel, designada de **modelo de riscos proporcionais parciais irrestrito**. A interpretação do seu *output* (Tabela 7.6) é bastante idêntica à realizada anteriormente, sendo necessário avaliar, para além dos coeficientes $\beta's$, os coeficientes $\gamma's$ significativos (desvios à proporcionalidade), os quais indicam que o pressuposto da proporcionalidade de riscos é violado. Assim verifica-se, como seria de esperar, que existe apenas um coeficiente gama significativo, o da segunda categoria da variável n_hpv_alto_risco_cat.

O valor dos coeficientes obtidos para a terceira categoria do n_hpv_alto_risco_cat e para o hpv_risco_indet são os mesmos que os do modelo de riscos proporcionais (Tabela 7.6). Para a segunda categoria do n_hpv_alto_risco_cat é necessário somar os coeficientes β e γ, obtendo-se o valor de 4.138278 (2.860291+1.277987).

Na Tabela 7.7 e 7.8 estão resumidos os valores de *odds ratios* obtidos a partir de cada um dos modelos construídos. O risco associado ao hpv_risco_indet é muito semelhante entre as várias equações de regressão, sempre na ordem das três unidades. Para a segunda categoria do n_hpv_alto_risco_cat, o risco de vir a desenvolver lesões de maior severidade pode aumentar entre quatro e dez vezes mais. A primeira categoria do n_hpv_alto_risco_cat é a que apresenta maior variabilidade no risco, podendo estar associada a um aumento de risco de cerca de 20 vezes mais ou atingir as 60 vezes.

Pela análise realizada anteriormente, não se justifica a utilização do modelo de riscos proporcionais generalizados, uma vez que o modelo resultante é menos parcimonioso e, duas das variáveis mostraram nunca violar o pressuposto da proporcionalidade de riscos. Por outro lado, o modelo de riscos proporcionais deixa a dúvida se um dos níveis de uma variável viola o pressuposto da linearidade e, esse mesmo nível apresenta magnitudes bastante diferentes entre as várias equações: (1) sem lesão *versus* com lesão ligeira ou severa e (2) sem lesão ou lesão ligeira *versus* lesão severa. Por último, o modelo de riscos proporcionais parciais e o modelo de riscos proporcionais parciais irrestritos só diferem na estimativa de um dos níveis de uma variável, sendo a diferença pouco relevante.

Tabela 7.6 Resultados para a regressão do modelo de riscos proporcionais parciais irrestrito (go*logit*2 – auto gamma).

```
. xi: gologit2 lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, auto gamma
i.n_hpv_alto_~t   _In_hpv_alt_0-2     (naturally coded; _In_hpv_alt_0 omitted)
------------------------------------------------------------------------------
Testing parallel lines assumption using the .05 level of significance...

Step  1:  Constraints for parallel lines imposed for hpv_risco_indet (P Value = 0.8295)
Step  2:  Constraints for parallel lines imposed for _In_hpv_alt_2 (P Value = 0.4904)
Step  3:  Constraints for parallel lines are not imposed for
          _In_hpv_alt_1 (P Value = 0.01488)

Wald test of parallel lines assumption for the final model:

 ( 1)  [ausente]hpv_risco_indet - [ligeira]hpv_risco_indet = 0
       ( 2)  [ausente]_In_hpv_alt_2 - [ligeira]_In_hpv_alt_2 = 0

                chi2(  2) =    0.52
              Prob > chi2 =    0.7707
An insignificant test statistic indicates that the final model does not violate the
proportional odds/ parallel lines assumption
      If you re-estimate this exact same model with gologit2, instead of autofit you can save
      time by using the parameter
      pl(hpv_risco_indet _In_hpv_alt_2)
      ------------------------------------------------------------------------------
      Generalized Ordered Logit Estimates               Number of obs   =        417
                                                        Wald chi2(4)    =      76.46
                                                        Prob > chi2     =     0.0000
      Log likelihood = -142.26208                       Pseudo R2       =     0.2402

       ( 1)  [ausente]hpv_risco_indet - [ligeira]hpv_risco_indet = 0
       ( 2)  [ausente]_In_hpv_alt_2 - [ligeira]_In_hpv_alt_2 = 0
      ------------------------------------------------------------------------------
           lesao_i |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
      -------------+----------------------------------------------------------------
      ausente      |
      _In_hpv_al~1 |   2.860291   .3694912     7.74   0.000     2.136102     3.58448
      _In_hpv_al~2 |    2.26865    .578973     3.92   0.000     1.133883    3.403416
      hpv_risco_~t |   1.123707   .5528647     2.03   0.042     .0401124    2.207302
             _cons |  -3.079003   .2612538   -11.79   0.000    -3.591051   -2.566955
      -------------+----------------------------------------------------------------
      ligeira      |
      _In_hpv_al~1 |   4.138278   .6131227     6.75   0.000     2.936579    5.339976
      _In_hpv_al~2 |    2.26865    .578973     3.92   0.000     1.133883    3.403416
      hpv_risco_~t |   1.123707   .5528647     2.03   0.042     .0401124    2.207302
          _cons |  -5.031011   .5461517    -9.21    0.000    -6.101448   -3.960573
      ------------------------------------------------------------------------------

      Alternative parameterization: Gammas are deviations from proportionality
      ------------------------------------------------------------------------------
           lesao_i |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
      -------------+----------------------------------------------------------------
      Beta         |
      _In_hpv_al~1 |   2.860291   .3694912     7.74   0.000     2.136102     3.58448
      _In_hpv_al~2 |    2.26865    .578973     3.92   0.000     1.133883    3.403416
      hpv_risco_~t |   1.123707   .5528647     2.03   0.042     .0401124    2.207302
      -------------+----------------------------------------------------------------
      Gamma_2      |
      _In_hpv_al~1 |   1.277987   .5247532     2.44   0.015     .2494893    2.306484
      -------------+----------------------------------------------------------------
      Alpha        |
           _cons_1 |  -3.079003   .2612538   -11.79   0.000    -3.591051   -2.566955
           _cons_2 |  -5.031011   .5461517    -9.21   0.000    -6.101448   -3.960573
------------------------------------------------------------------------------
```

Tabela 7.7 *Odds Ratios* obtidos nas equações de regressão logística ordinal (modelo de riscos proporcionais, modelo de riscos proporcionais generalizados, modelo de riscos proporcionais parciais e modelo de riscos proporcionais parciais irrestritos).

Modelo	Modelo Riscos Proporcionais		Modelo Riscos Proporcionais Generalizados		Modelo Riscos Proporcionais Parciais		Modelo Riscos Proporcionais Parciais irrestritos	
Equações	0 *vs* (1+2)	(0+1) *vs* 2	0 *vs* (1+2)	(0+1) *vs* 2	0 *vs* (1+2)	(0+1) *vs* 2	0 *vs* (1+2)	(0+1) *vs* 2
n_hpv_alto_risco_cat(1)	20,73		17,37	50,94	17,47	62,69	17,47	53,95
n_hpv_alto_risco_cat(2)	8,40		10,46	4,40	9,67		9,67	
hpv_risco_indet	3,05		3,00	3,63	3,08		3,08	

Legenda: 0 – sem lesão; 1 – com lesão ligeira e 2 – com lesão severa.

Tabela 7.8 Intervalos de confiança a 95% para *Odds Ratios* obtidos nas equações de regressão logística ordinal (modelo de riscos proporcionais, modelo de riscos proporcionais generalizados, modelo de riscos proporcionais parciais e modelo de riscos proporcionais parciais irrestritos).

Modelo	Modelo Riscos Proporcionais		Modelo Riscos Proporcionais Generalizados		Modelo Riscos Proporcionais Parciais		Modelo Riscos Proporcionais Parciais irrestritos	
Equações	0 *vs* (1+2)	(0+1) *vs* 2	0 *vs* (1+2)	(0+1) *vs* 2	0 *vs* (1+2)	(0+1) *vs* 2	0 *vs* (1+2)	(0+1) *vs* 2
n_hpv_alto_risco_cat(1)	[10,023;42,837]		[8,38;35,98]	[14,01;185,18]	[8,47;36,03]	[18,85;208,51]	[8,47;36,03]	[18,85;208,51]
n_hpv_alto_risco_cat(2)	[2,754;25,609]		[3,28;33,35]	[0,39;50,37]	[3,11;30,07]		[3,11;30,07]	
hpv_risco_indet	[1,0468;8,9070]		[0,92;9,77]	[0,67;19,69]	[1,04;9,09]		[1,04;9,09]	

Legenda: 0 – sem lesão; 1 – com lesão ligeira e 2 – com lesão severa.

8. Considerações Finais

O presente trabalho teve como principais objectivos a identificação dos factores de risco que levam à infecção do vírus do papiloma humano, bem como determinar os factores que contribuem para o desenvolvimento de lesões do colo do útero. Neste âmbito foram analisados questionários sobre aspectos demográficos e comportamento sexual das inquiridas, avaliaram-se os resultados histológicos, citológicos e biópsias realizadas, assim como os resultados efectuados para a genotipagem do HPV.

Dos 417 questionários analisados, foi confirmada a presença de HPV em 107 mulheres, que corresponde a 25.6% dos casos. Os vírus mais frequente são os HPV's de alto risco (16, 52, 31 e 58, por ordem decrescente), seguidos dos vírus de baixo risco, de risco indeterminado e por último dos de provável alto risco. A maioria das infecções é simples, tendo sido detectadas co-infecções em 36 indivíduos. De referir que as infecções múltiplas estão quase sempre associadas a um HPV de alto risco (94.44%).

Relativamente à existência de lesões, foram detectadas alterações de caracter patológico em 50 indivíduos (12.0%), dos quais 28 (6.7%) apresentam evidências de lesões ligeiras (condilomas, ASC-US e LSIL) e 22 (5.3%) evidências de condições mais severas (HSIL e carcinomas). A análise exploratória mostrou que as alterações patológicas são consistentes com a existência de HPV's, sendo que 86% dos indivíduos com lesões severas estão infectados com pelo menos um vírus.

A análise exploratória permitiu identificar algumas associações entre a infecção e características demográficas e hábitos sexuais. No entanto, algumas das análises inicialmente pretendidas não foram efectuadas, tanto pela representatividade dos indivíduos como pela consistência na informação.

Assim, alguns dos objectivos inicialmente traçados não foram conseguidos, nomeadamente a identificação das estripes de HPV mais frequentes nos vários grupos populacionais. Seria de esperar, dado o *pool* geográfico da população que reside nos concelhos de Sintra e Amadora, uma determinada distribuição das estripes de HPV entre vários grupos populacionais. No entanto, esta variável não foi identificada em 13 indivíduos e houve ainda classes constituídas por menos de cinco observações.

Outra variável que não foi incluída na análise, apesar de ser considerada fundamental na epidemiologia do HPV, foi a existência de actividade sexual. Não é claro que uma resposta negativa

indique que nunca tenha ocorrido contacto sexual ou que no momento do estudo a mulher seja sexualmente inactiva. E mesmo que a pergunta tenha sido respondia correctamente, não existe indicação de quanto tempo demora a abstinência sexual. Por outro lado, o número de parceiros sexuais foi desde o início uma variável categórica. Assim, propõe-se num trabalho futuro a utilização, sempre que possível, de variáveis de natureza continua, sendo a categorização feita à posteriori, tendo em conta as referencias bibliográficas, mas também a representatividade de indivíduos em cada classe. Pela bibliografia consultada, sugere-se também que seja incluída informação adicional sobre o comportamento sexual, nomeadamente a data de início da actividade sexual e o estado civil.

A identificação dos factores de risco na infecção do HPV foi realizada com recurso à regressão logística binária, onde se identificou como tendo um contributo relevante na infecção a idade, o número de parceiros sexuais, a infecção por HIV, o uso de preservativo e uso de dispositivo intra-uterino. Os resultados obtidos para a idade são semelhantes aos identificados em outras populações europeias, onde se verifica uma diminuição do risco à medida que aumenta a idade. Neste caso, o aumento de um ano de idade diminui em 6% o risco de infecção. Sendo o contacto a principal via de infecção, os resultados obtidos estão de acordo com o esperado, *i.e.*, mulheres que usam preservativo estão menos expostas à infecção, sendo o seu risco 59% mais baixo que nas restantes mulheres. Por outro lado, o número de parceiros sexuais é um factor determinante na infecção. Como seria de esperar, quanto maior for o contacto com diferentes parceiros, maior é a janela de abertura para a infecção. Nesta dissertação verificou-se que mulheres que já tiveram mais do que cinco parceiros têm o triplo de possibilidades de se tornarem infectadas. O uso de dispositivo intra-uterino foi identificado como factor de risco, aumentando as possibilidades de infecção em 2.8 vezes. Não existe suporte bibliográfico para esta associação, na medida em que possa existir uma característica que torne as mulheres mais frágeis e por isso mais susceptíveis ao contágio. O aumento da infecção deste grupo de mulheres deverá estar associado à não utilização de preservativo como método preferencial de contracepção. Por último, a infecção por HPV mostrou ser um dos factores mais importantes no contágio do HPV com cerca de 14 vezes mais de possibilidades. Esta associação é identificada na bibliografia, na medida em que a existência de co-infecções comprometem o sistema imunitário, tornando-o mais frágil e tornando a infecção persistente, sendo assim detectável em programas de rastreio.

Relativamente à análise das lesões, a dimensão da amostra não permitiu fazer uma análise individual ao tipo de lesões, tendo-se optado por agrupá-las em classes de severidade (ligeiras e severas).

A primeira abordagem a este tópico foi efectuada pela aplicação da regressão multinomial que comparou mulheres com lesões ligeiras e severas com a classe de referência, sem lesão. O modelo

resultante é constituído pelos efeitos principais da variável idade e da infecção de HPV de alto risco. A idade, que mostrou cumprido o pressuposto da linearidade do logit, apenas foi significativa na primeira equação (sem lesão *versus* lesão ligeira), onde foi classificada como tendo um efeito protector, sendo que o aumento de um ano de idade está associado a uma diminuição de cerca de 6% no desenvolvimento de lesões ligeiras. Por outro lado, e como seria de esperar, a infecção por vírus de alto risco é um factor determinante no desenvolvimento de lesões. Mulheres com este tipo de infecção têm vezes mais de possibilidade de desenvolverem lesões ligeiras e 72 vezes mais de possibilidade de adquirem uma lesão severa.

A outra abordagem efectuada ao desenvolvimento de lesões foi considerar que a variável resposta teria os seus níveis ordenados, realizando-se a modelação via modelos ordinais. Dentro destes, e dada a natureza da variável resposta, optou-se pelo modelo de riscos proporcionais. O modelo resultante tem como variáveis explicativas o número de HPV's de alto risco (em categorias) e a existência de HPV's de risco indeterminado. A análise aos *odds ratios* permitiu concluir que mulheres infectadas por um único vírus de alto risco têm 20 vezes mais de possibilidades de adquirem uma lesão severa do que desenvolverem lesões ligeiras ou mesmo não terem qualquer alteração de caracter patológico. Da mesma forma, e dada a proporcionalidade de riscos, o risco de ter lesão severa ou ligeira aumenta 20 vezes comparativamente ao risco de não ter lesão. Para ser possível tecer estas considerações é necessário ser válido o pressuposto de riscos proporcionais. A validação deste pressuposto deixou algumas dúvidas, por um lado, o teste de razão de verosimilhança e o teste aproximado de verosimilhança de Hosmer rejeitam a proporcionalidade, por outro, o teste de Brant aceita como válido este pressuposto. Assim, foram analisados outros modelos ordinais, nomeadamente o modelo ordinal generalizado, o modelo de riscos proporcionais parciais e o modelo de riscos proporcionais parciais irrestritos.

O modelo ordinal generalizado não necessita que nenhuma variável cumpra a proporcionalidade de riscos, no entanto, tem a desvantagem de ser um modelo mais pesado relativamente ao número de parâmetros do que o modelo anterior, com uma complexidade semelhante ao modelo multinomial. A comparação de modelos, via razão de verosimilhança, indica que o modelo ordinal generalizado é preferível neste caso ao modelo ordinal de riscos proporcionais.

Por outro lado, os modelos ordinais de riscos proporcionais parciais permitem ter um modelo onde apenas algumas variáveis cumprem o pressuposto, sem no entanto, comprometer a proporcionalidade de risco no modelo final. Neste modelo, apenas a segunda categoria do número de HPV de alto risco compromete a proporcionalidade, obtendo-se assim dois coeficientes para este nível. A partir da interpretação dos seus coeficientes, conclui-se que a infecção por um vírus de alto

risco aumenta em 17 vezes a possibilidade de desenvolver lesões ligeiras ou severas, enquanto que a possibilidade de desenvolver lesões severas comparativamente às outras duas classes, aumenta quase 63 vezes. Relativamente às infecções múltiplas, a possibilidade de terem lesões com maior severidade aumenta praticamente 10 vezes. Por outro lado, os vírus classificados como risco indeterminado triplicam o risco de desenvolvimento de lesões mais graves. A comparação deste modelo com o anterior, via razão de verosimilhança, indica que o modelo mais parcimonioso é o mais indicado para descrever os dados em análise.

Por último, o modelo de riscos proporcionais parciais irrestritos origina estimativas muito próximas ao modelo anterior, variando apenas na variável que não cumpre a proporcionalidade, onde o risco de desenvolver lesões severas é cerca de 54 vezes mais do que de desenvolver lesões ligeiras ou não ter qualquer tipo de lesão.

Em conclusão, esta dissertação permitiu obter os resultados esperados e habitualmente encontrados na bibliografia. A infecção por HPV foi explicada considerando a idade, o número de parceiros sexuais, a infecção por HIV, o uso de preservativo e uso de dispositivo intra-uterino. Relativamente ao desenvolvimento de lesões, a bibliografia de suporte é bastante mais reduzida, em especial na população portuguesa. Os modelos encontrados sugerem que a idade, a infecção de HPV de alto risco e de provável alto riscos são factores importantes para o aparecimento de alterações de caracter patológico.

9. Bibliografia

Abreu, MNS, Siqueira, AL e Caiaffa, WT 2009, 'Regressão Logística ordinal em estudos epidemiológicos', *Revista de Saúde Pública*, vol.43, no1, pp.183-194.

Agresti, A 2007, *An introduction to categorical data analysis*, 2nd ed, New Jersey: Wiley-Interscience.

Casella, G e Berger, R L 2002, *Statistical Inference*, 2nd ed, Duxburu, Thomson Learning Inc.

Castellsagué, X 2008, 'Natural History and Epidemiology of HPV Infection and Cervical Cancer', *Gynecological Oncology*, vol. 110, pp.S4-S7.

Clifford, GM, Smith, JS, Plummer, M, Muñoz, N e Franceschi, S 2003, 'Human Papillomavirus types in invasive Cervical Cancer Worldwide: a meta-analysis, *British Journal of Cancer*, vol. 88, pp. 63-73.

Dunne, EF, Unger, ER, Sternberg, M, McQuillan, G, Swan, DC, Patel, SS, Markowitz, LE 2007, 'Prevalence of HPV infection among females in the United States', Journal of the American Medical Association, vol.297, no.8, pp. 813-819.

Hameed, M, Fernandes, H, Skurnick, J, Moore, D, Kloser, P e Heller, D 2001, 'Human papillomavirus typing in HIV-positive women', Infectious Diseases in Obstetrics Gynecology, vol.9, pp. 89–93.

Holly, EA 1996, 'Cervical Intraepithelial Neoplasia, Cervical Cancer and HPV', *Annual Review of Public Health*, vol. 17, pp. 69-84.

Hosmer, DW e Lemeshow, S 2000, *Applied Logistic Regression*, 2nd ed, Wiley, John Wiley & Sons Inc.

Long, JS e Cheng, S 2004, 'Regression Models for Categorical Outcomes', in Hardy, M e Bryman, A (ed.), The Hand Book of Data Analysis, Sages Publications Inc, pp. 259-284.

Long, S e Freese, J 2001, *Regression Models for Categorical Outcomes Using Stata*, Stata Press, Texas.

McCullagh, P e Nelder, JA 1989, *Generalized Linear Models*, 2nd ed, New York: Chapman & Hall.

Montgomery, DC 2005, *Introduction to Statistical Quality Control*, 5th ed, Wiley, John Wiley & Sons, Inc.

Myers, JL e Well, AD 2003, *Research Design and Statistical Analysis*, 2nd ed, Lawrence Erlbaum Associates, Publishers, London.

Oehlert, GW 2010, *A First Course in Design and Analysis of Experiments*, Library of Congress Cataloging-in-Publication Data.

Papachristou, E, Sypsa, V, Paraskevis, D, Gkekas, A, Politi, E, Nicolaidou, E, Anifantis, I, Psichogiou, M, Tsitsika, A, Hadjivassiliou, M, Petrikkos, G, Katsambas, A, Creatsas, G, Hatzakis A 2009, 'Prevalence of different HPV types and estimation of prognostic risk factors based on the linear array HPV genotyping test', *Journal of Medical Virology*, vol. 81, pp.2059-2065.

Pista, A, Oliveira, CF, Cunha, MJ, Paixão, MT, Real, O 2011, 'Prevalence of Human Papillomavirus Infeccion in Women in Portugal: the CLEOPATRE Study', *International Journal of Gynecological Cancer*, vol 21, no6, pp.1150-1158.

Pista, A, Oliveira, CF, Cunha, MJ, Paixão, MT, Real, O 2012, 'Risk factors for human papillomavirus infeccion among women in Portugal: the CLEOPATRE Portugal Study', *Internationa Journal of Gynecology and Obstetrics*, *in press*.

Powers, DA e Xie, Y 1999, *Statistical Methods for Categorical Data Analysis*, Academic Press, Inc.

Razali, NM e Wah, YB 2011, 'Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lillefors and Anderson-Darling tests', *Journal of Statistical Modeling and Analytics*, vol.2, no1, pp. 21-33.

Risser, JMH, Risser, WL, Risser, AL 2008, 'Epidemiology of Infections in Women', *Infectious Disease Clinics of North America*, vol. 22, pp. 581-599.

Rutherford, A 2001, *Introducing ANOVA and ANCOVA: A GLM Approach*, Thousand Oaks, CA: Sage Publications.

Sanjosé, S Diaz, M, Castellsagué, X, Clifford, G, Bruni, L Muñoz, N e Bosch FX 2007, 'Wordwide prevelence and genotype distribution of cervical human papillomavirus DNA in women with normal citology: a meta-analysis', *The Lancet Infectious Diseases*, vol. 7, no 7, pp. 453-459.

Siegel, S 1957, 'Nonparametric Statistics', *The American Statistics*, vol. 11, no.3, pp. 13-19.

Silva, J, Ribeiro, J, Sousa, H, Cerqueira, F, Teixeira, AL, Baldaque, I, Osório, T e Medeiros, R 2011, 'Oncogenic HPV Types Infeccion in Adolescents and University Women from North Portugal: From Self-Sampling to Cancer Prevention', *Journal of Oncology*, vol. 2011, pp.1-8.

Smith, JS, Melendy, A, Rana, RK e Pimenta, JM 2008, 'Age-Specific Prevalence of Infection with Human Papillomavirus in Females: A Global Review', *Journal of Adolescent Health*, vol. 43, pp. S5-S25.

Sokal, RR e Rohlf, FJ 2009, *Introduction to Biostatistics,* 2nd ed, Dover publications Inc., Mineola, New York.

Stata 9 2005, A Stata Press Publiation, College Station, Texas.

Turkman, MAA e Silva, GL 2000, *Modelos Lineares Generalizados – da Teoria à Prática*, Edições SPE, Lisboa.

Vermund, SH e Bhatta, MP 2004, 'Papillomavirus infections', In Cohen, J e Powderly, WG, (eds.), Infectious Diseases, 2nd ed, St Louis: Mosby.

Wheeler, CM 2008, 'Natural history of Human Papillomavirus Infections, Cytologic and Histologic Abnormalities and Cancer', *Obstetrics and Gynecology Clinics of North America*, vol. 35, pp-519-536.

Williams, R 2006, 'Generalized Ordered Logit/ Partial Proportional Odds Models for Ordinal Dependent Variables', *The Stata Journal*, vol. 6, no.1, pp. 49-52.

Wolfe, R e Gould, W 1998, 'An approximate likelihood-ratio test for ordinal response models', *Stata Technical Bulletin*, no 42, pp. 24-27.

10. Anexos

Anexo 1 – Inquerito realizado aos indivíduos integrantes no estudo.

ANEXO 1

Questionário epidemiológico para genotipagem de HPV

Processo nº Análise nº
Nome:
Idade:

Serviço/Centro de Saúde:

Seleccione a resposta correcta:

Raça: Negra ____
Caucasiana ____
Ameríndia ____
Asiática ____
Outra ____

Nacionalidade:
Portuguesa ____ Outra ____

Actividade sexual: Sim ___
Não ___

Nº parceiros sexuais: 1 a 5
5 a 10
> 10

DST: Sífilis / HIV / Chlamydia
Outra

Anticonc.: Pílula ___
D.I.U ___ Preservativo ___
Adesivo/implante ___
Barreira ____
Outro ____

Terapêutica local? ____
Qual? _____

Anexo 2 – Variáveis analisadas

Número Variável	Variável	Categorias	Observações	utilização na modelação
1	Processo	Variável que identifica a utente		
2	Proveniência	1 - HFF 2 - CC_Cacem 3 - CC_Brandoa 4 - CC_Damaia 5 - CC_Reboleira 6 - CC_A/MM 7 - CC_S. Marcos 8 - CC_Amadora 9 - CC_Venda Nova		
2.1	proveniência	1 - HFF 2 - Centro de Saúde		✓
3	Dt_Nascimento	dd-mm-aaaa		
3.1	Idade	(anos)		✓
4	Grupo Popolacional	1 - Negroide 2 - Caucasóide 3 - Ameríndea 4 - Asiática 5 - Outra		
4.1	grupo_pop	1 - Negroide 2 - Caucasóide		✓
4.2	grupo_pop2	1 - Negroide 2 - Caucasóide grupo_pop2 3 - Outra grupo_pop2(1))		✓
5	Nacionalidade	1 - Portuguesa 2 - Outra		
6	Perm_Portugal	(anos)		
7	Activ_Sexual	0 - Não 1 - Sim		
8	N_Parceiros_categ	1 - 1 a 5 2 - 6 a 10 3 - mais de 11		
8.1	N_Parceiros_categ_2	1 - até cinco; 2 - mais de cinco		✓
9	DST_Sífilis	0 - Não 1 - Sim		✓
10	DST_HIV	0 - Não 1 - Sim		✓
11	DST_Clamídea	0 - Não 1 - Sim		✓
12	DST_Outra	0 - Não 1 - Sim		✓
13	Anticoncept_Pílula	0 - Não 1 - Sim 8 - Não aplicável		✓
14	Anticoncept_DIU	0 - Não 1 - Sim 8 - Não aplicável		✓
15	Anticoncept_Preserv	0 - Não 1 - Sim 8 - Não aplicável		✓
16	Anticoncept_Adesivo	0 - Não 1 - Sim 8 - Não aplicável		✓
17	Anticoncept_Barreira	0 - Não 1 - Sim 8 - Não aplicável		✓

Número Variável	Variável	Categorias	Observações	utilização na modelação
18	Anticoncept_Outro	0 - Não 1 - Sim 8 - Não aplicável		✓
19	Terap Local	0 - Não 1 - Sim		
20	Dt_Citologia	dd-mm-aaaa		
21	Citologia_Res	0 - Sem lesão 1 - Com inflamação 2 - Com lesão 3 - Com Infl + lesão 8 - insatisfatório		
22	Infl_Cocos	0 - Não 1 - Sim		
23	Infl_Tricomonas	0 - Não 1 - Sim		
24	Infl_Fungos	0 - Não 1 - Sim		
25	Infl_Actinomices	0 - Não 1 - Sim		
26	Les_ASC-US	0 - Não 1 - Sim		✓
27	Les_AGC	0 - Não 1 - Sim		✓
28	Les_LSIL	0 - Não 1 - Sim		✓
29	Les_HSIL	0 - Não 1 - Sim		✓
30	Les_ASC-H	0 - Não 1 - Sim		✓
31	Les_Carc_Pav_Cel	0 - Não 1 - Sim		✓
32	Les_Car_NOS	0 - Não 1 - Sim		✓
33	Les_Adenocarc	0 - Não 1 - Sim		✓
34	HPVar_16	0 - Negativo 1- Positivo		✓
35	HPVar_18	0 - Negativo 1- Positivo		✓
36	HPVar_31	0 - Negativo 1- Positivo		✓
37	HPVar_33	0 - Negativo 1- Positivo		✓
38	HPVar_35	0 - Negativo 1- Positivo		✓
39	HPVar_39	0 - Negativo 1- Positivo		✓
40	HPVar_45	0 - Negativo 1- Positivo		✓
41	HPVar_51	0 - Negativo 1- Positivo		✓
42	HPVar_52	0 - Negativo 1- Positivo		✓
43	HPVar_56	0 - Negativo 1- Positivo		✓
44	HPVar_58	0 - Negativo 1- Positivo		✓

Número Variável	Variável	Categorias	Observações	utilização na modelação
45	HPVar_59	0 - Negativo 1- Positivo		✓
46	HPVar_68	0 - Negativo 1- Positivo		✓
47	HPVpar_26	0 - Negativo 1- Positivo		✓
48	HPVpar_53	0 - Negativo 1- Positivo		✓
49	HPVpar_66	0 - Negativo 1- Positivo		✓
50	HPVpar_73	0 - Negativo 1- Positivo		✓
51	HPVpar_82	0 - Negativo 1- Positivo		✓
52	HPVri_62	0 - Negativo 1- Positivo		✓
53	HPVri_71	0 - Negativo 1- Positivo		✓
54	HPVri_83	0 - Negativo 1- Positivo		✓
55	HPVri_84	0 - Negativo 1- Positivo		✓
56	HPVri_85	0 - Negativo 1- Positivo		✓
57	HPVbr_6	0 - Negativo 1- Positivo		✓
58	HPVbr_11	0 - Negativo 1- Positivo		✓
59	HPVbr_40	0 - Negativo 1- Positivo		✓
60	HPVbr_42	0 - Negativo 1- Positivo		✓
61	HPVbr_43	0 - Negativo 1- Positivo		✓
62	HPVbr_44	0 - Negativo 1- Positivo		✓
63	HPVbr_54	0 - Negativo 1- Positivo		✓
64	HPVbr_61	0 - Negativo 1- Positivo		✓
65	HPVbr_70	0 - Negativo 1- Positivo		✓
66	HPVbr_72	0 - Negativo 1- Positivo		✓
67	HPVbr_81	0 - Negativo 1- Positivo		✓
68	HPVbr_89	0 - Negativo 1- Positivo		✓

Número Variável	Variável	Categorias	Observações	utilização na modelação
69	hpv	0 - Negativo 1- Positivo	(reune informação das variáveis 48 - 82)	✓
69.1	hpv_multiplos	0 - Não 1 - Sim	(reune informação das variáveis 48 - 82)	✓
69.2	n_hpv_cat	0 - 0 1 - 1 n_hpv_cat(1) 2 - dois ou mais n_hpv_cat(2)	(reune informação das variáveis 48 - 82)	✓
69.3	hpv_alto_risco	0 - Não 1 - Sim	(reune informação das variáveis 48 - 60)	✓
69.4	n_hpv_alto_risco_cat	0 - 0 1 - 1 2 - dois ou mais	(reune informação das variáveis 48 - 60)	✓
69.5	hpv_prov_alto_risco	0 - Não 1 - Sim	(reune informação das variáveis 61-65)	✓
69.6	hpv_risco_indet	0 - Não 1 - Sim	(reune informação das variáveis 66-70)	✓
69.7	hpv_baixo_risco	0 - Não 1 - Sim	(reune informação das variáveis 71-82)	✓
70	lesão_i	0 - ausente 1 - ligeira 2 - alta	(reune informação das variáveis 30-46)	✓

Legenda: As variáveis sublinhadas a rosa correspondem ao evento de interesse e foram incluídas como variáveis respostas nos capítulos 5, 6 e 7. Para efeitos de modelação, o STATA utiliza a primeira classe como a classe de referência. No caso de variáveis constituídas por três ou mais níveis, o Stata ao apresenta o valor dos coeficientes, acrescenta o sufixo “(1)” na segunda categoria, “(2)” na terceira e assim sucessivamente. São exemplo disso, as variáveis grupo_pop2 e n_hpv_cat.

Anexo 3 – Comandos Stata Utilizados

```
**********Estatística Descritiva**********
*****************************************
***1. Idade***
swilk idade
sfrancia idade
symplot idade
qnorm idade
quantile idade
sktest idade
histogram idade, normal
stem idade
dotplot idade
pnorm idade
qnorm idade

summarize idade
ksmirnov idade = normal((idade-40.86091)/12.46976)


***2. Idade vs activ_sexual***
summarize idade, by (activ_sexual)

histogram idade , normal by (activ_sexual)
sktest idade if activ_sexual==0
sktest idade if activ_sexual==1
swilk idade if activ_sexual==0
swilk idade if activ_sexual==1


tabulate activ_sexual, summarize(idade)
graph box idade, over(activ_sexual) title (activ_sexual)

*igualdade de variâncias
sdtest idade, by (activ_sexual)

*teste t quando existe homocedastcidade
ttest idade, by (activ_sexual)

*Mann-Whitney***
ranksum idade, by(activ_sexual)

***3. Idade vs  n_parceiros_categ_2***
tabulate n_parceiros_categ_2, summarize(idade)
graph box idade, over( n_parceiros_categ_2) title (n_parceiros_categ)

sdtest idade, by ( n_parceiros_categ_2)
*teste t quando não existe homocedastcidade
ttest idade, by (activ_sexual) welch

*Mann-Whitney***
ranksum idade, by(n_parceiros_categ_2)


***4.1 contraceptivos vs activ_sexual***
tabulate contraceptivos activ_sexual, chi2 exact


***4.2 contraceptivos vs idade***
tabulate contraceptivos, summarize(idade)
graph box idade, over(contraceptivos)

graph box idade, over(contraceptivos)  title(contraceptivos)

sdtest idade, by (contraceptivos)
ttest idade, by (contraceptivos) welch
ranksum idade, by(contraceptivos)
```

```
 ***4.3. HPV vs Idade***
tab  hpv
tab  n_hpv
tabulate hpv, summarize(idade)
graph box idade, over(hpv) title (HPV)

*igualdade de variâncias
sdtest idade, by (hpv)

*teste t quando existe homocedastcidade
ttest idade, by (hpv)

*Anova*
oneway idade n_hpv_cat, tab
oneway idade n_hpv_cat, tab scheffe
anova idade n_hpv_cat


gen id=_n
predict residuals
predict residuals, residuals
predict stdres, rstandard
predict yhat

twoway   (scatter   stdres   yhat),   graphregion(fcolor(white)   lcolor(white)   ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))

qnorm stdres

sktest  stdres
sort stdres
tab stdres
drop  residuals stdres p id yhat

sdtest idade, by (n_hpv_cat)
ttest idade, by (hpv)

***Kruskal-Wallis***
kwallis n_hpv_cat, by(idade)
ranksum idade if (n_hpv_cat==0|n_hpv_cat==1), by(n_hpv_cat)
ranksum idade if (n_hpv_cat==0|n_hpv_cat==2), by(n_hpv_cat)
ranksum idade if (n_hpv_cat==1|n_hpv_cat==2), by(n_hpv_cat)
di 0.1668/2

*** HPV vs nacional***

*** HPV vs activ_sexual e  n_parceiros_categ_2***
tabulate hpv activ_sexual, chi2 exact
tabulate hpv n_parceiros_categ_2, chi2 exact

findit tabchi
tabchi hpv n_parceiros_categ_2, raw pearson cont adjust noo noe
tabchi hpv n_parceiros_categ_2


findit tab3way
tab3way  hpv n_parceiros_categ_2  activ_sexual , coltot rowp
cc hpv  n_parceiros_categ_2 , by( activ_sexual)


*** HPV vs dst***
tabulate hpv dst, chi2 exact

*** HPV vs  contraceptivos***
tabulate hpv  contraceptivos, chi2 exact
tab3way  activ_sexual hpv contraceptivos, coltot rowp
cc hpv  contraceptivos , by( activ_sexual)

tabulate hpv  anticoncept_preserv, chi2 exact
tabulate hpv  anticoncept_pílula, chi2 exact
tabulate hpv  anticoncept_diu, chi2 exact
tabulate hpv  anticoncept_adesivo, chi2 exact
tabulate hpv  anticoncept_barreira, chi2 exact
tabulate hpv  anticoncept_outro, chi2 exact

***3. Idade vs tipo de lesão***
oneway idade lesao_i, tab
```

```
oneway idade lesao_i, tab scheffe
anova idade lesao_i
gen id=_n
predict residuals, residuals
predict stdres, rstandard
predict yhat
twoway (scatter stdres yhat), graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
sktest  stdres
sort stdres
tab stdres
drop  residuals stdres id yhat

anova idade lesao_i activ_sexual lesao_i#activ_sexual
anova idade lesao_i##activ_sexual
gen id=_n
predict residuals, residuals
predict stdres, rstandard
predict yhat
twoway (scatter stdres yhat), graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
sktest  stdres
sort stdres
tab stdres
drop  residuals stdres id yhat

graph box idade if lesao_i~=0, over( lesao_i) title (severidade_lesão)
sdtest idade if lesao_i~=0, by(lesao_i)
ttest idade if lesao_i~=0, by (lesao_i)

tabulate lesao_i hpv_alto_risco if lesao_i~=0, chi2 exact


**********Regressão Logística*************
*****************************************

*** Regressão Logistica - ODDS***
logistic hpv proveniência
logistic hpv idade
xi: logistic hpv i.grupo_etário_3
xi:logistic hpv  i.grupo_pop_2
logistic hpv grupo_pop_caucasoide
logistic hpv grupo_pop_negroide
logistic hpv nacionalidade
logistic hpv activ_sexual
xi: logistic hpv i.n_parceiros_categ
logistic hpv n_parceiros_categ_2
logistic hpv  dst
logistic hpv  dst_sífilis
logistic hpv dst_hiv
logistic hpv dst_clamídea
logistic hpv  dst_outra
logistic hpv contraceptivos
logistic hpv anticoncept_pílula
logistic hpv anticoncept_diu
logistic hpv anticoncept_preserv
logistic hpv anticoncept_adesivo
logistic hpv anticoncept_barreira
logistic hpv anticoncept_outro
logistic hpv terap_local

*** Regressão Logistica - COEFICIENTES***
logit hpv proveniência
logit hpv idade
xi: logit hpv i.grupo_etário_3
xi:logit hpv  i.grupo_pop_2
logit hpv grupo_pop_caucasoide
logit hpv grupo_pop_negroide
logit hpv nacionalidade
logit hpv activ_sexual
xi: logit hpv i.n_parceiros_categ
logit hpv n_parceiros_categ_2
logit hpv  dst
logit hpv  dst_sífilis
logit hpv dst_hiv
logit hpv dst_clamídea
logit hpv  dst_outra
```

```
logit hpv contraceptivos
logit hpv anticoncept_pílula
logit hpv anticoncept_diu
logit hpv anticoncept_preserv
logit hpv anticoncept_adesivo
logit hpv anticoncept_barreira
logit hpv anticoncept_outro
logit hpv terap_local

***Regressão Logistica Multivariada***

***1. Todas as variáveis significativas a 25%***
xi: logistic hpv idade i.grupo_pop_2 nacionalidade activ_sexual i.n_parceiros_categ_2 dst_hiv
anticoncept_pílula anticoncept_diu anticoncept_preserv anticoncept_adesivo anticoncept_outro
xi: logit hpv idade i.grupo_pop_2 nacionalidade activ_sexual i.n_parceiros_categ_2 dst_hiv
anticoncept_pílula anticoncept_diu anticoncept_preserv anticoncept_adesivo anticoncept_outro

***2. Exclusão das menos significativas***
***2.1 exclusão nacionalidade***
xi: logistic hpv idade i.grupo_pop_2 activ_sexual i.n_parceiros_categ_2 dst_hiv
anticoncept_pílula anticoncept_diu anticoncept_preserv anticoncept_adesivo anticoncept_outro
xi: logit hpv idade i.grupo_pop_2 activ_sexual i.n_parceiros_categ_2 dst_hiv
anticoncept_pílula anticoncept_diu anticoncept_preserv anticoncept_adesivo anticoncept_outro

***2.2 exclusão anticoncept_pílula***
xi: logistic hpv idade i.grupo_pop_2 activ_sexual i.n_parceiros_categ_2 dst_hiv
anticoncept_diu anticoncept_preserv anticoncept_adesivo anticoncept_outro
xi: logit hpv idade i.grupo_pop_2 activ_sexual i.n_parceiros_categ_2 dst_hiv anticoncept_diu
anticoncept_preserv anticoncept_adesivo anticoncept_outro

***2.3 exclusão activ_sexual***
xi: logistic hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu
anticoncept_preserv anticoncept_adesivo anticoncept_outro
xi: logit hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu
anticoncept_preserv anticoncept_adesivo anticoncept_outro


***2.4 exclusão anticoncept_outro***
xi: logistic hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu
anticoncept_preserv anticoncept_adesivo
xi: logit hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu
anticoncept_preserv anticoncept_adesivo

***2.5 exclusão anticoncept_adesivo***
xi: logistic hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu
anticoncept_preserv
xi: logit hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu
anticoncept_preserv


***3.Inclusão das variáveis que não tinham sido significativas em 2.***
***3.1 inclusao terap_local***
xi: logistic hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu
anticoncept_preserv terap_local
xi: logit hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu
anticoncept_preserv terap_local

***3.2 inclusão dst_sífilis***
xi: logistic hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu
anticoncept_preserv dst_sífilis
xi: logit hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu
anticoncept_preserv dst_sífilis

***3.3 inclusão dst***
xi: logistic hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu
anticoncept_preserv dst
xi: logit hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu
anticoncept_preserv dst

***3.4 inclusão ***
xi: logistic hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu
anticoncept_preserv dst_outra
xi: logit hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu
anticoncept_preserv dst_outra


***3.5 inclusão anticoncept_barreira***
```

```
xi: logistic hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv anticoncept_barreira
xi: logit hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv anticoncept_barreira

***4.Avaliação da variável grupo_pop_2***
xi: logistic hpv idade i.n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv
xi: logit hpv idade i.n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv

***Modelos finais***
xi: logistic hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv
xi: logit hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv

xi: logistic hpv idade i.n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv
xi: logit hpv idade i.grupo_pop_2 dst_hiv anticoncept_diu anticoncept_preserv

xi: stepwise,pr(.10) lr:logit hpv idade grupo_pop_2 nacionalidade activ_sexual n_parceiros_categ_2 dst_sífilis dst dst_hiv dst_clamídea dst_outra anticoncept_pílula anticoncept_diu anticoncept_preserv anticoncept_adesivo anticoncept_barreira anticoncept_outro terap_local

***5. Interacções***

xi: logistic hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv idade*grupo_pop_2

**Modelo com Efeitos Principais***

xi: logistic hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv
xi: logit hpv idade i.grupo_pop_2 i.n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv

xi: logistic hpv idade i.n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv
xi: logit hpv idade i.n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv
predict prlogit
lab var prlogit "logit prob(hpV)"
dotplot prlogit, ylabel(0 .2 to 1)


***Pressuposto da Linearidade do Logit***

***1. Método fráfico - Univariable smoothes scatterplot***
lowess hpv idade, gen(var3) logit nodraw
graph twoway line var3 idade, sort xlabel(20(10)80)

***2. Método gráfico - quartis***
sort idade
tabstat idade, statistics( min p25 p50 p75 max) columns(variables)
centile idade, centile (25 50 75)
xtile quart = idade, nq(4)
xi:logit hpv i.quart  n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv

***input idade_percentil

di (31+16)/2
di (41+31)/2
di (50+41)/2
di (79+50)/2

input idade_quartil or
23.5 0
36 -0.5947919
45.5 -1.191847
64.5 -2.254573
end

twoway (connected idade_quartil or), caption(, color(white)) note(, color(white)) scheme(s2color) graphregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))

***3. Método Analítivo - Polinómios Fraccionários***

fracpoly logit hpv idade n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv, degree(2) compare
```

```
*** Modelo com Interação***

logit hpv idade  n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv


gen idade_n_parceiros_categ_2 = idade*n_parceiros_categ_2
gen idade_dst_hiv = idade*dst_hiv
gen idade_anticoncept_diu = idade*anticoncept_diu
gen idade_anticoncept_preserv= idade*anticoncept_preserv

gen n_parceiros_categ_2_dst_hiv = n_parceiros_categ_2*dst_hiv
gen n_parceiros_categ_2_ant_diu = n_parceiros_categ_2*anticoncept_diu
gen n_parceiros_categ_2_ant_preserv = n_parceiros_categ_2*anticoncept_preserv

gen dst_hiv_anticoncept_diu = dst_hiv*anticoncept_diu
gen dst_hiv_anticoncept_preserv= dst_hiv*anticoncept_preserv

gen anticoncept_diu_preserv= anticoncept_diu*anticoncept_preserv

***não sig***
logit   hpv   idade      n_parceiros_categ_2   dst_hiv   anticoncept_diu   anticoncept_preserv
idade_n_parceiros_categ_2,or
logit   hpv   idade      n_parceiros_categ_2   dst_hiv   anticoncept_diu   anticoncept_preserv
idade_n_parceiros_categ_2
logit   hpv   idade      n_parceiros_categ_2   dst_hiv   anticoncept_diu   anticoncept_preserv
idade_dst_hiv,or
logit hpv idade  n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv idade_dst_hiv
logit   hpv   idade      n_parceiros_categ_2   dst_hiv   anticoncept_diu   anticoncept_preserv
idade_anticoncept_diu,or
logit   hpv   idade      n_parceiros_categ_2   dst_hiv   anticoncept_diu   anticoncept_preserv
idade_anticoncept_diu
logit   hpv   idade      n_parceiros_categ_2   dst_hiv   anticoncept_diu   anticoncept_preserv
idade_anticoncept_preserv,or
logit   hpv   idade      n_parceiros_categ_2   dst_hiv   anticoncept_diu   anticoncept_preserv
idade_anticoncept_preserv

***não sig***
xi:  logit  hpv  idade     n_parceiros_categ_2  dst_hiv  anticoncept_diu  anticoncept_preserv
i.n_parceiros_categ_2_dst_hiv,or
xi:  logit  hpv  idade     n_parceiros_categ_2  dst_hiv  anticoncept_diu  anticoncept_preserv
i.n_parceiros_categ_2_dst_hiv
xi:  logit  hpv  idade     n_parceiros_categ_2  dst_hiv  anticoncept_diu  anticoncept_preserv
i.n_parceiros_categ_2_ant_diu,or
xi:  logit  hpv  idade     n_parceiros_categ_2  dst_hiv  anticoncept_diu  anticoncept_preserv
i.n_parceiros_categ_2_ant_diu
xi:  logit  hpv  idade     n_parceiros_categ_2  dst_hiv  anticoncept_diu  anticoncept_preserv
i.n_parceiros_categ_2_ant_preserv,or
xi:  logit  hpv  idade     n_parceiros_categ_2  dst_hiv  anticoncept_diu  anticoncept_preserv
i.n_parceiros_categ_2_ant_preserv

***não sig***
xi:  logit  hpv  idade     n_parceiros_categ_2  dst_hiv  anticoncept_diu  anticoncept_preserv
i.dst_hiv_anticoncept_diu,or
xi:  logit  hpv  idade     n_parceiros_categ_2  dst_hiv  anticoncept_diu  anticoncept_preserv
i.dst_hiv_anticoncept_diu
xi:  logit  hpv  idade     n_parceiros_categ_2  dst_hiv  anticoncept_diu  anticoncept_preserv
i.dst_hiv_anticoncept_preserv,or
xi:  logit  hpv  idade     n_parceiros_categ_2  dst_hiv  anticoncept_diu  anticoncept_preserv
i.dst_hiv_anticoncept_preserv

***não sig***
xi:   logit   hpv   idade   n_parceiros_categ_2   dst_hiv   anticoncept_diu   anticoncept_preserv
i.dst_hiv_anticoncept_preserv,or
xi:   logit   hpv   idade   n_parceiros_categ_2   dst_hiv   anticoncept_diu   anticoncept_preserv
i.dst_hiv_anticoncept_preserv

***Modelo1***
tab n_parceiros_categ_2, miss
tab dst_hiv, miss
tab anticoncept_diu, miss
tab anticoncept_preserv, miss

xi: logistic hpv idade i.n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv
xi: logit hpv idade i.n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv
```

******Avaliação do Modelo********

```
estat gof, table group(10)
estat classification
lsens

lstat, cutoff(.6)
lstat, cutoff(.55)
lstat, cutoff(.50)
lstat, cutoff(.45)
lstat, cutoff(.40)
lstat, cutoff(.35)
lstat, cutoff(.30)
lstat, cutoff(.25)
lstat, cutoff(.20)
lstat, cutoff(.15)
lstat, cutoff(.10)
lstat, cutoff(.05)

lsens
lsens, graphregion(fcolor(white) ifcolor(white)) plotregion(fcolor(white) ifcolor(white))

lroc
lroc, graphregion(fcolor(white) ifcolor(white)) plotregion(fcolor(white) ifcolor(white))

*IC
predict p
roctab hpv p

*** criar variável com padrão de covariadas identicos***
predict n, n

***Criar uma variável com números sequenciais***
gen id=_n

***Probabilidades previstas***
predict p

***Resíduos de Pearson Standartizados***
predict stdres, rstand

***Resíduos Deviance***
predict dv, dev

*** Leverage***
predict hat, hat

predict h, h

***residuos Pearson standartizados vs valores previstos***
twoway (scatter stdres p), graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter stdres p, mlabel(n) graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))

***residuos Pearson standartizados vs observações***
twoway (scatter stdres id), graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter stdres id, mlabel(n) graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))

***residuos Deviance vs valores previstos***
twoway (scatter dv p), graphregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter dv p, mlabel(n) graphregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))

***residuos Deviance vs observações***
twoway (scatter dv id), graphregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter dv id, mlabel(n) graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))

***residuos Leverage vs valores previstos***
twoway (scatter hat p), graphregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
```

```
scatter hat p, mlabel(n) graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))


predict dx, dx2
twoway (scatter dx p), graphregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter dx p, mlabel(n) graphregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
gsort -dx
list id dx hpv idade  n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv in 1/20


predict dd, dd
twoway (scatter dd p), graphregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter dd p, mlabel(n) graphregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
gsort -dd
list id dd hpv idade  n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv in 1/20


predict db, db
twoway (scatter db p), graphregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter db p, mlabel(n) graphregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
gsort -db
list id db hpv idade  n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv in 1/20


graph twoway scatter dx p [weight=db], xlabel(0(.2)1) ylabel(0 15 30)
msymbol(oh)graphregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))

format n hpv idade n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv %2.0f
sort n
list n hpv idade n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv p  db dx dd h
if (n==51|n==76|n==115|n==88)
list n  p  db dx dd h if (n==51|n==76|n==115|n==88)
clist n  p  db dx dd h if (n==51|n==76|n==115|n==88)


xi: logit hpv idade i.n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv
lfit

logit hpv idade  n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv if n~=51,
nolog
lfit

logit hpv idade  n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv if n~=76,
nolog
lfit

logit hpv idade  n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv if n~=115,
nolog
lfit

logit hpv idade  n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv if n~=88,
nolog
lfit

logit hpv idade  n_parceiros_categ_2 dst_hiv anticoncept_diu anticoncept_preserv if
(n~=51&n~=76&n~=115&n~=88), nolog
lfit


********Regressão Multinomial************
****************************************
***Análise Univariada***

mlogit lesao_i  hpv, rrr
mlogit lesao_i hpv_multiplos, rrr
xi:mlogit lesao_i i.n_hpv_cat, rrr
mlogit lesao_i idade, rrr
mlogit lesao_i hpv_alto_risco, rrr
xi: mlogit lesao_i i.n_hpv_alto_risco_cat , rrr
```

```
mlogit lesao_i hpv_prov_alto_risco, rrr
mlogit lesao_i hpv_risco_indet, rrr
mlogit lesao_i hpv_baixo_risco, rrr
xi: mlogit lesao_i i.n_parceiros_categ_2, rrr
xi: mlogit lesao_i i.n_hpv_cat,rrr
mlogit lesao_i activ_sexual, rrr
mlogit lesao_i dst, rrr
mlogit lesao_i dst_sífilis, rrr
mlogit lesao_i dst_hiv, rrr
mlogit lesao_i dst_clamídea, rrr
mlogit lesao_i dst_outra, rrr


***Análise Multivariada***
xi:   mlogit   lesao_i   idade   i.n_hpv_alto_risco_cat   hpv_prov_alto_risco   hpv_risco_indet
hpv_baixo_risco i.n_parceiros_categ_2, rrr
lrtest, saving (0)
xi: mlogit lesao_i i.n_hpv_alto_risco_cat hpv_prov_alto_risco hpv_risco_indet hpv_baixo_risco
i.n_parceiros_categ_2, rrr
lrtest
xi:    mlogit    lesao_i    idade    hpv_prov_alto_risco    hpv_risco_indet    hpv_baixo_risco
i.n_parceiros_categ_2, rrr
lrtest
xi:    mlogit    lesao_i    idade    i.n_hpv_alto_risco_cat    hpv_risco_indet    hpv_baixo_risco
i.n_parceiros_categ_2, rrr
lrtest
xi:   mlogit   lesao_i   idade   i.n_hpv_alto_risco_cat   hpv_prov_alto_risco   hpv_baixo_risco
i.n_parceiros_categ_2, rrr
lrtest
xi:   mlogit   lesao_i   idade   i.n_hpv_alto_risco_cat   hpv_prov_alto_risco   hpv_risco_indet
hpv_baixo_risco i.n_parceiros_categ_2, rrr
lrtest

xi:   mlogit   lesao_i   idade   i.n_hpv_alto_risco_cat   hpv_prov_alto_risco   hpv_risco_indet
i.n_parceiros_categ_2, rrr
xi: mlogit lesao_i idade i.n_hpv_alto_risco_cat hpv_risco_indet i.n_parceiros_categ_2, rrr
xi: mlogit lesao_i idade i.n_hpv_alto_risco_cat hpv_risco_indet, rrr
xi: mlogit lesao_i idade i.n_hpv_alto_risco_cat, rrr
lrtest, saving (1)
xi: mlogit lesao_i i.n_hpv_alto_risco_cat, rrr
xi: mlogit lesao_i idade, rrr
lrtest

xi: mlogit lesao_i idade hpv_alto_risco, rrr
xi: mlogit lesao_i hpv_alto_risco, rrr
xi: mlogit lesao_i idade , rrr

xi: mlogit lesao_i idade hpv_alto_risco hpv_prov_alto_risco, rrr
xi: mlogit lesao_i idade hpv_alto_risco hpv_baixo_risco, rrr
xi: mlogit lesao_i idade hpv_alto_risco hpv_risco_indet , rrr


gen idade_n_hpv_alto_risco_cat=idade*n_hpv_alto_risco_cat
gen  idade_hpv_alto_risco=idade*hpv_alto_risco
gen  idade_hpv_prov_alto_risco=idade*hpv_prov_alto_risco
gen hpv_prov_alto_alto_risco=hpv_prov_alto_risco*hpv_alto_risco

xi: mlogit lesao_i idade hpv_alto_risco idade_hpv_alto_risco , rrr
lrtest, using (final)


**************MODELO**************************************
xi: mlogit lesao_i idade hpv_alto_risco, rrr
lrtest, saving(final)
**********************************************************


***Pressuposto da Linearidade do Logit***

***1. Método fráfico - Univariable smoothes scatterplot***
lowess hpv idade if lesao_i!=1, gen(aux1) logit nodraw
lowess hpv idade if lesao_i!=2, gen(aux2) logit nodraw
graph twoway line aux2 idade, sort xlabel(20(10)80)
graph twoway line aux1 idade, sort xlabel(20(10)80)
gen idade2=idade*idade

xi: logit lesao_i idade hpv_alto_risco idade
```

```
lrtest, saving (3)
xi: logit lesao_i idade hpv_alto_risco idade2
lrtest, using (3)
xi: logit lesao_i idade hpv_alto_risco idade2 if lesao_i!=1

***2. Método gráfico - quartis***

**1º logit
sort idade
tabstat idade if lesao_i!=2, statistics( min p25 p50 p75 max) columns(variables)
centile idade, centile (25 50 75)
xtile quart = idade, nq(4)
xi: logit lesao_i i.quart hpv_alto_risco if lesao_i!=2

di (16+31)/2
di (31+41)/2
di (41+50)/2
di (50+79)/2

input idade_quartil coef1 coef2
23.5 0 0
36  .8199542 -0.9437977
45.5 0.9335787 -1.03249
64.5 1.161444 -17.3194
end
twoway (connected coef1 coef2 idade_quartil ), caption(, color(white)) note(, color(white))
scheme(s2color)   graphregion(fcolor(white)   lcolor(white)   ifcolor(white)   ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
       twoway (connected coef2 idade_quartil), caption(, color(white)) note(, color(white))
scheme(s2color)   graphregion(fcolor(white)   lcolor(white)   ifcolor(white)   ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))


**2º logit
sort idade
tabstat idade if lesao_i!=1, statistics( min p25 p50 p75 max) columns(variables)
centile idade, centile (25 50 75)
xtile quart2 = idade, nq(4)
xi: logit lesao_i i.quart2 hpv_alto_risco if lesao_i!=1

di (16+32)/2
di (32+41)/2
di (41+51)/2
di (51+79)/2

input idade_quartil2 coef1 coef2
24 0 0
36.5  .8199542 0.9335787
45.5 0.9335787 1.161444
64.5 1.161444 -4.192742
end
twoway (connected coef1 coef2 idade_quartil2 ), caption(, color(white)) note(, color(white))
scheme(s2color)   graphregion(fcolor(white)   lcolor(white)   ifcolor(white)   ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
twoway  (connected  coef2  idade_quartil2),  caption(,  color(white))  note(,  color(white))
scheme(s2color)   graphregion(fcolor(white)   lcolor(white)   ifcolor(white)   ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))

***3. Método Analítico - Polinómios Fraccionários***
tab. 6.5 e 6.6
fracpoly logit lesao_i idade hpv_alto_risco if lesao_i~=1, degree(2) compare
fracpoly logit lesao_i idade hpv_alto_risco if lesao_i~=2, degree(2) compare




*** Bondade de Ajustamento ***

***1º logit***
logit lesao_i idade hpv_alto_risco if lesao_i!=2
lfit
estat gof, group (10) table
estat gof
lfit

generate lesao_iaux1=0
replace lesao_iaux1=1 if lesao_i==1
```

```
replace lesao_iaux1=. if lesao_i==2
generate lesao_iaux2=0
replace lesao_iaux2=1 if lesao_i==2
replace lesao_iaux2=. if lesao_i==1

quietly logit lesao_iaux1 idade hpv_alto_risco
predict p1
generate g1=ln(p1/(1-p1))
generate z11=0.5*g1^2
replace z11=0 if p1<0.5
generate z12=-0.5*g1^2
replace z12=0 if p1>=0.5
quietly logit lesao_iaux1 idade hpv_alto_risco
estimates store reduced
quietly logit lesao_iaux1 idade hpv_alto_risco z11 z12
estimates store full
lrtest reduced full



***2° logit***
logit lesao_i idade hpv_alto_risco if lesao_i!=1
estat gof, group (10) table
estat gof
lfit

quietly logit lesao_iaux2 idade hpv_alto_risco
predict p2
generate g2=ln(p2/(1-p2))
generate z21=0.5*g2^2
replace z21=0 if p2<0.5
generate z22=-0.5*g2^2
replace z22=0 if p2>=0.5
quietly logit lesao_iaux2 idade hpv_alto_risco
estimates store reduced
quietly logit lesao_iaux2 idade hpv_alto_risco z21 z22
estimates store full
lrtest reduced full

drop  _est_full _est_reduced z22 z21 g2 p2 z12 z11 g1 p1


***Diagnóstico do Modelo***
**************************
gen id=_n

***1° logit***
logit lesao_i idade hpv_alto_risco if lesao_i!=2

predict p1
predict dx1, dx2
predict db1, db
predict dd1, dd
predict h1, h
predict n1, n
predict stdres1, rstand
predict dv1, dev

list n1 id lesao_i idade hpv_alto_risco p1 dx1 db1 dd1 h1 stdres1 dv1 if (dv1>2 | dv1<-2) &
lesao_i!=2


*residuos Pearson standartizados vs valores previstos*
twoway   (scatter   stdres1   p1),   graphregion(fcolor(white)   lcolor(white)   ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter   stdres1   p1,   mlabel(n1)   graphregion(fcolor(white)   lcolor(white)   ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
gsort  stdres1
list n1 id stdres1 lesao_i idade hpv_alto_risco if (stdres1>2 | stdres1<-2) & lesao_i!=2

*residuos Pearson standartizados vs observações*
twoway   (scatter   stdres1   id),   graphregion(fcolor(white)   lcolor(white)   ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter   stdres1   id,   mlabel()   graphregion(fcolor(white)   lcolor(white)   ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))

*residuos Deviance vs valores previstos*
```

```
twoway (scatter dv1 p1), graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter dv1 p1, mlabel(n1) graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))

*residuos Deviance vs observações*
twoway (scatter dv1 id), graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter dv id, mlabel(id) graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
list n1 id dv1 lesao_i idade hpv_alto_risco if (dv1>2 | dv1<-2) & lesao_i!=2

***residuos Leverage vs valores previstos***
twoway (scatter h1 p1), graphregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter h1 p1, mlabel(n1) graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
list n1 id h1 lesao_i idade hpv_alto_risco if (h1>0.1 | h1<-0.1) & lesao_i!=2

format n1 idade hpv_alto_risco %2.0f
clist n1 lesao_i idade hpv_alto_risco p1 h1 dv1 db1 dx1 dd1 h1 if (n1==4 | n1==50)
&lesao_i~=2

***ddx vs valores previstos***
twoway (scatter dx1 p1), graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter dx1 p1, mlabel(n1) graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
gsort dx1
list n1 dx1 in 1/3
list dx1 n1 if n1==4 | n1==50

***dd vs valores previstos***
twoway (scatter dd1 p1), graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter dd1 p1, mlabel(n1) graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
list dd1 n1 if n1==4 | n1==50

***Pregibon's dbeta vs valores previstos***
twoway (scatter db1 p1), graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter db1 p1, mlabel(n1) graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))


graph twoway scatter dx1 p1 [weight=db1], xlabel(0(.2)1) ylabel(0 15 30)
msymbol(oh)graphregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))

mlogit lesao_i idade hpv_alto_risco, rrr
mlogit lesao_i idade hpv_alto_risco if n1~=4, rrr
mlogit lesao_i idade hpv_alto_risco if n1~=50, rrr
mlogit lesao_i idade hpv_alto_risco if (n1~=50 & n1~=4), rrr

logit lesao_i idade hpv_alto_risco if (n1~=50 | n1~=4  &  lesao_i!=2)
lfit


***2º logit***
logit lesao_i idade hpv_alto_risco if lesao_i!=1

predict p2
predict dx2, dx2
predict db2, db
predict dd2, dd
predict h2, h
predict n2, n
predict stdres2, rstand
predict dv2, dev


*residuos Pearson standartizados vs valores previstos*
twoway (scatter stdres2 p2), graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter stdres2 p2, mlabel(n2) graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
```

```
gsort - stdres2
list stdres2 n2 in 1/20
list n2 id stdres2 lesao_i idade hpv_alto_risco if (stdres2>2 | stdres2<-2) & lesao_i!=1

*residuos Pearson standartizados vs observações*
twoway (scatter stdres2 id), graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter stdres2 id, mlabel(id) graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))

*residuos Deviance vs valores previstos*
twoway (scatter dv2 p2), graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter dv2 p2, mlabel(n2) graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
gsort - dv2
list dv2 n2 in 1/20

*residuos Deviance vs observações*
twoway (scatter dv1 id), graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter dv id, mlabel(id) graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
list n2 id dv2 lesao_i idade hpv_alto_risco if (dv2>2 | dv2<-2) & lesao_i!=1

***residuos Leverage vs valores previstos***
twoway (scatter h2 p2), graphregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
scatter h2 p2, mlabel(id) graphregion(fcolor(white) lcolor(white) ifcolor(white)
ilcolor(white)) plotregion(fcolor(white) lcolor(white) ifcolor(white) ilcolor(white))
list n2 id h2 lesao_i idade hpv_alto_risco if (h2>0.1 | h2<-0.1) & lesao_i!=1

format n2 idade hpv_alto_risco %2.0f
list n2 id lesao_i idade n_hpv_alto_risco p2  db2 dx2 dd2 h2 if (n2==42|n2==67|n2==77|n2==28)
& lesao_i!=1


mlogit lesao_i idade hpv_alto_risco, rrr
mlogit lesao_i idade hpv_alto_risco if n1~=42, rrr
mlogit lesao_i idade hpv_alto_risco if n1~=67, rrr
mlogit lesao_i idade hpv_alto_risco if n1~=77, rrr
mlogit lesao_i idade hpv_alto_risco if n1~=28, rrr
mlogit lesao_i idade hpv_alto_risco if (n1~=42 & n1~=67 & n1~=77 & n1~=28), rrr

logit lesao_i idade hpv_alto_risco if (lesao_i!=1)
logit lesao_i idade hpv_alto_risco if ((n1~=42 & n1~=67 & n1~=77 &n1~=28 ) & lesao_i!=1)
lfit


**********Regressão Ordinal**********
****************************************

***construção do modelo***
ologit lesao_i  hpv, or
ologit lesao_i hpv_multiplos, or
xi:ologit lesao_i i.n_hpv_cat, or
ologit lesao_i idade, or
ologit lesao_i hpv_alto_risco, or
xi: ologit lesao_i i.n_hpv_alto_risco_cat , or
ologit lesao_i hpv_prov_alto_risco, or
ologit lesao_i hpv_risco_indet, or
ologit lesao_i hpv_baixo_risco, or
xi: ologit lesao_i i.n_parceiros_categ_2, or
xi: ologit lesao_i i.n_hpv_cat,or
ologit lesao_i activ_sexual, or
ologit lesao_i dst, or
ologit lesao_i dst_sífilis, or
ologit lesao_i dst_hiv, or
ologit lesao_i dst_clamídea, or
ologit lesao_i dst_outra, or


***remoção de variáveis***
xi:  ologit  lesao_i  i.n_hpv_cat  idade  i.n_hpv_alto_risco_cat  hpv_prov_alto_risco
hpv_risco_indet hpv_baixo_risco i.n_parceiros_categ_2 i.n_hpv_cat dst_hiv, or
xi:  ologit  lesao_i  idade  i.n_hpv_alto_risco_cat  hpv_prov_alto_risco  hpv_risco_indet
hpv_baixo_risco i.n_parceiros_categ_2 i.n_hpv_cat dst_hiv, or
```

```
xi:  ologit  lesao_i  idade  i.n_hpv_alto_risco_cat  hpv_prov_alto_risco  hpv_risco_indet
hpv_baixo_risco i.n_parceiros_categ_2 dst_hiv, or
xi:  ologit  lesao_i  idade  i.n_hpv_alto_risco_cat  hpv_prov_alto_risco  hpv_risco_indet
i.n_parceiros_categ_2 dst_hiv, or
xi: ologit lesao_i idade i.n_hpv_alto_risco_cat hpv_risco_indet i.n_parceiros_categ_2 dst_hiv,
or
xi: ologit lesao_i idade i.n_hpv_alto_risco_cat hpv_risco_indet i.n_parceiros_categ_2, or
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet i.n_parceiros_categ_2, or
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, or



xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet i.n_parceiros_categ_2 dst_hiv, or

gen par_alto=n_parceiros_categ_2*n_hpv_alto_risco_cat
gen par_indet=n_parceiros_categ_2*hpv_risco_indet
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet i.n_parceiros_categ_2 par_alto, or
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet i.n_parceiros_categ_2 par_indet, or

gen hiv_alto=dst_hiv*n_hpv_alto_risco_cat
gen hiv_indet=dst_hiv*n_parceiros_categ_2
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet dst_hiv hiv_alto, or
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet dst_hiv n_parceiros_categ_2, or

gen par_hiv=n_parceiros_categ_2*dst_hiv
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet n_parceiros_categ_2 dst_hiv par_hiv,
or


***inclusaão de variáveis***
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet i.n_hpv_cat, or
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet hpv_baixo_risco, or
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet hpv_prov_alto_risco, or
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet dst_hiv, or
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet idade, or
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet i.n_parceiros_categ_2, or
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, or
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet hpv_multiplos, or


***pesquisa de interações***
gen n_hpv_alto_risco_cat_risc_ind=n_hpv_alto_risco_cat* hpv_risco_indet
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet i.n_hpv_alto_risco_cat_risc_ind, or

gen idade_alto=n_hpv_alto_risco_cat*idade
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet idade_alto idade, or



*********************************MODELO*****************************************
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, or
************************************************************************

***validação do pressuposto de riscos proporcionais
findit omodel
xi:omodel logit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet
brant, detail


xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, or
scalar m1 = e(ll)
xi: mlogit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, rrr
scalar m2 = e(ll)
di "chi2(3) = " 2*(m2-m1)
di "Prob > chi2 = "chi2tail(3, 2*(m2-m1))

xi: mlogit lesao_i idade hpv_alto_risco, nolog
quietly fitstat, saving(mod1)
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, nolog
fitstat, using(mod1)force

***Comparação entre modelo ordinal e multinomial***
xi: mlogit lesao_i idade hpv_alto_risco, rrr
quietly fitstat, saving(mult)
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, or
fitstat, using(mult)force
```

```
***Modelo ordinal generalizado***
findit gologit2

xi:gologit2  lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet
xi: gologit2 lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, or

***Comparação entre modelo ordinal e ordinal generalizado***
xi: ologit lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, or
lrtest, saving (o)
xi: gologit2 lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, or
lrtest, using (o) force

***Modelo ordinal generalizado de riscos parciais

xi: gologit2 lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, autofit
xi: gologit2 lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, autofit lrforce
xi: gologit2 lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, or autofit lrforce

***Modelo de riscos proporcionais irrestrito
xi: gologit2 lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, auto gamma lrforce
xi: gologit2 lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, auto gamma
xi: gologit2 lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, or auto gamma


***Comparação entre modelo generalizado e o modelo de riscos parciais***
xi: gologit2 lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, or
lrtest, saving (g)
xi: gologit2 lesao_i i.n_hpv_alto_risco_cat hpv_risco_indet, autofit lrforce
lrtest, using (g) force
```