UNIVERSIDADE DA BEIRA INTERIOR

Introdução à Regressão Categorial
Aplicação a dados de escolas angolanas, com utilização do SPSS

Versão Final Após Defesa

Evaristo José das Mangas

Dissertação para obtenção do Grau de Mestre em
Matemática para Professores
(2º ciclo de estudos)

Orientador: Prof. Doutor Jorge Manuel dos Reis Gama

Covilhã, junho de 2019

Dedicatória

Dedico este trabalho à rima da minha vida, à Joaquina, Josina e à Helena. Esposa, Filha e Mãe, respetivamente!

Agradecimentos

Muitas outras pessoas tiveram importância relevante em termos dos estímulos, desafios e críticas que me proporcionaram ao longo dos anos para frequentar este Mestrado. Agradeço sobretudo ao Professor Eugénio Manuel, ao Professor Carlos Pinto, à Professora Cláudia Pinto, ao Luaty Beirão, ao Professor Paulo Kassavela, ao Professor Vladi Pereira, ao Professor Bernardo Filipe Matias, à Liliana Brás, ao Higino Nazaré Intya, ao Mateus Santiago, ao José Francisco das Mangas, ao Adelino Natalício Kumena, à Fernanda Maria José da Rocha, ao Augusto Mayambi Tyilunda, ao Manuel Bule das Mangas, à Sónia Robalo e ao Júlio Padez.
Ao Governo Provincial da Huila, muito em especial, à Professora Maria João Chipalavela!
Ao Gabinete Provincial da Educação da Huila. À Direcção do Colégio Nº 852 "11 de Novembro"-Lubango, ao Dr. Benjamim António, à Dra Catarina Romão e à todos os colegas.
Aos meus Padrinhos, Augusto Moura Rasga e Isabel Carla Rasga!
Cada um dos meus amigos ocupa uma linha de centro nas páginas da minha vida e não um rodapé, por isso, os meus agradecimentos são extensíveis aos amigos: Abel Zacarias, Américo Rocha, Dinis Amaro, Mateus Katucha Akuaki, Nicodemos Sipata Dumbo, Vadilson Jacob, Jurandir Jacob, Nhenze Abias, Yohelns Bahu, Tomás Hambili, Antero Neves, Luís Máximo, Allison Rodrigues, Juliana Nunes, Manuel Teixeira, Ngaiele Fundão e João Lazaro!
Aos meus colegas, Elias Chiwaia, Emílio Evaristo e Fonseca André!
À dona Branca Regina Fiadeiro (in memorian), ao Sérgio Fiadeiro Guerra Carneiro e à Dora Nogueira!
Ao Professor Jorge Gama, que muito pacientemente soube orientar-me e potenciar-me a cada dia.
Aos Professores Pedro Morais, Henrique Cruz, Ilda Inácio, Rogério Serôdio, Nuno Correia, César Silva, Fernando Pereira, Célia Nunes, Hélder Vilarinho, Rui Pacheco e à todos os Professores do Departamento de Matemática da UBI.
Ao Colégio Reitoral da UBI pela oportunidade concedida ao permitirem a abertura do curso quando era o único aluno do 1.º ano do curso. Um Agradecimento especial aos Amigos, António Carreto Fidalgo, João Manuel Messias Canavilhas, Ana Silva e toda equipa do GISP e aos amigos(as) dos serviços académicos.
Ao Professor Sílvio Filipe Velosa da Universidade da Madeira pelo incentivo, ao oferecer-me o livro de Estatística doou-me parte de si, muito obrigado pela amizade e consideração.
Aos meus amigos da portaria da Biblioteca Central da UBI, o João Baptista, Marco Figueiredo e ao Victor Santos. Às minhas amigas da sexta fase, em particular, a Joana Raquel Farias e a Alice Gomes pela amizade e pelo carinho.
À Cruz Vermelha da Covilhã e ao Refood da Covilhã!
A todos os covilhanenses que nunca deixaram que me sentisse só diante da multidão! A UBI mudou a minha vida!
Somos todos UBI!
Da UBI para o Mundo!
Evaristo José das Mangas

Resumo

A regressão categorial constitui uma forma linearizável de regressão não linear, por meio de uma ou mais variáveis explicativas, permitindo predizer a probabilidade de ocorrência das várias categorias da variável dependente. O objetivo deste trabalho consistiu em apresentar uma introdução aos diferentes tipos de regressão categorial, como a regressão logística, a regressão multinomial e a regressão ordinal. Abordamos os diferentes modelos de regressão categorial com a função de ligação *logit*, permitindo esta a linearização de cada modelo, onde aparece, naturalmente, a medida de associação epidemiológica *odds ratio*, sendo esta medida interpretada para variáveis categóricas e quantitativas. Realçamos alguns pormenores fundamentais da estimação dos coeficientes de regressão do *logit*, dos pressupostos e avaliação da qualidade de cada um dos modelos: logístico, multinomial e ordinal. Cada um destes modelos foram aplicados a dados colhidos em escolas secundárias angolanas, com utilização do SPSS. Os modelos estimados evidenciaram um bom ajuste e poderão contribuir para um melhor entendimento do processo ensino-aprendizagem da disciplina de Matemática em Angola. No entanto, estes modelos de regressão categorial ajustados carecem de validação.

Palavras-chave

Regressão logística, multinomial, ordinal, odds ratio, ROC, verosimilhança, Wald, *score*, SPSS.

Abstract

Categorical regression is a linearized form of non-linear regression, using one or more explanatory variables, allowing prediction of the probability of occurrence of the various categories of the dependent variable. The objective of this work is to present an introduction to the different types of categorial regression, such as logistic, multinomial and ordinal regression. We approach the different models of categorial regression with the logit link function, allowing the linearization of each model, where the odds ratio epidemiological association measure appears naturally, being this measure interpreted for categorical and quantitative variables. We highlight some fundamental details of the estimation of logit regression coefficients, assumptions and quality evaluation of each of the regression models: logistic, multinomial and ordinal. Each of these regressions was applied to data collected in Angolan secondary schools, using SPSS. The estimated models showed a good fit and could contribute to a better understanding of the teaching-learning process of Mathematics in Angola. However, these adjusted categorical regression models need to be validated.

Keywords

Logistic regression, multinomial, ordinal, odds ratio, ROC, Likelihood, Wald, score, SPSS.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

IC	Intervalo de confiança
LR	Likelihood Ratio
OR	Odds ratio
p-value, p	Valor de prova
ROC	Receiver Operating Characteristic
SPSS	Statistical Package for the Social Sciences
AIC	Akaike's Information Criteria
BIC	Bayesian Information Criterion
IBM	International Business Machines Corporation

Capítulo 1

Introdução

Os modelos de regressão, linear ou não linear, são ferramentas estatísticas indispensáveis na análise de dados para descrever relações entre variáveis resposta (dependente) e explicativas (independentes), com uma enorme aplicabilidade e de grande relevância em todas as áreas de investigação, como, por exemplo, nas ciências sociais e do comportamento, nas ciências da vida, economia, pesquisas de mercado, engenharias, etc.

A regressão categorial é constituída, em geral, por modelos de regressão não linear, mas linearizáveis, adequada em cenários em que a variável dependente é qualitativa e assume apenas categorias mutuamente exclusivas. Por esta razão, a estimação destas categorias não é efetuada diretamente, como, por exemplo, no caso da regressão linear, mas são estimadas recorrendo-se ao cálculo da sua probabilidade de ocorrência.

No caso particular da regressão categorial em que a variável dependente é dicotómica, policotómica ou ordinal, existem vários modelos, que utiliza uma função, conhecida por função de ligação, que permite linearizar um modelo de probabilidades para a estimação da ocorrência das duas ou mais categorias. Entre essas funções de ligação destacamos as funções *logit*, *probit*, *log-log* negativa e complementar *log-log*. Neste trabalho iremos abordar principalmente a aplicação da função de ligação *logit*, quando a variável dependente é dicotómica, policotómica ou ordinal.

Deste modo, o presente trabalho, que se enquadra no âmbito da obtenção do grau de Mestre em Matemática para Professores, teve como objetivo introduzirmos os conceitos básicos da regressão logística, da regressão multinomial e da regressão logística ordinal e fazer a sua aplicação a dados observados em três escolas angolanas com utilização do programa estatístico IBM SPSS (Statistical Package for the Social Sciences).

Assim, no segundo capítulo deduziremos o modelo de regressão logística, abordaremos o método de estimação dos coeficientes de regressão, a sua interpretação e os pressupostos do modelo: análise de resíduos, diagnósticos de outliers e de observações influentes e o poder de discriminação do modelo com uma análise ROC (Receiver Operating Characteristic). E, finalmente, aplicaremos o modelo de regressão logística, implementado no programa estatístico IBM SPSS na sua versão 25, a dados obtidos por aplicação de um inquérito a uma amostra de alunos do ensino secundário de três escolas da cidade de Luena, município do Moxico, província do Moxico em Angola, durante o mês de fevereiro de 2017, com a prévia autorização do Gabinete Provincial de Educação, Ciência e Tecnologia do Moxico. Os dados que resultaram da aplicação desse inquérito foram-nos gentilmente disponibilizados pelo Mestre em Matemática para Professores da UBI, Ngaiele Muecheno Fundão, que os utilizou na sua dissertação de mestrado intitulada “Modelo de Regressão Linear: Fatores que influenciam o aproveitamento escolar dos alunos Angolanos”.

No terceiro capítulo apresentaremos os conceitos relacionados com a regressão multinomial

como, por exemplo, a linearização deste modelo, a estimação e a interpretação dos coeficientes, testes do rácio de verosimilhanças e o teste de Wald, bem como a sua aplicação aos mesmos dados provenientes das três escolas angolanas.

Finalmente, no quarto capítulo apresentaremos os conceitos relacionados com o modelo de regressão logística ordinal, que permite aproveitar a ordinalidade quando a variável dependente é ordinal, ao contrário do modelo multinomial. Neste trabalho iremos abordar, exclusivamente, o modelo ordinal assumindo-se a homogeneidade dos declives e este será aplicado aos mesmos dados provenientes das três escolas angolanas.

Capítulo 2

O Modelo de Regressão Logística

2.1 Introdução

Considere-se uma amostra de dimensão n obtida numa dada população. Nesta população, seja Y uma variável aleatória com distribuição de Bernoulli de parâmetro π, com $0 < \pi < 1$, sendo esta probabilidade, $\pi = P(Y = 1)$, conhecida por probabilidade do acontecimento "sucesso" e $1 - \pi = P(Y = 0)$ a probabilidade do acontecimento "insucesso". Considere-se, em relação a Y, uma amostra aleatória $(Y_1, Y_2, \cdots, Y_n)$.

Para a probabilidade de ocorrência de $\{Y_i = 1\}$, $\pi_i = P(Y_i = 1)$, $i = 1, \ldots, n$, estamos interessados em analisar como variam estas probabilidades do acontecimento sucesso em função de uma amostra aleatória $(X_1, X_2, \cdots, X_n)$ de uma variável aleatória X, observada na mesma população, que é conhecida por variável independente ou covariável, isto é, as probabilidades condicionadas:

$$\pi_i = \pi(X_i) = P(Y_i = 1|X_i), i = 1, \cdots, n,$$

onde $P(Y_i = 1|X_i)$ é uma notação abreviada de $P(Y_i = 1|X_i = x_i)$, que usaremos muitas vezes ao longo do texto. Devido ao facto de a variável Y ser dicotómica, não é de esperar que um modelo linear seja adequado para estabelecer a associação com a variável X. Observe-se que a variância de Y_i, $var(Y_i|X_i) = \pi_i(1 - \pi_i)$, depende de X_i e, por esta razão, não é constante.

A figura 2.1 ilustra esta inadequabilidade, onde se considerou, como exemplo, uma base de dados relativa a 799 partos, em que $(y_1, y_2, \cdots, y_{799})$ corresponde a uma amostra observada de Y, considerando-se como "sucesso" o peso "normal" (não baixo) à nascença de um bebé (peso mínimo de 2500 gramas) [Wayne W. Daniel, 2013]. Como covariável, considerou-se o tempo de gravidez, em semanas.

Deste modo, a abordagem à criação de um modelo preditivo para a variável Y terá que ser diferente. Em vez da abordagem "usual" proporcionada pelo modelo linear, irá estimar-se a probabilidade da ocorrência do acontecimento sucesso em função da covariável X.

Para se ilustrar essa abordagem, observe-se a tabela 2.1, relativa às probabilidades estimadas para o peso não baixo à nascença em cada uma das classes do tempo de gravidez, consideradas como exemplo, e a respetiva representação destes dados (figura 2.2).

Uma curva que permite um ajuste adequado, amplamente usada e de fácil interpretação, é a bem conhecida função logística:

$$\pi = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}$$

Tabela 2.1: Probabilidades estimadas do peso normal à nascença em cada classe do tempo de gravidez, em semanas

Classes	x_i	$\#(Y=0)$	$\#(Y=1)$	$\hat{\pi}=\hat{P}(Y=1)$
≤ 24	24	5	0	0
$[25-27]$	26	6	0	0
$[28-30]$	29	4	0	0
$[31-33]$	32	6	5	0,48
$[34-36]$	35	22	44	0,67
$[37-39]$	38	21	392	0,95
$[40-42]$	41	5	256	0,98
≥ 43	43	1	32	0,97

Como veremos mais adiante, esta função é linearizável e permite uma fácil interpretação nas aplicações, evidenciando-se assim o porquê de ser a mais popular entre os métodos de regressão quando a variável dependente é dicotómica.

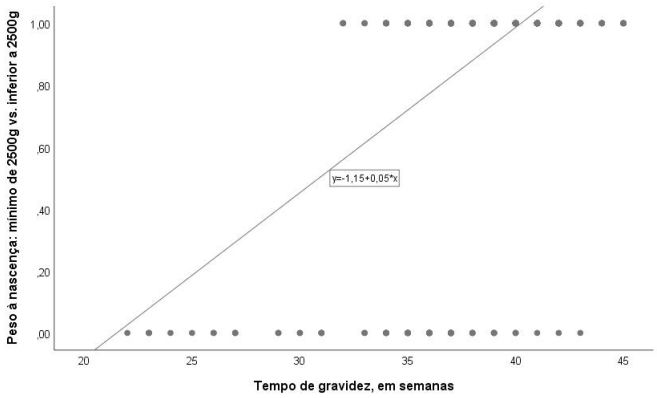

Figura 2.1: Reta de regressão ajustada à variável dependente peso normal à nascença em função do tempo de gravidez.

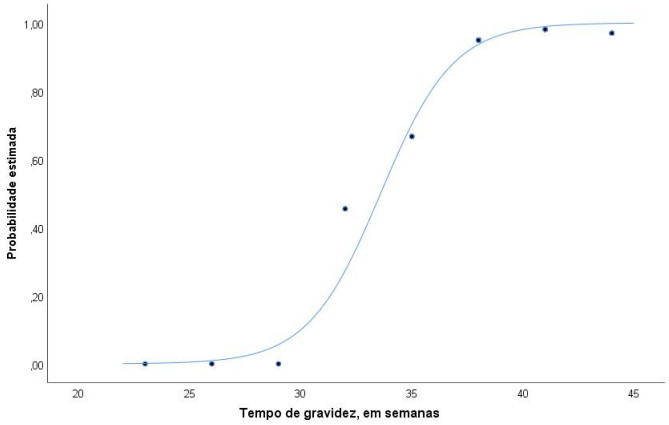

Figura 2.2: Probabilidades estimadas de peso normal à nascença em função do tempo de gravidez (representação dos pontos da tabela 1) e o ajuste com a função $\pi = \frac{1}{1+e^{-x}}$

2.2 *Odds e Linearização da Função Logística*

Defina-se *odds* (chances ou rácio de verosimilhanças) por

$$odds = \frac{\pi}{1-\pi},$$

que traduz a razão entre as probabilidades do sucesso e de insucesso, isto é, a possibilidade de se observar o sucesso, $\{Y = 1\}$, relativamente à possibilidade de se observar o insucesso, $\{Y = 0\}$. Defina-se ainda a função que é o logaritmo neperiano do odds:

$$logit(\pi) = ln\left(\frac{\pi}{1-\pi}\right),$$

conhecido por transformação *logit*. Se assumirmos que esta função é linear:

$$logit(\pi) = \beta_0 + \beta_1 X,$$

se tomarmos uma única variável independente (covariável) X, onde a constante β_0 é a ordenada na origem e β_1 é o coeficiente de regressão linear (declive), neste caso da função *logit*, ou

$$logit(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k,$$

se tomarmos várias variáveis independentes $X_1, X_2, \cdots, X_k$ (k covariáveis), onde, novamente, β_0 é a constante e $\beta_1, \cdots, \beta_k$ são os coeficientes de regressão linear múltipla da função *logit*, facilmente se deduz que esta transformação *logit* lineariza a função logística:

$$\begin{aligned} \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X \quad &\Leftrightarrow \quad \frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 X} \\ &\Leftrightarrow \quad \pi\left(1 + e^{\beta_0+\beta_1 X}\right) = e^{\beta_0+\beta_1 X} \\ &\Leftrightarrow \quad \pi = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \\ &\Leftrightarrow \quad \pi = \frac{1}{1 + e^{-(\beta_0+\beta_1 X)}} \end{aligned} \tag{2.1}$$

ou, de modo similar,

$$\pi = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}} \tag{2.2}$$

A transformação *logit* oferece assim maiores facilidades do ponto de vista matemático, mas também do ponto de vista da interpretação do próprio modelo, como veremos na secção seguinte.

2.3 Interpretação dos coeficientes de regressão do *logit*

É sabido que uma variável independente qualitativa com j categorias, com $j \leq 7$, somente poderá ser utilizada num modelo linear desde que seja recodificada em $j-1$ variáveis dummy (variáveis auxiliares indicadoras) [Maroco, 2018, p. 786].
Suponhamos, por agora, que a variável independente X é qualitativa dicotómica, em que as categorias estão codificadas como 0 e 1, isto é, X é uma variável *dummy*. Facilmente se verifica

que:

$$logit\,(\pi(1)) - logit\,(\pi(0)) = \beta_1,$$

evidenciando que o coeficiente de regressão β_1 tem no *logit* a mesma interpretação que no contexto do modelo linear: para $X = 1$, o *logit* da probabilidade de sucesso aumenta β_1, se $\beta_1 > 0$, ou diminui $|\beta_1|$, se $\beta_1 < 0$. No entanto, esta interpretação é para o *logit* da probabilidade de sucesso e não diretamente para a probabilidade de sucesso, perdendo-se assim a sua fácil interpretação em aplicações.
Deste modo, será necessário uma outra abordagem, que ficará evidente se considerarmos a razão entre os $odds$ quando $X = 1$ e $X = 0$, isto é, a razão das chances do sucesso *versus* insucesso quando $X = 1$ *versus* $X = 0$. Esta razão é a bem conhecida medida de associação epidemiológica *odds ratio*, OR (razão de chances ou razão de possibilidades), com ampla aplicação, por exemplo, nas ciências da vida:

$$OR = \frac{\dfrac{\pi(1)}{1-\pi(1)}}{\dfrac{\pi(0)}{1-\pi(0)}}.$$

Com o modelo logístico deduzido em 2.1, obtém-se de imediato que a razão de chances coincide com a exponencial do coeficiente de regressão do *logit*:

$$\begin{aligned} OR &= \frac{\dfrac{\pi(1)}{1-\pi(1)}}{\dfrac{\pi(0)}{1-\pi(0)}} \\ &= \frac{\left(\dfrac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right) / \left(\dfrac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\dfrac{e^{\beta_0}}{1+e^{\beta_0}}\right) / \left(\dfrac{1}{1+e^{\beta_0}}\right)} \\ &= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} \\ &= e^{\beta_1} \end{aligned} \tag{2.3}$$

Se a variável X tiver $j > 2$ categorias, obter-se-á a razão de chances para cada uma das $j-1$ categorias, recodificadas cada uma como variável *dummy*, em relação à categoria fixada como referência, que toma o valor 0 em todas essas $j-1$ variáveis *dummy*.
Se a variável X é quantitativa, a interpretação da razão de chances é efetuada por unidade dessa variável, isto é, para a variação de uma unidade quando $X = x$, pois, similarmente, tem-se que:

$$\begin{aligned} OR &= \frac{\left(\dfrac{e^{\beta_0+\beta_1(x+1)}}{1+e^{\beta_0+\beta_1(x+1)}}\right) / \left(\dfrac{1}{1+e^{\beta_0+\beta_1(x+1)}}\right)}{\left(\dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\right) / \left(\dfrac{1}{1+e^{\beta_0+\beta_1 x}}\right)} \\ &= \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} \\ &= e^{\beta_1} \end{aligned} \tag{2.4}$$

Consequentemente, a razão das chances do sucesso *versus* insucesso quando a variável inde-

pendente quantitativa X aumenta uma unidade *versus* manter-se constante é a exponencial do respetivo coeficiente de regressão do *logit*.
Convém recordar que a medida de efeito OR é um número real estritamente positivo e:

- se $OR < 1$, significa que as chances do evento de interesse (categoria em numerador) diminui em relação ao grupo que se está a analisar;
- se $OR = 1$, significa que não há associação entre ambas categorias ou que as duas categorias têm as mesmas chances de ocorrer;
- se $OR > 1$, significa que as chances do evento de interesse (da categoria em numerador) aumenta em relação ao grupo que se está a analisar.

2.4 Estimação dos coeficientes de regressão

O método de ajustamento usado na regressão logística é o Método da Máxima Verosimilhança (Maximum Likelihood) para estimar os parâmetros (β_0,β_1). Este método estima os coeficientes de regressão que maximizam a probabilidade de encontrar as realizações da variável dependente $(y_1,y_2,\cdots,y_n)$ amostradas, isto é, que maximizem a verosimilhança desses valores [Maroco, 2018, p. 793-794].
Consequentemente, a função de verossimilhança é da seguinte forma:

$$\begin{aligned} L = L(y_1,y_2,\cdots,y_n) &= P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) \\ &= P(Y_1 = y_1)P(Y_2 = y_2)\cdots P(Y_n = y_n) \\ &= \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{(1-y_i)} \end{aligned} \tag{2.5}$$

E para o modelo logístico 2.2 a ajustar, a função L é dada por:

$$L = L(\beta_1,\beta_2,\cdots,\beta_n) = \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki}}} \right)^{y_i} \left(\frac{1}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki}}} \right)^{(1-y_i)}, \tag{2.6}$$

onde $(x_{j1},\ldots,x_{jn})$ é uma amostra observada da variável independente X_j, com $j = 1,\ldots,k$.
Como 2.6 é linearizável com uma transformação logarítmica, que tomaremos de base natural, então 2.6 é equivalente a

$$\begin{aligned} LL = \ln(L) &= \sum_{i=1}^{n} \big[y_i(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}) \\ &- \ln(1 + e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki}}) \big]. \end{aligned}$$

Como o logaritmo é uma função contínua e estritamente crescente no seu domínio, o vetor $(\beta_0,\beta_1,\ldots,\beta_k)$ que maximiza a função L é o mesmo que maximiza a função LL, mas esta segunda forma permite calcular as derivadas parciais em relação a cada coeficiente β_j, $j = 0,\ldots,k$, mais facilmente:

$$\begin{aligned}\frac{\partial}{\partial\beta_0}LL &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki}}};\\ \frac{\partial}{\partial\beta_j}LL &= \sum_{i=1}^{n} y_i x_{ji} - \sum_{i=1}^{n} x_{ji}\frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki}}},\ j=1,\ldots,k. \qquad (2.7)\end{aligned}$$

O máximo da função LL ocorre quando $\frac{\partial}{\partial\beta_j}LL = 0$ e $\frac{\partial^2}{\partial\beta_j}LL < 0$, $j = 0,1,\ldots,k$, obtendo-se os estimadores $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k\right)$ dos coeficientes $\beta_0,\beta_1,\ldots,\beta_k$, respetivamente. As equações que permitem maximizar LL:

$$\begin{aligned}&\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\hat{\beta}_2 x_{2i}+\cdots+\hat{\beta}_k x_{ki}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\hat{\beta}_2 x_{2i}+\cdots+\hat{\beta}_k x_{ki}}} = 0,\\ &\sum_{i=1}^{n} y_i x_{ji} - \sum_{i=1}^{n} x_{ji}\frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\hat{\beta}_2 x_{2i}+\cdots+\hat{\beta}_k x_{ki}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\hat{\beta}_2 x_{2i}+\cdots+\hat{\beta}_k x_{ki}}} = 0,\ j=1,\ldots,k.\end{aligned}$$

não têm uma solução analítica, implicando a necessidade de se recorrer a métodos numéricos iterativos, como o método de Newton-Raphson, e o modelo de regressão logística fica então da seguinte forma:

$$\hat{\pi} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2+\cdots+\hat{\beta}_k X_k}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2+\cdots+\hat{\beta}_k X_k}} \qquad (2.8)$$

2.5 Estimação da variância de $\hat{\beta}$

Do anteriormente exposto, já sabemos que os coeficientes que maximizam a função de verosimilhança obtêm-se pela diferenciação da função de verosimilhança em relação aos $k+1$ coeficientes e as equações de verosimilhança resultantes podem ser escritas da seguinte forma:

$$\frac{\partial}{\partial\beta_0}LL = 0, \qquad (2.9)$$

$$\frac{\partial}{\partial\beta_j}LL = 0. \qquad (2.10)$$

Sendo $\hat{\beta} = \left(\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k\right)$ o vector solução dessas equações, os estimadores dos coeficientes da regressão logística são funções das amostras, $\hat{\beta}(x_i)$.

Usaremos a notação matricial para melhor explicitar o modelo de regressão logística. Portanto, seja X a matriz de ordem $n \times (k+1)$ que contém as observações para cada variável explicativa e para cada um dos sujeitos, onde a primeira coluna é composta por 1's, que corresponde ao elemento independente (constante):

$$\mathsf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}. \tag{2.11}$$

Seja V a matriz diagonal quadrada de ordem n de elemento geral $\hat{\pi}_i (1 - \hat{\pi}_i)$ que é dada por:

$$\mathsf{V} = \begin{pmatrix} \hat{\pi}_1 (1 - \hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2 (1 - \hat{\pi}_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\pi}_n (1 - \hat{\pi}_n) \end{pmatrix}. \tag{2.12}$$

Sendo X' a matriz transposta de X, o produto destas três matrizes dão-nos a chamada matriz de informação de Fisher, que é dada por:

$$I\left(\hat{\beta}\right) = X'VX.$$

As variâncias-covariância dos estimadores $\hat{\beta}_j$ são obtidas a partir da inversa da matriz $I\left(\hat{\beta}\right)$, ou seja, $Var\left(\hat{\beta}\right) = I^{-1}\left(\hat{\beta}\right)$. Nalguns casos excecionais não é possível expressar de uma forma explícita a matriz dos estimadores dos coeficientes.

Observe-se que a variância de $\hat{\beta}_j$, $Var\left(\hat{\beta}_j\right)$, é o j-ésimo elemento da diagonal principal da matriz $Var\left(\hat{\beta}\right)$ e a $cov\left(\hat{\beta}_j, \hat{\beta}_l\right)$ é o elemento da linha j e coluna l, dessa mesma matriz, ou linha l e coluna j, já que $cov\left(\hat{\beta}_j, \hat{\beta}_l\right) = cov\left(\hat{\beta}_l, \hat{\beta}_j\right)$.

A forma mais fácil de se obter a matriz de informação de Fisher é a partir da sua relação com a matriz Hessiana. A matriz Hessiana é, por definição, a matriz quadrada cujos elementos são as derivadas de segunda ordem da função de verosimilhança, isto é:

$$H = \begin{pmatrix} \dfrac{\partial^2 L}{\partial \beta_1^2} & \dfrac{\partial^2 L}{\partial \beta_1 \partial \beta_2} & \cdots & \dfrac{\partial^2 L}{\partial \beta_1 \partial \beta_k} \\ \dfrac{\partial^2 L}{\partial \beta_2 \partial \beta_1} & \dfrac{\partial^2 L}{\partial \beta_2^2} & \cdots & \dfrac{\partial^2 L}{\partial \beta_2 \partial \beta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 L}{\partial \beta_k \partial \beta_1} & \dfrac{\partial^2 L}{\partial \beta_k \partial \beta_2} & \cdots & \dfrac{\partial^2 L}{\partial \beta_k^2} \end{pmatrix}.$$

Assim, a matriz de informação de Fisher é a oposta da matriz Hessiana, isto é,

$$I(\beta) = -H(\beta),$$

e, consequentemente, a matriz de variâncias-covariâncias vem dada por

$$Var\left(\hat{\beta}\right) = I^{-1}\left(\hat{\beta}\right) = -H^{-1}.$$

O desvio padrão de cada $\hat{\beta}_j$ é estimado por:

$$\widehat{\sigma}\left(\beta_j\right) = \sqrt{Var\left(\hat{\beta}_j\right)}. \tag{2.13}$$

Estes resultados são úteis para a construção de intervalos de confiança (IC) e testes de hipóteses para β_i.

2.6 Testes de significância sobre os coeficientes do *logit*

2.6.1 Estatística *Deviance* e os Pseudo-R^2

Uma função de verosimilhança, L, é inferior a 1 e, em geral, um valor pequeno, pois é um produto de probabilidades, o que implica que a transformação logarítmica, LL, seja um número negativo. É usual multiplicar-se LL por -2 para se obter um número positivo e maior. A estatística *deviance*, D, na regressão logística é definida por $D = 2\left(LL_C - LL_0\right)$, onde LL_C e LL_0 representam o logaritmo da função de verosimilhança dos modelos completo (saturado) e nulo, respetivamente. A *deviance* constitui a quantidade de informação não explicada pelo modelo e quanto menor, melhor será o modelo. Isto é, D é um indicador da mediocridade do ajustamento do modelo aos dados; se $-2LL = 0$ o ajustamento é perfeito. [Maroco, 2018, p. 797].
É por meio da *deviance* que se obtém o teste do rácio de verosimilhança, como veremos mais adiante, que é definido como sendo a diferença entre $-2LL_0$, que é relativa ao modelo nulo ou reduzido, e $-2LL_C$, que é relativa ao modelo saturado ou com todas as variáveis. Os pseudo-R^2, também são expressos em função da *deviance*, como veremos.
Um pseudo-R^2 $\left(0 \leq R^2 \leq 1\right)$, ou "pseudo coeficiente de determinação", tem por objetivo medir a qualidade do modelo de regressão logística ajustado, à semelhança do coeficiente de determinação na regressão linear.
No entanto, os pseudo-R^2 não são de fácil interpretação devido à natureza da variável dependente, que na regressão logística é dicotómica e cuja variância é máxima quando há uma distribuição equitativa de frequências entre as duas categorias. Ou seja, a variância não sendo constante, torna difícil a interpretação dos pseudo-R^2.
Os autores D. R. Cox e E. J. Snell propuseram o seguinte pseudo-R^2 [Cox and Snell, 1989]:

$$R^2_{CS} = 1 - e^{\frac{2\left(LL_C - LL_0\right)}{n}}.$$

Verifica-se que este pseudo-R^2 nunca atinge o valor 1, contrariamente ao que acontece com o coeficiente de determinação na regressão linear, tornando mais subjetiva a sua interpretação. Posteriormente, o autor N. Nagelkerke propôs uma correção ao pseudo-R^2 de Cox & Snell de modo a permitir atingir o valor 1. Este pseudo-R^2 é definido por [Nagelkerke, 1991]:

$$R^2_N = \frac{R^2_{CS}}{1 - e^{\frac{2LL_0}{n}}}.$$

Há autores [Shtatland et al., 2002] que afirmam que o pseudo-R^2 que apresenta uma melhor interpretabilidade é o que foi proposto por D. McFadden, que é definido por [McFadden, 1973]:

$$R^2_{MF} = 1 - \frac{LL_C}{LL_0}.$$

Este pode ser interpretado como a proporção da redução do LL do modelo nulo, relativamente

ao modelo completo. Isto é, o rácio do ganho de informação estimada pelo modelo completo em comparação com o modelo nulo, com o ganho de informação, potencialmente, recuperável por um modelo saturado [Maroco, 2018, p. 802].
Em virtude dos pseudo-R^2 na regressão logística serem, em geral, baseados na comparação do modelo completo com o modelo nulo, será necessário ter um grande cuidado na interpretação desses valores, pois nem sempre conseguem explicar o poder preditivo de um modelo.
Para um estudo mais aprofundado acerca dos pseudo-R^2 aconselha-se o artigo de [Hu et al., 2006].

2.6.2 O Teste do Rácio de Verosimilhanças

O teste do rácio de verosimilhanças visa a significância dos coeficientes de regressão das covariáveis, permitindo determinar se existe pelo menos uma covariável que pode ser incluída no modelo logístico. Assim, o teste do rácio de verosimilhanças permite testar as hipóteses:

$$\begin{aligned} &H_0 : \beta_j = 0,\ j = 1,\ldots,k; \\ &H_1 : \exists_j : \beta_j \neq 0,\ \text{para algum } j = 1,\ldots,k. \end{aligned}$$

Se, para um nível de significância α, não se rejeitar a hipótese H_0, então nenhuma das covariáveis permite predizer as probabilidades de ocorrência de cada uma das categorias da variável dependente. Isto é, nenhuma covariável será explicativa da variável dependente.
O teste do rácio de verosimilhanças consiste em comparar as funções de verosimilhança de dois modelos, um modelo com as covariáveis incluídas (completo) e um modelo simples (nulo ou reduzido), somente com uma constante (ordenada na origem do *logit*). Assim, a estatística de teste é dada por [Agresti and Finlay, 2009, p. 493]

$$G^2 = -2LL_0 - (-2LL_C) = -2\ln\left(\frac{L_0}{L_C}\right), \tag{2.14}$$

onde LL_0 e LL_C representam as transformações logarítmicas das funções de verosimilhança do modelo nulo (L_0) e do modelo com todas as covariáveis incluídas (L_C), respetivamente.
Sob H_0, G^2 tem distribuição assintótica do qui-quadrado com k graus de liberdade, sendo k o número de covariáveis incluídas no modelo. Pois, verifica-se, para o modelo nulo, que

$$-2LL_0 \overset{a}{\sim} {\chi^2}_{n-1}$$

e, para o modelo com as covariáveis incluídas,

$$-2LL_C \overset{a}{\sim} {\chi^2}_{n-1-k}.$$

Deste modo, rejeita-se H_0, se p-value $\approx P(\chi^2_k \geq g) \leq \alpha$, para o nível de significância α especificado (por exemplo, $\alpha = 0{,}05$), sendo g o valor da estatística de teste calculado a partir da amostra disponível.
O teste do rácio de verosimilhanças permite ser adaptado de modo a testar-se individualmente o coeficiente de regressão de cada uma das covariáveis, condicionado pelos restantes coeficientes estarem incluídos no modelo. Isto é, para a covariável X_j, $j = 1,\ldots,k$, permite testar as hipóteses:

$$H_0 : \beta_j = 0 \mid \beta_0, \beta_1; \ldots; \beta_{j-1}; \beta_{j+1}; \ldots; \beta_k;$$
$$H_1 : \beta_j \neq 0 \mid \beta_0, \beta_1; \ldots; \beta_{j-1}; \beta_{j+1}; \ldots; \beta_k.$$

Sob H_0, a estatística G^2 é a diferença entre $-2LL_0$, que inclui todas as covariáveis exceto a covariável em teste (modelo aninhado), e $-2LL_C$, que inclui todas as covariáveis (modelo completo). Novamente, G^2 tem distribuição assintótica do qui-quadrado, mas agora com 1 grau de liberdade. Observe-se que 1 grau de liberdade resulta da diferença $1 = (n-1-(k-1)) - (n-1-k)$, onde $(n-1-(k-1))$ e $(n-1-k)$ são os graus de liberdades da distribuição do qui-quadrado assintótica para $-2LL_0$ e $-2LL_C$, respetivamente.

2.6.3 O Teste de Wald

O teste de Wald consiste em testar a significância dos parâmetros num modelo de regressão de modo a aferir se o coeficiente de uma determinada covariável é ou não explicativa da variável dependente condicionado pelos outros coeficientes das restantes covariáveis.

O teste de Wald é semelhante ao teste do rácio de verosimilhanças, mas agora pretende-se, para a covariável X_j, $j = 1, \ldots, k$, testar as hipóteses:

$$H_0 : \beta_j = 0 \mid \beta_0, \beta_1; \ldots; \beta_{j-1}; \beta_{j+1}; \ldots; \beta_k;$$
$$H_1 : \beta_j \neq 0 \mid \beta_0, \beta_1; \ldots; \beta_{j-1}; \beta_{j+1}; \ldots; \beta_k.$$

A estatística de teste do teste de Wald é, sob H_0, dada por ([Maroco, 2018, p. 800]):

$$T_j = \frac{\hat{\beta}_j}{\widehat{\sigma}\left(\hat{\beta}_j\right)} \tag{2.15}$$

Nessa expressão, $\hat{\beta}_j$ é o estimador de β_j e $\widehat{\sigma}\left(\hat{\beta}_j\right)$ é o estimador do erro padrão de $\hat{\beta}_j$ (ver 2.13). A estatística de Wald tem uma distribuição t de Student, que é assintoticamente normal padrão para amostras de grande dimensão. É usual a estatística de Wald ser dada por:

$$T_j^2 = \left(\frac{\hat{\beta}_j}{\widehat{\sigma}\left(\hat{\beta}_j\right)}\right)^2.$$

Consequentemente, esta estatística tem distribuição assintótica do qui-quadrado com 1 grau de liberdade, para amostras de grande dimensão.

Para amostras de pequena dimensão ou quando o coeficiente de regressão em teste é grande, o teste de Wald é menos potente que o teste do rácio de verosimilhanças, devido ao facto de o erro padrão do respetivo estimador do coeficiente de regressão tender a inflacionar ([Maroco, 2018, p. 801]). Para amostras de grande dimensão os testes de Wald e do rácio de verosimilhanças geralmente fornecem resultados semelhantes. No entanto, o teste de Wald é computacionalmente menos exigente que o teste do rácio de verosimilhanças (este último necessita estimar um outro modelo).

De 2.15 deduz-se facilmente que os extremos de um intervalo de confiança (IC) a $(1-\alpha)\times 100\%$ para β_j são dados por:

$$\hat{\beta}_j \pm t_{1-\frac{\alpha}{2}} \widehat{\sigma}\left(\hat{\beta}_j\right),$$

onde $t_{1-\frac{\alpha}{2}}$ representa o quantil $1-\frac{\alpha}{2}$ da distribuição t de Student com 1 grau de liberdade.

2.6.4 O Teste *Score*

O teste *score* tem como objetivo testar individualmente cada um dos coeficientes de regressão do modelo logístico, isto é, pretende-se, para um coeficiente β_j, $j = 1, \ldots, k$, testar:

$$H_0 : \beta_j = 0;$$
$$H_1 : \beta_j \neq 0.$$

A estatística de teste *score* resulta da derivada parcial do logaritmo da função de verosimilhança, LL, em relação ao coeficiente em teste. Sob H_0, mostra-se que esta estatística é dada por:

$$Z = \frac{U\left(\hat{\beta}_j\right)}{\hat{\sigma}\left(\hat{\beta}_j\right)} \tag{2.16}$$

seguindo uma distribuição assintótica normal padrão [Dobson and Barnett, 2018, p. 81], onde

$$U\left(\hat{\beta}_j\right) = \frac{\partial LL}{\partial \beta_j},$$

$\hat{\sigma}\left(\hat{\beta}_j\right)$ resulta da equação 2.13 e observe-se que o valor esperado de $U\left(\hat{\beta}_j\right)$ é zero. Das equações 2.7 e 2.8 resulta que a estatística 2.16 é também dada por:

$$Z = \frac{\sum_{i=1}^{n} X_i \left(Y_i - \hat{\pi}\right)}{\sqrt{\hat{\pi}(1-\hat{\pi}) \sum_{i=1}^{n} (X_i - \bar{x})^2}}.$$

É vulgar usar-se o quadrado da estatística Z, verificando-se facilmente, neste caso, que tem distribuição assintótica do qui-quadrado com 1 grau de liberdade.
A hipótese H_0 será rejeitada se p-value $= 2P\left(Z \geq |z|\right) \leq \alpha$, para o nível de significância α especificado, onde z representa o valor da estatística de teste *score* obtido a partir da amostra disponível.
É possível generalizar o teste *score* para testar todos os coeficientes de regressão do modelo logístico em simultâneo:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0,$$
$$H_1 : \exists_j : \beta_j \neq 0, \text{ para algum } j = 1 \ldots, k,$$

possibilitando, assim, verificar-se a existência de pelo menos uma covariável significativa no modelo logístico ajustado. Para este teste, tem-se que a matriz das primeiras derivadas parciais

do logaritmo da função de verosimilhança em relação aos parâmetros β_j, $j = 1, \ldots, k$, $U\left(\hat{\beta}\right)$, tem uma distribuição assintótica normal multivariada de média nula e matriz de variâncias-covariâncias $Var(\hat{\beta}) = I^{-1}(\beta)$ ou, equivalentemente,

$$U^T\left(\hat{\beta}\right) I^{-1}\left(\hat{\beta}\right) U\left(\hat{\beta}\right) \overset{a}{\sim} \chi^2_k,$$

para amostras de grandes dimensões.

2.7 Teste ao Ajustamento do Modelo

Os testes de qui-quadrado de Pearson e *deviance* [Hosmer et al., 2013, p. 155] não serão aqui apresentados como medidas de distância ou diferença entre os dados observados e ajustados que visam avaliar a qualidade do modelo, pelo facto de ambos os testes não serem adequados quando existem covariáveis contínuas e por dependerem da dimensão da amostra de modo que as frequências esperadas sejam no mínimo 5, isto é, dependem do Teorema do Limite Central. Um teste que fornece uma indicação do quão bem o modelo se ajusta aos dados é o teste de Hosmer e Lemeshow que se apresenta a seguir.

2.7.1 O Teste de Hosmer e Lemeshow

O teste de Hosmer e Lemeshow é uma variante do teste de ajustamento do qui-quadrado, que consiste em comparar as discrepâncias entre as frequências observadas e as frequências que se esperam observar se o modelo de regressão logística se ajusta aos dados, isto é, pretende-se testar:

H_0 : O modelo de regressão logística ajusta-se aos dados;

H_1 : O modelo de regressão logística não se ajusta aos dados.

Hosmer e Lemeshow [Hosmer and Lemeshow, 1980] propuseram, para tal, que os dados observados e ajustados (estimados) sejam agrupados e ordenados em dez grupos, $g = 10$, e nunca menos do que três (no caso das variáveis categóricas poderá ter menos do que 10 grupos), com base nos decis das probabilidades estimadas pelo modelo logístico ajustado. Em cada grupo (em geral, obtido pelos decis) são determinadas as frequências observados e as esperadas para cada uma das categorias da variável dependente dicotómica. Para cada uma das categorias, $Y = 1$ e $Y = 0$, da variável dependente, as respetivas frequências esperadas para um dado grupo resultam do produto do número total de sujeitos nesse grupo e as médias das probabilidades estimadas, pelo modelo logístico ajustado, para esses sujeitos. Sob H_0, Hosmer e Lemeshow deduziram a seguinte estatística [Hosmer et al., 2013, p.158]:

$$\chi^2_{HL} = \sum_{i=1}^{g} \frac{(O_i - n_i\overline{\pi}_i)^2}{n_i\overline{\pi}_i(1 - \overline{\pi}_i)}, \tag{2.17}$$

onde, para cada grupo $i = 1, \ldots, g$, n_i representa o número total de sujeitos observados, O_i a soma das frequências observadas e $\overline{\pi}$ é a média das probabilidades estimadas de todos os sujeitos.

A estatística 2.17 não é mais do que a soma das estatísticas de teste de ajustamento do quiquadrado para cada uma das categorias da variável dependente. Isto é, a estatística 2.17 foi deduzida da estatística:

$$\chi^2_{HL} = \sum_{i=1}^{g} \left[\frac{(o_{1k} - \hat{e}_{1k})^2}{\hat{e}_{1k}} + \frac{(o_{0k} - \hat{e}_{0k})^2}{\hat{e}_{0k}} \right],$$

onde, para cada grupo $i = 1, \ldots, g$, o_{1k} e $\hat{e}_{1k}$ representam as frequências observadas e esperadas, respetivamente, da categoria $Y = 1$; o_{0k} e $\hat{e}_{0k}$ representam as frequências observadas e esperadas, respetivamente, da categoria $Y = 0$.
Hosmer e Lemeshow [Hosmer and Lemeshow, 1980] demonstraram com simulações para amostras de grande dimensão que a estatística 2.17 é bem aproximada pela distribuição do quiquadrado com $g - 2$ graus de liberdade. Assim, para a amostra disponível, rejeita-se H_0 se

$$p\text{-value} \approx P\left(\chi^2_{g-2} \geq \chi^2_{HL}\right) \leq \alpha,$$

para o nível de significância α especificado.

2.8 Diagnóstico de Outliers e de Observações Influentes

Os diagnósticos da regressão logística, baseados na análise de resíduos, têm os mesmos objetivos dos diagnósticos utilizados na regressão linear. A análise de resíduos permite identificar outliers e observações influentes na estimação do modelo ou na estimação de algum coeficiente em particular, e as medidas usadas são similares às da regressão linear (resíduos, *leverage*, distância de Cook, e *DfBetas*). Contudo, como referido anteriormente, na regressão logística as variâncias das observações não são constantes, como acontece na regressão linear, obrigando a que as fórmulas de cálculo destas medidas necessitem ser adaptadas.
Os resíduos, não estandardizados, são definidos para a observação j como sendo a diferença entre y_j e a probabilidade de sucesso estimada, $\hat{y}_j$:

$$e_j = y_j - \hat{y}_j = y_j - \hat{\pi}_j$$

Para obter resíduos estandardizados, também designados por resíduos de Pearson, é necessário dividir os resíduos não-estandardizados pela estimativa do desvio padrão dos valores estimados:

$$e'_j = \frac{e_j}{\sqrt{\hat{\pi}_j(1 - \hat{\pi}_j)}}$$

Estes resíduos apresentam média 0 e desvio-padrão aproximadamente 1. Para amostras de grande dimensão, a distribuição de e'_j é assintoticamente normal padrão. Assim, 95% dos valores $|e'_j|$ devem ser inferiores a 1,96 e qualquer observação com $|e'_j|$ superior a $1{,}96 \approx 2$ pode ser classificado como um outlier para o nível de significância de 5% (o quantil $z_{0{,}975} = 1{,}96$). Convém referir que estes resíduos são de difícil interpretação, porque a variância varia com as observações e da sua influência na estimação dos coeficientes, implicando que a estandardização destes resíduos não tem o mesmo sentido que a estandardização dos resíduos na regressão linear. Por esta razão, é preferível usar-se um outro tipo de resíduos, resíduos *deviance*, definidos anteriormente neste trabalho (ver, por exemplo, [Hosmer et al., 2013, p. 155-156]), que, segundo [Simonoff, 2003, p. 133], são mais similares com os resíduos provenientes do método dos mínimos quadrados.

A influência de uma observação (ou categoria, no caso das variáveis categóricas) pode ser medida pela respetiva *leverage*, b_j, que é o j-ésimo elemento da diagonal principal da matriz "chapéu". Esta matriz "chapéu" para a regressão logística, que foi estimada por [Pregibon, 1981] usando o método dos mínimos quadrados ponderados, é dada por:

$$H_L = V^{1/2} X \left(X'VX\right)^{-1} X' V^{1/2}, \tag{2.18}$$

onde V e X são as matrizes 2.12 e 2.11, respetivamente.

Uma observação j é influente na estimação do modelo se tiver uma *leverage* próxima de 1 e é considerada não influente se for próxima de 0. Em [Hosmer et al., 2013, p. 191] demonstra-se que a variação da estatística do χ^2 de Pearson, que pode ser usada para testar o ajustamento do modelo, por eliminação da observação (ou categoria) j, é dada por:

$$\Delta\chi_j^2 = \frac{(e'_j)^2}{1-b_j} = r_j^2,$$

que corresponde ao quadrado do resíduo "estudentizado" da observação j:

$$r_j = \frac{e'_j}{\sqrt{1-b_j}}.$$

Como critério empírico, verifica-se que valores da variação $\Delta\chi_j^2$ superiores a $3{,}84$ indicam que a observação j influenciou no ajustamento do modelo, para o nível de significância de 5%.

A influência de uma determinada observação j na estimação dos coeficientes do modelo é estimada por uma medida análoga à distância de Cook (DC_j). Esta medida indica a variação dos resíduos quando a observação j é eliminada do ajustamento do modelo. Usando a *leverage* e os resíduos de Pearson estandardizados, esta medida pode calcular-se como (ver [Maroco, 2018, p. 806-807]):

$$DC_j = {r_j}^2 \frac{b_j}{1-b_j}.$$

Os valores da DC_j superiores a 1 indicam observações influentes na estimação dos coeficientes do modelo [Maroco, 2018, p. 807].

A influência de uma dada observação j na estimação de cada um dos coeficientes de regressão pode ser estimada pela medida *DfBetas*, à semelhança do modelo de regressão linear, em que na regressão logística são calculados aproximadamente por:

$$\mathit{DfBeta}_{ij} = \hat{\beta}_i - \hat{\beta}_{i(-j)}$$

Onde $\hat{\beta}_i$ é a estimativa do coeficiente de regressão do *logit*, ajustado com todas as observações, e $\hat{\beta}_{i(-j)}$ é a estimativa do coeficiente de regressão do *logit* ajustado sem a observação j. À semelhança da regressão linear, pode-se considerar que valores de *DfBetas* superiores a 2 (ou em rigor superiores a $2\times\sqrt{\frac{k+1}{n}}$, onde $k+1$ é o número de coeficientes do modelo e n é a dimensão da amostra) são observações influentes, sugerindo que a observação j deve ser examinada com precaução já que afeta a estimativa de β_j [Maroco, 2018, p. 807].

2.9 Outros Modos de Avaliar a Qualidade do Modelo de Regressão Logística

Como vimos anteriormente, existem vários testes de hipóteses que permitem avaliar a qualidade do modelo de regressão logística ajustado aos dados, como, por exemplo, os testes do rácio de verosimilhanças ou de Hosmer-Lemeshow. E uma medida para a avaliação da qualidade do ajuste consiste em utilizar-se um pseudo-R^2, como o de Nagerkerke, que sugere-se ser interpretado como o coeficiente de determinação da regressão linear, mas com maior subjetividade.
As razões da subjetividade dos diferentes pseudo coeficientes de determinação foram apresentadas anteriormente na secção 2.6.1. Para não ficarmos limitados à utilização subjetiva destes coeficientes, vejamos agora uma outra abordagem para se determinar a qualidade de um modelo logístico, que consiste em se determinar o seu poder de discriminação dos sujeitos.

2.9.1 Poder de Classificação (ou Discriminação) do Modelo de Regressão Logística

Depois de obtidas as estimativas dos coeficientes do *logit*, a partir da amostra observada disponível, é possível estimar a probabilidade de "sucesso", $\hat{\pi}_j$, para cada observação j, $j = 1,2,\cdots,n$, com o modelo 2.8 e, consoante a grandeza dessa probabilidade estimada, o sujeito relativo à observação j irá pertencer ao grupo "1: sucesso" ou ao grupo de referência "0: insucesso". Para tal, será necessário usar-se um valor de corte. Por exemplo, a probabilidade $0{,}5$. Isto é, se $\hat{\pi}_j \leq 0{,}5$, o sujeito será classificado no grupo "0", caso contrário, o sujeito será classificado no grupo "1".
A escolha da probabilidade de corte é arbitrária e outro valor pode ser escolhido de forma a tornar a classificação mais (por exemplo, $0{,}7$) ou menos (por exemplo, $0{,}4$) rigorosa. A classificação dos sujeitos da amostra disponível, usados na obtenção do modelo, é geralmente enviesada a favor das taxas de classificação corretas mais elevadas. Para resolver-se este problema, deveria usar-se uma parte da amostra disponível ($1/2$ a $3/4$ da amostra) para a criação do modelo e a outra parte (ou mesmo uma nova amostra) para a validação do modelo, isto é, para testar o poder classificativo do modelo. Para amostras pequenas deverá usar-se algum método de reamostragem (Jackknife, bootstrap ou validação cruzada).
Para a avaliação do poder classificativo efetuado pelo modelo é usual comparar a percentagem de sujeitos corretamente classificados com o modelo com a percentagem proporcional de classificações corretas por acaso. Esta percentagem é calculada a partir do número de sujeitos observados, n_i, em cada uma das 2 categorias da variável dependente Y pela expressão:

$$\text{Classificação Correta Proporcional por Acaso (\%)} = \sum_{i=1}^{2} \left(\frac{n_i}{n}\right)^2.$$

Se a percentagem de casos classificados corretamente pelo modelo for superior em pelo menos 25% à percentagem de classificação proporcional por acaso, considera-se que o modelo tem boas propriedades classificativas [Maroco, 2018, p. 808].

2.9.2 Sensibilidade e Especificidade

A eficiência classificativa do modelo logístico pode também ser avaliada pela sensibilidade (s) e especificidade (e) apresentada pelo modelo.

A sensibilidade é a percentagem de classificações corretas no grupo "1-sucesso" da variável dependente (isto é, o modelo prevê corretamente para um dado sujeito a característica que se quer modelar). A especificidade é a percentagem de classificações corretas no grupo "0-insucesso" (isto é, o modelo prevê corretamente para um dado sujeito que não tem a característica que se quer modelar).
Considera-se que um modelo com boas capacidades preditivas apresenta sensibilidade e especificidade superiores a 80%. Para percentagens entre 50% e 80% as capacidade preditivas são consideradas razoáveis. Abaixo de 50% as capacidades preditivas são consideradas medíocres. [Maroco, 2018, p. 808]

2.9.3 Curva ROC

O grande problema da utilização da sensibilidade e da especificidade é que estas medidas dependem do valor de corte, cuja escolha é arbitrária, como referido anteriormente, implicando aumentar-se a sensibilidade em detrimento da especificidade, ou vice-versa. O modo que permite resolver este problema, que não dependa do valor de corte, consiste em utilizar-se uma análise ROC (Receiver Operating Characteristic).
A análise ROC consiste em determinar-se uma curva que corresponde à razão entre a sensibilidade e um menos a especificidade (proporção de falsos positivos, isto é, proporção de incorretamente classificados como tendo a característica a modelar, em relação ao total de sujeitos que não têm a característica a modelar).
A curva ROC é desenvolvida num plano unitário de abcissa $1 - e$ e ordenada s, sobre o qual são traçados os diferentes pontos de corte das probabilidades relativas à sensibilidade e à especificidade. Deve-se escolher como ponto de corte aquele que mais se aproxima do canto superior esquerdo do plano unitário, ou seja, aquele que maximiza a sensibilidade e a especificidade. Quanto mais próximo do canto superior esquerdo o ponto de corte for, maior será o poder de discriminação do modelo.
Sobre o plano unitário é traçado também uma bissectriz ($y = x$), quanto mais próxima a curva estiver da bissectriz, menor é o poder de discriminação do modelo.
A área debaixo da curva ROC, varia, obviamente, entre 0 e 1 e é utilizada para medir a eficiência de como o modelo consegue efetuar a classificação ou discriminação dos sujeitos. Se a área for igual a 0,5, o modelo não consegue discriminar os sujeitos com *versus* sem a característica melhor do que a escolha por mero acaso (isto é, acerta com probabilidade 0,5). Quanto mais próximo a área estiver de 1 (isto é, da razão entre a sensibilidade igual a 1 e especificidade igual a 0), maior é a capacidade do modelo para discriminar os sujeitos que apresentam a característica modelada (sucesso) dos indivíduos que não apresentam essa característica (insucesso). Na tabela seguinte encontram-se os valores indicativos da área ROC que, de acordo com os autores Hosmer e Lemeshow (ver [Hosmer et al., 2013, p. 177]), servem como critério de aplicação geral para a descrição do poder discriminante apresentado pelo modelo de regressão logística:

Tabela 2.2: Valores indicativos da área ROC

Área ROC	Poder discriminante do modelo
$0{,}5$	Sem poder discriminante
$]0{,}5; 0{,}7[$	Discriminação fraca
$[0{,}7; 0{,}8[$	Discriminação aceitável
$[0{,}8; 0{,}9[$	Discriminação boa
$\geq 0{,}9$	Discriminação excecional

2.10 Pressupostos do Modelo Logístico

Tendo em conta as condições impostas anteriormente na criação do modelo de regressão logística, podemos resumir os seus pressupostos da seguinte forma [Maroco, 2018, p. 793]:

- Linearidade e aditividade: a escala *logit* é aditiva e linear (mas a de π não);
- Proporcionalidade: a contribuição de cada X_j, $j = 1,2,\cdots,k$, é proporcional ao seu valor com um factor β_i;
- Constância de efeito: a contribuição de uma variável independente é constante e independente da contribuição das outras variáveis independentes;
- Os resíduos são independentes e binomialmente distribuídos;
- As variáveis X_j, $j = 1,2,\cdots,k$ não são multicolineares (à semelhança da regressão linear múltipla).

A validação dos pressupostos do modelo pode fazer-se graficamente pela análise de resíduos e a multicolinearidade pode ser diagnosticada calculando a tolerância, T, a partir de $R^2 (T = 1-R^2)$ obtido pela regressão linear múltipla entre cada uma das variáveis independentes e as restantes variáveis independentes no modelo.

2.11 A Procura do Melhor Modelo

À semelhança da regressão linear múltipla, na regressão logística múltipla é também possível utilizar algoritmos de seleção de variáveis com poder preditivo. Existem métodos do tipo *backward* e *forward*, que, simplificadamente, passaremos a descrever, mas poderão ser encontrados mais pormenores acerca deste assunto em [Maroco, 2018, p. 802-804].

- Métodos de Seleção do Tipo *Backward*:

  Backward Baseada no Rácio de Verosimilhanças (Likelihood Ratio) (Backward LR): Num primeiro passo todas as variáveis independentes são adicionadas ao modelo e, nos passos seguintes, são removidas as variáveis que não são significativas no modelo (em geral adota-se o nível de significância de 5%), considerando o teste do rácio de verosimilhanças baseado nas estimativas parciais de máxima verosimilhança do modelo.

  Seleção Backward Baseada no Rácio de Verosimilhanças Condicionada (Backward Condicional): No primeiro passo todas as variáveis independentes são adicionadas ao modelo e, nos passos seguintes, são removidas as variáveis cuja probabilidade do rácio de verosimilhanças baseada nas estimativas condicionais dos coeficientes do modelo é superior ao p-value de remoção selecionado (em geral adota-se o nível de significância de 5%). Neste método, os novos coeficientes sem uma dada variável são estimados a partir dos coeficientes originais e das covariâncias entre estes e o coeficiente da variável eliminada do modelo.

  Backward Baseada no Teste de Wald (Backward Wald): Consiste na remoção das variáveis do modelo a partir da significância do teste de Wald (em geral adota-se o nível de significância de 5%).

- Métodos de Seleção do Tipo *Forward*: De um modo geral, este tipo de método é similar ao método *stepwise* usado na regressão linear, que consiste na adição de uma variável no modelo se é significativa para 5%, (isto é, a variável entra no modelo se o seu p-value no teste de adição não exceder 0,05) e a remoção para o nível de significância de 10% (isto é, uma variável é removida se o seu p-value no teste de remoção exceder 0,10).

  Forward Baseada no Rácio de Verosimilhanças (Forward LR): A entrada de uma variável independente no modelo é feita em função da significância da estatística *Score* e a remoção de uma variável no modelo é feita a partir do teste do rácio de verosimilhanças baseados nas estimativas parciais de máxima verosimilhança do modelo.

  Seleção Forward Baseada no Rácio de Verosimilhanças Condicionada (Forward Conditional): A entrada de uma variável independente no modelo é feita em função da significância da estatística *Score* do modelo e a remoção de uma variável do modelo é feita em função da significância do teste do rácio de verosimilhanças baseado nas estimativas condicionadas dos coeficientes do modelo.

  Seleção Forward Baseada no Teste de Wald (Forward Wald): é um método de selecção stepwise, em que a entrada de uma variável independente no modelo é feita em função da significância da estatística Score e a remoção de uma variável do modelo é feita em função da significância do teste de Wald.

2.12 Exemplo de Aplicação a Dados Angolanos

2.12.1 Introdução

Nesta secção aplicaremos os conhecimentos expostos anteriormente e utilizaremos o programa estatístico IBM SPSS, versão 25, na análise de dados proveniente de escolas angolanas. Os dados utilizados nesta aplicação foram recolhidos em 3 escolas angolanas do ensino secundário, todas da cidade do Luena, município do Moxico, província do Moxico em Angola:

1. 11 de Novembro Periurbana;
2. 338 Tchifuchi - Urbana;
3. 4 de Abril- Suburbana.

Para a recolha dos dados foi utilizado como instrumento um inquérito (ver apêndice A), elaborado para a dissertação de mestrado em Matemática para Professores da UBI do mestre Ngaiele Muecheno Fundão, que amavelmente nos disponibilizou os dados para serem utilizados no presente trabalho.
Nos resultados do estudo, que a seguir apresentaremos, todos os testes de hipóteses foram considerados significativos quando o respetivo valor de prova não excedeu o nível de significância de 5% e os intervalos de confiança foram considerados a 95%.

2.12.2 Resultados

Dos 350 alunos inquiridos, aleatoriamente escolhidos, 170 (48,6%) pertenciam à escola 11 de Novembro Periurbana (73 (42,9%) do sexo feminino e 97 (57,1%) do sexo masculino), 140 (40,0%)

à escola 338 Tchifuchi - Urbana (52 (37,1%) do sexo feminino e 88 (62,9%) do sexo masculino) e os restantes 40 à escola 4 de Abril- Suburbana (14 (35,0%) do sexo feminino e 26 (65,0%) do sexo masculino).

A variável dicotómica escolhida para variável dependente deste estudo foi a classificação atribuída ao professor de Matemática (CAP), com as categorias "Bom" (código 1) e "Mau" (código 0). Na amostra, 321 (91,7%) dos alunos classificaram o professor de Matemática como "Bom" e os restantes 29 (8,3%) dos alunos como "Mau". As associações, obtidas por regressão logística, entre esta variável e cada uma das variáveis consideradas neste estudo (versão univariada) encontram-se na tabela 2.3.

Pode observar-se na tabela 2.3 que a escola, o estado civil, a renda familiar e as notas a Matemática e Física mostraram-se significativamente associadas à classificação atribuída ao professor de Matemática, enquanto para a idade essa associação foi marginalmente significativa. Observe-se, em particular, que as chances de um aluno atribuir a classificação "Bom" ao professor de Matemática aumenta cerca de 32% com o aumento de 1 valor na nota a Matemática, quando comparada com as chances de um aluno atribuir a classificação "Mau" ($OR = 1{,}324; IC\ 95\% : (1{,}037; 1{,}691); p = 0{,}024$). Para o caso da Física, essas chances aumentaram cerca de 26,1% ($OR = 1{,}261; IC\ 95\% : (1{,}032; 1{,}541); p = 0{,}023$).

Em relação à escola, as chances de um aluno da escola 338 Tchifuchi - Urbana atribuir a classificação "Bom" ao professor de Matemática estimou-se ser cerca de 4 vezes maiores, quando comparada com as chances para um aluno da escola 11 de Novembro Periurbana ($OR = 4{,}014; IC 95\% : (1{,}478; 10{,}895); p = 0{,}006$); na comparação entre as escolas 4 de Abril- Suburbana e 11 de Novembro Periurbana, o OR não se mostrou significativamente diferente de 1 ($OR = 2{,}824; IC\ 95\% : (0{,}626; 12{,}541); p = 0{,}172$). Os alunos solteiros apresentaram chances 2,8 vezes maiores de atribuírem a classificação "Bom" ao professor de Matemática, quando comparadas com as chances dos alunos casados ou que estão numa relação união de facto ($OR = 2{,}830; IC 95\% : (1{,}268; 6{,}317); p = 0{,}011$). Os que tinham rendimento familiar considerado médio apresentam chances cerca de 4,5 vezes maior de atribuir a classificação "Bom" ao professor de Matemática, quando comparadas com as chances daqueles que tinham um baixo rendimento familiar ($OR = 4{,}520; IC\ 95\% : (2{,}026; 10{,}083); p < 0{,}001$).

No entanto, quando se utiliza um método de seleção de variáveis, muitas das covariáveis deixam de ser significativas num modelo logístico múltiplo. Ao efetuar-se um diagnóstico de multicolinearidade, verifica-se a existência deste problema nas quatro variáveis quantitativas presentes neste estudo: idade e notas a Matemática, Física e Lingua Portuguesa, para as quais se observa um índice de condição (condition index) superior a 30 e com contribuições para a variância superiores a 50% (ver apêndice B) [Maroco, 2018].

Com o método de seleção *Forward LR* (seleção *Forward* baseada no rácio de verosimilhanças), onde se usou como critério o nível de significância de 5% para a inclusão de uma variável no modelo e 10% para a exclusão, obteve-se o seguinte modelo logístico:

$$\hat{P}(CAP = 1) = \frac{1}{1 + e^{-(0{,}485 + 1{,}539RF + 0{,}846EC + 1{,}346E_1 + 1{,}077E_2)}}, \tag{2.19}$$

onde E_1 e E_2 representam as variáveis *dummy* para as escolas 338 Tchifuchi - Urbana e 4 de Abril - Suburbana, respetivamente. Na tabela 2.4 pode verificar-se que no modelo de regressão logística múltipla 2.19, ajustado com todas as covariáveis disponíveis, a escola manteve-se como um fator importante para a classificação atribuída ao professor de Matemática, em particular para a escola 338 Tchifuchi - Urbana, quando comparada com a escola 11 de Novembro Periurbana. O

Tabela 2.3: Associações entre a variável dependente (CAP) e cada covariável considerada no estudo.

	CAP			
	Bom	Mau	OR	**Teste de Wald**
Variável	N (%)	N (%)	(IC 95%)	p-value
Escola (E)				0,015
11 de Novembro Periurbana	148 (46,1)	22 (75,9)	referência	
338 Tchifuchi - Urbana	135 (42,1)	5 (17,2)	4,014 (1,478; 10,895)	0,006
4 de Abril - Suburbana	38 (11,8)	2 (6,9)	2,824 (0,626; 12,541)	0,172
Sexo				
Feminino	127 (39,6)	12 (41,4)	referência	
Masculino	197 (60,4)	17 (58,6)	1,078 (0,498; 2,334)	0,848
Estado civil (EC)				
Casado/união de facto	57 (17,8)	11 (37,9)	referência	
Solteiro	264 (82,2))	18 (62,1)	2,830 (1,268; 6,317)	0,011
Renda familiar (RF)				
Baixa	95 (29,6)	19 (65,5)	referência	
Média	226 (70,4)	10 (34,5)	4,520 (2,026; 10,083)	<0,001
Situação laboral do aluno				
Não tem emprego	294 (91,6)	25 (86,2)	1,742 (0,565; 5,375)	0,334
Tem emprego	27 (8,4)	4 (13,8)	referência	
Grau de satisfação escolar				0,338
Muito insatisfeito	8 (2,5)	0 (0,0)	-	-
Insatisfeito	13 (4,0)	2 (6,9)	0,539 (0,111; 2,613)	0,443
Parcialmente satisfeito	54 (16,8)	9 (31,0)	0,497 (0,206; 1,199)	0,120
Satisfeito	65 (20,2)	3 (10,3)	1,796 (0,503; 6,404)	0,367
Muito satisfeito	181 (56,4)	15 (51,7)	referência	
Idade (em anos)			0,916 (0,838; 1,000)	0,051
Média± SD	21,07± 3,74	22,55± 4,76		
Mediana (mínimo; máximo)	20 (16; 32)	21 (16; 32)		
Notas a Matemática (0-20)			1,324 (1,037; 1,691)	0,024
Média± SD	12,58± 1,76	11,80± 1,85		
Mediana (mínimo; máximo)	12,2 (9; 17)	11,0 (9,8; 17)		
Notas a Língua Portuguesa (0-20)			1,072 (0,892; 1,288)	0,458
Média± SD	12,96± 2,15	12,65± 2,09		
Mediana (mínimo; máximo)	13 (8; 19)	12 (10; 19)		
Notas a Física (0-20			1,261 (1,032; 1,541)	0,023
Média± SD	12,68± 2,12	11,73± 2,17		
Mediana (mínimo; máximo)	12,3 (7; 18)	11 (8,8; 17,6)		

estado civil passou a ser um fator marginalmente significativo para a classificação atribuída ao professor de Matemática (p-value=0,051) e o fator renda familiar reforça a sua importância na classificação atribuída ao professor de Matemática ($p < 0{,}001$). Em particular para este último fator, as chances de um aluno com rendimento considerado médio atribuir a classificação "Bom" ao professor de Matemática foi cerca de 4,66 vezes maior do que as chances para um aluno com rendimento considerado baixo ($OR = 4{,}659; IC\ 95\% : (2{,}045; 10{,}615); p < 0{,}001$).

Na figura 2.3 pode observar-se a curva ROC do modelo de regressão logística 2.19. A área ROC foi aproximadamente igual a $0{,}770$ ($IC95\%$ para a verdadeira área: $(0{,}675; 0{,}865)$; p-value $< 0{,}001$), que evidencia uma discriminação aceitável dos alunos [Hosmer et al., 2013]. Adicionalmente, com a curva ROC verifica-se que a sensibilidade e a especificidade apresentada pelo modelo logístico foram aproximadamente iguais 78,5% e 69,0%, para a probabilidade de corte 0,9, que permite maximizar simultaneamente estas duas medidas (ver anexo C).

O modelo 2.19 apresentou 13 outliers, mas nenhuma destas observações, ou qualquer outra, influenciou a estimação do modelo, como se poderá constatar pelo gráfico 2.4

Nos diagramas de dispersão 2.5, 2.6, 2.7, 2.8 e 2.9 encontram-se representados os valores da medida de influência *DFbeta* de cada observação sobre os coeficientes do *logit*. Utilizando-se

Tabela 2.4: Regressão logística múltipla preditiva da classificação atribuída ao professor de Matemática.

	CAP			
Variável	Boa N (%)	Má N (%)	*OR* (*IC* 95%)	**Teste de Wald** p-value
Escola (E)				0,022
11 de Novembro Periurbana	148 (46,1)	22 (75,9)	referência	
338 Tchifuchi - Urbana	135 (42,1)	5 (17,2)	3,842 (1,375; 10,734)	0,010
4 de Abril - Suburbana	38 (11,8)	2 (6,9)	2,935 (0,641; 13,442)	0,166
Estado civil (EC)				
Casado/união de facto	57 (17,8)	11 (37,9)	referência	
Solteiro	264 (82,2))	18 (62,1)	2,331 (0,996; 5,457)	0,051
Renda familiar (RF)				
Baixa	95 (29,6)	19 (65,5)	referência	
Média	226 (70,4)	10 (34,5)	4,659 (2,045; 10,615)	<0,001

Teste do Rácio de Verosimilhanças, $p < 0{,}001$; teste de Hosmer-Lemeshow, $p = 0{,}805$.

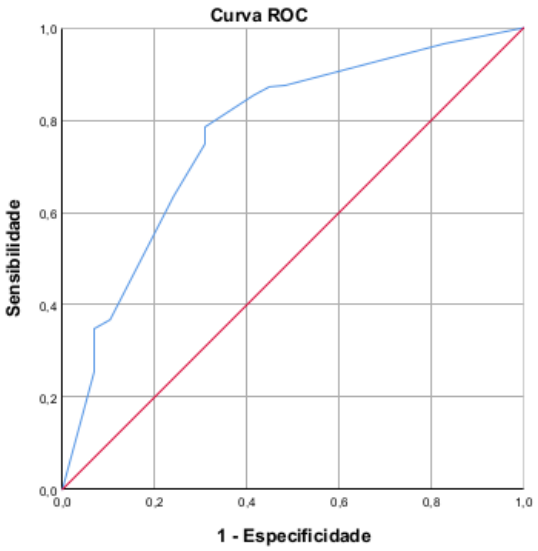

Figura 2.3: Curva ROC do modelo de regressão logística múltipla.

como critério que uma observação influenciou a estimação de um dado coeficiente se o respetivo *DFbeta* é superior a $2\sqrt{\frac{k+1}{n}}$, onde k representa o número de covariáveis inseridas no modelo e n é, obviamente, a dimensão da amostra, verificou-se que duas observações (uma delas é um outlier) influenciaram a estimação do coeficiente de regressão do logit da escola 4 de Abril - Suburbana (ver no diagrama de dispersão 2.7 os dois pontos assinalados: observações 321 e 340). Esta influência explica-se pelo facto de estas duas observações corresponderem aos únicos dois alunos da escola 4 de Abril - Suburbana que atribuíram a classificação "Mau" ao professor de Matemática. Assim, a exclusão desses dois alunos do estudo implicaria a retirada da escola 4 de Abril - Suburbana do modelo. Por esta razão, decidiu-se não se proceder a essa exclusão.

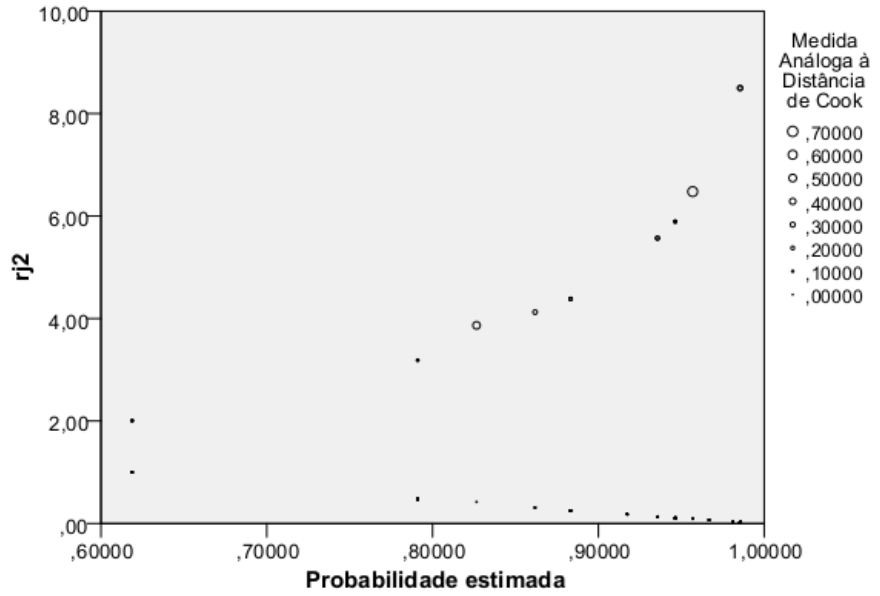

Figura 2.4: Probabilidades estimadas vs. quadrados dos resíduos "estudentizados" e medida análoga à distância de Cook para cada observação.

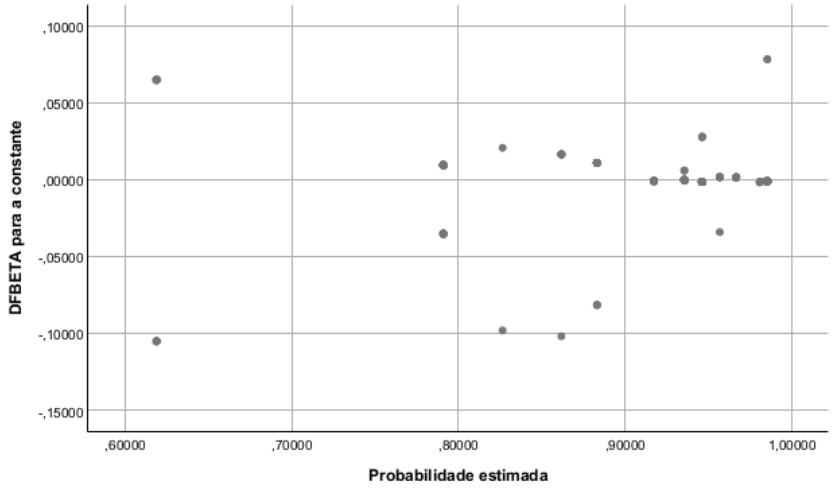

Figura 2.5: Probabilidades estimadas pelo modelo vs. valores da medida DFbeta para a constante para cada observação.

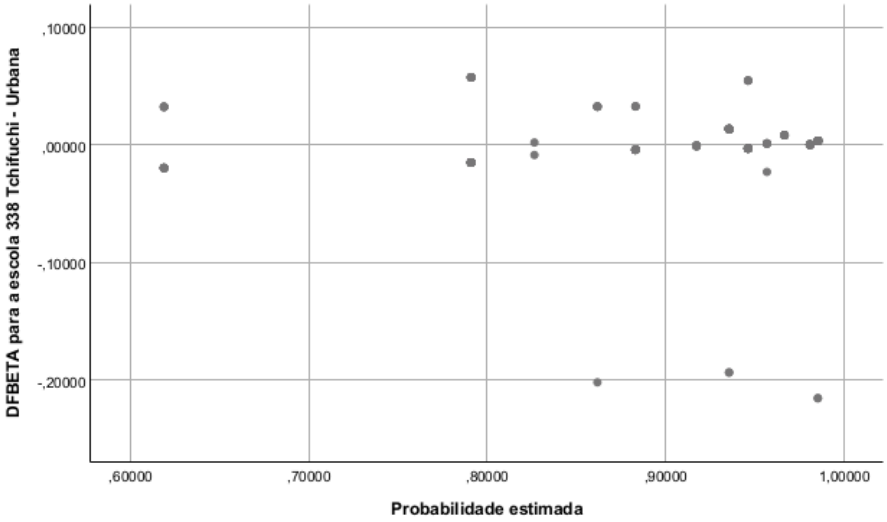

Figura 2.6: Probabilidades estimadas pelo modelo vs. valores da medida DFbeta para a escola Tchifuchi - Urbana para cada observação.

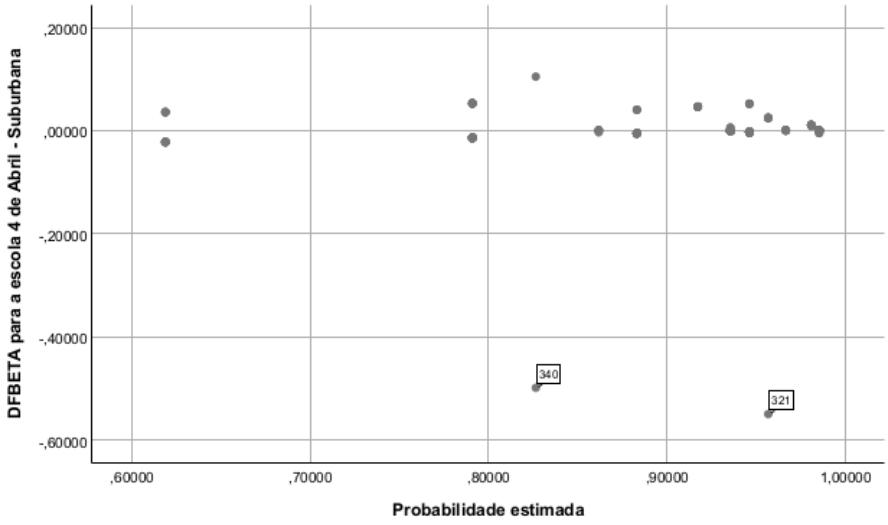

Figura 2.7: Probabilidades estimadas pelo modelo vs. valores da medida DFbeta para a escola 4 de Abril - Suburbana para cada observação.

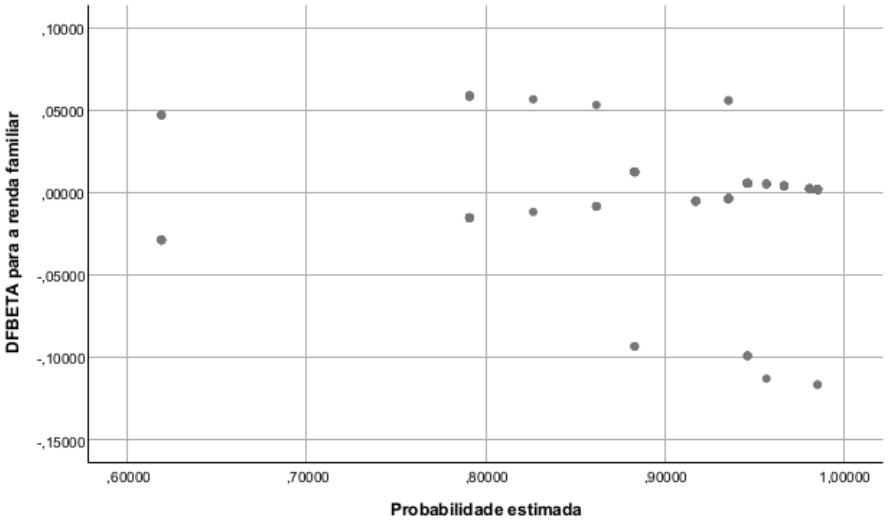

Figura 2.9: Probabilidades estimadas pelo modelo vs. valores da medida DFbeta para a renda familiar para cada observação.

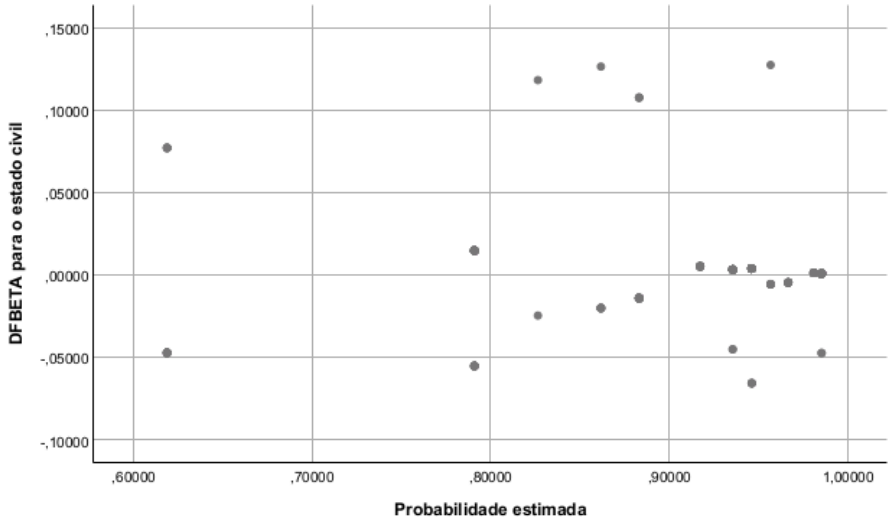

Figura 2.8: Probabilidades estimadas pelo modelo vs. valores da medida DFbeta para o estado civil para cada observação.

2.12.3 Conclusão

Com este estudo, foi possível relacionar a classificação atribuída ao professor de Matemática com a escola, o estado civil e a renda familiar.

O modelo ajustado evidenciou que o rendimento familiar dos alunos propiciou a classificação de "Bom" atribuída ao professor de Matemática e, de igual modo, a geografia da escola. Ou seja, os alunos das escolas urbanas tendem a atribuir uma melhor classificação aos professores de Matemática e, consequentemente, uma melhor relação aluno/professor, quando comparado aos alunos de escolas periurbanas. Tal pode ser verdade, visto que muitas vezes os alunos não dominam a língua oficial (português) e o professor não domina os diversos dialetos locais, que permitem servir de vetor de proximidade.

Os alunos solteiros tenderam a atribuir uma melhor classificação ao professor de Matemática, mas não está claro o porquê desta tendência. Provavelmente estes sentem-se mais à vontade ao lado do professor, quando comparado com os casados.

Capítulo 3

O Modelo de Regressão Logística Multinomial

3.1 Introdução

Os métodos de regressão, linear e não linear, têm uma grande importância devido à sua grande aplicabilidade nos mais diversos campos do saber humano. Neste capítulo a atenção será voltada para um dos métodos de regressão não linear, a regressão logística multinomial, que é uma extensão do modelo de regressão logística. A regressão logística multinomial é adequada em cenários em que a variável dependente é qualitativa (ou quantitativa agrupada) e apresenta mais de duas categorias mutuamente exclusivas, ou seja, é um modelo de regressão logística que consiste num conjunto de modelos logísticos corrigidos de acordo com as várias categorias da variável dependente.

É escopo deste capítulo a apresentação de uma Introdução aos conceitos utilizados na regressão logística multinomial, como o método da máxima verosimilhança, e as medidas de qualidade de ajustamento, como o teste do rácio de verosimilhança, o teste de Wald e os critérios de informação AIC e BIC.

Consta ainda deste capítulo a aplicação do modelo de regressão logística multinomial, com utilização do programa estatístico IBM SPSS na sua versão 25, aos mesmos dados provenientes de escolas angolanas, utilizados no capítulo anterior.

3.2 O Modelo de Regressão Logística Multinomial

O modelo de regressão logística é um modelo de regressão não linear, adequado em cenários em que a variável dependente é qualitativa dicotómica, que assume apenas valores discretos de categorias mutuamente exclusivas. Quando a variável dependente toma mais do que duas categorias mutuamente exclusivas, será necessário tomar uma forma generalizada do modelo de regressão logística, que é conhecido por modelo de regressão logística multinomial ou, simplesmente, regressão multinomial. Na regressão logística multinomial o método de estimação das probabilidades para a variável dependente, condicionadas por covariáveis $X = (X_1, \ldots, X_k)$, e os pressupostos são similares ao modelo de regressão logística com a transformação *logit*.
Considere-se a variável dependente Y com $m+1$ categorias codificadas por $0, 1, \ldots, m$. À semelhança do modelo de regressão logística, no modelo de regressão multinomial uma das categorias da variável dependente será escolhida como categoria de referência e, na estimação dos *odds ratios*, cada uma das outras categorias é comparada com essa referência. A escolha da classe de referência é arbitrária e é um critério do investigador. Frequentemente, toma-se como categoria de referência a primeira ou a última categoria. A classe de referência pode ser alterada de acordo com o objetivo do estudo. A alteração da categoria de referência não altera o modelo em si, mas altera a interpretação dos coeficientes do modelo. De qualquer modo, será

possível determinar à posteriori os *odds ratio* usando-se como referência qualquer outra categoria. Assim, o modelo de regressão multinomial consiste num conjunto de $m+1$ modelos de regressão logística corrigidos, um para cada uma das $m+1$ categorias da variável dependente [Maroco, 2018, p. 842].

3.2.1 Odds e Linearização do Modelo Multinomial

Defina-se **odds** (chances ou rácio de verosimilhanças) para cada categoria $j = 1, \ldots, m$ em relação à categoria de referência ($Y = 0$) por

$$odds_j = \frac{P(Y = j|X)}{P(Y = 0|X)}. \tag{3.1}$$

Cada um destes *odds* representam as razões entre as probabilidades condicionadas por X de cada uma das categorias $Y = j$, $j = 1, \ldots, m$, em relação à categoria tomada como referência, $Y = 0$.

À semelhança do modelo de regressão logística, defina-se o *logit* para cada categoria j, com $j = 1, \ldots, m$, como o logaritmo natural de cada *odd* e assuma-se que são lineares:

$$f_j(x) = \textbf{\textit{logit}}\left(odds_j\right) = \ln\left(\frac{P(Y = j|X)}{P(Y = 0|X)}\right) = \beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2 + \cdots + \beta_{jk}X_k, \tag{3.2}$$

onde $\beta_{j0}, \beta_{j1}, \ldots, \beta_{jk}$ representam os coeficientes de regressão linear.

Vejamos como poderemos deduzir o modelo de regressão multinomial para o caso particular em que $m = 2$. Com o pressuposto de que o *logit* é linear, obtém-se de 3.2 as expressões seguintes:

$$\begin{aligned} f_1(x) = \ln\left(\frac{P(Y = 1|X)}{P(Y = 0|X)}\right) &\Leftrightarrow \frac{P(Y = 1|X)}{P(Y = 0|X)} = e^{f_1(x)} \\ &\Leftrightarrow P(Y = 1|X) = P(Y = 0|X)\, e^{f_1(x)}; \end{aligned} \tag{3.3}$$

$$\begin{aligned} f_2(x) = \ln\left(\frac{P(Y = 2|X)}{P(Y = 0|X)}\right) &\Leftrightarrow \frac{P(Y = 2|X)}{P(Y = 0|X)} = e^{f_2(x)} \\ &\Leftrightarrow P(Y = 2|X) = P(Y = 0|X)\, e^{f_2(x)}. \end{aligned} \tag{3.4}$$

Pelo teorema da probabilidade total, sabe-se que:

$$P(Y = 0|X) + P(Y = 1|X) + P(Y = 2|X) = 1. \tag{3.5}$$

Substituindo-se 3.3 e 3.4 em 3.5, obtém-se:

$$P(Y = 0|X)[1 + e^{f_1(x)} + e^{f_2(x)}] = 1,$$

e, consequentemente,

$$P(Y = 0|X) = \frac{1}{1 + e^{f_1(x)} + e^{f_2(x)}}. \tag{3.6}$$

Agora, das equações 3.3, 3.4 e 3.6, facilmente se deduz que:

$$P(Y=1|X) = \frac{e^{f_1(x)}}{1+e^{f_1(x)}+e^{f_2(x)}} \tag{3.7}$$

e

$$P(Y=2|X) = \frac{e^{f_2(x)}}{1+e^{f_1(x)}+e^{f_2(x)}}, \tag{3.8}$$

As expressões 3.6, 3.7 e 3.8 também podem ser escritas da seguinte forma:

$$P(Y=0|X) = \frac{1}{1+e^{\beta_{10}+\sum_{i=1}^{k}\beta_{1i}X_i}+e^{\beta_{20}+\sum_{i=1}^{k}\beta_{2i}X_i}}; \tag{3.9}$$

$$P(Y=1|X) = \frac{e^{\beta_{10}+\sum_{i=1}^{k}\beta_{1i}X_i}}{1+e^{\beta_{10}+\sum_{i=1}^{k}\beta_{1i}X_i}+e^{\beta_{20}+\sum_{i=1}^{k}\beta_{2i}X_i}}; \tag{3.10}$$

$$P(Y=2|X) = \frac{e^{\beta_{20}+\sum_{i=1}^{k}\beta_{2i}X_i}}{1+e^{\beta_{10}+\sum_{i=1}^{k}\beta_{1i}X_i}+e^{\beta_{20}+\sum_{i=1}^{k}\beta_{2i}X_i}}. \tag{3.11}$$

Similarmente, pode deduzir-se a expressão geral para a probabilidade de se observar uma determinada categoria j, com $j = 0,1,\ldots,m$, onde todos os coeficientes de regressão da categoria tomada como referência (j=0) são nulos:

$$P(Y=j|X) = \frac{e^{\beta_{j0}+\sum_{i=1}^{k}\beta_{ji}X_i}}{1+\sum_{l=1}^{m}e^{\beta_{l0}+\sum_{i=1}^{k}\beta_{li}X_i}},\ j = 0,1,\ldots,m. \tag{3.12}$$

Cada uma das funções de probabilidade representa uma curva e as curvas correspondentes, a cada uma das m categorias de uma variável dependente, têm a mesma forma que no modelo de regressão logística quando a variável dependente é dicotómica.
Com este modelo, um sujeito será classificado na categoria que apresentar maior probabilidade e esta classificação, aplicada aos sujeitos de uma amostra, permitirá determinar o poder de classificação dos sujeitos pelo modelo logístico multinomial ajustado.

3.2.2 Estimação dos Coeficientes de Regressão

À semelhança da regressão logística, os estimadores dos coeficientes de regressão do *logit* do modelo de regressão logística multinomial também são deduzidos pelo método da máxima verosimilhança.
Façamos esta dedução para o caso particular em que $m = 2$. Recodifique-se a variável dependente Y em três variáveis *dummy* Y_0, Y_1 e Y_2, tais que:

- $Y_0 = 1$, $Y_1 = 0$ e $Y_2 = 0$, se $Y = 0$;
- $Y_0 = 0$, $Y_1 = 1$ e $Y_2 = 0$, se $Y = 1$;
- $Y_0 = 0$, $Y_1 = 0$ e $Y_2 = 1$, se $Y = 2$.

Tome-se uma amostra com n observações independentes da variável Y, recodificadas como anteriormente: $(y_{01},y_{02},\ldots,y_{0n})$, $(y_{11},y_{12},\ldots,y_{1n})$ e $(y_{21},y_{22},\ldots,y_{2n})$, e uma amostra, também com n observações independentes, de cada uma das covariáveis $X_1, X_2, \ldots, X_k$: $(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_k)$, onde $\mathbf{x}_i = (x_{i1},x_{i2},\ldots,x_{in})$, para $i = 1,2,\ldots,k$.
A função de verosimilhança de Y condicionada por X será dada por:

$$L(\beta) = \prod_{i=1}^{n} [\pi_0(\mathbf{x}_i)^{y_{0i}} \pi_1(\mathbf{x}_i)^{y_{1i}} \pi_2(\mathbf{x}_i)^{y_{2i}}], \tag{3.13}$$

onde $\pi_{0i}(\mathbf{x}_i) = P(Y = 0|X_i = \mathbf{x}_i)$, $\pi_{1i}(\mathbf{x}_i) = P(Y = 1|X_i = \mathbf{x}_i)$ e $\pi_{2i}(\mathbf{x}_i) = P(Y = 2|X_i = \mathbf{x}_i)$.
Linearizando-se a função 3.13 com o logaritmo neperiano, observando-se que $\sum_j y_{ji} = 1$, para cada $i = 1,\ldots,n$, e atendendo-se ao pressuposto que o *logit* é linear, obtém-se o logaritmo da função de verosimilhança:

$$LL(\beta) = \sum_{i=1}^{n} \left[y_{1i} f_1(x_i) + y_{2i} f_2(x_i) - \ln\left(1 + e^{f_1(x_i)} + e^{f_2(x_i)}\right)\right].$$

Para cada um dos $2(k+1)$ parâmetros desconhecidos, o máximo do logaritmo da função de verosimilhança ocorre quando

$$\frac{\partial LL(\beta)}{\partial \beta_{jk}} = 0$$

e

$$\frac{\partial^2 LL(\beta)}{\partial \beta_{jk} \partial \beta_{jk}} < 0,$$

onde a forma geral das primeiras derivadas parciais de $LL(\beta)$, em relação a cada um dos parâmetros desconhecidos, é dada por:

$$\frac{\partial LL(\beta)}{\partial \beta_{jl}} = \sum_{i=1}^{n} x_{li}\left(y_{ji} - \pi_{ji}\right),\ j = 1,2, \text{ e } l = 0,1,\ldots,k,$$

em que $x_{0i} = 1$, para cada $i = 1,\ldots,n$.

3.2.3 Interpretação dos Coeficientes do *Logit* Multinomial

Assim como na regressão logística, na regressão logística multinomial os coeficientes de regressão não são diretamente interpretados, mas sim a exponencial dos coeficientes de regressão, conhecidas por *odds ratio* (OR) ou razão de chances. No modelo de regressão logística multinomial, a estimativa e interpretação do OR para cada uma das m categorias, tomando-se a categoria restante como referência, que assumiremos, como anteriormente, ser a categoria $Y = 0$, é uma generalização do OR do modelo de regressão logística. Assim, o *odds ratio* da categoria $Y = j$, com $j = 1,\ldots,m$, *versus* a categoria tomada como referência, $Y = 0$, para valores das covariáveis de $X = a$ *versus* $X = b$ é dada por [Hosmer et al., 2013, p. 273]:

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

Com a notação utilizada em 3.6—3.8, resulta de imediato, para qualquer m, que

$$OR_j(a,b) = e^{f_j(a)-f_j(b)},\ j = 1,\ldots,m.$$

Agora, atendendo-se a 3.12, facilmente se estabelece que a razão de chances para a categoria j, $j = 1,\ldots,m$, quando se aumenta uma unidade na covariável quantitativa X_i, isto é, $X_i = x_i + 1$ *versus* $X_i = x_i$, mantendo-se constante os valores nas restantes covariáveis, será dada por

$$\begin{aligned} OR_j(x_i+1,x_i) &= \frac{\dfrac{P(Y=j|X_1=x_1,\ldots,X_{i-1}=x_{i-1},X_i=x_i+1,X_{i+1}=x_{i+1},\ldots,X_k=x_k)}{P(Y=0|X_1=x_1,\ldots,X_{i-1}=x_{i-1},X_i=x_i+1,X_{i+1}=x_{i+1},\ldots,X_k=x_k)}}{\dfrac{P(Y=j|X_1=x_1,\ldots,X_{i-1}=x_{i-1},X_i=x_i,X_{i+1}=x_{i+1},\ldots,X_k=x_k)}{P(Y=0|X_1=x_1,\ldots,X_{i-1}=x_{i-1},X_i=x_i,X_{i+1}=x_{i+1},\ldots,X_k=x_k)}} \\ &= \frac{e^{\beta_{j0}+\beta_{j1}x_1+\cdots+\beta_{j,i-1}x_{i-1}+\beta_{ji}(x_i+1)+\beta_{j,i+1}x_{i+1}+\cdots+\beta_{jk}x_k}}{e^{\beta_{j0}+\beta_{j1}x_1+\cdots+\beta_{j,i-1}x_{i-1}+\beta_{ji}x_i+\beta_{j,i+1}x_{i+1}+\cdots+\beta_{jk}x_k}} \\ &= e^{\beta_{ji}}, \end{aligned}$$

para $j = 1,\ldots,m$ e $i = 1,\ldots,k$.

Similarmente, obtém-se o mesmo resultado para o *odds ratio* de uma categoria j quando a covariável X_i é uma variável *dummy* ($X_i = 1$ *versus* $X_i = 0$), mantendo-se constante os valores nas restantes covariáveis.

Sempre que seja necessário estimar alguma razão de chances alterando a categoria de referência, facilmente se deduz que o quociente entre duas razões de chances com a mesma categoria de referência é a razão de chances da categoria em numerador em relação à categoria em denominador, que passará a ser a referência. Por exemplo, se OR_2 e OR_1 são as razões de chances para as categorias $Y = 2$ e $Y = 1$, respetivamente, em relação à categoria de referência $Y = 0$, o quociente OR_2/OR_1 é a razão de chances para a categoria $Y = 2$ em relação à referência $Y = 1$.

3.3 Teste de significância sobre os coeficientes do *logit* multinomial

3.3.1 O Teste do Rácio de Verosimilhanças

Apresentaremos agora o teste do rácio de verosimilhanças, generalizado para o modelo multinomial, que, novamente, permite determinar a existência de pelo menos uma covariável a incluir no modelo. Assim, o teste do rácio de verosimilhanças permite testar as hipóteses:

$$\begin{aligned} &H_0 : \beta_{ji} = 0,\ j = 1,\ldots,m,\ \ i = 1,\ldots,k; \\ &H_1 : \exists_i : \beta_{ji} \neq 0,\ i = 1,\ldots,k, \text{ para algum } j = 1,\ldots,m. \end{aligned}$$

Se para um nível de significância α, não se rejeitar a hipótese H_0, então nenhuma das covariáveis permite predizer as probabilidades de ocorrência de cada uma das categorias da variável dependente. Isto é, nenhuma covariável será explicativa da variável dependente.

Como vimos na subsecção 2.6.2, o teste do rácio de verosimilhanças consiste em comparar as funções de verosimilhança de dois modelos, um modelo com as covariáveis incluídas (modelo completo) e um modelo simples (nulo ou reduzido), somente com uma constante (ordenada na

origem do *logit*). Assim, a estatística de teste é dada por [Agresti and Finlay, 2009, p. 493]:

$$G^2 = -2LL_0 - (-2LL_C) = -2\ln\left(\frac{L_0}{L_C}\right), \tag{3.14}$$

onde LL_0 e LL_C representam as transformações logarítmicas das funções de verosimilhança do modelo nulo (L_0) e do modelo completo (L_C), respetivamente.
Sob H_0, G^2 tem distribuição assintótica do qui-quadrado com $k \times m$ graus de liberdade, sendo k o número de covariáveis incluídas no modelo e $m+1$ é o número de categorias da variável dependente. Pois, verifica-se, para o modelo nulo, que

$$-2LL_0 \overset{a}{\sim} \chi^2_{(n-1)\times m}$$

e, para o modelo completo,

$$-2LL_C \overset{a}{\sim} \chi^2_{(n-1-k)\times m}.$$

Deste modo, rejeita-se H_0, se p-value $\approx P(\chi^2_{k\times m} \geq g) \leq \alpha$, para o nível de significância α especificado (por exemplo, $\alpha = 0{,}05$), sendo g o valor da estatística de teste calculado a partir da amostra disponível.
O teste do rácio de verosimilhanças permite ser adaptado de modo a testar-se individualmente o coeficiente de regressão de cada uma das covariáveis, condicionado pelos restantes coeficientes estarem incluídos no modelo. Isto é, para a covariável X_i, $i = 1,\ldots,k$, permite testar as hipóteses:

$$H_0 : \beta_{ji} = 0 \mid \beta_{j0}, \beta_{j1}; \ldots; \beta_{j,i-1}; \beta_{j,i+1}; \ldots; \beta_{jk},\ j = 1,\ldots,m;$$
$$H_1 : \beta_{ji} \neq 0 \mid \beta_{j0}, \beta_{j1}; \ldots; \beta_{j,i-1}; \beta_{j,i+1}; \ldots; \beta_{jk},\ \text{para algum } j = 1,\ldots,m.$$

Para este teste, a estatística G^2 é a diferença entre $-2LL_0$, que inclui todas as covariáveis exceto a covariável em teste (modelo aninhado), e $-2LL_C$, que inclui todas as covariáveis (modelo completo). Novamente, G^2 tem distribuição assintótica do qui-quadrado, mas agora com m graus de liberdade. Observe-se que os m graus de liberdade resultam da diferença $m = (n-1-(k-1)) \times m - (n-1-k) \times m$, onde $(n-1-(k-1)) \times m$ e $(n-1-k) \times m$ são os graus de liberdades da distribuição do qui-quadrado assintótica para $-2LL_0$ e $-2LL_C$, respetivamente.

3.3.2 O Teste de Wald

Como referido no capítulo anterior, o teste de Wald permite determinar se uma covariável é, ou não, significativa para explicar a variável dependente.
Na regressão multinomial, para testar a significância de uma dada covariável com o teste de Wald é necessário generalizar o teste de Wald da regressão logística ao número de categorias da variável dependente. Assim, para se determinar a significância de uma covariável X_i, $i = 1,\ldots,k$, para uma categoria j, $j = 1,\ldots,m$, da variável dependente Y, testam-se as hipóteses:

$$H_0 : \beta_{ji} = 0 \mid \beta_{j0}, \beta_{j1}; \ldots; \beta_{j,i-1}; \beta_{j,i+1}; \ldots; \beta_{jk};$$
$$H_1 : \beta_{ji} \neq 0 \mid \beta_{j0}, \beta_{j1}; \ldots; \beta_{j,i-1}; \beta_{j,i+1}; \ldots; \beta_{jk}.$$

Sob H_0, a estatística de Wald é dada por ([Maroco, 2018, p. 800]):

$$T_j = \frac{\hat{\beta}_{ji}}{\widehat{\sigma}\left(\hat{\beta}_{ji}\right)}.$$

Nessa expressão, $\hat{\beta}_{ji}$ é o estimador de β_{ji} e $\widehat{\sigma}\left(\hat{\beta}_{ji}\right)$ é o estimador do erro padrão de $\hat{\beta}_{ji}$ ([Maroco, 2018, p. 800]). A estatística de Wald tem uma distribuição t de Student, que é assintoticamente com distribuição normal padrão, para amostras de grande dimensão. É usual tomar-se o quadrado da estatística de Wald:

$$T_j^2 = \left(\frac{\hat{\beta}_{ji}}{\widehat{\sigma}\left(\hat{\beta}_{ji}\right)}\right)^2,$$

Consequentemente, esta estatística tem distribuição assintótica do qui-quadrado com 1 grau de liberdade, para amostras de grande dimensão.

Para amostras de pequena dimensão ou quando o coeficiente de regressão em teste é grande, o teste de Wald é menos potente que o teste do rácio de verosimilhança, devido ao facto do erro padrão do respetivo estimador do coeficiente de regressão tender a inflacionar ([Maroco, 2018, p. 801]). Para amostras de grande dimensão os testes de Wald e do rácio de verosimilhança geralmente fornecem resultados semelhantes. No entanto, o teste de Wald é computacionalmente menos exigente que o teste do rácio de verosimilhanças (este último necessita estimar um outro modelo).

Os extremos de um intervalo de confiança para β_{ji} é dado por:

$$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}}\widehat{\sigma}\left(\hat{\beta}_{ji}\right),$$

onde $z_{1-\frac{\alpha}{2}}$ representa o quantil $1-\frac{\alpha}{2}$ da distribuição normal padrão.

3.3.3 Validação do Modelo de Regressão Multinomial

Qualquer modelo ajustado será necessário validá-lo antes de que possa ser usado para fazer inferências. O ajuste geral e a contribuição de cada sujeito deve ser avaliado por pelo menos um método estatístico de qualidade do ajustamento.

Na regressão logística multinomial as 3 ou mais categorias da variável dependente tornam o problema da validação mais difícil do que no caso da regressão logística. Quando modelamos uma variável dependente dicotómica, o modelo ajustado permite estimar um único valor para a probabilidade de ocorrência do acontecimento sucesso para cada sujeito (observação). Quando a variável dependente tem três ou mais categorias, três ou mais probabilidades são estimadas para cada sujeito. E, neste caso, cada sujeito será classificado na categoria da variável dependente onde se obteve a maior probabilidade.

A classificação dos sujeitos da amostra usada na obtenção do modelo é geralmente enviesada a favor das taxas de classificação corretas mais elevadas. Para resolver-se este problema, deveria usar-se uma parte da amostra disponível ($1/2$ a $3/4$ da amostra) para a criação do modelo e a outra parte (ou mesmo uma nova amostra) para a validação do modelo, isto é, para testar o poder classificativo do modelo. Para amostras pequenas deverá usar-se algum método de reamostragem (Jackknife, bootstrap ou validação cruzada).

Para a avaliação do poder classificativo efetuado pelo modelo é usual comparar a percentagem de sujeitos corretamente classificados com o modelo com a percentagem proporcional de classificações corretas por acaso. Esta percentagem é calculada a partir dos números de sujeitos observados, n_i, em cada uma das $m+1$ classes da variável dependente Y pela expressão:

$$\text{Classificação Correta Proporcional por Acaso (\%)} = \sum_{i=0}^{m} \left(\frac{n_i}{n}\right)^2.$$

Se a percentagem de casos classificados corretamente pelo modelo for superior em pelo menos 25% à percentagem de classificação proporcional por acaso, considera-se que o modelo tem boas propriedades classificativas [Maroco, 2018, p. 808].

Outra forma de avaliar a qualidade do modelo logístico multinomial consiste em usarem-se os critérios de informação de Akaike (AIC) e Bayesiano de Schwarz (BIC), que passaremos a descrever a seguir.

3.3.3.1 O critério de informação de Akaike

O critério de informação de Akaike (Akaike's Information Criteria, AIC), desenvolvido e proposto pelo japonês Hirotugu Akaike em 1973, representa uma extensão do logaritmo natural da razão de verosimilhança (Log-likelihood). É uma medida relativa da qualidade do ajuste de um modelo estatístico estimado. Com fortes fundamentos no conceito de entropia, oferece uma medida relativa das informações perdidas, quando um determinado modelo é usado para descrever a realidade.
Se o modelo ajustado explicar melhor os dados do que o modelo nulo, o denominador será maior do que o numerador e a razão de verosimilhança será menor que 1. Quanto menor a razão, maior a evidência contra a hipótese que o modelo não se ajusta aos dados [Gouveia de Oliveira, 2014, p. 181]. Assim, o critério de informação de Akaike (AIC) é uma medida de mediocridade do modelo, isto é, quanto maior o AIC, pior é o ajustamento do modelo. Este é calculado, para o modelo logístico multinomial com $m+1$ categorias e k covariáveis, por:

$$AIC = -2LL + 2(k+1)m.$$

Esta medida penaliza a *deviance*, $-2LL$, em função do número de parâmetros do modelo $(k+1)m$.

3.3.3.2 O Critério de informação Bayesiano de Schwarz

O critério de Informação Bayesiano (Bayesian Information Criterion, BIC), conhecido também por critério de Schwarz, foi proposto por Schwarz em 1978, é um critério de avaliação de modelos definido em termos de probabilidade à posteriori, sendo assim chamado porque, para prová-lo, Schwarz valeu-se do teorema das probabilidades condicionadas proposto por Bayes.
O critério de informação Bayesiano de Schwarz penaliza ainda mais a *deviance* do que o AIC, em função da dimensão da amostra, n:

$$BIC = -2LL + (k+1)m\ln(n).$$

Ambos os critérios definidos anteriormente tentam estimar a inverosimilhança do modelo e a

regra é: quanto menor o AIC (ou o BIC), melhor o ajustamento do modelo. Porém, como não é possível decidir quão pequeno o AIC (ou BIC) tem de ser para que o ajustamento seja bom, estes critérios são utilizados apenas na comparação de diferentes modelos (aninhados ou não). O melhor modelo, entre todos os testados, é o que tiver menor AIC e/ou BIC.

3.4 Exemplo de Aplicação a Dados Angolanos

3.4.1 Introdução

Nesta secção aplicaremos os conhecimentos expostos anteriormente sobre a regressão multinomial e utilizaremos o programa estatístico IBM SPSS, versão 25, na análise de dados provenientes de escolas angolanas. Os dados são os mesmos que foram utilizados como aplicação do modelo de regressão logística, que foram recolhidos em 3 escolas angolanas do ensino secundário, todas da cidade do Luena, município do Moxico, província do Moxico em Angola:

1. 11 de Novembro Periurbana;
2. 338 Tchifuchi - Urbana;
3. 4 de Abril- Suburbana.

Nos resultados deste estudo, que a seguir apresentaremos, todos os testes de hipóteses foram considerados significativos quando o respetivo valor de prova não excedeu o nível de significância de 5% e os intervalos de confiança foram considerados a 95%.

3.4.2 Resultados

Entre os 350 alunos inquiridos, aleatoriamente escolhidos, 170 ($48{,}6\%$) pertenciam à escola 11 de Novembro Periurbana (73 ($42{,}9\%$) do sexo feminino e 97 ($57{,}1\%$) do sexo masculino), 140 ($40{,}0\%$) à escola 338 Tchifuchi - Urbana (52 ($37{,}1\%$) do sexo feminino e 88 ($62{,}9\%$) do sexo masculino) e os restantes 40 à escola 4 de Abril- Suburbana (14 ($35{,}0\%$) do sexo feminino e 26 ($65{,}0\%$) do sexo masculino).
Para esta aplicação foi escolhida como variável dependente a Classificação a Matemática com os dados agrupados em três classes:

- < 10 valores (classificação: insuficiente, representada pelo código 2);
- $[10, 14[$ valores (classificação: suficiente, representada pelo código 1);
- ≥ 14 valores (classificação: bom, representada pelo código 0, que se tomará como referência).

Deste modo, representaremos por OR_1 e OR_2 as razões de chances para as classificações suficiente e insuficiente a Matemática, respetivamente, em relação à classificação bom.
Na amostra, 94 ($26{,}9\%$) alunos obtiveram uma boa classificação a Matemática, 252 ($72{,}0\%$) obtiveram uma classificação suficiente e os restantes 4 ($1{,}1\%$) uma classificação insuficiente. Observe-se que as notas mínima e máxima a Matemática obtidas nas escolas em estudo foram 9 e 17 valores, respetivamente.

As associações, obtidas por regressão logística multinomial, entre a variável dependente anteriormente definida, que representaremos por CM, e cada uma das covariáveis consideradas neste estudo (versão univariada) são dadas na tabela 3.1.

Tabela 3.1: Associações entre a variável dependente e cada covariável considerada no estudo.

	Classificação a Matemática						
	Insuficiente	Suficiente	Bom	OR_1	**Wald**	OR_2	**Wald**
Covariável	N (%)	N (%)	N (%)	(IC 95%)	p	(IC 95%)	p
Escola							
11 de Novembro Periurbana	3 (75,0)	130 (51,6)	37 (39,4)	2,108 (1,009; 4,405)	0,047	–	–
338 Tchifuchi - Urbana	1 (25,0)	97 (38,5)	42 (44,7)	1,386 (0,664; 2,891)	0,385	–	–
4 de Abril - Suburbana	0 (0,0)	25 (9,9)	15 (16,0)	referência	–	referência	
Sexo							
Feminino	3 (75,0)	98 (38,9)	38 (40,4)	0,938 (0,578; 1,521)	0,795	4,421 (0,443; 44,112)	0,205
Masculino	1 (25,0)	154 (61,1)	56 (59,6)	referência		referência	
Estado civil							
Casado/união de facto	2 (50,0)	55 (21,8)	11 (11,7)	2,107 (1,050; 4,227)	0,036	7,545 (0,963; 59,106)	0,054
Solteiro	2 (50,0))	197 (78,2)	83 (88,3)	referência		referência	
Renda familiar							
Baixa	4 (100,0)	101 (40,1)	9 (9,6)	6,317 (3,039; 13,132)	<0,001	–	–
Média	0 (0,0)	151 (59,9)	85 (90,4)	referência		referência	
Situação laboral do aluno							
Não tem emprego	4 (100,0)	235 (93,3)	80 (85,1)	2,419 (1,141; 5,129)	0,021	–	–
Tem emprego	0 (0,0)	17 (6,7)	14 (14,9)	referência		referência	
Classificação atribuída ao professor de Matemática							
Má	2 (50,0)	23 (9,1)	4 (4,3)	2,260 (0,760; 6,717)	0,142	22,500 (2,491; 203,271)	0,006
Boa	2 (50,0)	229 (90,9)	90 (95,7)	referência		referência	
Grau de satisfação escolar				0,338			
Insatisfeito	1 (25,0)	16 (6,3)	6 (6,4)	1,011 (0,381; 2,684)	0,983	6,000 (0,473; 76,141)	0,167
Parcialmente satisfeito	1 (25,0)	46 (18,3)	16 (17,0)	1,089 (0,580; 2,046)	0,790	2,250 (0,192; 26,357)	0,518
Satisfeito	2 (50,0)	190 (75,4)	72 (76,6)	referência		referência	
Idade (em anos)				1,385 (1,243; 1,542)	<0,001	2,144 (1,527; 3,011)	<0,001
Média±DP	29,50±4,36	21,89±3,85	18,99±2,38				
Mediana (mínimo; máximo)	31,5 (23; 32)	21,0 (16; 32)	19,0 (16; 26)				
Notas a Português (0-20)				0,369 (0,297; 0,459)	<0,001	0,110 (0,042; 0,292)	<0,001
Média±DP	10,08±0,70	12,13±1,65	15,19±1,64				
Mediana (mínimo; máximo)	10,0 (9,3; 11)	12 (8; 17,9)	15 (10; 19)				
Notas a Física (0-20)				0,367 (0,296; 0,454)	<0,001	0,227 (0,107; 0,482)	<0,001
Média±DP	10,71±1,21	11,79±1,70	14,87±1,48				
Mediana (mínimo; máximo)	10,2 (10; 12,5)	11,9 (7; 17,6)	15 (10,4; 18)				

Pode observar-se na tabela 3.1 que foram encontradas várias variáveis significativamente associadas à classificação a Matemática, para o nível de significância de 5%. Verificou-se, em particular, que as chances de um aluno ter uma classificação suficiente ou insuficiente aumenta cerca de $38{,}5\%$ ou $114{,}4\%$, respetivamente, com o aumento de 1 ano na sua idade, quando comparada com as chances de um aluno ter uma boa nota a Matemática ($OR_1 = 1{,}385; IC\ 95\% : (1{,}243; 1{,}542)\,; p < 0{,}001$) e ($OR_2 = 2{,}144; IC\ 95\% : (1{,}527; 3{,}011)\,; p < 0{,}001$) respetivamente. Mas, com o aumento de 1 valor na nota a Português, as chances de um aluno ter uma nota suficiente ou insuficiente diminuiu cerca de $63{,}1\%$ ou $89{,}0\%$, respetivamente, quando comparadas com as chances de um aluno ter uma boa nota a Matemática ($OR_1 = 0{,}369; IC\ 95\% : (0{,}297; 0{,}459); p < 0{,}001$ e $OR_2 = 0{,}110; IC\ 95\% : (0{,}042; 0{,}292); p < 0{,}001$). Algo similar foi verificado com o aumento de 1 valor na nota a Física, evidenciando uma diminuição das chances de um aluno ter uma nota suficiente ou insuficiente a Matemática em cerca de $63{,}3\%$ ou $77{,}3\%$, respetivamente, quando comparadas com as chances de ter uma boa nota a Matemática ($OR_1 = 0{,}367; IC\,95\% : (0{,}296; 0{,}454); p < 0{,}001$ e $OR_2 = 0{,}227; IC\,95\% : (0{,}107; 0{,}482); p < 0{,}001$). Em três variáveis categóricas, "Escola", "Renda familiar" e "Situação laboral do aluno", não foi possível estimar OR_2, devido ao facto de existirem frequências nulas em relação à classificação insuficiente a Matemática. No entanto, foi possível encontrar-se chances aproximadas de $2{,}11$ e $1{,}39$ vezes maiores para os alunos das escolas 11 de Novembro Periurbana e 338 Tchifuchi - Urbana, respetivamente, conseguirem uma classificação suficiente a Matemática, quando comparadas com os alunos da escola 4 de Abril - Suburbana, do que conseguirem uma boa nota a Matemática ($OR_1 = 2{,}108; IC\ 95\% : (1{,}009;\ 4{,}405);\ p = 0{,}047$) e

($OR_1 = 1{,}386$; $IC\ 95\% : (0{,}664;\ 2{,}891)$; $p = 0{,}385$) respetivamente. Quando comparadas as escolas 11 de Novembro Periurbana e 338 Tchifuchi - Urbana, as chances para uma classificação suficiente a Matemática é cerca de $1{,}52$ ($2{,}108/1{,}386 \approx 1{,}521$) vezes maior para a primeira, quando comparada com as chances para uma boa classificação a Matemática.

Os alunos casados ou unidos de facto, quando comparados com os alunos solteiros, apresentaram maiores chances de conseguirem uma classificação mais baixa a Matemática do que a obtenção de uma boa classificação, sendo as razões de chances aproximadamente iguais a $2{,}11$ ($OR_1 = 2{,}107; IC\ 95\% : (1{,}050;\ 4{,}227)$; $p = 0{,}036$, para uma classificação suficiente, e $7{,}55$ $OR_2 = 7{,}545;\ IC\ 95\% : (0{,}963;\ 59{,}106)$; $p = 0{,}054$), para uma classificação insuficiente.

Rendimentos familiares mais baixos e alunos sem emprego apresentam maiores chances de conseguirem uma classificação suficiente a Matemática, quando comparadas com as chances de conseguirem uma boa classificação a Matemática, sendo essas razões de chances aproximadamente iguais a $6{,}32$ e $2{,}42$, respetivamente ($OR_1 = 6{,}317;\ IC\ 95\% : (3{,}039;\ 13{,}132)$; $p < 0{,}001$) e ($OR_1 = 2{,}419;\ IC\ 95\% : (1{,}141;\ 5{,}129)$; $p = 0{,}021$), respetivamente.

Estimou-se em cerca de $2{,}26$ e $22{,}5$ vezes maiores as chances dos alunos conseguirem uma classificação suficiente e insuficiente, respetivamente, quando comparadas com as chances de conseguirem uma boa classificação a Matemática, se atribuíram uma má classificação ao professor de Matemática. No entanto, para a classificação suficiente, não se obteve uma significância estatística ($p = 0{,}142$), para o nível de significância de 5% ($OR_1 = 2{,}260;\ IC\ 95\% : (0{,}760;\ 6{,}717)$; $p = 0{,}142$ e $OR_2 = 22{,}5;\ IC\ 95\% : (2{,}491;\ 203{,}271)$; $p = 0{,}006$).

A utilização de um método de seleção de variáveis do tipo *stepwise*, baseado no rácio de verosimilhanças (onde se usou como critério o nível de significância de 5% para a inclusão de uma variável no modelo e 10% para a exclusão), evidenciou que somente as covariáveis classificação atribuída pelo aluno ao professor de Matemática, idade e notas a Português e Física mostraram-se preditivas da classificação a Matemática.

Ao efetuar-se o diagnóstico de multicolinearidade, verifica-se existir um problema de multicolinearidade entre as covariáveis idade e notas a Português e Física, para as quais se observa um índice de condição (condition index) superior a 30 na dimensão 5 e com contribuições para a variância superiores a 50% (ver apêndice B) [Maroco, 2018]. Verifica-se ainda que o modelo com estas três covariáveis quantitativas tende a sobrestimar os valores das razões de chances, em particular da variável categórica classificação atribuída ao professor de Matemática. Com a exclusão de qualquer uma destas três covariáveis quantitativas, o modelo obtido continua a sobrestimar os valores das razões de chances. Assim, é preferível ajustar-se um modelo mais simples.

O modelo mais simples, que minimiza os valores AIC e BIC, inclui somente a nota a Português e a classificação atribuída pelo aluno ao professor de Matemática, sendo as probabilidades estimadas para cada categoria da classificação a Matemática em função destas duas covariáveis dadas por:

$$\hat{P}(\mathsf{CM}=0) = \frac{1}{1 + e^{14{,}535-1{,}001\mathsf{NP}+1{,}038\mathsf{CAP}} + e^{26{,}536-2{,}568\mathsf{NP}+4{,}141\mathsf{CAP}}}, \tag{3.15}$$

$$\hat{P}(\mathsf{CM}=1) = \frac{e^{14{,}535-1{,}001\mathsf{NP}+1{,}038\mathsf{CAP}}}{1 + e^{14{,}535-1{,}001\mathsf{NP}+1{,}038\mathsf{CAP}} + e^{26{,}536-2{,}568\mathsf{NP}+4{,}141\mathsf{CAP}}}, \tag{3.16}$$

$$\hat{P}(\text{CM}=2)=\frac{e^{26{,}536-2{,}568\text{NP}+4{,}141\text{CAP}}}{1+e^{14{,}535-1{,}001\text{NP}+1{,}038\text{CAP}}+e^{26{,}536-2{,}568\text{NP}+4{,}141\text{CAP}}}, \tag{3.17}$$

onde NP representa a nota a Português e CAP representa a variável *dummy* correspondente à categoria "Má" da classificação atribuída pelo aluno ao professor de Matemática. Este modelo de regressão logística multinomial é significativamente melhor que o modelo nulo (teste do rácio de verosimilhanças, $\chi^2(4)=181{,}233$, $p<0{,}001$) e cada uma das covariáveis incluídas também são significativas no modelo (teste do rácio de verosimilhanças, $\chi^2(2)=173{,}832$, $p<0{,}001$, para a nota a Português; teste do rácio de verosimilhanças, $\chi^2(2)=8{,}452$, $p=0{,}015$, para a classificação atribuída pelo aluno ao professor de Matemática). Verificou-se que, com o aumento de 1 valor na nota a Português, as chances de um aluno obter uma classificação suficiente ou insuficiente a matemática são $63{,}3\%$ ou $92{,}3\%$ menores, respetivamente, quando comparado com as chances de obter uma boa classificação a Matemática ($OR_1=0{,}367$; $IC95\%:(0{,}295;0{,}458)$; $p<0{,}001$ e $OR_2=0{,}077$; $IC\,95\%:(0{,}023;0{,}255)$; $p<0{,}001$). A atribuição de uma má classificação ao professor de Matemática aumentou as chances de um aluno ter uma classificação mais baixa a Matemática, sendo as razões de chances aproximadamente iguais a $2{,}82$, para uma classificação suficiente, e $62{,}86$, para uma classificação insuficiente ($OR_1=2{,}823$; $IC95\%:(0{,}686;11{,}608)$; $p=0{,}150$) e ($OR_2=62{,}856$; $IC\ 95\%:(3{,}787;\ 1043{,}169)$; $p=0{,}004$). No entanto, a razão de chances para a classificação suficiente a Matemática não é significativamente diferente de 1 ($p=0{,}150$), para o nível de significância de 5%. O modelo aqui descrito encontra-se resumido na tabela 3.2.

Tabela 3.2: Regressão logística multinomial múltipla para a classificação a Matemática.

	Classificação a Matemática						
Covariável	Insuficiente N (%)	Suficiente N (%)	Bom N (%)	OR_1 (IC 95%)	**Wald** p	OR_2 (IC 95%)	**Wald** p
Classificação atribuída ao professor de Matemática							
Má	2 (50,0)	23 (9,1)	4 (4,3)	2,823 (0,686; 11,608)	0,150	62,856 (3,787; 1043,169)	0,004
Boa	2 (50,0)	229 (90,9)	90 (95,7)	referência		referência	
Notas a Português (0-20)				0,367 (0,295; 0,458)	<0,001	0,077 (0,023; 0,255)	<0,001
Média±DP	10,08±0,70	12,13±1,65	15,19±1,64				
Mediana (mínimo; máximo)	10,0 (9,3; 11)	12 (8; 17,9)	15 (10; 19)				

A maior probabilidade estimada 3.15, 3.16 ou 3.17 permite classificar um aluno na categoria 0 (Bom), 1 (Suficiente) ou 2 (insuficiente) da classificação a Matemática e, com a amostra observada disponível, verifica-se que $84{,}3\%$ dos alunos foram corretamente classificados por este modelo de regressão logística multinomial, em que $62{,}8\%$ dos alunos foram corretamente classificados com Bom, $93{,}7\%$ corretamente classificados com Suficiente e $0{,}0\%$ corretamente classificados com Insuficiente. Verificou-se assim que o modelo de regressão logística multinomial ajustado apresentou boas propriedades classificativas já que classificou $34{,}2\%$ melhor que a percentagem proporcional de classificações corretas por acaso, que é neste caso dada por: $\frac{4^2+252^2+94^2}{350^2}\times 100\%\approx 50{,}1\%$. No entanto, o modelo tem dificuldades em classificar corretamente os alunos com classificação insuficiente a Matemática (ver tabela 3.3).

Na figura 3.1 encontram-se representadas as probabilidades estimadas pelo modelo de regressão logística multinomial para cada aluno da amostra. Verifica-se facilmente que a probabilidade estimada para uma classificação insuficiente a Matemática é sempre inferior à respetiva probabilidade estimada para uma classificação suficiente a Matemática, o que evidencia a dificuldade de o modelo classificar corretamente os alunos com classificação insuficiente a Matemática. Com a resolução das equações:

Tabela 3.3: Classificação a Matemática observada versus predita dos alunos da amostra pelo modelo logístico multinomial ajustado com as covariáveis relação aluno/professores e nota a Português.

	Predita			
Observada	Insuficiente	Suficiente	Bom	% correta
Insuficiente	0	4	0	0,0
Suficiente	0	236	16	93,7
Bom	0	35	59	62,8
% global	0,0	78,6	21,4	84,3

$$e^{14,535-1,001\text{NP}+1,038\times 1} = 1 \tag{3.18}$$

e

$$e^{14,535-1,001\text{NP}+1,038\times 0} = 1, \tag{3.19}$$

que resultam de se igualarem as equações 3.15 e 3.16 para $CAP = 1$ e $CAP = 0$, respetivamente, facilmente se verifica que as notas estimadas a Português que servem de valores de corte de passagem da classificação "suficiente" para a classificação "bom" a Matemática são aproximadamente iguais a 15,6 e 14,5 valores, para os alunos que atribuíram uma "Má" ou "Boa" classificação ao professor de Matemática, respetivamente.

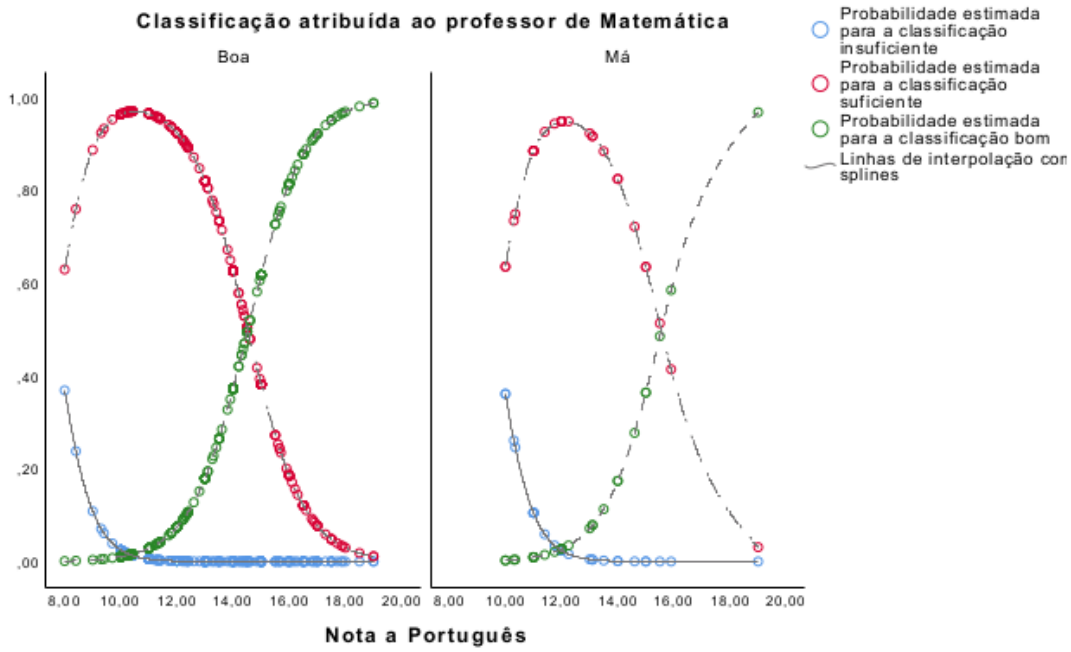

Figura 3.1: Variação das probabilidades estimadas das classificações a Matemática com a nota a Português e para cada uma das categorias da classificação atribuída pelo aluno ao professor de Matemática.

3.4.3 Conclusão

O enquadramento sociodemográficos dos estudantes do ensino secundário angolanos da província do Moxico, tais como a escola, o estado civil, o rendimento familiar, a situação laboral (ser ou não estudante-trabalhador) e a idade poderão ter alguma influência no sucesso a Matemática.

Mas, parece que todos estes fatores são de menor importância na obtenção de melhores notas a Matemática. Os resultados deste estudo evidenciaram que a obtenção de melhores notas a Matemática pelos estudantes angolanos do ensino secundário da província do Moxico dependem principalmente do domínio da língua portuguesa e do professor de Matemática. No entanto, esta conclusão carece de validação. Esta não foi possível efetuar-se no presente estudo, pois requer uma outra amostra de maior dimensão que possibilite ter mais estudantes com classificação "insuficiente" a Matemática.

Capítulo 4

O Modelo de Regressão Logística Ordinal

4.1 Introdução

Em muitas situações de regressão categorial quando a variável dependente assume mais de duas categorias mutuamente exclusivas ordinais, a regressão multinomial não tem em conta essa relação de ordem entre elas. A modelação que consiste na utilização da ordinalidade da variável dependente para fazer inferências, que colmata a insuficiência da regressão multinomial, chama-se regressão logística ordinal ou, simplesmente, regressão ordinal.
Exemplos de variáveis categóricas ordinais encontram-se em pesquisas em epidemiologia, onde se deseja predizer o nível de severidade de uma doença (leve, moderada, grave), pesquisas em educação, onde procura aferir-se o grau de proficiência numa língua (elementar, independente, proficiente), e pesquisas em educação, onde se deseja predizer o grau de soluções de problemas aritméticos (fraco, aceitável, excelente).
A regressão ordinal, à semelhança do modelo de regressão multinomial, é um modelo de probabilidades, mas agora a ocorrência de uma categoria é expressa em termos de probabilidades acumuladas ou cumulativas das várias categorias da variável dependente. Assim, este modelo é também chamado de modelo de probabilidades cumulativas.

4.2 Odds e linearização da função logística ordinal

Dada uma variável dependente Y com as categorias j, $j = 1,2,\cdots,m$, e um conjunto de variáveis explicativas $X = (X_1, X_2, \cdots, X_k)$, define-se a probabilidade comulativa de uma determinada categoria j como:

$$\pi_j(X) = P(Y \leq j|X), j = 1,\ldots,m, \tag{4.1}$$

cuja probabilidade de ocorrência do acontecimento complementar é $P(Y > j|X) = 1-P(Y \leq j|X)$ e, obviamente, $P(Y \leq m|X) = 1$.
Defina-se *odds* como:

$$odds_j = \frac{P(Y \leq j|X)}{1 - P(Y \leq j|X)} = \frac{P(Y \leq j|X)}{P(Y > j|X)} \tag{4.2}$$

Cada um desses *odds* estima as chances de um sujeito pertencer a uma categoria igual ou inferior a j comparativamente a pertencer a uma categoria superior.
Assuma-se que o logaritmo natural de cada *odd*, isto é, o *logit*, seja linear:

$$logit\left(\pi_j(X)\right) = \ln\left(odds_j\right) = \ln\left(\frac{P(Y \leq j|X)}{P(Y > j|X)}\right) = \alpha_j + X\beta, \tag{4.3}$$

onde α_j representa o ponto de corte entre as categorias j, $\beta = [\beta_1\,\beta_2\,\cdots\,\beta_k]'$ e X é a matriz das variáveis independentes, sem a coluna composta pelo elemento geral 1, para que o mo-

delo não seja indeterminado, uma vez que essa coluna é redundante com os valores dos pontos de cortes (α_j) entre as diferentes categorias . Assim, o modelo de regressão logística ordinal tem como pressuposto básico que cada variável independente tem o mesmo efeito sobre as várias categorias da variável dependente, ou seja, para cada variável independente só há um único coeficiente do *logit* para todas as categorias (pressuposto da homogeneidade dos declives) [Liu, 2015, p. 198]. Consequentemente, o modelo logístico ordinal permite determinar as probabilidades acumuladas em todas as categorias da variável dependente. As curvas das probabilidades acumuladas têm exatamente a mesma forma, apenas estão deslocadas horizontalmente em função do parâmetro α_j, isto é, têm a mesma variação da probabilidade em função de X.

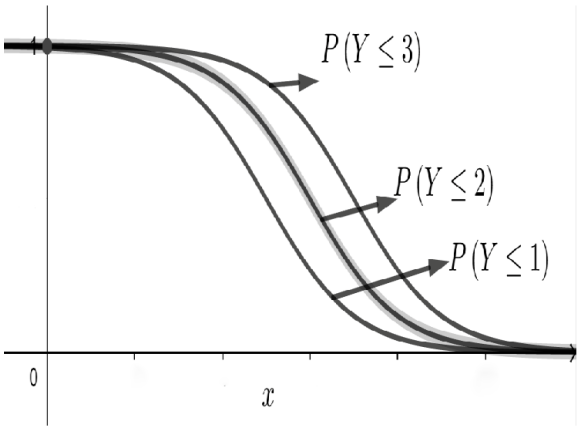

Figura 4.1: Curvas de probabilidades cumulativas para uma variável dependente com 3 categorias.

Da expressão 4.3 verifica-se que a relação entre o sinal de β e a variação de Y é contrária à interpretação geral do sinal de β em regressão. Porque se $\beta > 0$, quando X aumenta, a resposta em Y é mais provável que ocorra para níveis mais baixos da escala ordinal. Similarmente, se $\beta < 0$, então quando X aumenta, a resposta em Y é mais provável que ocorra para níveis mais elevados da escala ordinal. Para que se cumpra a interpretação do sinal de β de tal modo que se $\beta > 0$, quando X aumenta, aumente a probabilidade de Y tomar valores superiores, é necessário reescrever o modelo como [Maroco, 2018, p. 860]:

$$logit\left(\pi_j(X)\right) = \ln\left(\frac{P(Y \leq j|X)}{P(Y > j|X)}\right) = \alpha_j - X\beta \tag{4.4}$$

4.3 Interpretação dos coeficientes do logit ordinal

À semelhança da regressão logística, na regressão ordinal os coeficientes não são diretamente interpretados, mas sim com base no *logit* através da medida conhecida como *odds ratio* (OR), que é uma razão de dois *odds*, isto é, a razão entre os *odds* de um sujeito pertencer a uma categoria igual ou inferior a j *versus* estar acima dessa categoria. Assim, para uma dada categoria j e incremento $d > 0$ num valor x da variável independente X, tem-se que

$$
\begin{aligned}
OR_j &= \frac{P(Y \leq j|X = x+d)/P(Y > j|X = x+d)}{P(Y \leq j|X = x)/P(Y > j|X = x)} \\
&= \frac{\left(\frac{e^{\alpha_j-(x+d)\beta}}{1+e^{\alpha_j-(x+d)\beta}}\right) / \left(1 - \frac{e^{\alpha_j-(x+d)\beta}}{1+e^{\alpha_j-(x+d)\beta}}\right)}{\left(\frac{e^{\alpha_j-x\beta}}{1+e^{\alpha_j-x\beta}}\right) / \left(1 - \frac{e^{\alpha_j-x\beta}}{1+e^{\alpha_j-x\beta}}\right)} \\
&= \frac{\frac{e^{\alpha_j-(x+d)\beta}}{1+e^{\alpha_j-(x+d)\beta}} / \frac{1}{1+e^{\alpha_j-(x+d)\beta}}}{\frac{e^{\alpha_j-x\beta}}{1+e^{\alpha_j-x\beta}} / \frac{1}{1+e^{\alpha_j-x\beta}}} \\
&= e^{-d\beta} \qquad (4.5)
\end{aligned}
$$

Consequentemente, para todas as categorias da variável dependente Y, o *odds ratio* é β-proporcional à distância d entre dois pontos para uma variável independente X. Neste modelo, dos *odds* proporcionais, o rácio das chances acumuladas é igual para todas as categorias da variável dependente.

4.4 Estimação dos coeficientes do logit ordinal

Para um conjunto de observações, $(x_1, \ldots, x_n)$, de uma variável X, a função de verosimilhança na regressão logística ordinal tem a seguinte forma [Agresti, 2013, p. 303]:

$$
\begin{aligned}
L &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{m} [P(Y \leq j|X_i = x_i) - P(Y \leq j-1|X_i = x_i)]^{Y_{ij}} \right\} \\
&= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{m} \left[\frac{e^{\alpha_j - x_i\beta}}{1+e^{\alpha_j-x_i\beta}} - \frac{e^{\alpha_{j-1}-x_i\beta}}{1+e^{\alpha_{j-1}-x_i\beta}} \right]^{Y_{ij}} \right\}, \qquad (4.6)
\end{aligned}
$$

que é uma função dos parâmetros α_j e β, que são estimados pelo método da máxima verosimilhança. As equações resultantes não têm soluções diretas porque são não-lineares. Por esta razão, é necessário a utilização de um método iterativo, como por exemplo, o método de Newton-Raphson. Para facilitar a implementação do método, utiliza-se a transformação logarítmica da função de verosimilhança, LL, que é dada por:

$$
\begin{aligned}
LL &= \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij} [P(Y \leq j|X_i = x_i) - P(Y \leq j-1|X_i = x_i)] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij} \left[\frac{e^{\alpha_j - x_i\beta}}{1+e^{\alpha_j-x_i\beta}} - \frac{e^{\alpha_{j-1}-x_i\beta}}{1+e^{\alpha_{j-1}-x_i\beta}} \right]. \qquad (4.7)
\end{aligned}
$$

4.5 Modelo ordinal de variável latente

A análise de variáveis categóricas é muitas vezes dificultada pela construção de instrumentos com variáveis quantitativas ou qualitativas, cujo objeto de mensuração não é diretamente observável. A existência de uma variável latente ou fantasma, η, que o observador não controla, condiciona a discriminação correta dos indivíduos.

A variável latente, η, assume-se como sendo a função que resulta da combinação linear das variáveis explicativas e do respetivo resíduo:

$$\eta_i = x_i\beta + \varepsilon_i \qquad (i = 1, \cdots, n)$$

A variável dependente Y assume uma determinada categoria j quando a variável latente se encontra entre dois pontos de corte (*thresholds*) α_{j-1} e α_j $(-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_m = +\infty)$. Os pontos de corte $\alpha_1 < \alpha_2 < \cdots < \alpha_{j-1}$ dividem a variável latente contínua, η, em j regiões, que correspondem aos valores da variável observada Y [Fox, 2016, p. 401]. Essa relação é expressa por meio do modelo estrutural que relaciona Y com η através da equação:

$$Y_i = j \qquad se \qquad \alpha_{j-1} \leq \eta \leq \alpha_j$$

Ou de forma geral:

$$Y_i = \begin{cases} 1 & \text{se} \quad \eta_i \leq \alpha_1 \\ 2 & \text{se} \quad \alpha_1 < \eta_i \leq \alpha_2 \\ \vdots & \\ j-1 & \text{se} \quad \alpha_{j-2} < \eta_i \leq \alpha_{j-1} \\ j & \text{se} \quad \alpha_{j-1} < \eta_i \end{cases}$$

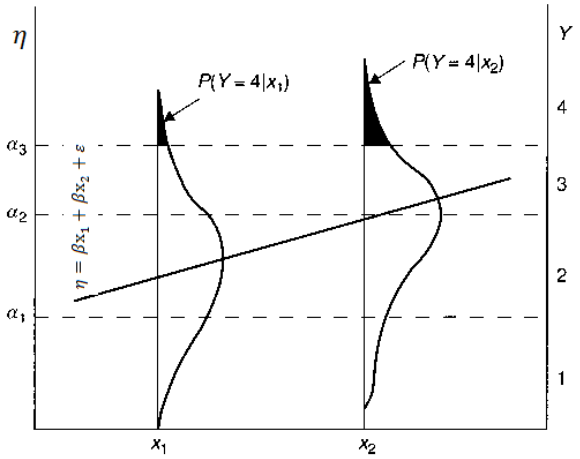

Figura 4.2: Relação entre a variável latente e as categorias da variável dependente com 4 categorias. A. Agresti,2013.p.304

As curvas em sino em torno de cada um dos pontos (x_i, η_i) são as curvas de densidade de probabilidade dos erros do modelo. A figura 4.2 mostra que a probabilidade de se observar uma determinada categoria para um determinado valor de X é dada pela área da curva entre dois pontos de corte.

No modelo ordinal de variável latente, a probabilidade de Y tomar uma determinada categoria j é obtida pela expressão:

$$P(Y = j|X_i = x_i) = P\left(\alpha_{j-1} \leq \eta \leq \alpha_j|X_i = x_i\right). \tag{4.8}$$

Se $\eta_i = \beta_0 + \beta X_i + \varepsilon_i$ e fazendo um arranjo dos termos vem:

$$\begin{aligned} P(Y = j|X_i = x_i) &= P\left(\alpha_{j-1} \leq \beta_0 + \beta x_i + \varepsilon_i \leq \alpha_j | X_i = x_i\right) \\ &= P\left(\alpha_{j-1} - \beta_0 - \beta x_i \leq \varepsilon_i \leq \alpha_j - \beta_0 - \beta x_i | X_i = x_i\right). \end{aligned}$$

Eliminando a constante β_0, que é redundante com α, e representando por F a função de distribuição dos erros do modelo linear da variável latente, tem-se que:

$$P(Y_i = j|X_i = x_i) = F\left(\alpha_j - x_i\beta\right) - F\left(\alpha_{j-1} - x_i\beta\right). \tag{4.9}$$

Logo, para a transformação *logit*, tem-se que:

$$F\left(\alpha_j - x_i\beta\right) = \frac{1}{1 + e^{-\left(\alpha_j - x_i\beta\right)}}, \tag{4.10}$$

de onde resulta que $F(-\infty - x_i\beta) = 0$ e $F(\infty - x_i\beta) = 1$.
De 4.9, a função de verosimilhança, 4.6, pode ser escrita da seguinte forma:

$$L = \prod_{i=1}^{n}\prod_{j=1}^{m} P(Y_i = j|X_i = x_i)^{Y_{ij}} = \prod_{i=1}^{n}\prod_{j=1}^{m}\left[F\left(\alpha_j - x_i\beta\right) - F\left(\alpha_{j-1} - x_i\beta\right)\right]^{Y_{ij}},$$

e a função *LL* por:

$$LL = \sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij} \ln\left[F\left(\alpha_j - x_i\beta\right) - F\left(\alpha_{j-1} - x_i\beta\right)\right] \tag{4.11}$$

4.6 Outras funções de ligação

A função F^{-1} em 4.10 é conhecida por função de ligação (*link function*) *logit*, que é a função inversa da função de distribuição cumulativa, F, dos erros num modelo, que, neste caso, é a função de distribuição logística. Em geral, uma função de ligação faz a ligação linear entre a parte aleatória do modelo ($P(Y \leq j)$) e a sua parte sistemática ($X^*\beta$) [Maroco, 2018, p. 862]:

$$F^{-1}[P(Y \leq j|X)] = \alpha_j - \beta' X \tag{4.12}$$

São várias as funções de ligação que se podem usar para a aproximação de um modelo, em que a sua escolha depende do tipo de resposta e dos objetivos do estudo particular que se pretende desenvolver. A escolha da função de ligação constitui, em geral, o passo inicial no processo de ajustamento de um modelo logístico aos dados. Neste trabalho, até agora, apenas foi usada a função de ligação *logit*.
A escolha de uma função de ligação deve estar de acordo com a forma da distribuição de frequências das categorias da variável dependente. Esta escolha é importante, porque a significância e a capacidade preditiva do modelo adotado pode ser comprometida por uma escolha incorreta da função de ligação.
Na tabela 4.1 estão apresentadas, além da função de ligação $logit$, outras funções de ligação que podem ser usadas quando a variável dependente é categórica ordinal [Maroco, 2018, p. 863].

Tabela 4.1: Funções de ligação que podem ser usadas quando a variável dependente é categórica

Função de Ligação	F^{-1}	Usar quando
Logit	$\ln\left[\frac{P(Y \leq j)}{P(y > j)}\right]$	As classes de Y apresentam distribuição uniforme.
Log-log Complementar	$\ln(-\ln(1 - P[Y \leq j]))$	As classes de Y de maior ordem são as mais frequentes.
Log-log negativo	$-\ln(-\ln(P[Y \leq j]))$	As classes de Y de menor ordem são as mais frequentes.
Cauchit	$\tan\left(\pi P[Y \leq j] - \frac{\pi}{2}\right)$	As classes de Y de menores e de maiores ordens são as mais frequentes.
Probit	$\Phi^{-1}(P[Y \leq j])$, onde Φ é a função de distribuição normal padrão.	A variável latente tem distribuição normal (pressuposto).

As funções de distribuição, associadas a cada função de ligação da tabela anterior, são dadas por:

$$F\left(\alpha_j - x\beta\right) = \frac{1}{1 + e^{-\left(\alpha_j - x\beta\right)}}, j = 1, \ldots, m-1,$$

para a função de ligação *logit*;

$$F\left(\alpha_j - x\beta\right) = 1 - e^{-e^{\left(\alpha_j - x\beta\right)}}, j = 1, \ldots, m-1,$$

para a função de ligação *log-log* complementar;

$$F\left(\alpha_j - x\beta\right) = e^{-e^{-\left(\alpha_j - x\beta\right)}}, j = 1, \ldots, m-1;$$

para a função de ligação *log-log* negativa;

$$F\left(\alpha_j - x\beta\right) = \frac{1}{\pi}\tan^{-1}\left(\alpha_j - x\beta\right) + \frac{1}{2}, j = 1, \ldots, m-1;$$

para a função de ligação de ligação *cauchit*;
e, obviamente, a função de distribuição normal padrão no caso da função de ligação *probit*:

$$F\left(\alpha_j - x\beta\right) = \Phi\left(\alpha_j - x\beta\right), j = 1, \ldots, m.$$

A interpretação dos modelos ordinais, onde não se usa a função de ligação *logit*, que, como sabemos, tem uma medida de efeito associada (*odds ratio*), é feita a partir dos coeficientes de regressão estimados e das probabilidades estimadas para cada categoria.

4.7 Testes à significância do modelo de regressão ordinal

Tendo em conta como foram definidas os *odds* na regressão ordinal, que compara as probabilidades de ocorrer uma categoria inferior ou igual a j com ocorrer uma categoria superior a j, não é mais do que uma extensão da versão dicotómica da regressão categorial, ou seja, da regressão logística. Assim, todas as ferramentas da qualidade do ajuste usadas na regressão logística, como o Pseudo-R^2, a *Deviance*, o teste do rácio de verosimilhança, o critério de informação de Akaike e o Bayesiano de Schwarz, são facilmente generalizados para avaliar a qualidade do ajuste do modelo de regressão ordinal.

4.8 Teste de homogeneidade dos declives ou das linhas paralelas

O teste de homogeneidade dos declives é usado para verificar o pressuposto de que na regressão ordinal a influência das variáveis independentes na função de ligação têm o mesmo efeito sobre todas as categorias da variável dependente Y, ou seja, que as linhas das funções de ligação são paralelas para as m categorias de Y. O teste de homogeneidade dos declives permite testar as hipóteses:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{m-1} = \beta$$
$$H_1 : \exists j : \beta_j \neq \beta_l,\ j \neq l,\ j,l = 1,\cdots,m-1$$

A estatística de teste é o rácio de verosimilhança entre dois modelos, o primeiro assumindo que os declives são iguais (isto é, sob H_0 verdadeira) e o segundo assumindo que os declives possam ser diferentes (isto é, sob H_0 falsa e que $F^{-1}[P(Y_i \leq j)] = \alpha_j - x_i\beta_k$). Os $-2LL$ dos dois modelos são usados para averiguar se o ganho de $-2LL_{H_1}$ (com declives livres) relativamente ao $-2LL_{H_0}$ (com declives homogéneos) é, ou não, significativo. A estatística de teste é então dada por [Maroco, 2018, p. 865]:

$$X^2_{LP} = -2LL_{H_1} - (-2LL_{H_0}) = -2\ln\left[\frac{L_{H_1}}{L_{H_0}}\right] \overset{a}{\sim} \chi^2_k \tag{4.13}$$

Rejeita-se H_0, para um nível de significância α, se o p-value $= P\left(\chi^2_k \geq X^2_{LP}\right) \leq \alpha$, e, neste caso, considera-se que os declives não são homogéneos. Observe-se que os k graus de liberdade da distribuição do qui-quadrado corresponde ao total de variáveis independentes, em que o contributo para esse total de cada variável categórica consiste no número de variáveis *dummy* em que ela é transformada, isto é, o número de categorias menos uma.

Como se referiu anteriormente, a escolha inapropriada da função de ligação pode comprometer a eficiência do modelo e, desse modo, a significância do teste das linhas paralelas pode ser influenciada pela escolha da função de ligação.

4.9 Classificação com o modelo de regressão ordinal

Entre os vários modelos que se podem ajustar a um conjunto de dados, o melhor será aquele que garante um bom equilíbrio entre um bom ajustamento, a parcimónia e a interpretação dos dados.

No modelo logístico ordinal, estima-se a probabilidade de Y estar numa dada categoria ou abaixo relativamente a estar acima dessa categoria j, para se determinar a probabilidade de Y pertencer a uma dada categoria, por exemplo, para se determinar a $P(Y = 2|x_i)$, para um sujeito $i = 1,\ldots,n$, será necessário usar-se a relação $P(Y = 2|X_i = x_i) = P(Y \leq 2|X_i = x_i) - P(Y \leq 1|X_i = x_i)$. Em geral tem-se o seguinte:

$$P(Y = j|X_i = x_i) = P(Y \leq j|X_i = x_i) - P(Y \leq j-1|X_i = x_i), \tag{4.14}$$

ou, por 4.9, a probabilidade de que um sujeito i, $i = 1,\ldots,n$, pertença a uma dada categoria j, $j = 1,\ldots,m$, de Y é dada por:

$$P(Y_i = j|X_i = x_i) = \begin{cases} F(\alpha_1 - \beta x_i) & j = 1 \\ F(\alpha_j - \beta x_i) - F(\alpha_{j-1} - \beta x_i) & 1 < j \leq m-1 \\ 1 - F(\alpha_{j-1} - \beta x_i) & j = m \end{cases}$$

À semelhança da regressão multinomial, na regressão ordinal cada sujeito da amostra será classificado na categoria da variável dependente onde a sua probabilidade de ocorrência é maior.

De modo semelhante à regressão logística e à regressão multinomial, a qualidade do modelo ordinal também pode ser avaliada comparando a percentagem de classificações corretas obtidas pelo modelo, com a percentagem de classificações corretas proporcional por acaso. Se a percentagem de classificações corretas pelo modelo exceder, em pelo menos, 25% a percentagem de classificações corretas proporcional por acaso, considera-se que o modelo apresenta boas propriedades discriminatórias.

Observe-se a representação gráfica geral (figura 4.3) das probabilidades individuais de cada sujeito da amostra estar classificado em alguma das categorias de Y, para se visualizar a evolução das mesmas e os diferentes pontos de interseção entre as várias categorias. Assim, em geral, as estimativas dos pontos de interseção x entre as probabilidades de categorias consecutivas j e $j+1$ de Y, resultam da resolução das equações:

$$P(Y = j|X = x) = P(Y = j+1|X = x), j = 1, \ldots, m-1. \tag{4.15}$$

E, por 4.9, as equações 4.15 são equivalentes a

$$F(\alpha_{j+1} - x\beta) - 2F(\alpha_j - x\beta) + F(\alpha_{j-1} - x\beta) = 0, j = 1, \ldots, m-1.$$

Obviamente, as soluções das equações anteriores dependem da função de distribuição associada à função de ligação adotada no modelo ajustado. Na secção 4.6 estão definidas as funções de distribuição associadas às funções de ligação mais frequentemente utilizadas.

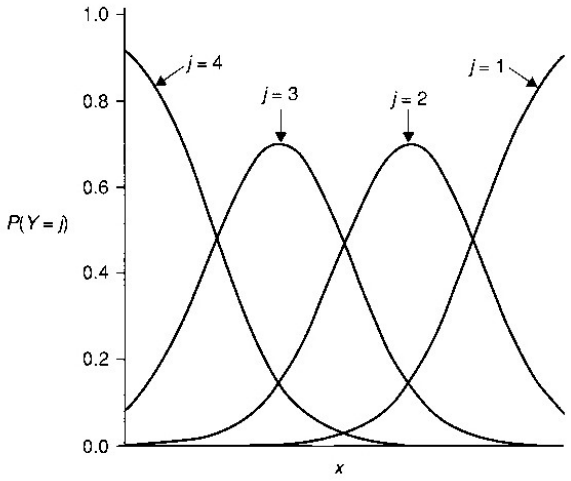

Figura 4.3: Curvas logit cumulativas de probabilidades individuais para uma variável dependente com 4 categorias. (Retirada de [Agresti, 2013, p. 302])

4.10 Aplicação a dados de escolas angolanas

4.10.1 Introdução

Nesta secção aplicaremos os conhecimentos expostos anteriormente sobre a regressão logística ordinal e utilizaremos o programa estatístico IBM SPSS, versão 25, na análise de dados provenientes de escolas angolanas. Os dados utilizados são os mesmos que foram usados como aplicação do modelo de regressão logística e multinomial.

Nos resultados deste estudo, que a seguir apresentaremos, todos os testes de hipóteses foram considerados significativos quando o respetivo valor de prova não excedeu o nível de significância de 5% e os intervalos de confiança foram considerados a 95%.

4.10.2 Resultados

Entre os 350 alunos, aleatoriamente inquiridos, 170 ($48{,}6\%$) pertenciam à escola 11 de Novembro Periurbana (73 ($42{,}9\%$) do sexo feminino e 97 ($57{,}1\%$) do sexo masculino), 140 ($40{,}0\%$) à escola 338 Tchifuchi - Urbana (52 ($37{,}1\%$) do sexo feminino e 88 ($62{,}9\%$) do sexo masculino) e os restantes 40 à escola 4 de Abril- Suburbana (14 ($35{,}0\%$) do sexo feminino e 26 ($65{,}0\%$) do sexo masculino).

Para esta aplicação foi escolhida como variável dependente a classificação a Matemática (CM), que resultou da nota a Matemática agrupada em três classes:

- < 10 valores (classificação: insuficiente);
- $[10, 14[$ valores (classificação: suficiente);
- ≥ 14 valores (classificação: bom).

Na amostra, 94 ($26{,}9\%$) alunos obtiveram uma boa classificação a Matemática, 252 ($72{,}0\%$) obtiveram uma classificação suficiente e os restantes 4 ($1{,}1\%$) uma classificação insuficiente. Observe-se que as notas mínima e máxima a Matemática obtidas nas escolas em estudo foram 9 e 17 valores, respetivamente.

As associações, obtidas por regressão logística ordinal, entre a variável dependente anteriormente definida e cada uma das covariáveis consideradas neste estudo (versão univariada) são dadas na tabela 4.2. Os *odds ratios* (OR) foram estimados para classificações inferiores em comparações com classificações superiores.

Tabela 4.2: Associações entre a classificação a Matemática e cada covariável considerada no estudo e os *odds ratios* estimados para classificações inferiores a Matemática (análise univariada).

	Classificação a Matemática				
	Insuficiente	Suficiente	Bom	*OR*	**Wald**
Covariável	N (%)	N (%)	N (%)	(IC 95%)	p
Escola					
11 de Novembro Periurbana	3 (75,0)	130 (51,6)	37 (39,4)	2,228 (1,074; 4,624)	0,032
338 Tchifuchi - Urbana	1 (25,0)	97 (38,5)	42 (44,7)	1,415 (0,682; 2,933)	0,351
4 de Abril - Suburbana	0 (0,0)	25 (9,9)	15 (16,0)	referência	—
Sexo					
Feminino	3 (75,0)	98 (38,9)	38 (40,4)	1,022 (0,635; 1,646)	0,928
Masculino	1 (25,0)	154 (61,1)	56 (59,6)	referência	
Estado civil					
Casado/união de facto	2 (50,0)	55 (21,8)	11 (11,7)	2,301 (1,167; 4,538)	0,016
Solteiro	2 (50,0))	197 (78,2)	83 (88,3)	referência	
Renda familiar					
Baixa	4 (100,0)	101 (40,1)	9 (9,6)	7,376 (3,553; 15,309)	<0,001
Média	0 (0,0)	151 (59,9)	85 (90,4)	referência	
Situação laboral do aluno					
Não tem emprego	4 (100,0)	235 (93,3)	80 (85,1)	2,492 (1,179; 5,267)	0,017
Tem emprego	0 (0,0)	17 (6,7)	14 (14,9)	referência	
Classificação atribuída ao professor de Matemática					
Má	2 (50,0)	23 (9,1)	4 (4,3)	3,406(1,126; 10,297)	0,030
Boa	2 (50,0)	229 (90,9)	90 (95,7)	referência	
Grau de satisfação escolar					
Insatisfeito	1 (25,0)	16 (6,3)	6 (6,4)	1,218 (0,458; 3,239)	0,692
Parcialmente satisfeito	1 (25,0)	46 (18,3)	16 (17,0)	1,131 (0,609; 2,101)	0,697
Satisfeito	2 (50,0)	190 (75,4)	72 (76,6)	referência	
Idade (em anos)				1,412 (1,277;1,561)	<0,001
Média±DP	29,50±4,36	21,89±3,85	18,99±2,38		
Mediana (mínimo; máximo)	31,5 (23; 32)	21,0 (16; 32)	19,0 (16; 26)		
Notas a Português (0-20)				0,363 (0,293; 0,449)	<0,001
Média±DP	10,08±0,70	12,13±1,65	15,19±1,64		
Mediana (mínimo; máximo)	10,0 (9,3; 11)	12 (8; 17,9)	15 (10; 19)		
Notas a Física (0-20)				0,379 (0,310; 0,462)	<0,001
Média±DP	10,71±1,21	11,79±1,70	14,87±1,48		
Mediana (mínimo; máximo)	10,2 (10; 12,5)	11,9 (7; 17,6)	15 (10,4; 18)		

Entre as 10 covariáveis analisadas, 8 estavam significativamente associadas à classificação a Matemática. Como se pode verificar na tabela 4.2, essas covariáveis são: escola, estado civil, renda familiar, situação laboral do aluno, classificação atribuída ao professor de Matemática, idade, e notas a Português e a Física. Somente o sexo e o grau de satisfação escolar não se mostraram significativamente associados à classificação a Matemática.

Para a Escola 11 de Novembro Periurbana as chances de um aluno apresentar uma classificação inferior foram cerca de 2,2 vezes maiores quando comparadas com as chances de um aluno da Escola 4 de Abril - Suburbana apresentar uma classificação inferior ($OR = 2{,}228$; $IC\ 95\% : (1{,}074;\ 4{,}624)$; $p = 0{,}032$). Similarmente, as chances para um aluno da Escola 338 Tchifuchi - Urbana foram aproximadamente 1,4 vezes maiores, mas não se apresentou significativa ($OR = 1{,}415$; $IC\ 95\% : (0{,}682;\ 2{,}933)$; $p = 0{,}351$).

Em relação ao estado civil, este mostrou-se significativamente associado à classificação a matemática e os alunos casados/união de facto apresentaram aproximadamente chances 2,3 vezes de obterem uma classificação mais baixa a Matemática, quando comparadas com as chances de um aluno solteiro obter uma classificação mais baixa ($OR = 2{,}301$; $IC 95\% : (1{,}167;\ 4{,}538)$; $p = 0{,}016$).

A classificação a Matemática também se mostrou significativamente associada à renda familiar (p<0,001) e à situação laboral do aluno ($p = 0{,}017$), onde se encontraram chances 7,4 ($OR =$

7,376; IC 95% : (3,553; 15,309)) e 2,5 ($OR = 2{,}492$; IC 95% : (1,179; 5,267)) vezes maiores de um aluno com baixo rendimento e sem emprego, respetivamente, obter uma classificação mais baixa a Matemática.

A classificação atribuída ao professor de Matemática também se mostrou significativamente associada à classificação obtida a Matemática ($p = 0{,}030$), estimando-se para os alunos que atribuíram uma má classificação ao Professor de Matemática chances de 3,4 vezes maiores de obter uma classificação mais baixa a Matemática ($OR = 3{,}406$; IC 95% : (1,126; 10,297)).

A classificação a Matemática verificou-se estar significativamente associada às três covariáveis quantitativas, idade, notas a Português e notas a Física ($p < 0{,}001$, para as três associações). Estimou-se que as chances de um aluno apresentar uma classificação mais baixa a Matemática aumentou com o aumento da idade, sendo essas chances de 41,2% por cada ano de idade. Para as notas a Português e a Física, aconteceu o oposto. As chances de um aluno obter uma classificação mais baixa a Matemática diminuíram em 63,7% ($OR = 0{,}363$; IC95% : (0,293; 0,449)) e 62,1% ($OR =$ 0,379; IC 95% : (0,310; 0,462)) com o aumento de um valor nas notas a Português e a Física, respetivamente.

No entanto, a versão ajustada do modelo logístico ordinal com todas as covariáveis disponíveis (versão múltipla), somente a classificação atribuída ao professor de Matemática, a idade e as notas a Português e a Física mostraram poder preditivo da classificação a Matemática, obtendo-se o seguinte modelo ordinal:

$$\hat{P}(CM \leq j|CAP, NP, NF, I) = \frac{1}{1 + e^{\alpha_j + 2{,}021CAP - 1{,}116NP - 1{,}144NF + 0{,}223I}},\ j = 1{,}2$$

onde $\alpha_1 = 13{,}915$ (IC 95% : (7,110; 20,721)), $\alpha_2 = 27{,}284$ (IC 95% : (19,264; 35,304)), NP representa a nota a Português, NF a nota a Física, I a idade e CAP representa a variável *dummy* correspondente à categoria "Má" da classificação atribuída pelo aluno ao professor de Matemática. Este modelo logístico ordinal mostrou-se significativamente melhor que o modelo nulo (teste do rácio de verosimilhanças, $\chi^2_4 = 292{,}795$, $p < 0{,}001$), com uma dimensão de efeito elevado (pseudo-R^2 Nagelkerke=0,785) e verificou-se que o pressuposto das linhas paralelas não foi violado ($p = 0{,}188$). O modelo encontra-se resumido na tabela 4.3.

Tabela 4.3: Regressão logística ordinal múltipla para a classificação a Matemática.

	Classificação a Matemática				
Covariável	Insuficiente N (%)	Suficiente N (%)	Bom N (%)	*OR* (IC 95%)	**Wald** *p*
Classificação atribuída ao professor de Matemática					
Má	2 (50,0)	23 (9,1)	4 (4,3)	7,549(1,718; 33,164)	0,007
Boa	2 (50,0)	229 (90,9)	90 (95,7)	referência	
Idade (em anos)				1,250 (1,066;1,467)	0,006
Média±DP	29,50±4,36	21,89±3,85	18,99±2,38		
Mediana (mínimo; máximo)	31,5 (23; 32)	21,0 (16; 32)	19,0 (16; 26)		
Notas a Português (0-20)				0,328 (0,242; 0,443)	<0,001
Média±DP	10,08±0,70	12,13±1,65	15,19±1,64		
Mediana (mínimo; máximo)	10,0 (9,3; 11)	12 (8; 17,9)	15 (10; 19)		
Notas a Física (0-20)				0,318 (0,229; 0,443)	<0,001
Média±DP	10,71±1,21	11,79±1,70	14,87±1,48		
Mediana (mínimo; máximo)	10,2 (10; 12,5)	11,9 (7; 17,6)	15 (10,4; 18)		

Este modelo ordinal ajustado permitiu estimar que as chances de um aluno obter uma classificação a Matemática mais baixa foi cerca de 7,5 vezes maior se ele atribuiu uma má classificação

ao professor de Matemática ($OR = 7{,}549;\ IC\ 95\% : (1{,}718;\ 33{,}164);\ p = 0{,}007$). Com aumento de 1 ano na idade, as chances de um aluno obter um classificação a Matemática mais baixa aumentam cerca de 25% ($OR = 1{,}250;\ IC\ 95\% : (1{,}066;\ 1{,}467); p = 0{,}006$). Para o aumento de 1 valor nas notas a Português e a Física, as chances de um aluno obter uma classificação a Matemática mais baixa diminuiu cerca de 67,2% ($OR = 0{,}328;\ IC\ 95\% : (0{,}242;\ 0{,}443);\ p < 0{,}001$) e 68,2% ($OR = 0{,}318;\ IC\ 95\% : (0{,}229; 0{,}443);\ p < 0{,}001$), respetivamente.
Verificou-se ainda que o modelo ajustado classificou corretamente 90,6% dos alunos da amostra 4.4. Entre os alunos que tiverem uma classificação de bom, o modelo ordinal classificou corretamente 85,1% e para aqueles que tiveram uma classificação suficiente, 94,0% foram corretamente classificados pelo modelo ordinal. No entanto, o modelo não conseguiu classificar qualquer aluno que teve classificação insuficiente, o que é natural, dado o número reduzido de alunos nessa categoria.
Observe-se ainda que este modelo apresentou boas propriedades classificativas, pois classificou 40,5% melhor que a percentagem proporcional de classificações corretas por acaso, que é neste caso dada por: $\frac{4^2+252^2+94^2}{350^2} \times 100\% \approx 50{,}1\%$.

Tabela 4.4: Classificação a Matemática observada versus predita dos alunos da amostra pelo modelo logístico ordinal ajustado com as covariáveis classificação atribuída ao professor de Matemática, idade e notas a Português e a Física.

	Predita			
Observada	Insuficiente	Suficiente	Bom	% correta
Insuficiente	0	4	0	0,0
Suficiente	0	237	15	94,0
Bom	0	14	80	85,1
% global	0,0	72,9	27,1	90,6

4.10.3 Conclusão

Os resultados obtidos evidenciaram que a classificação a Matemática é essencialmente influenciada pela idade, pelas notas obtidas a Português e a Física e pela percepção que o aluno teve do professor de Matemática e não outros fatores sociodemográficos, como a escola, rendimento familiar, estado civil e situação laboral do aluno.

Bibliografia

[Agresti, 2013] Agresti, A. (2013). *Categorical Data Analysis, Third Edition*. ISBN: 978-0-470-46363-5. John Wiley & SONS, Inc. xiii, 43, 48

[Agresti and Finlay, 2009] Agresti, A. and Finlay, B. (2009). *Statistical Methods for the Social Sciences, Fourth Edition*. ISBN: 987-0-13-027295-9. Pearson Hall, Inc. 11, 32

[Cox and Snell, 1989] Cox, D. R. and Snell, E. J. (1989). *The analysis of binary data, 2nd Edition*. ISBN: 0-412-30620-4. Chapman and Hall. 10

[Dobson and Barnett, 2018] Dobson, A. J. and Barnett, A. G. (2018). *An Introduction to Generalized Linear Models, 4th Edition*. ISBN: 978-1-1387-4168-3. CRC Press. 13

[Fox, 2016] Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models, Third Edition*. ISBN: 978-1-4522-0566-3. SAGE Publications, Inc. 44

[Gouveia de Oliveira, 2014] Gouveia de Oliveira, A. (2014). *Bioestatística Descodificada: Bioestatística, Epidemiologia e Investigação, 2.ª edição*. ISBN: 978-989-752-044-0. LIDEL - Edições Técnicas, Lda. 34

[Hosmer and Lemeshow, 1980] Hosmer, D. W. and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, A10:1043–1069. 14, 15

[Hosmer et al., 2013] Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression, 3nd Edition*. ISBN: 978-0-47-058247-3. Wiley Series in Probability and Statistics, John Wiley & Sons, Inc. 14, 15, 16, 18, 22, 30

[Hu et al., 2006] Hu, B., Shao, J., and Palta, M. (2006). Pseudo-r^2 in logistic regression model. *Statistica Sinica*, 16:691–692. 11

[Liu, 2015] Liu, X. (2015). Applied ordinal logistic regression using stata. page 195. 42

[Maroco, 2018] Maroco, J. (2018). *Análise Estatística Com o SPSS Statistics, 7ª Edição*. ISBN: 978-989-96763-5-0. Gráfica Manuel Barbosa & Filhos: ReportNumber. 5, 7, 10, 11, 12, 16, 17, 18, 19, 21, 28, 33, 34, 37, 42, 45, 47

[McFadden, 1973] McFadden, D. (1973). *Conditional logit analysis of qualitative choice behavior*. Frontiers in Econometrics (Edited by P. Zarembka). 10

[Nagelkerke, 1991] Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78:691–692. 10

[Pregibon, 1981] Pregibon, D. (1981). Logistic regression diagnostics. *Annals os Statistics*, 9:705–724. 16

[Shtatland et al., 2002] Shtatland, E. S., Kleinman, K., and Cain, E. M. (2002). One more time about R^2 measures of fit in logistic regression. In *NESUG 15 - Statistics, Data Analysis & Econometrics*. 10

[Simonoff, 2003] Simonoff, J. S. (2003). *Analyzing Categorical Data*. Springer , New York. 15

[Wayne W. Daniel, 2013] Wayne W. Daniel, C. L. C. (2013). *Biostatistics: A Foundation for Analysis in the Health Sciences, 10th*. ISBN: 978-1-118-30279-8 (cloth). Library of Congress Catologing. 3

Apêndice A

Anexo

CARIMBO DA ESCOLA PARA VALIDAÇÃO DA FICHA DE INQUÉRITO

UNIVERSIDADE DA BEIRA INTERIOR

FICHA DE INQUÉRITO PARA RECOLHA DE DADOS DE ESTUDANTES DE TRÊS ESCOLAS DO 2º CICLO DO ENSINO SECUNDÁRIO E FORMAÇÃO DE PROFESSORES DA CIDADE DO LUENA, PROVÍNCIA DO MOXICO EM ANGOLA, PARA ELABORAÇÃO DA BASE DE DADOS, QUE PODERÁ SER USADA NA PARTE PRÁTICA DA CADEIRA DE PROJECTO DE ENSINO II E DA MONOGRAFIA DO CURSO DO 2º CICLO (MESTRADO) EM MATEMÁTICA PARA PROFESSORES NA UNIVERSIDADE DA BEIRA INTERIOR EM PORTUGAL.

EM PRIMEIRO LUGAR AGRADECEMOS DESDE JÁ PELA SUA COLABORAÇÃO NESTE TRABALHO.
RESPONDA O QUESTIONÁRIO ASSINALANDO EM CADA PERGUNTA APENAS UM CAMPO COM X

DADOS PESSOAIS E SITUAÇÃO ACADÉMICA DO ESTUDANTE

SEXO: MASCULINO ☐ FEMININO ☐ | ESTADO CIVIL: SOLTEIRO ☐ CASADO ☐ VIVE MARITALMENTE COM ALGUÉM ☐ | IDADE: 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 M

RESULTADO OBTIDO NO ANO LECTIVO 2016: APROVADO ☐ REPROVADO ☐ |

MÉDIAS FINAIS OBTIDAS EM MATEMÁTICA, LÍNGUA PORTUGUESA E FÍSICA NO ANO LECTIVO ANTERIOR (2016):

MATEMÁTICA: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

LÍNGUA PORTUGUESA: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

FÍSICA: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

COMO CLASSIFICA O TEU PROFESSOR DE MATEMÁTICA DO ANO PASSADO: BOM ☐ MAU ☐ |

QUAL É O TEU NIVÉL DE SATISFAÇÃO COM A TUA ESCOLA: 0 1 2 3 4 5 6 7 8 9 10

COMO CLASSIFICA AS RELAÇÕES NO AMBIENTE ESCOLAR?

RELAÇÃO ENTRE ALUNOS E PROFESSORES: 0 1 2 3 4 5 6 7 8 9 10

RELAÇÃO ENTRE ALUNOS: 0 1 2 3 4 5 6 7 8 9 10

RELAÇÃO ENTRE ALUNOS E A DIRECÇÃO DA ESCOLA: 0 1 2 3 4 5 6 7 8 9 10

SITUAÇÃO SOCIAL DO ESTUDANTE

CONDIÇÃO ECONÓMICA DA FAMÍLIA (RENDA): ALTA ■ MÉDIA ☐ BAIXA ☐

TEM UM EMPREGO: SIM ☐ NÃO ☐

DISTÂNCIA DA CASA À ESCOLA EM QUILÓMETROS: -1 1 2 3 4 M

ZONA EM QUE HABITA: URBANA ☐ NÃO URBANA ☐

OBS.: ESTES DADOS TÊM FINS MERAMENTE ACADÉMICOS.

MAIS UMA VEZ, OBRIGADO PELA COLABORAÇÃO.

NOTA EXPLICATIVA DA FICHA DE INQUÉRITO

Quanto ao estado civil quem vive maritalmente com alguém deve assinar na última opção, porque neste inquérito é considerado casado que casou oficialmente e solteiro aquele que vive sozinho.

No tocante a idade, os estudantes devem assinalar apenas o número correspondente a sua idade, e caso tenham mais de 30 anos devem escrever os anos que têm.

No que concerne às médias finais obtidas no ano lectivo anterior, a intenção é que o estudante coloque o resultado que foi lançado na pauta final das disciplinas descritas na ficha de inquérito.

Quanto ao nível de satisfação, caso o aluno o considere Muito alto o estudante deve assinalar 9 ou 10, se for Alto deve assinalar 7 ou 8, Se for Médio deve assinalar 5 ou 6, se for suficiente deve assinalar 3 ou 4, se for Baixo deve assinalar 1 ou 2 e se for muito baixo deve assinalar 0 (zero).

Quanto ao relacionamento entre o aluno e professores, entre alunos e entre alunos a direcção da escola, caso o aluno o considere Muito Bom o estudante deve assinalar 9 ou 10, se for Bom deve assinalar 7 ou 8, Se for Medíocre deve assinalar 5 ou 6, se for Mau deve assinalar 3 ou 4, se for muito mau deve assinalar 0 ou 2.

E quanto à distância entre a casa e a escola, caso vivam há menos de 1 quilómetro da escola devem assinalar -1 e os que vivem muito distante da escola ou seja há mais de 4 quilómetros, devem assinalar M.

DESCRIÇÃO DA BASE DE DADOS RESULTANTE DA PESQUISA

Estes dados contêm os resultados de uma pesquisa efetuada em três escolas do ensino secundário da cidade do Luena, município do Moxico, província do Moxico em Angola, durante o mês de Fevereiro de 2017, com a prévia autorização da Direção Provincial de Educação Ciência e Tecnologia do Moxico. Durante a pesquisa, foram selecionados aleatoriamente por meio de rifas dos números de alunos em cada turma, 450 alunos que correspondiam a igual número de observações previstas para a amostra inicial. Porém, pelo facto de muitos alunos não terem preenchido corretamente as fichas de inquérito, vimo-nos obrigados a reduzir o tamanho da amostra durante o processo de tratamento de dados para 350 observações, que representam os 14.007[1] estudantes matriculados no ensino médio na província do Moxico no ano letivo 2017.

A pesquisa foi realizada com o intuito de elaborar uma base de dados para fins académico, concretamente para servir de suporte da abordagem a realizar, na cadeira de projeto de ensino 2 e na Dissertação do curso do 2º ciclo, conducente ao grau de Mestre em Matemática para professores, na Universidade da Beira Interior, em Portugal, nos anos académicos de 2016/2017 e 2017/2018. Eis a descrição dos dados fundamentais que constam na base de dados do ficheiro do SPSS.

Nome da Variável (Name)	Tipo (Type)	Descrição (Label)	Medida (Measure)	Rótulos (Values)
ID	Numeric	Número identificador	Ordinal	
CÓD_ESCOLA	String	Código de escola	Nominal	
ESCOLA_LOCALIZAÇÃO	Numeric	Escola do aluno e localização	Nominal	1 - 11 de Novembro Periurbana 2 – 338 Tchifuchi - Urbana 3 – 4 de Abril- Suburbana
SEXO	Numeric	Sexo do aluno	Nominal	0 – Feminino 1 – Masculino
ESTADO_CIVIL	Numeric	Estado civil	Nominal	0 – Solteiro 1– Casado/União de facto
MATEMÁTICA	Numeric	Notas em Matemática	Scale	
LÍNGUA_PORTUGUESA	Numeric	Notas em Língua Portuguesa	Scale	
FÍSICA	Numeric	Notas em Física	Scale	
IDADE	Numeric	Idade	Scale	
RENDA_FAMILIAR	Numeric	Renda familiar	Nominal	0 – Baixa 1 – Média
SITUAÇÃO_LABORAL	Numeric	Situação laboral do aluno	Nominal	0 – Tem emprego 1– Não tem emprego
RELAÇÃO_ALU_PROF	Numeric	Relação aluno/Professores	Nominal	0 – Má 1 – Boa
SATISFAÇÃO	Numeric	Grau de satisfação com o ambiente escolar	Ordinal	
GRAU_DE_SATISFAÇÃO (*Com base nas 5 categorias desta variável, foram criadas as últimas 4 variáveis dummies da base de dados*)	Numeric	Grau de satisfação com o ambiente escolar (Likert 5)	Ordinal	1 - Muito insatisfeito 2 - Insatisfeito 3 – Parcialmente satisfeito 4 - Satisfeito 5 - Muito satisfeito

Por: Ngaiele M. Fundão

[1] *Fonte dos dados: Direção Provincial da Educação, Ciência e Tecnologia do Moxico – Angola.*

Apêndice B

Anexo

Collinearity Diagnostics[a]

				Variance Proportions										
Model	Dimension	Eigenvalue	Condition Index	(Constant)	Escola do aluno e localização	Sexo do aluno	Estado civil	Notas em Matemática	Notas em Língua Portuguesa	Notas em Física	Idade	Renda familiar	Situação laboral do aluno	Grau de satisfação com ambiente escolar
1	1	9,013	1,000	,00	,00	,00	,00	,00	,00	,00	,00	,00	,00	,00
	2	,917	3,135	,00	,00	,00	,00	,00	,00	,00	,00	,00	,86	,00
	3	,418	4,644	,00	,00	,60	,00	,00	,00	,00	,00	,18	,00	,00
	4	,249	6,020	,00	,03	,38	,00	,00	,00	,00	,01	,54	,04	,00
	5	,167	7,347	,00	,09	,01	,85	,00	,00	,00	,00	,00	,04	,00
	6	,125	8,503	,00	,75	,00	,05	,00	,00	,00	,02	,01	,03	,03
	7	,052	13,190	,00	,03	,00	,04	,01	,02	,02	,01	,11	,01	,54
	8	,040	15,057	,00	,01	,00	,04	,00	,01	,01	,30	,13	,00	,38
	9	,014	25,081	,00	,08	,01	,00	,00	,31	,24	,01	,00	,00	,00
	10	,005	44,630	,87	,01	,00	,01	,00	,12	,10	,58	,00	,01	,03
	11	,002	70,957	,12	,00	,00	,01	,99	,54	,63	,07	,02	,00	,01

a. Dependent Variable: Relação aluno/Professores

Apêndice C

Anexo

CAP=1, se maior ou igual a	Sensibilidade (S)	1 - Especificidade	Especificidade (E)	Distância ao ponto (S=1,1-E=0)
0,000	1,000	1,000	0,000	1,000
0,705	0,966	0,828	0,172	0,828
0,809	0,875	0,483	0,517	0,499
0,844	0,872	0,448	0,552	0,466
0,873	0,854	0,414	0,586	0,439
0,900	0,785	0,310	0,690	0,378
0,927	0,748	0,310	0,690	0,400
0,941	0,636	0,241	0,759	0,437
0,952	0,368	0,103	0,897	0,641
0,962	0,349	0,069	0,931	0,655
0,974	0,315	0,069	0,931	0,689
0,983	0,255	0,069	0,931	0,748
1,000	0,000	0,000	1,000	1,000