

Analise de Regressão Categórica: Uma abordagem na Admissão dos Estudantes na Universidade de São Tomé e Príncipe

Wadiley Sousa do Nascimento ¹

¹ Pós-Graduação em Estatística, Matemática e Computação (ramo: Estatística Computacional)
E-mail: wadmiguel547@yahoo.com

Sumário

Os modelos de regressão logística ordinal têm vindo a ser utilizados na análise de dados cuja variável de resposta se apresenta em categorias com ordenação.

Procurou-se apresentar os aspetos teóricos dos principais modelos de regressão categórica ordinal e bem como as condições de utilização e/ou aplicação.

Na componente prática, foram utilizados dados de 932 candidaturas dos diversos cursos oferecidos pelo ISCSVSM, onde foram consideradas 12 variáveis, com objectivo de analisar a admissão de novos candidatos.

Com recurso as ferramentas do IBM SPSS Statistics (versão 28), o modelo de regressão logística apresentou adequados indicadores de ajuste aos dados, podendo ser tomado como referência nas próximas candidaturas, ou seja, aplicação na admissão de novos alunos.

I. INTRODUÇÃO

Geralmente associa-se causa-efeito a todos fenómeno que nos rodeiam, quer naturais quer artificiais, sem, ao certo ponto, conhecer os factos associados à sua origem ou relação entre estes. Ou, procuramos descobrir as relações entre os factos de maneira a justificar certas atitudes ou acontecimentos.

Neste último, diversos investigadores associaram esta possibilidade a técnica denominada de regressão linear. A técnica de regressão linear oferece a possibilidade de que seja estudada a relação entre uma ou mais variáveis explicativas, que se apresentam na forma linear, e uma variável dependente. ^[1]

Assim, Marôco, J., (2021) e Pestana, M. H., & Gageiro, J. N. (2020) conceituam a regressão como modelo estatístico usado para prever o comportamento de

uma variável a partir de uma ou mais variáveis relevantes de natureza essencialmente quantitativa.

Neste trabalho, pretende-se explorar o tipo de análise em que a variável endógena é qualitativa, chamada de Regressão Categórica com destaque para Regressão Ordinal.

Este estudo apresenta na sua estrutura, duas partes principais, sendo a primeira, a contextualização e fundamentação teórica dos Modelos de Regressão Ordinal, destacando os aspectos relevantes e etapas fundamentais para sua construção, salientando os aspectos que o distingue dos Modelos de Regressão Linear Múltipla e Multivariada.

E na segunda parte é apresentado um caso prática da aplicação do modelo de Regressão Categórica, concretamente a Regressão Logística Binária.

II. MÉTODOS

O tipo de pesquisa utilizado para concepção deste trabalho é um estudo de caso, onde o método científico utilizado é indutivo, com características exploratórias e descritiva onde procura-se a maior familiaridade com o problema em estudo ^[6].

O método estatístico utilizado é de natureza multivariada com vista a entender a relação ou pesos que as notas do teste de aferição e as da candidatura têm na admissão dos candidatos aos cursos oferecidos pelo Instituto Superior de Ciências da Saúde Victor Sá Machado da Universidade de São Tomé e Príncipe (ISCSVSM-USTP).

O conjunto de dados utilizados neste artigo foram obtidos nos Serviços Académicos do referido Instituto, cujo levantamento foi realizado no ano lectivo

2020/2022, por meio de dados das candidaturas feitas nesse período.

Os dados contêm informações de candidaturas feitas em todos os cursos oferecidos no referido ano lectivo.

Todas as análises dos dados foram feitas no software IBM SPSS Statistics (versão 28).

Todavia, antes da apresentação dos resultados tem-se a discussão alguns termos e conceitos da análise de regressão categorial e a sua distinção/relação com outras análises de regressão.

III. ANÁLISE DE REGRESSÃO CATEGORIAL

Os modelos de **regressão** mais conhecidos são aqueles em que a variável endógena é quantitativa. Porém, quando esta não é quantitativa, emergiu a **Regressão Categorial** que carrega consigo os mesmos propósitos da regressão linear que são a inferência e estimação, onde a variável dependente é do tipo qualitativo, pois, são muitas as situações em que se pretende realizar uma análise de regressão, mas a variável dependente é qualitativa com valores de classe discreta e mutuamente exclusivos.^[3]

A Regressão Categorial subdivide-se ou toma designações diferentes segundo as características da variável dependente em análise. Se esta for dicotómica (assume apenas dois valores como 1 - Sucesso, 0 - Insucesso) a regressão categorial é chamada de **Regressão Logística**. Se esta for policotómica (assume mais dois valores como 0 – Não Sabe/Não Responde, 1 - Concordo, 2 - Discordo, ...) a regressão categorial é chamada de **Regressão Multinomial**. Observa-se que esta é uma extensão da Regressão Logística e ambas são aplicadas à variável dependente nominais. E se as classes das variáveis puderem ser ordenadas ou for uma variável medida numa escala ordinal, a regressão categorial é chamada de **Regressão Ordinal**, que será discutido em pormenor, já a seguir.

3.1. Regressão Ordinal

Na análise de dados em diversas áreas de estudo, com frequência nas ciências sociais, as variáveis dependentes apresentam valores discretos e/ou ordenáveis cujas distâncias não são calculáveis nem tampouco constantes, onde Marôco, J. (2021) exemplifica com o estudo de

opinião, que se recorre a escala de Likert ou grau de risco numa aplicação financeira.

Recorda-se que a **Regressão Ordinal** é uma regressão categorial onde a variável explicada é medida numa escala ordinal.

A regressão ordinal apresenta os seguintes modelos:^[3]

3.1.1. Modelos de Regressão Ordinal

1) De forma **Probabilística Cumulativa** não linear, que é dado por:

$$\text{Logit}[P(Y_j \leq k|x)] = \text{Ln} \left(\frac{P(Y_j \leq k|x)}{P(Y_j > k|x)} \right) = \alpha_k + X^* \beta$$

Caso exista apenas uma variável independente, o modelo é dado por:

$$\text{Logit}[P(Y_j \leq k|x)] = \text{Ln} \left(\frac{P(Y_j \leq k|x)}{P(Y_j > k|x)} \right) = \alpha_k + \beta X_j$$

Porém, para $\beta > 0$, quando X aumenta Y diminui. De forma a obter o modelo contrário, ou seja, quando X aumenta Y também aumenta, viu-se necessário reescrever o modelo como mostrado a seguir, salientado ser modelado por diversos softwares de análise estatística.

$$\text{Logit}[P(Y_j \leq k|x)] = \text{Ln} \left(\frac{P(Y_j \leq k|x)}{P(Y_j > k|x)} \right) = \alpha_k - \beta X_j$$

onde α_k é parâmetro da localização para as $k = 1, 2, \dots, K - 1$ classes da variável dependente, β é o vector dos coeficientes de regressão (declives) e X é a matriz das variáveis independente.^[3]

Este autor acrescenta ainda que a variação da probabilidade em função do X é constante em todas as classes, cujo rácio de deslocamento é dado por:

$$OR_k = e^{-d\beta}$$

Onde d é a distância entre os pontos.^[3]

2) Através da operacionalização de uma **Variável Latente** não medida/observada directamente, que é dado por:

$$\text{Link}(P[Y \leq k]) = \alpha_k + X^* \beta$$

Este modelo assume a existência de uma variável latente e que a variável dependente resulta do corte da variável latente em k -classes ordinais e mutuamente exclusivas. Porém, o autor adverte para levar em consideração o tipo

de distribuição das classes da variável dependente, como apresentado no Quadro 1, pois “a escolha inapropriada pode comprometer a significância do modelo e sua capacidade preditiva”.

Link(F⁻¹)		Usar quando
<i>Logit</i>	$Ln\left(\frac{P[Y \leq k]}{P[Y > k]}\right)$	As classes de Y apresentam Distribuição Uniforme
<i>Log-Log Complementar</i>	$Ln(-Ln(1 - P[Y \leq k]))$	As classes de Y de maior ordem são as mais frequentes
<i>Log-Log Negativo</i>	$-Ln(-Ln(1 - P[Y \leq k]))$	As classes de Y de menor ordem são as mais frequentes
<i>Cauchit</i>	$Tan(\pi(P[Y \leq k] - 0,5))$	As classes de Y de maior e menor ordens são as mais frequentes
<i>Probit</i>	$\Phi^{-1}(P[Y \leq k])$ onde Φ função distribuição $N(0, 1)$	A variável latente é do tipo Normal (assunção)

Fonte: Adaptado de Marôco, J. (2021)

Onde F^{-1} é a função de ligação linear entre a parte aleatória do modelo ($P[Y \leq k]$) e a parte sistemática ($X^*\beta$).

3) De Escala

O SPSS Statistics apresenta outra vantagem que permite definir um modelo ordinal mais flexível e realista que, na sua opinião, “permite controlar diferenças de distribuição de probabilidades das classes de variáveis dependente em função das variáveis independentes” [3], dado por:

$$Link(P[Y \leq k]) = \frac{\alpha_k + X^*\beta}{e^{X\gamma}}$$

Onde γ descreve o padrão de dispersão nas observados em X . Se $\gamma = 0$ obtém-se o modelo original definido anteriormente. Se $X\gamma > 0$ o modelo tende de aproximar as probabilidades cumulativas.

3.1.2. Pressuposto do Modelo de Regressão Ordinal

No modelo de regressão ordinal, conforme definidos anteriormente, observa-se que o coeficiente de regressão (β) não apresenta índice k , obrigando a pressuposição de que o efeito da(s) variável(eis) independente(s) sobre o modelo ser igual em todas as classes. Esse efeito é conhecido como **Homogeneidade de Declives** onde a hipótese do teste é definida por: [1]

$$\begin{cases} H_0: \beta_1 = \beta_2 = \dots = \beta_{K-1} = \beta \\ H_1: \exists k, l: \beta_k \neq \beta_l \quad (k \neq l; k, l = 1, 2, \dots, K) \end{cases}$$

E a estatística de teste é dada pelo rácio de verosimilhança de dois modelos ordinais dado por:

$$X_{LP}^2 = -2LL_{H_1} - (-2LL_{H_0}) = -2Ln\left[\frac{L_{H_1}}{L_{H_0}}\right] \sim \chi_{(k-2)p}^2$$

3.1.3. Avaliação e Classificação com Modelo de Regressão Ordinal

A avaliação do modelo é feita comparando os valores de $-2LL$, sendo o melhor aquele que apresenta o menor $-2LL$. [2]

A sua significância é avaliada pelo teste do rácio da verosimilhança e significância prática avaliada pelo teste *pseudo-R²* de MacFadden que é dado por: $R_{MF}^2 = 1 - \frac{LL_c}{LL_0}$, onde LL_c é a verosimilhança do modelo completo ajustado. [3]

Os coeficientes do modelo de Regressão Ordinal, podem ser utilizados para classificar novos casos em estudo, recorrendo ao inverso da função para estimar a probabilidade acumulada de cada uma das classes da variável ordinal conforme descritos pelo Marôco, J. (2021) sendo:

Se a *Link* for *Logit*, tem-se:

$$F(\alpha_k + X^*\beta) = P[Y \leq k] = \frac{1}{1 + e^{-(\alpha_k + X^*\beta)}}$$

Se a *Link* for *Log-log* complementar, tem-se:

$$P[Y \leq k] = 1 - e^{-e^{-(\alpha_k + X^*\beta)}}$$

Se a *Link* for *Log-log* negativa, tem-se:

$$F(\alpha_k + X^*\beta) = P[Y \leq k] = e^{-e^{-(\alpha_k + X^*\beta)}}$$

Para determinar a função densidade da classe k , basta subtrair a classe anterior, pois as funções anteriores são da classe inferior ou igual a k e é dado por:

$$P[Y_j = k] = 1 - F(\alpha_{k-1} + X_j^* \beta)$$

O acrescenta que o caso/sujeito j é classificado na classe k da variável dependente onde a sua probabilidade de ocorrência for maior.

3.2. Regressão Categórica vs Regressão Linear Múltipla vs Regressão Multivariada

Os modelos de Regressão Linear Múltiplas são aqueles em que a variável endógena é quantitativa e a variável preditor é superior a uma, ou seja, mais de uma variável independente. [5]

Observa-se que tanto na regressão Linear Simples como regressão Linear Múltipla, a variável dependente ou endógena é apenas uma ao passo que na regressão Linear Múltipla se configura com uma extensão da primeira com mais de uma variável independente. Já a Regressão Multivariada, que também se configura com extensão da Regressão Linear Múltipla, é aquela em que tem-se mais de uma variável endógena.

Para a sua aplicação e/ou validação são requeridos a satisfação dos pressupostos da Normalidade dos resíduos, ausência de multicolinearidade, heterocedasticidade, autocorrelação e a especificação do modelo. [1]

Observa-se que o Regressão Categórica usa a variável qualitativa, o que não possibilita a verificação dos pressupostos acima mencionado, que de certa maneira enfraquece a análise estatística.

Porém, convém destacar, que a Regressão Categórica tem uma aplicação mais extensa em relação à outros modelos de regressão, pois *i*) aceita preditores ou variáveis independentes do tipo qualitativos e/ou quantitativos, *ii*) não assume relações lineares entre a variável dependente e as covariáveis, *iii*) a normalidade não se constitui um exigência nas variáveis preditoras e *iv*) é menos sensível ao *outliers*. [3]

IV. RESULTADOS

A Universidade de São Tomé e Príncipe (USTP), institucionalizada pelo Decreto-Lei n.º 9/2014 - Estatuto da Universidade de São Tomé e Príncipe (USTP),

publicado no Diário da República n.º 41 de 22 de Maio do mesmo ano, é um estabelecimento público de ensino superior.

Tem como missão criação, difusão e promoção da cultura, ciência e tecnologia, articulando o estudo e a investigação, de modo a potenciar o desenvolvimento humano, como factor estratégico do desenvolvimento sustentável do país.

A USTP ministra o ensino e organiza a investigação científica em torno de grandes áreas científicas, designadamente: Ciências da Natureza, da Vida e do Ambiente; Ciências Humanas, Sociais e Artes; Ciências Exactas, Tecnologias e Engenharias e; Ciências Económicas, Jurídicas e Políticas.

No que se refere a sua estrutura, abrangência e fins, a USTP integra, a Faculdade de Ciências e das Tecnologias (FCT-USTP), o Instituto Superior de Educação e Comunicação (ISEC-USTP) e o Instituto Superior de Ciências de Saúde Victor Sá Machado (ISCSVSM-USTP).

Em cada ano lectivo, a USTP disponibiliza a lista de cursos para ingresso de novos alunos, onde os mesmos procedem a inscrição.

Neste artigo, pretende-se prever a admissão de novos alunos baseados nos dados recolhidos no Instituto Superior de Ciências da Saúde Victor Sá Machado relativo ao processo do ano lectivo 2021/2022.

Todo o processo tem como base as: *i*) Média Final do Ensino Secundário; *ii*) Notas de Língua Portuguesa; *iii*) Notas de Matemática; *iv*) Notas de Física; *v*) Notas de Química; *vi*) Notas de Biologia; *vii*) Notas de Economia; *viii*) Notas de Direito; *ix*) Notas de Contabilidade; *x*) Notas de Geografia; *xi*) Notas de Psicologia e as *xii*) Notas do Teste de Aferição.

Com recurso ao IBM SPSS Statistics (versão 28), realizou-se a análise descritiva dos dados recolhidos onde foram eliminados todos os valores discrepantes, os *outliers*, de forma a evitar uma incorrecta classificação no modelo.

Neste processo, constatou-se que a variabilidades das notas finais do ensino secundário (Média12Ano) dos alunos em torno da média é ligeiramente pequena, com um erro padrão próximo de unidade, Tabela 1.

Observa-se ainda uma média negativa de 5,128 nos Teste de Aferição com variabilidade considerável.

Importa salientar que as notas de Matemática, Física, Química, Biologia, Economia, Direito, Contabilidade, Geografia e Psicologia não têm o mesmo peso em todos os candidatos pois, nem todos os cursos têm as mesmas como requisitos para admissão do candidato. Nisto

observa-se, por exemplo, uma média inferior a unidade nas notas de *Contabilidade* pois, apenas o curso de Gestão e Administração de Serviços de Saúde tem esta como um dos requisitos na admissão do candidato.

No que se refere a simetria e achatamento, verifica-se a predominância de uma distribuição de platicúrtica com assimetria positivas e negativas nas diferentes variáveis.

Tabela 1 - Análise Descritiva dos dados

	N	Média		Desvio padrão	Variância	Assimetria		Curtose	
		Estatística	Erro Padrão			Estatística	Erro Padrão	Estatística	Erro Padrão
Média12Ano	932	11,7912	,02828	,86329	,745	,631	,080	,120	,160
L.Portuguesa	932	10,6731	,03022	,92253	,851	1,452	,080	1,977	,160
Matemática	932	6,4763	,16893	5,15724	26,597	-,414	,080	-1,721	,160
Física	932	1,5119	,12821	3,91406	15,320	2,262	,080	3,310	,160
Química	932	4,3692	,18886	5,76577	33,244	,619	,080	-1,496	,160
Biologia	932	3,9516	,18386	5,61289	31,505	,769	,080	-1,297	,160
Economia	932	2,5805	,16431	5,01630	25,163	1,475	,080	,289	,160
Direito	932	1,7682	,14197	4,33430	18,786	2,093	,080	2,523	,160
Contabilidade	932	,1556	,04305	1,31440	1,728	8,420	,080	69,822	,160
Geografia	932	1,5837	,14037	4,28527	18,364	2,395	,080	3,919	,160
Psicologia	932	2,1659	,15364	4,69044	22,000	1,746	,080	1,162	,160
Teste de Aferição	932	5,1281	,11368	3,47049	12,044	-,468	,080	-1,116	,160

Tendo em conta que a variável endógena “Aprovação” é do tipo qualitativa nominal e dicotómica (1 – “Sim”, 0 – “Não”), o modelo de regressão potencial é a Regressão Categórica, concretamente Regressão Logística.

Nestes termos, a partir dos resultados da tabela seguinte

Variáveis na equação

	B	S.E.	Wald	df	Sig.	Exp(B)
Etapa 0 Constante	-1,266	,079	256,26	1	<,001	,282

Constata-se que a constante do modelo é dada por: $\beta_0 = -1,266$, onde a estimativa da probabilidade de admissão no modelo é $\hat{\pi}_0 = \frac{\sum_{i=1}^n Y_i}{n} = \frac{205}{932} = 0,22$.

O coeficiente é significativo pelo teste de Wald, pois $p\text{-value} < 0,05$, embora Marôco, J. (2021) aconselha à aplicação do teste de rácio de verossimilhança.

Como o teste *omnibus* apresenta $p\text{-value} < 0,05$,

Testes de Omnibus do Modelo de Coeficientes

	Qui-quadrado	df	Sig.
Etapa 1 Etapa	959,728	12	<,001
Bloco	959,728	12	<,001
Modelo	959,728	12	<,001

pode-se afirmar que existe evidências de que existe pelo menos uma variável independente com poder preditivo.

Além disso, o teste de rácio de verossimilhança indica a presença de um ajustamento perfeito do modelo aos

Resumo do modelo

Etapa	Verossimilhança de log -2	R quadrado Cox & Snell	R quadrado Nagelkerke
1	22,327 ^a	,643	,987

a. Estimação finalizada no número de iteração 16 porque as estimativas de parâmetro mudaram foram alteradas para menos de ,001.

dados $p\text{-value} = 1,00$ (usando a função 1 – CDF.CHISQ(22.327, 918) do menu Transform-Compute) corroborado com o teste $R_{MF}^2 = 1 - \frac{-2LL_c}{G^2 + (-2LL_c)} = 0,98$, confirmado pelo pseudo- R^2 de Nagelkerke.

No que respeita aos valores estimados do modelo as estatísticas de Teste de Hosmer e Lemeshow informam

Teste de Hosmer e Lemeshow

Etapa	Qui-quadrado	df	Sig.
1	,009	8	1,000

que os valores estimados pelo modelo estão/são próximos dos valores originais, pois $p\text{-valor} > 0,05$.

Verificou-se, no quadro *Tabela de Classificação* a presença de dois candidatos falso-positivos (candidatos

Tabela de Classificação^a

Observado	Previsto		Porcentagem correta	
	Admissão Não	Admissão Sim		
Etapa 1 Admissão	Não	725	2	99,7
	Sim	2	203	99,0
Porcentagem global				99,6

a. O valor de recorte é ,500

não admitidos e que o modelo afirma que admitiram) e

dois positivo-falsos (candidatos admitidos que o modelo nega).

Embora com os resultados, o modelo apresenta 99,6% (um acréscimo de 21,6% em relação ao modelo nulo) de casos correctamente classificados. Esta classificação é superior em mais de 33,9% em relação a classificação correcta por acaso.

No quadro seguinte tem-se as informações do modelo completo, permitindo a escrita da mesma da seguinte forma:

Variáveis na equação

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. para EXP(B)	
							Inferior	Superior
Etapa 1 ^a Média12Ano	37,282	11,530	10,456	1	,001	1,554E+16	2385512,124	1,013E+26
L.Portuguesa	7,714	2,600	8,800	1	,003	2240,126	13,700	366281,618
Matemática	6,272	1,952	10,321	1	,001	529,461	11,536	24299,669
Física	5,786	1,779	10,577	1	,001	325,661	9,964	10643,611
Química	5,509	1,701	10,490	1	,001	246,916	8,804	6924,931
Biologia	5,330	1,593	11,197	1	<,001	206,486	9,099	4685,874
Economia	4,611	1,427	10,447	1	,001	100,584	6,141	1647,531
Direito	6,911	2,090	10,934	1	<,001	1002,805	16,685	60270,431
Contabilidade	6,393	3,738	2,926	1	,087	597,757	,394	907913,017
Geografia	6,256	1,927	10,537	1	,001	520,889	11,922	22757,720
Psicologia	5,293	1,598	10,973	1	<,001	198,856	8,680	4555,786
Teste de Aferição	39,125	11,614	11,349	1	<,001	9,815E+16	12771439,100	7,543E+26
Constante	-967,142	292,87	10,905	1	<,001	,000		

a. Variável(is) inserida(s) no passo 1: Média12Ano, L.Portuguesa, Matemática, Física, Química, Biologia, Economia, Direito, Contabilidade, Geografia, Psicologia, Teste de Aferição.

$$\begin{aligned} \text{Logit}(\hat{\pi}) = & -967,142 + 37,282 * \text{Média12Ano} + 7,714 * \text{L.Portuguesa} + 6,272 * \text{Matemática} \\ & + 5,786 * \text{Física} + 5,509 * \text{Química} + 5,330 * \text{Biologia} + 4,611 * \text{Economia} \\ & + 6,911 * \text{Direito} + 6,393 * \text{Contabilidade} + 6,256 * \text{Geografia} + 5,293 * \text{Psicologia} \\ & + 39,125 * \text{TestedeAferição} \end{aligned}$$

Observa-se ainda que apenas a variável *Contabilidade* não é significativo no modelo, pois, pelo teste de Wald apresenta *p-value* < 0,05. Neste sentido, o modelo pode ser reajustado apenas com as variáveis significativas, como:

$$\begin{aligned} \text{Logit}(\hat{\pi}) = & -967,142 + 37,282 * \text{Média12Ano} + 7,714 * \text{L.Portuguesa} + 6,272 * \text{Matemática} \\ & + 5,786 * \text{Física} + 5,509 * \text{Química} + 5,330 * \text{Biologia} + 39,125 * \text{TestedeAferição} \\ & + 6,911 * \text{Direito} + 6,256 * \text{Geografia} + 5,293 * \text{Psicologia} + 4,611 * \text{Economia} \end{aligned}$$

Relativamente às chances de aproveitamento, o modelo apresenta elevados rácio das chances de aproveitamento em todas as variáveis.

E em termo das probabilidades estimadas, o modelo apresenta-se da seguinte forma:

$$\hat{\pi} = \frac{1}{1 + e^{-(-967,142 + 37,282 * \text{Média12Ano} + 7,714 * \text{L.Portuguesa} + \dots + 5,293 * \text{Psicologia} + 4,611 * \text{Economia})}}$$

Portanto, com esta função o candidato pode estimar as probabilidades de ser ou não admitidos.

V. CONCLUSÃO

O IBM SPSS Statistics, embora pago, tem confirmado o seu potencial para a análise e interpretação de dados características exploratórias e confirmatórias ou preditivas da população em estudo.

Este trabalho permitiu desenvolver o raciocínio no que respeita a regressão categórica e interligar ao modelo de *workflow* que pressupõe basicamente *i*) carregar os dados; *ii*) limpar os dados; *iii*) transformar, visualizar e modelar; e *iv*) comunicar o resultado. ^[4]

O modelo mostrou-se óptimo, pois a estatística de MacFadden $R_{MF}^2 = 1$.

Uma observação importante, com exceção da variável contabilidade, a totalidade dos coeficientes do modelo final são significativos, pois pelo teste de Wald, apresentam $p\text{-valor} < 0,05$.

Outro ponto importante no modelo é que a Média Final do Ensino Secundário e o resultado do Teste de Aferição têm grande impacto/influência no modelo e contribui para o aumento das chances de aproveitamento do candidato, pois apresentam além dos maiores coeficientes, elevados rácios das chances de admissão dos candidatos.

VI. REFERÊNCIAS

- [1] Fávero, L. P., & Belfiore, P. (2017). *Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®*. Elsevier Brasil.
- [2] Hair, J. F. Jr et al. (2006). *Multivariate Data Analysis*, 6th Edition. Tradução: Adonai Schlup Sant'Anna (2009). Análise multivariada de dados. Porto Alegre: Bookman. ISBN 978-85-7780-534-1
- [3] Marôco, J., (2021). *Análise estatística com o SPSS Statistics*. 8. ed. Lisboa: Edições Sílabo.
- [4] Oliveira, P. D., Guerra, S., & McDonnell, R. (2018). *Ciência de Dados com R: Introdução*. Brasília: IBPAD.
- [5] Pestana, M. H., & Gageiro, J. N. (2020). *Análise de dados para ciências sociais: a complementaridade do SPSS*, 6ª Edição, 2ª Impressão – Lisboa, Outubro.
- [6] Pradanov, C. C e Freitas, E. C. (2013). *Metodologia do trabalho Científico: Métodos e Técnicas da Pesquisa e do Trabalho Académico*. Novo Hamburg Feevalle. 2.ª Edição