UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

ÍTALO LUIGI SANTA CAPITA

REGRESSÃO LOGÍSTICA APLICADA EM LEAGUE OF LEGENDS

PRESIDENTE PRUDENTE
2023

ÍTALO LUIGI SANTA CAPITA

REGRESSÃO LOGÍSTICA APLICADA EM LEAGUE OF LEGENDS

Relatório Final de Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/UNESP para aproveitamento na disciplina Trabalho de Conclusão de Curso.
Orientadora: Profa. Dra. Miriam Rodrigues Silvestre.

PRESIDENTE PRUDENTE
2023

C244r

Capita, Ítalo Luigi Santa
Regressão logística aplicada em league of legends / Ítalo Luigi Santa Capita. -- Presidente Prudente, 2023
41 p. : tabs.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientadora: Miriam Rodrigues Silvestre

1. regressão logística. 2. league of legends. 3. modelos de classificação. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

Ítalo Luigi Santa Capita

REGRESSÃO LOGÍSTICA APLICADA EM LEAGUE OF LEGENDS

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Orientadora: _____

Profa. Dra. Miriam Rodrigues Silvestre
Departamento de Estatística

_____

Prof. Dr. Edilson Ferreira Flores
Departamento de Estatística

_____

Prof. Dr. Fernando Antônio Moala
Departamento de Estatística

Presidente Prudente, 28 de janeiro de 2023.

RESUMO

Analisar os dados dos jogadores e equipes que participaram do Worlds de 2019 requer avaliação do desempenho e identificação de padrões de vitória ou derrota. O objetivo é prever o resultado de uma partida baseando-se nas estatísticas individuais de cada jogador, especificamente o score, através de um modelo de classificação, como o de regressão logística. Este modelo compara diferentes combinações de variáveis, escolhendo o que apresenta o melhor desempenho. A regressão logística estabelece uma relação entre as variáveis independentes e a variável dependente, a vitória ou derrota. A técnica é eficaz para prever resultados em jogos de League of Legends, avaliando a influência de variáveis como o score individual dos jogadores e obtendo resultados precisos e confiáveis. Em resumo, a regressão logística é uma ferramenta valiosa dentro do modelo de classificação para compreender e prever o desempenho em jogos de League of Legends.

Palavras-chave: regressão logística; league of legends; modelos de classificação.

ABSTRACT

Analyzing the data of players and teams that participated in the 2019 Worlds requires performance evaluation and identification of victory or defeat patterns. The objective is to predict the outcome of a match based on the individual statistics of each player, specifically the score. The logistic regression model compares different combinations of variables, choosing the one that presents the best performance. The logistic regression model establishes a relationship between independent variables and the dependent variable, victory or defeat. The technique is effective in predicting results in League of Legends games, evaluating the influence of variables such as individual player score and obtaining accurate and reliable results. In summary, the logistic regression model is a valuable tool for understanding and predicting performance in League of Legends games.

Keywords: logistic regression; league of legends; classification models.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O League of Legends, comumente referido como LOL, é um jogo eletrônico do tipo MOBA (*Multiplayer Online Battle Arena*), desenvolvido pela empresa Riot Games, no ano de 2009. O jogo possui referências ao Dota (*Defense Of The Ancients*), o que faz com que tenham muita semelhança no método de jogabilidade.

Para jogar, o usuário precisa criar uma conta no próprio site do desenvolvedor, de forma totalmente gratuita. Pós criação, inicia-se do nível um, e conforme vai ganhando experiência (ganha-se experiência jogando, e quando vence, o jogador ganha mais pontos em relação a derrota) o jogador eleva o seu nível.

Dentro do jogo, o confronto ocorre entre duas equipes, contendo cada uma cinco jogadores, nas seguintes posições: Topo, Caçador, Meio, Atirador e Suporte. O intuito da partida, é uma das equipes destruir a base (conhecida como Nexus) inimiga da outra, em um tempo indeterminado, pois o jogo só acaba quando essa condição estabelecida é concretizada.

Durante as partidas, os dez jogadores utilizam campeões (nome referido aos personagens) com habilidades e funcionalidades diferentes, para buscar recurso dentro do mapa. Recurso este que o tornará mais forte que o inimigo, aumentando as chances de vitória em relação ao oponente.

O jogo é tão cativante, que a contagem de jogadores do League of Legends atingiu um total de 115 milhões de jogadores mensais em 2021 de acordo com dados recentes e se tornou o número um jogado por horas de jogo (RUNAS, 2021). Com números tão representativos, o cenário competitivo passou a ter mais investimentos, e de acordo com o relatório anual desenvolvido pela especializada SuperData, League of Legends alcançou a expressiva marca de US$ 1.75 bilhão de receita em 2020. Trata-se do único título gratuito (*free-to-play*) e não-móvel que obteve tal marca (MKT ESPORTIVO, 2021).

O League of Legends deixou de ser apenas um jogo distrativo, para potencializar um cenário regado de entretenimento, competitividade e investimentos. O LOL é responsável por gerar uma das competições e-sports mais consolidadas, o Worlds de League of Legends (comumente chamada de Mundial). A competição reúne equipes e espectadores de todo o mundo.

Dentro da competição, as partidas geram em forma de score a performance

individual de cada jogador. Portanto, o estudo em questão tem como objetivo aplicar uma modelagem estatística onde será possível classificar estas pontuações em vitória ou derrota.

Segundo Mesquita (2014), um modelo é uma versão simplificada, ou seja, generalizadora dos aspectos no mundo real. Portanto, são construídos para realizar inferências e encontrar padrões nas variáveis independentes (preditoras), de modo que elas justifiquem o evento ocorrido ou de interesse.

Sendo assim, Gonzalez (2018) complementa que a modelagem de previsão, tem como objetivo a criação de um modelo para a variável dependente, baseado em suas variáveis independentes. Portanto, a regressão logística é uma técnica perfeita para este estudo, dado que queremos prever uma variável dependente categórica (Binária para o estudo específico), indicada por 1 ou 0.

No estudo, o modelo de predição utilizado será a regressão logística, que tem como objetivo usar variáveis independentes para ajustar o melhor modelo que descreve a variável dependente. Batista (2015) ainda diz que a popularidade da regressão logística é devido a aplicação da função logística, uma função matemática que descreve o modelo logístico.

OBJETIVOS

O objetivo geral deste trabalho será obter um modelo de Regressão Logística Binária baseado em scores (pontuações individuais) de jogadores no mundial de League of Legends 2019, que permita calcular as probabilidades de vitória ou derrota em determinada partida, e que apresente o melhor desempenho possível.

Para alcançar o objetivo geral será necessário atingir os seguintes objetivos específicos:

- Construir modelos para as combinações possíveis de variáveis. Seja usando todas as variáveis (abate, mortes e assistências), duas variaveis (abate e assistencias, por exemplo) e apenas o intercepto;
- Comparar os modelos através de métricas e selecionar o de melhor desempenho de classificação.

JUSTIFICATIVA

Vivenciando o constante crescimento tecnológico, e a inclusão que ele nos permite, seja social, profissional, otimização ou produtividade, temos como interesse, verificar se ele atende as expectativas de modelagem, pois o mesmo se prova cada dia mais presente no cotidiano dos usuários e amantes de tecnologia. Portanto, estamos interessados em poder incluir o universo dos e-Sports na estatística.

2 REFERENCIAL TEÓRICO

De acordo com Batista (2015), a expressão “regressão” em estatística significa a dependência funcional entre duas ou mais variáveis aleatórias, correspondendo em termos matemáticos à obtenção de uma função que melhor represente a dependência entre aquelas variáveis.

A regressão logística é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomador por uma variável categórica, frequentemente binária, em função de uma ou mais variáveis independentes contínuas e/ou binárias (GONZALEZ, 2018). Ele ainda completa, dizendo, que a técnica se difere das demais, principalmente pelo fato de sua variável dependente ser categórica, e mesmo quando ela não é dicotômica, é possível torná-la.

A regressão logística é um método amplamente utilizado para classificar objetos em duas categorias. No entanto, ela também pode ser aplicada de outras maneiras. Existem três tipos principais de regressão logística: binominal, ordinal e multinomial.

A regressão logística binominal é um modelo de classificação de objetos em apenas dois grupos ou categorias. Por exemplo, pode ser utilizado para determinar se uma mensagem é spam ou não, se uma imagem é colorida ou não, ou se uma célula é cancerígena ou não. Esta é a abordado no estudo.

A regressão logística ordinal é um modelo de classificação de objetos em três ou mais categorias com uma ordem estabelecida. Por exemplo, pode ser utilizado para classificar o desempenho de um atleta como "ruim", "justo" ou "excelente" ou para classificar o grau de satisfação de um paciente com o tratamento como "insatisfeito", "satisfeito" ou "muito satisfeito".

A regressão logística multinomial é um modelo de classificação de objetos em três ou mais categorias sem ordem estabelecida. Por exemplo, pode ser utilizado para classificar um animal como "gato", "leão" ou "tigre" ou para classificar uma fruta como "maçã", "pera", "manga" ou "maracujá".

Mesquita (2014), observa que a regressão logística, desenvolvida no século XIX, obteve maior visibilidade após 1950 ficando então, mais conhecida. Ainda diz que houve grande avanço a partir dos trabalhos de Cox e Snell (1989) e Hosmer e

Lemeshow (2000).

Cabral (2013), informa que em qualquer regressão a quantidade chave é o valor médio da variável resposta dado o valor da variável independente. Ela ainda complementa dizendo, que está quantidade é chamada valor médio condicional e é expressa como E[Y|X] onde Y representa a variável resposta e X a variável explicativa. Está expressão pode ser interpretada como o valor esperado de Y dado X.

Hosmer e Lemeshow (2000) exemplificam o racional expondo que em um modelo de regressão linear, têm-se que o valor médio condicional, é dado por uma equação linear em x, e pode ser representado por:

$$E[Y|X = x] \; = \; \beta_0 \; + \; \beta_1 x \; . \tag{4.1}$$

Assim, esta expressão poderá assumir valores de menos infinito à mais infinito para qualquer valor que x representar.

Logo, a quantidade que representa a regressão logística é dada por:

$$\pi(x) \; = \; \frac{e^{\beta_0+\beta_1 x}}{1 \; + \; e^{\beta_0+\beta_1 x}} \; . \tag{4.2}$$

Batista (2015), afirma que a popularidade da regressão logística undamenta-se na função logística que descreve a forma matemática, na qual o modelo logístico se baseia. O modelo logit foi inserido na terminologia estatístico médica por Berkson, em 1944, que batizou o referido termo, por analogia com o modelo desenvolvido por Bliss em 1934, designado por probit.

Cabral (2013), diz que o objetivo da aplicação da função logit é linearizar o modelo, aplicando o logaritmo. Logo, temos a seguinte expressão:

$$\begin{aligned} g(x) = \; & ln\left(\frac{\pi(x)}{1 \; - \; \pi(x)}\right) \\ & = \; \beta_0 \; + \; \beta_1 x. \end{aligned} \tag{4.3}$$

onde, a função logit é linear nos parâmetros, é contínua e seus valores são variados em R.

2.1 REGRESSÃO LOGÍSTICA SIMPLES

Na regressão logística, a variável dependente Y corresponde a dois valores, eles são:

$$Y = \begin{cases} 1, para\ sucesso \\ 0, para\ fracasso \end{cases}.$$

Cabral (2013), afirma que em qualquer regressão a quantidade chave é o valor médio da variável resposta dado o valor da variável independente. Está quantidade é chamada valor médio condicional e é expressa como $E[Y|X]$ onde Y representa a variável resposta e X a variável explicativa.

Observa-se que em um modelo de regressão linear, o valor médio condicional, pode ser representado com a seguinte equação linear para x:

$$E[Y|X = x] = \beta_0 + \beta_1 x\ , \qquad (4.1.1)$$

Onde, esse valor médio pode assumir valores dentro do intervalo $[-\infty, +\infty]$.

A notação pode ser simplificada quando se considera $\pi(x) = E[Y|X = x]$.

Conclui-se dizendo que o modelo de regressão logística, pode ser dado por:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}\ . \qquad (4.1.2)$$

2.2 TRANSFORMAÇÃO LOGIT

Batista (2015), afirma que a popularidade da regressão logística se fundamenta na função logística que descreve a forma matemática, na qual o modelo logístico se baseia. O modelo logit foi inserido na terminologia estatístico médica por Berkson, em 1944, que batizou o referido termo, por analogia com o modelo desenvolvido por Bliss em 1934, designado por probit.

Cabral (2013), diz que o objetivo da aplicação da função logit é linearizar o modelo, aplicando o logaritmo. Logo, temos a seguinte expressão:

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

tal que:

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

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

Está é a transformação logit para a quantidade $\pi(x)$.

2.3 AJUSTAMENTO DO MODELO

Por ter como objetivo prever uma variável Y dicotômica, no modelo de regressão logística simples, a variável dependente segue uma distribuição de Bernoulli. A distribuição de Bernoulli é representada por:

$$f(y_i) = \pi(x_i)^{y_i}(1 - \pi(x_i))^{1 - y_i},\ y_i \in \{0, 1\}. \qquad (4.3.1)$$

Para ajustar o modelo, é usado o método de máxima verossimilhança, para estimar os valores de βo e β_1, que até então, são desconhecidos.

Supondo que os dados possuem independência dentre suas observações, ou seja, não há dados repetidos, a estimação pode ser obtida da seguinte forma:

$$L(\beta) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1 - y_i}.$$

Está expressão da log-verossimilhança pode ser representada por:

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

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

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

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

Substituindo, tem-se:

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

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

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

$$= \sum_{i=1}^{n} [y_i\beta_0 + y_i\beta_1 x_i - ln(1 + e^{\beta_0 + \beta_1 x_i})]\ . \qquad (4.3.2)$$

Para obter o valor de β que maximiza $ln[L(\beta)]$, é necessário derivar $l(\beta)$, em relação aos parâmetros β_0 e β_1.

Derivando em função de β_0 , têm-se:

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

$$= \sum_{i=1}^{n}[y_i - \pi(x_i)] \ . \tag{4.3.3}$$

Derivando em função de β_1 , têm-se:

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

$$= \sum_{i=1}^{n} x_i\,[y_i - \pi(x_i)] \ . \tag{4.3.4}$$

Cabral (2013), observa que são equações não lineares nos parâmetros, portanto são necessários métodos iterativos para sua resolução. O método mais utilizado, na maioria dos softwares estatísticos é o Newton-Raphson.

2.4 REGRESSÃO LOGÍSTICA MÚLTIPLA

Em regressão logística simples, o estudo contava apenas com uma variável independente, o que tornava o cenário univariado. Consideremos então, que existe um conjunto de variáveis independentes, onde podem ser denotadas por um vetor $x^t \equiv (x_1, x_2, \ldots, x_p)$.

Logo, a quantidade $E[Y|X) = \pi(x)$ pode ser expressa por:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}} \ . \tag{4.4.1}$$

E seu logit, respectivamente por:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \tag{4.4.2}$$

2.5 AJUSTAMENTO DO MODELO MÚLTIPLO

Para o caso múltiplo, é bem semelhante ao simples, ou seja, utiliza-se o método de máxima verossimilhança, para um vetor de parâmetros $\beta^t = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)$.

Supondo que os dados possuem independência dentre suas observações, ou seja, não há dados repetidos, a estimação pode ser obtida da seguinte forma:

$$L(\beta) = ln[L(\beta)] = \sum_{i=1}^{n} \{y\ ln[\pi(x)]\ (1 - y)\ ln[1 - \pi(x)]\}$$

$$= \cdots$$

$$= \sum_{i=1}^{n} \left[y\ \beta_0 + y\ \beta_1 x_1 + \cdots + y\beta_p x_p - ln(1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p})\right] . \quad (4.5.1)$$

2.6 TESTE DA RAZÃO DE VEROSSIMILHANÇA

Pós ajuste do modelo, é necessário checar a sua significância. Uma técnica para testar a significância é o teste da razão de verossimilhança.

O teste mostrará que o modelo ajustado obteve êxito, se quando testadas simultaneamente, os coeficientes da regressão associados a β são todos nulos, com exceção ao intercepto (β_0).

O teste pode ser expresso por:

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

Onde, o modelo saturado contém todas as variáveis e interações, enquanto o modelo corrente contém apenas as variáveis desejadas para o estudo.

Logo, o teste pode ser representado por:

$$D = -2 \sum_{i=1}^{n} \left[y_i \, ln \left(\frac{\widehat{\pi}_i}{y_i} \right) + (1 - y_i) \, ln \left(\frac{1 - \widehat{\pi}_i}{1 - y_i} \right) \right] . \quad (4.6.1)$$

O teste possui as seguintes hipóteses:

$$\begin{cases} Ho: \beta_1 = \cdots = \beta_p = 0 \\ Ha: Existe \; algum \; \beta \neq 0 \end{cases} .$$

A estatística do teste é definida por:

$$G = -2ln \left[\frac{modelo \; corrente}{modelo \; saturado} \right] \sim sob \; Ho \; \chi^2(p - 1) \; .$$

Ao rejeitar a hipótese nula, assumimos que estatisticamente, existe pelo menos um dos coeficientes diferentes de zero.

2.7 TESTE DE WALD

O teste de Wald testa cada coeficiente de forma particular, diferente do teste de razão de verossimilhança, que testa todos simultaneamente. Este teste é poderoso para checar se o coeficiente específico possui uma relação significativa com a variável dependente.

O teste possui as seguintes hipóteses:

$$\begin{cases} Ho: \beta_j = 0 \\ Ha: \beta_j \neq 0 \end{cases} .$$

A estatística do teste é definida por:

$$W_j = \frac{\widehat{\beta}_j}{var \left(\widehat{\beta}_j \right)} \sim sob \; Ho, \qquad W_j \sim N(0,1). \quad (4.7.1)$$

2.8 AVALIAÇÃO DO MODELO

De acordo com Cabral (2013), em qualquer modelo de regressão, é necessário proceder à análise dos resíduos para validação da qualidade do modelo estimado. Em outras palavras, é de interesse saber a distância entre o modelo observado e os seus valores estimados.

Portanto, temos a técnica de Deviance, que pode ser representada por:

$$d(y,\hat{\pi}) = dj \pm \left\{2\left[y\, ln\left(\frac{y}{\hat{\pi}}\right) + (1 - y)\, ln\left(\frac{1-y}{1-\hat{\pi}}\right)\right]\right\}^{1/2}. \quad (4.8.1)$$

O teste estatístico é dado por:

$$D = \sum_{i=1}^{n} d(y,\hat{\pi})^2 \sim sob\ Ho\ \chi^2(n - p - 1)\,. \quad (4.8.2)$$

De acordo com Mesquita (2014), a deviance sempre é positiva e quanto menor, melhor é o ajuste do modelo. Para estimar a significância de uma variável independente, comparam-se o valor de D com e sem a variável independente na equação.

Outro teste, é o de Hosmer e Lemeshow, que responde por meio de hipóteses se o modelo encontrado é explicativo para os dados. Hosmer e Lemeshow (2000), apontam que está estatística corresponde a um teste qui-quadrado que consiste em dividir o numero de observações em dez classes, e comparar as frequências preditas com as observadas.

Logo, o teste possui as seguintes hipóteses:

$$\begin{cases} Ho\text{: } O\ modelo\ encontrado\ explica\ bem\ os\ dados \\ Ha\text{: } O\ modelo\ não\ explica\ bem\ os\ dados \end{cases}.$$

A estatística do teste é dada por:

$$\chi 2\ HL = \sum_{k=1}^{g} \frac{(o_k - n'k\bar{\pi}k)}{n'k\bar{\pi}k\ (1 - \bar{\pi}k)} \sim sob\ Ho\ \chi^2\ (g-2)\ , \quad (4.8.3)$$

em que,

$n'k$ é o número de indivíduos no *k*-ésimo grupo;

$c'k$ é o número de covariáveis padrão no *k*-ésimo grupo;

$o_k = \sum_{i=1}^{c'k} y$ é o número de respostas ao longo de $c'k$ classes de variáveis

$$\bar{\pi}k = \sum_{i=1}^{c'k} \frac{\bar{\pi}\mu}{n'i} .$$

3 METODOLOGIA

O estudo em questão começou com o entendimentos dos dados, ou seja, quais variáveis farão parte da pesquisa, como se comportam, e se atendem os seguintes pressupostos:

- análise exploratória: entendimento dos dados, para checar se há discrepâncias ou incoerências no estudo;
- amostral: se o espaço amostral é adequado para a análise de Regressão Logística;
- outliers: descobrir se há valores extremos que possam interferir em resultados

independência dos dados e resíduos: os dados devem e resíduo devem ser independentes;

- multicolinearidade: não existencia de correlação entre as variáveis independentes;
- variáveis independentes estão linearmente relacionadas ao log das probabilidades: linearidade dos dados.

Pós validação dos tópicos acima, será feito a formalização dos modelos que vão aderir à aplicação da Regressão Logística binária para o conjunto de dados, e checada a sua performance por:

- Teste de Wald e Odds Ratio.

Dado estes passos, é crucial a comparação entre os modelos, para entender qual teve a melhor performance, e ajustou-se melhor nos dados.

4 ANÁLISE EXPLORATÓRIA

O presente estudo extraiu dados do Worlds (Campeonato Mundial de League of Legends) no ano de 2019. Devido à complexidade da análise será utilizado o programa R para extrair todos os dados necessários, tanto nas hipóteses da construção do modelo quanto no resultado das razões de chances. Esse banco de dados contém em torno de 770 observações dividido em oito colunas sendo elas:

- Id: Número de registro;
- Rota: Posição em que ele atua no mapa;
- Abate: Quantidade de abates que o jogador teve no jogo;
- Morte: Quantidade de mortes que o jogador teve no jogo;
- Ass: Quantidade de assistências que o jogador teve no jogo;
- Nome: Nome do jogador;
- Time: Nome do time em que o jogador atua;
- Default: Resultado da partida.

Dentre essas colunas, 4 são variáveis numéricas, 2 de textos e uma variável dicotômica que é exatamente nossa variável dependente (VD). Nas Tabelas 1 e 2 são apresentados resumos descritivos das variáveis quantitativas e categóricas, respectivamente.

Tabela 1 – Estatísticas descritivas das variáveis quantitativas

Descrição	rota	abate	morte	ass
Min.	1	0,000	0,000	0,000
1st Qu	2	1,000	1,000	3,000
Median	3	2,000	2,000	5,000
Mean	3	2,694	2,696	5,891
3rd Qu	4	4,000	4,000	8,000
Max.	5	12,000	10,000	22,000
Length	770	770	770	770

Fonte: Elaborado pelo autor.

Tabela 2 - Estatísticas descritivas das variáveis categóricas

Descrição	nome	time	default
Length	770	770	770
Class	character	character	-
Mode	character	character	-
GANHOU	-	-	385
PERDEU	-	-	385

Fonte: Elaborado pelo autor.

Será utilizado como variável depende a coluna default para realizar a regressão logística binaria. Para facilitar a análise no R foi criado uma coluna a mais contendo o texto Ganhou ou Perdeu, sendo assim a tabela informa qual o resultado temos como sucesso e qual resultado iremos atribuir como fracasso. Na Tabela 3, podemos ver como será o reconhecimento da ocorrência do evento.

Tabela 3 - Descrição da variável de ocorrência

Situação	Descrição	Nº
Sucesso	Ganhou	1
Fracasso	Perdeu	0

Fonte: Elaborado pelo autor.

Como variável independente (VI), será testado a coluna abate, morte e ass. Nessa análise partindo pela teoria que tanto abate quanto ass influenciam positivamente para uma chance maior de vitória e morte negativamente, visto que, se o outro time possui mais mortes demostra que ele tem mais chances devido ao fato de estar perdendo no controle do jogo. Mas essa teoria será testada na prática e verificada se esse pressuposto inicial será verificado na análise.

Tendo em vista isso, foram utilizados 7 modelos nessa análise, objetivando utilizar o melhor modelo para análise. Cada modelo terá que atender os pressupostos para que as informações contidas no resultado não estejam em desacordo com a realidade. Na Tabela 4 abaixo mostra esses modelos e a diferença em cada um deles.

Tabela 4 - Combinação de variáveis dos Modelos

Modelos	Dependente	Independente
Modelo 1	default	abate + morte + ass
Modelo 2	default	abate + ass
Modelo 3	default	abate + morte
Modelo 4	default	morte + ass
Modelo 5	default	morte
Modelo 6	default	ass
Modelo 7	default	abate

Fonte: Elaborado pelo autor.

A variação desse modelo ocorre com a utilização de uma ou mais variáveis independentes no modelo tendo em vista que existe um resultado esperado que é de forma binária.

4.1 AMOSTRA

Para uma regressão logística se faz necessários uma quantidade de dados razoáveis para que a construção da regressão seja a mais próxima possível da realidade, não se tem uma quantidade exata de dados necessários, mas existe um mínimo de pelo menos 20 amostras de cada dos eventos analisados para que se faça a regressão logística (VIEIRA, 2011). Na Tabela 5, podemos ver a parte amostral dos modelos.

Tabela 5 - Quantidade amostral para cada Modelo

Modelos	Dependente	Independente	Dados utilizados na análise
Modelo 1	default	abate + morte + ass	770
Modelo 2	default	abate + ass	770
Modelo 3	default	abate + morte	770
Modelo 4	default	morte + ass	770
Modelo 5	default	morte	770
Modelo 6	default	ass	770
Modelo 7	default	abate	770

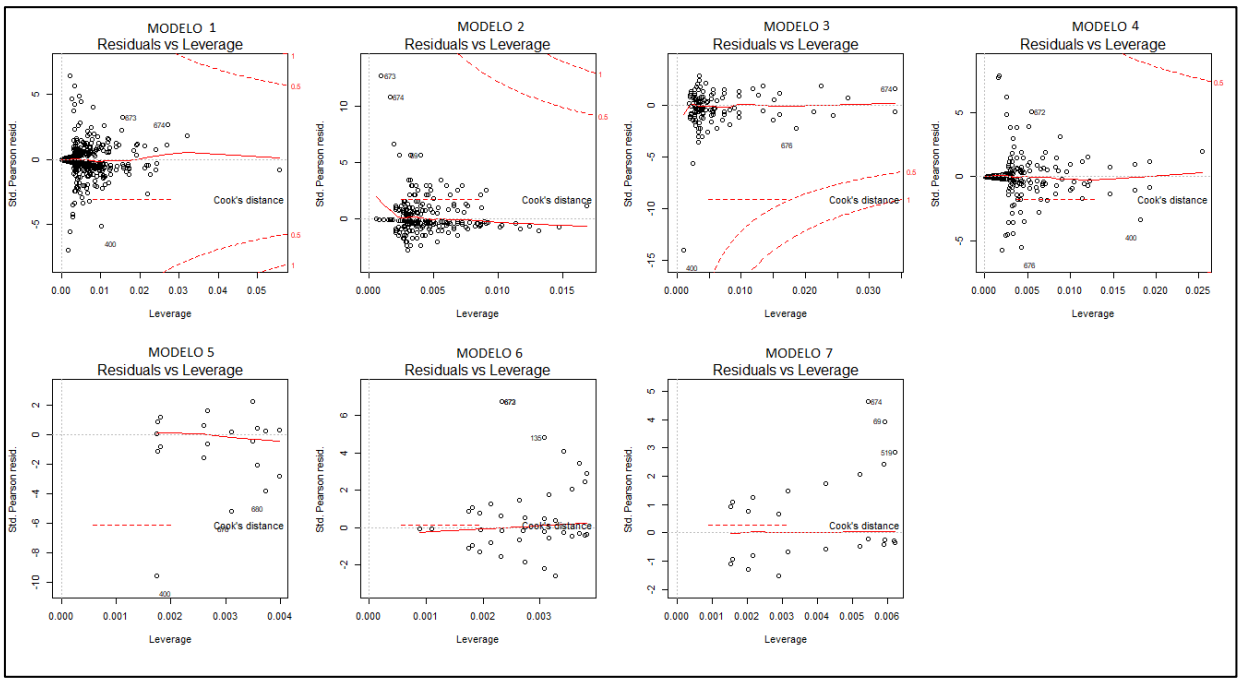
Fonte: Elaborado pelo autor.

Levando em conta que nossos dados possuem 770 observações conforme tabela acima, observa se que o primeiro pressuposto foi atendido. A tabela também informa que em todos os modelos apresentados não se verificou a necessidade de retirar nenhum dado do modelo.

4.2 OUTLIERS

Buscando evitar que o modelo seja influenciado por variáveis que estão fora do normal, se torna necessário analisar se existe outliers em cada modelo. Com isso, foram elaborados gráficos que verificam a influência de cada resíduo no modelo. Se existir resíduos dentro das linhas pontilhadas na parte esquerda como mostrado abaixo na Figura 1, que não existem resíduos no modelo.

Figura 1 - Teste de Leverage para detecção de outliers



Fonte: Elaborado pelo autor.

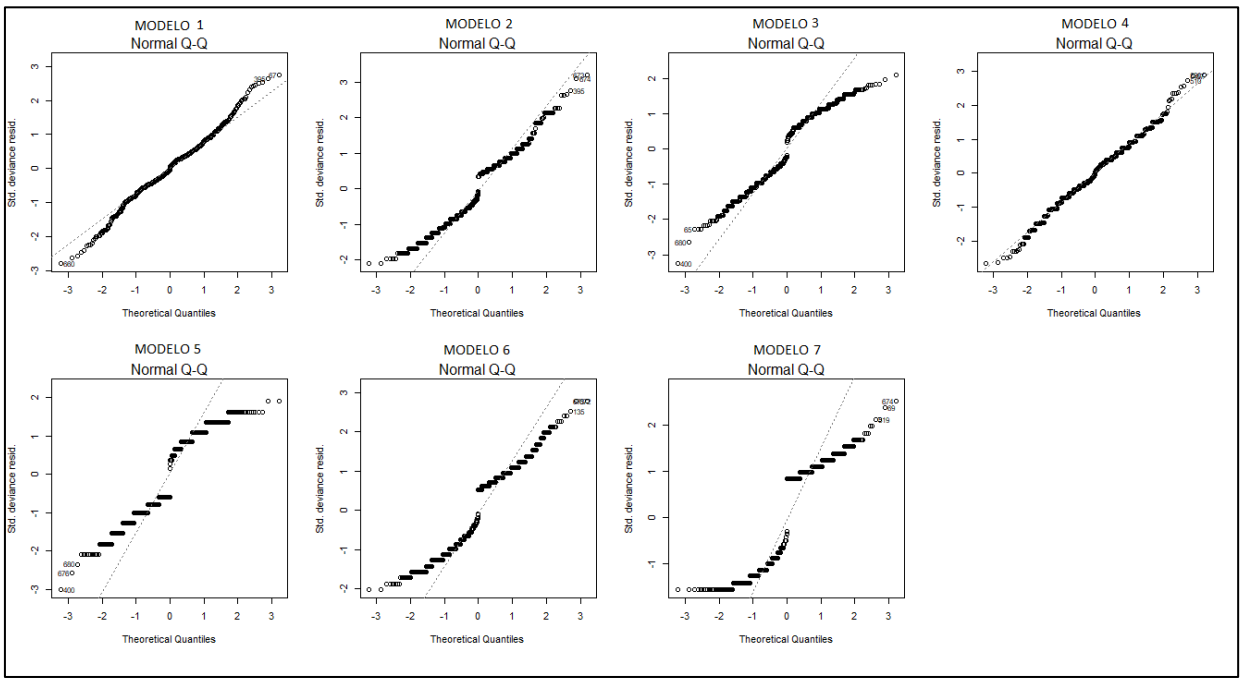
Visto que, na Figura 1, nenhum dos modelos analisados indicou a presença de outliers, logo fica evidentemente que o pressuposto de não outliers nos modelos foi atendida.

4.3 INDEPENDÊNCIA DOS DADOS E RESÍDUOS

Independência dos dados é um pressuposto base para diversos testes que se refere de não haver informações duplicadas ou repetidas, como o estudo trata de eventos diferentes dado a um número de registo especifico do Worlds, assim, o modelo atende esse pressuposto inicial.

Para analisar a independência dos resíduos se torna necessário analisar a informação graficamente para uma melhor interpretação dos dados. A Figura 2 avalia o quantil de cada modelo e analise se os resíduos estão distribuídos normalmente. A independência dos resíduos se verifica quanto mais perto os resíduos estiver da linha tracejada.

Figura 2 - QQ-Plot para detecção de outliers



Fonte: Elaborado pelo autor.

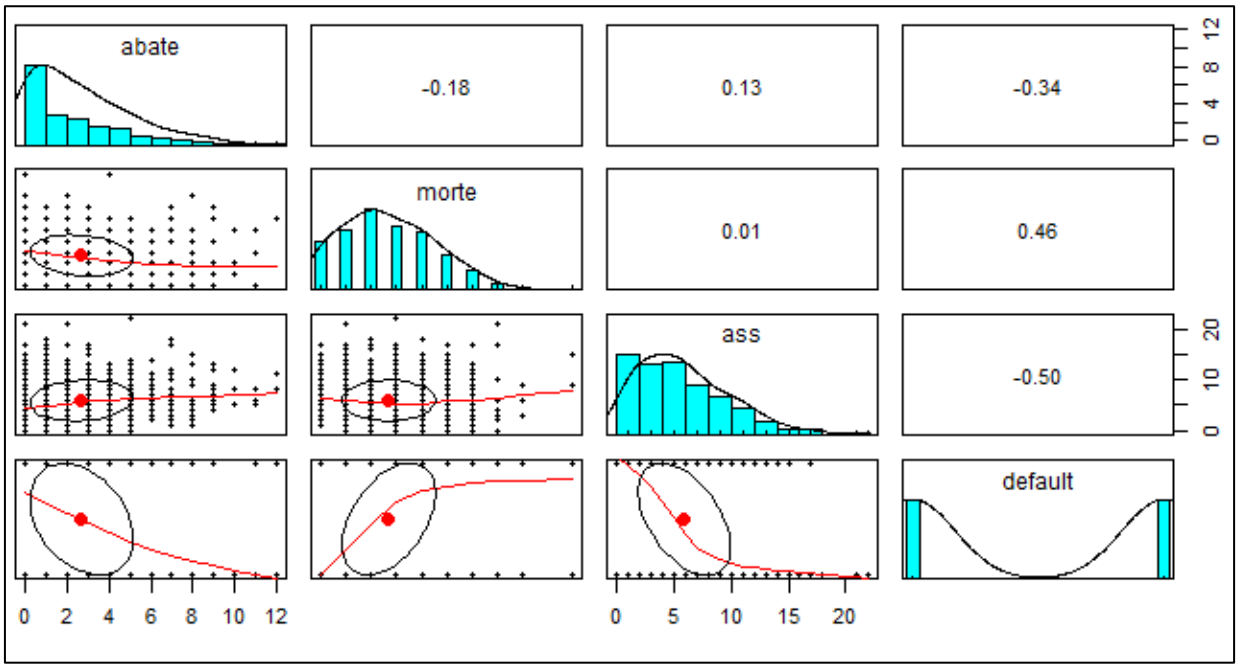
Sabendo disso, se verifica que o Modelo 1 se percebe uma distribuição normal entre os pontos de -2 e 2 quando se verifica o intervalo até -3 ou 3, se observa que os resíduos começam a se distanciar da linha pontilhada, mas até então nada que influencia o modelo. O Modelo 4 também se verifica uma normalidade até melhor do que a do Modelo 1. O Modelo 2, 3, 5, 6 e 7. Indicam um grau de dispersão maior, o que gera dúvidas referente a independia dos dados nesses modelos. Logo os Modelos 1 e 4 atendem os critérios de independência já os

2, 3, 5, 6 e 7 não atenderam esses critérios portanto podem influenciar na análise.

4.4 MULTICOLINEARIDADE

Se existir certo grau de correlação entre as variáveis independente o modelo poderá ser afetado, influenciando negativamente no resultado, sendo assim, foi extraído o grau de correlação de cada elemento dos dados representado pela matriz de SPLOM apresentada na Figura 3.

Figura 3 - Análise de multicolinearidade



Fonte: Elaborado pelo autor.

Assim, as informações numéricas na Figura 3, representam a correlação de Pearson que está analisando o pressuposto de não existir Multicolinearidade entre as variáveis independentes, visto, se o valor for maior que 0,8 ou menos que -0,8 se constataria certo grau de correlação entre as variáveis, mas conforme a tabela nenhum dado apresentou quase nenhum grau de correlação.

Para constatar de vez essa hipótese, será avaliado também o VIF (fator de inflação da variância) de cada modelo que também mostra se existe certo grau de avaria dentro de cada modelo. Sendo que nesse modelo vou considerar se VIF maior que 5 existe certo grau de correlação nos dados, como podemos ver na Tabela 6.

Tabela 6 - Fator de Inflação da Variância (VIF)

Modelos	Dependente	Independente	VIF
Modelo 1	default	abate + morte + ass	abate = 1,00717 morte = 1,372919 ass = 1,364756
Modelo 2	default	abate + ass	abate = 1,007038 ass = 1,007038
Modelo 3	default	abate + morte	abate = 1,018974 morte = 1,018974
Modelo 4	default	morte + ass	morte = 1,40506 ass = 1,40506
Modelo 5	default	morte	-
Modelo 6	default	ass	-
Modelo 7	default	abate	-

Fonte: Elaborado pelo autor.

Extraindo o VIF em cada modelo não se verificou também grau de correlação entre as variáveis do modelo. Comprovando que os dados não contêm grau significativos de Multicolinearidade dos dados.

4.5 VARIÁVEIS INDEPENDENTES ESTÃO LINEARMENTE RELACIONADAS AO LOG DAS PROBABILIDADES

O último pressuposto é da linearidade entre os dados. Esse pressuposto foi analisado de duas formas, a primeira foi realizada o teste de Box Tidwell, onde é necessário realizar a observação dos modelos com uma interação com o log de cada variável independente e assim aplicando regressão para obter as informações. Partindo pelo conceito que o nível de significância de cada interação não pode ser significante para o modelo, logo se aceita o modelo para um nível maior que 0,05.

Na Tabela 7, podemos ver os resultados para o teste de Wald.

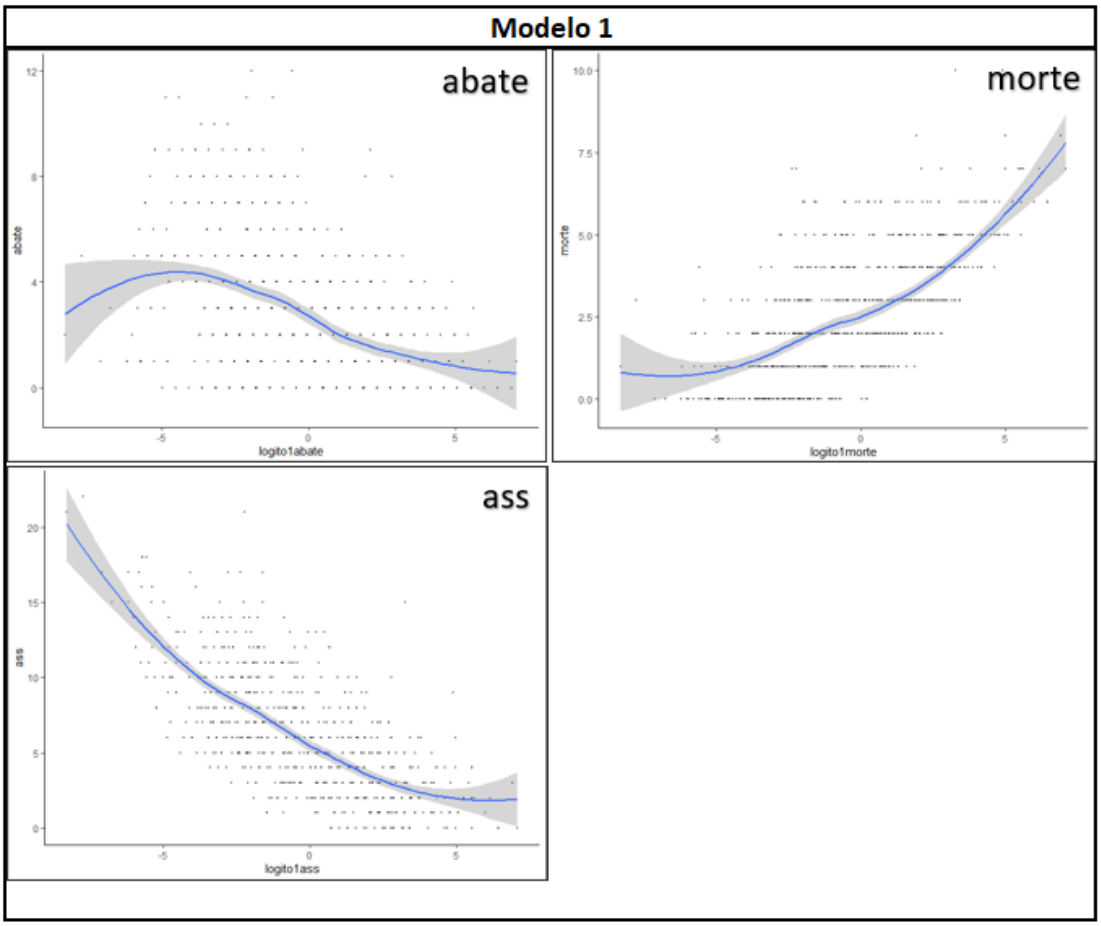
Tabela 7 - Teste De Wald com interação de Log das variáveis independentes

Modelos	Dependente	Independente	p value
Modelo 1	default	abate + morte + ass	abate = 0,94 morte = 0,016 ass = < 0,001
Modelo 2	default	abate + ass	abate = 0,91 ass = < 0,001
Modelo 3	default	abate + morte	abate = 0,93 morte = 0,012
Modelo 4	default	morte + ass	morte = < 0,001 ass = < 0,001
Modelo 5	default	morte	morte = < 0,001
Modelo 6	default	ass	ass = < 0,001
Modelo 7	default	abate	abate = 0,75

Fonte: Elaborado pelo autor

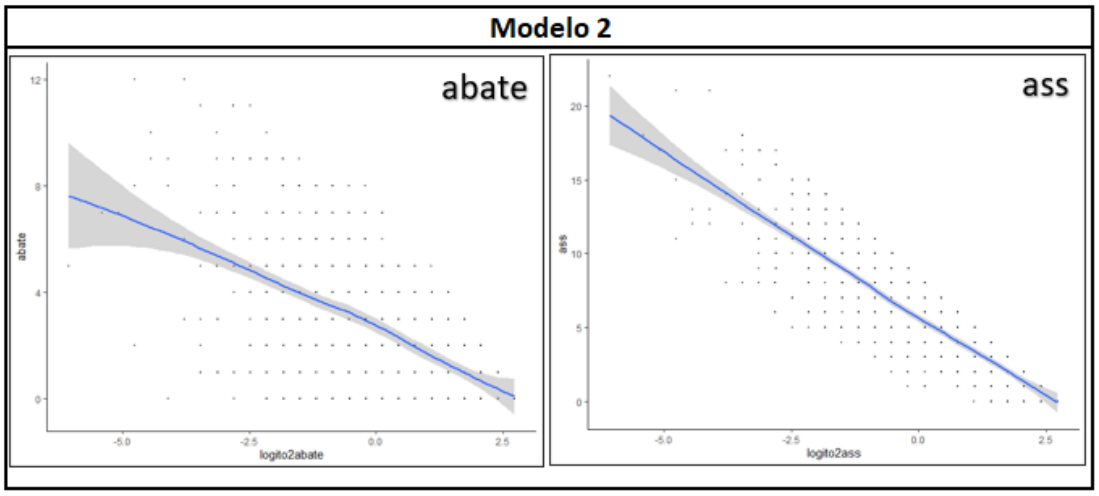
Tendo em vista isso, os modelos que contém a variável “abate” não seriam aceitos se partissem pelo pressuposto somente desse teste. Tendo em vista que o Modelo 1 o Modelo 2 foram influenciados na interação de ass e o Modelo 3 apresentando dúvida em relação a morte, foi necessário analisar essa reação graficamente para analisar se existe uma discrepância entre os dados e o logito de cada modelo. Não será analisado graficamente o Modelo 5, 6 e 7 pois como existe somente uma variável independente no Modelo o logito terá uma reação de 1 para 1 nesses modelos. Os resultados podem ser observados nas Figuras 4, 5 e 6, que encontram-se abaixo.

Figura 4 – Relação linear entre morte e assistência



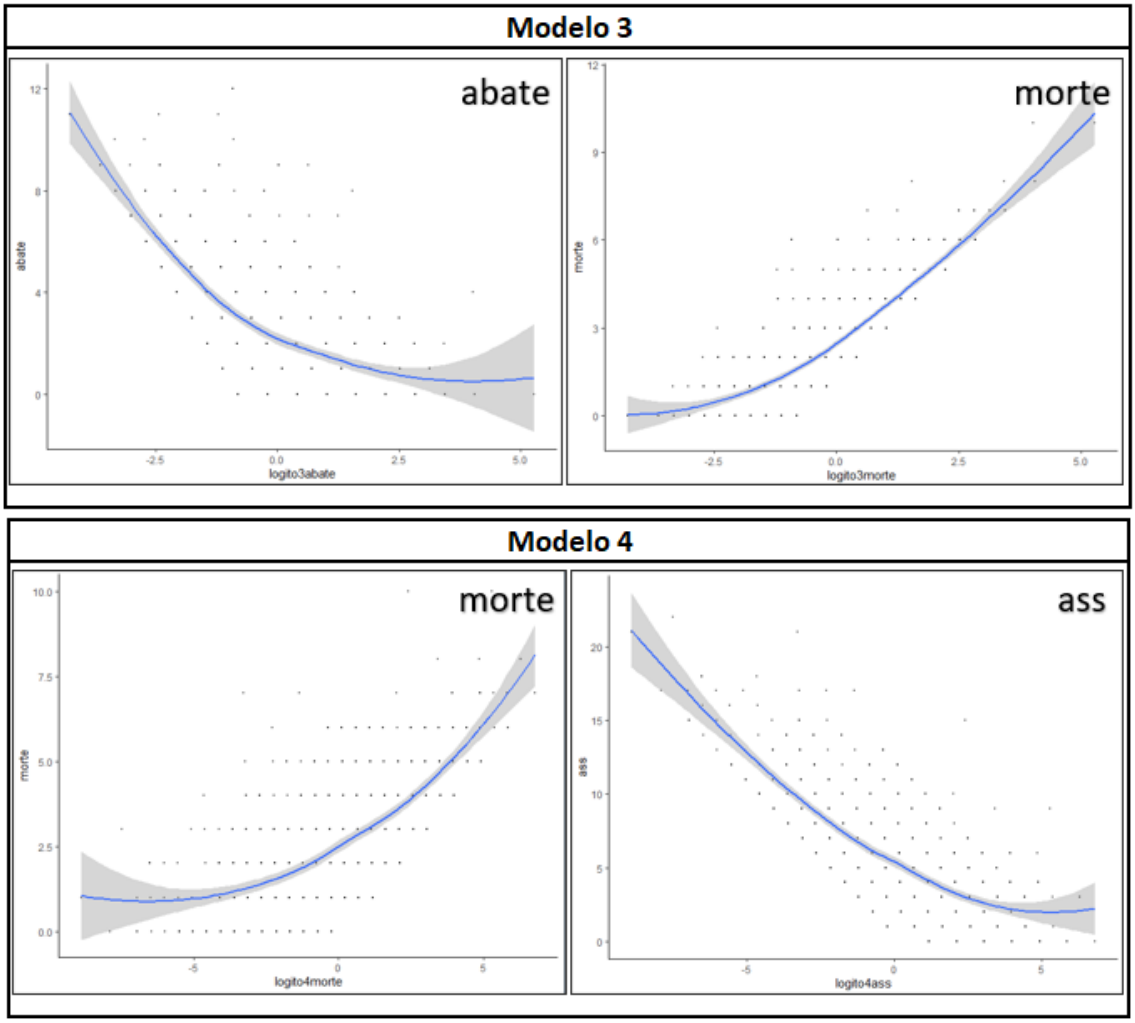
Fonte: Elaborado pelo autor.

Figura 5 – Linearidade entre as variáveis



Fonte: Elaborado pelo autor.

Figura 6 - Relações lineares entre abate e morte no Modelo 3 e morte e assistência no Modelo 4



Fonte: Elaborado pelo autor.

Assim, analisando o Modelo 1, se verifica certa relação linear entre morte e ass, o que gera dúvida graficamente, é o abate que no p-value tinha sido aceito. Portanto para o Modelo 1, mesmo apresentando informações assimétricas, vou considerar que o modelo atende essa hipótese de linearidade com o logito.

O Modelo 2, demostra linearidade entre as variáveis, o que garante o pressuposto, o Modelo 3 e 4 mesmo com a curva sendo quadrática aparentando relação linear, portanto, considera-se o aceite para o pressuposto destas variáveis.

5 MODELOS

Tendo em vista todas os pressupostos a serem atendidos até agora, a Tabela 8 contém o resumo de cada modelo informando se foi aceito ou não.

Tabela 8 - Resumo de pressupostos por modelo

Modelos	Dependente	Independente	Outliers	Independência dos Dados e Resíduos	Multicolinearidade	Variáveis Independentes estão linearmente relacionadas ao log das probabilidades
Modelo 1	default	abate + morte + ass	aceito	aceito	aceito	aceito
Modelo 2	default	abate + ass	aceito	aceito	aceito	aceito
Modelo 3	default	abate + morte	aceito	aceito	aceito	aceito
Modelo 4	default	morte + ass	aceito	aceito	aceito	aceito
Modelo 5	default	morte	aceito	aceito	aceito	aceito
Modelo 6	default	ass	aceito	aceito	aceito	aceito
Modelo 7	default	abate	aceito	aceito	aceito	aceito

Fonte: Elaborado pelo autor.

Analisando os dados apresentados na Tabela 8, se verifica que os únicos modelos que foram aceitos em todos os pressupostos foram os Modelos 1 e 4, o Modelo 2, 3 e 7 se mostraram influenciados por conta dos resíduos que podem influenciar nesses modelos. Já os 5 e 6 além dos resíduos também não apresentam linearidade com os dados o que pode acarretar resultados insatisfatórios.

5.1 APLICAÇÃO DA REGRESSÃO LOGÍSTICA BINÁRIA NOS DADOS

5.1.1 Teste de Wald e OR

Sabendo quais os modelos que podem estar discrepantes na análise vamos analisar os resultados de cada modelo tendo em vista que alguns podem sofrer alteração devido a rejeição entre um dos critérios pressupostos.

A Tabela 9 abaixo mostra a relação dos dados aplicando o Teste De Wald na regressão de cada modelo, extraindo dos dados a informação se o valor z é significativo para o modelo.

Tabela 9 - Teste De Wald

Modelos	Dependente	Independente	Estimate	Std. Error	Z value	p value
Modelo 1	default	Intercept	-0,94708	0,25726	-3,681	< 0.001
		abate	0,27929	0,04681	5,966	< 0.001
		morte	-0,91712	0,07891	-11,62	< 0.001
		ass	0,4567	0,0374	12,21	< 0.001
Modelo 2	default	Intercept	-2,71949	0,21391	-12,71	< 0.001
		abate	0,32562	0,04098	7,946	< 0.001
		ass	0,32483	0,02779	11,69	< 0.001
Modelo 3	default	Intercept	0,81346	0,18578	4,379	< 0.001
		abate	0,3133	0,03958	7,916	< 0.001
		morte	-0,60899	0,05614	-10,85	< 0.001
Modelo 4	default	Intercept	-0,24032	0,21782	-1,103	0,27
		morte	-0,93632	0,07594	-12,33	< 0.001
		ass	0,47929	0,0376	12,75	< 0.001
Modelo 5	default	Intercept	1,61992	0,159	10,19	< 0.001
		morte	-0,61349	0,05312	-11,55	< 0.001
Modelo 6	default	Intercept	-1,8959	0,16999	-11,15	< 0.001
		ass	0,33592	0,02759	12,17	< 0.001
Modelo 7	default	Intercept	-0,84275	0,11742	-7,177	< 0.001
		abate	0,32522	0,03646	8,92	< 0.001

Fonte: Elaborado pelo autor.

Dados a um nível de significância de 95%, ou seja, se espera um p-value menor que 0,05 para rejeitar a hipótese nula em favor da hipótese alternativa de que a variável é significante para o modelo. Conforme apresentado na Tabela 9 praticamente todas das opções foram aceitas em suma maioria com p-value menor

que 0,001, com exceção do intercepto do Modelo 4. Se observa também todos os modelos com a informação morte foram dadas como valor negativo o que afirma a tese inicial que a morte do jogador contribui negativamente para o resultado vitorioso da partida. Na Tabela 10, podemos ver os resultados do teste de razão de verossimilhança.

Tabela 10 - Teste Razão de Verossimilhança

Modelos	Dependente	Independente	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
Modelo 1	default	NULL		769	1067.45	< 0.001
		abate	98.047	768	969.40	< 0.001
		morte	153.687	767	815.71	< 0.001
		ass	259.745	766	555.97	< 0.001
Modelo 2	default	NULL		769	1067.45	< 0.001
		abete	98.047	768	969.40	< 0.001
		ass	197.647	767	771.75	< 0.001
Modelo 3	default	NULL		769	1067.45	< 0.001
		abate	98.047	768	969.40	< 0.001
		morte	153.687	767	815.71	< 0.001
Modelo 4	default	NULL		769	1067.45	< 0.001
		morte	178.88	768	888.57	< 0.001
		ass	291.99	767	596.57	< 0.001
Modelo 5	default	NULL		769	1067.45	< 0.001
		morte	178.88	768	888.57	< 0.001
Modelo 6	default	NULL		769	1067.45	< 0.001
		ass	219.98	768	847.47	< 0.001
Modelo 7	default	NULL		769	1067.45	< 0.001
		abate	98.047	768	969.4	< 0.001

Fonte: Elaborado pelo autor.

A razão demostra o quanto a variável é importante na explicação do modelo, ela mostra se da primeira variável independente até a última sequencialmente se ele ajuda a explicar mais o modelo, essa análise ajuda aqueles regressão que existe muitas variáveis independentes que podem até mesmo algumas não ser tão relevante para o resultado. Nos modelos se verifica que todas as variáveis independentes adotadas têm grau de significância no teste, ou seja, as 3 variáveis ajudam a explicar o modelo. Na Tabela 11, podemos ver os resultados para a Odds Ratio.

Tabela 11 - Odds Ratios

Modelos	Dependente	Independente	OR	2.5 %	97.5 %
Modelo 1	default	Intercept	0,3878715	0,232426	0,6381733
		abate	1,322194	1,2090504	1,4531628
		morte	0,3996703	0,3402056	0,4637672
		ass	1,5788496	1,4717111	1,7045489
Modelo 2	default	Intercept	0,06590842	0,04278039	0,09905224
		abate	1,38488394	1,2804713	1,5038924
		ass	1,38379457	1,31261934	1,46389346
Modelo 3	default	Intercept	2,2556933	1,5721037	3,2593516
		abate	1,3679291	1,2680395	1,4812269
		morte	0,5438984	0,4857261	0,6054379
Modelo 4	default	Intercept	0,7863774	0,5121602	1,2043696
		morte	0,3920687	0,3358196	0,4524643
		ass	1,6149293	1,5046356	1,7440122
Modelo 5	default	Intercept	5,0526645	3,7239262	6,9495399
		morte	0,5414589	0,4865214	0,5992735
Modelo 6	default	Intercept	0,1501834	0,1067378	0,2079789
		ass	1,3992282	1,3277203	1,4795555
Modelo 7	default	Intercept	0,430524	0,3410063	0,5405202
		abate	1,38433	1,2910869	1,4896309

Fonte: Elaborado pelo autor.

Quanto as razões de chance em cada modelo, se verificam que a morte definitivamente contribui negativamente para o resultado da partida como vencedor, visto que os valores estão menores que 1 nos modelos. Quanto a abate não se vê tanta modificação entre os modelos, já o ass apresenta certa discrepância quando analisa cada uma das observações.

5.1.2 Comparação dos Modelos

Nessa parte iremos analisar qual modelo melhor acertou mais tanto estatisticamente quanto na simulação nos dados. Na Tabela 12, podemos ver os resultados do ajuste do modelo.

Tabela 12 - Pseudo R^2, AIC E BIC

Modelos	Dependente	Independente	Pseudo R^2 (Nagelkerke)	df	AIC
Modelo 1	default	abate + morte + ass	0,65	4	563,97
Modelo 2	default	abate + ass	0,43	3	777,75
Modelo 3	default	abate + morte	0,37	3	821,71
Modelo 4	default	morte + ass	0,61	3	602,57
Modelo 5	default	morte	0,28	2	892,57
Modelo 6	default	ass	0,33	2	851,47
Modelo 7	default	abate	0,16	2	973,40

Fonte: Elaborado pelo autor.

O Pseudo R^2 dessa análise buscou utilizar o modelo de Nagelkerke ele também oferece uma situação de qual modelo atende mais levando em conta que, quanto maior melhor. Nesse pressuposto os Modelos 1 (0,65) e 4 (0,61) foram os que mais atenderam no Pseudo R^2.

O AIC também busca avaliar qual modelo explica mais o resultado obtido sendo assim, o teste mostra que os modelos melhores são também os Modelos 1 e 4 devido ao fato que apresentarem os valores mais baixos, como podem ser observados na Tabela 13.

Tabela 13 - Teste Hosmer e Lemeshow

Modelos	Dependente	Independente	X-squared	df	p-value
Modelo 1	default	abate + morte + ass	5,7899	8	0,6707
Modelo 2	default	abate + ass	51,284	8	< 0,001
Modelo 3	default	abate + morte	26,815	8	< 0,001
Modelo 4	default	morte + ass	16,478	8	0,03603
Modelo 5	default	morte	35,707	8	< 0,001
Modelo 6	default	ass	36,141	8	< 0,001
Modelo 7	default	abate	2,9815	8	0,9355

Fonte: Elaborado pelo autor.

O teste de Hosmer e Lemeshow, apresentado na Tabela 13, indica qual é o modelo que mais acerta os dados, ou seja, o modelo mais adequado. Segundo essa análise os modelos são o Modelo 1 e o Modelo 7.

Como podemos ver na Tabela 14, obtemos valores de acertos e erros para o teste de Hosmer e Lemeshow.

Tabela 14 - Resumo de Acertos e Erros

Modelos	Dependente	Independente	ACERTOS		ERROS	
			GANHAR	PERDER	GANHAR	PERDER
Modelo 1	default	abate + morte + ass	326	331	59	54
Modelo 2	default	abate + ass	292	309	93	76
Modelo 3	default	abate + morte	275	300	110	85
Modelo 4	default	morte + ass	321	328	64	57
Modelo 5	default	morte	275	274	110	111
Modelo 6	default	ass	270	294	115	91
Modelo 7	default	abate	228	271	157	114

Fonte: Elaborado pelo autor.

5.1.3 Resultado do Modelo

Tendo em vista que somente o Modelo 1 e o Modelo 4 atenderam todos os pressupostos iniciais, no teste de Wald todos as variáveis apresentaram significância para o teste, na razão de Verossimilhança não foi identificada nenhuma variável que não seja relevante ao modelo, o maior valor do Pseudo R2 foi exatamente nos Modelos 1 e 4, o AIC apontaram o Modelo 1 como o melhor modelo. O Teste Hosmer e Lemeshow apontou que o Modelo 1 e 7 são melhores previsões. No resumo de acertos e erros observamos que os maiores acertos foram do Modelo 1 e Modelo 4 com mais de 300 acertos tanto para ganhar quanto para perder sendo assim, a análise considera o primeiro modelo como o que mais explica a razão de chances.

Tabela 15 - Resultado do melhor Modelo

Modelos	Dependente	Independente	OR	2.5 %	97.5 %
Modelo 1	default	Intercept	0,3878715	0,232426	0,6381733
		abate	1,322194	1,2090504	1,4531628
		morte	0,3996703	0,3402056	0,4637672
		ass	1,5788496	1,4717111	1,7045489

Fonte: Elaborado pelo autor.

Assim, o modelo que mais acertou foi o primeiro (Modelo 1), logo ele pressupõe que, a cada 1 acréscimo de abate suas chances de ganhar a partida aumenta em 32,22%, também a cada 1 participação em ass você aumente suas chances de vitória em 57,88%. Somente a morte que apresentou uma relação negativa, logo a cada 1 participação do grupo de morte você diminui sua chance de vitória em quase -60%.

6 CONCLUSÃO

Os resultados dos testes apresentados foram bastante informativos e nos permitiram avaliar a qualidade dos modelos. No Teste de Razão de Verossimilhança, todos os modelos passaram com êxito, o que significa que as variáveis de scores foram aceitas para todos os modelos. Já no Teste de Wald, o intercepto do Modelo 4 foi rejeitado com significância de 5%, sugerindo que ajustes adicionais precisam ser feitos nesse modelo. Por fim, no Teste de Hosmer e Lemeshow, quando foi feita a simulação de resultados, os Modelos 1 e 4 foram os que performaram melhor, indicando que eles são os mais adequados para a previsão dos dados.

Nossa escolha do melhor modelo foi baseada em uma avaliação cuidadosa dos resultados obtidos a partir da checagem de pressupostos e das avaliações dos modelos. Quando examinamos o desempenho dos Modelos 1 e 4, levando em conta tanto os pressupostos quanto as avaliações, ficamos entre esses dois modelos. No entanto, o Teste de Wald apontou para a rejeição do intercepto do Modelo 4, o que levou a escolha do Modelo 1 como o melhor. Além disso, a simulação pelo Teste de Hosmer e Lemeshow mostrou que o modelo escolhido acertou aproximadamente 85% dos 770 registros, o que reforçou a nossa escolha do Modelo 1. Em geral, a combinação de pressupostos e avaliações dos modelos nos permitiu escolher o Modelo 1 como o mais preciso para prever os dados.

Em conclusão, para futuros trabalhos seria indicado dividir o conjunto de dados em duas partes, treino e teste, com proporções geralmente de 70% e 30%. Usar todo o conjunto de dados para construir o modelo resulta em uma taxa de acerto superestimada e falta de capacidade de generalização, ou seja, o modelo não será tão preciso em classificar dados não usados para sua construção. Além disso, seria benéfico realizar uma análise de diferentes pontos de corte para identificar o melhor ponto de equilíbrio entre sensibilidade e especificidade, o que resultaria em melhores taxas de classificação e menores erros.

REFERÊNCIAS

BATISTA, A. M. S. **Regressão Logística:** Uma Introdução ao Modelo Estatístico - Exemplo de Aplicação ao Resolving Credit. Porto: Vida Económica, 2015.

CABRAL, C. I. S. **Aplicação do modelo de regressão logística num estudo de mercado**. 2013. 58 f. Dissertação (Mestrado em Matemática Aplicada à Economia e à Gestão) - Universidade de Lisboa, Lisboa, 2013.

GONZALEZ, L. A. **Regressão logística e suas aplicações**. 2018. 45 f. Monografia (Graduação em Ciência da Computação) - Universidade Federal do Maranhão, São Luis, 2018.

HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**. New York: Wiley, 2000.

KAGGLE. **League of Legends:** Worlds. 2019. Disponível em: https://www.kaggle.com/datasets/talocapita/league-of-legends-worlds-2019 Acesso em: 18 mai. 2022.

MESQUITA, P. S. B. **Um modelo de regressão logística para avaliação dos programas de pós-graduação no Brasil**. 2014. 107 f. Dissertação (Mestrado em Engenharia de Produção) – Universidade Estadual do Norte Fluminense, Campo Goytacazes, RJ, 2014.

MKT ESPORTIVO. **League of Legends faturou US$ 1.75 bilhão em 2020**. 2021. Disponível em: https://www.mktesportivo.com/2021/01/league-of-legends-faturou-us-1-75-bilhao-em-2020/ Acesso em: 18 mai. 2022.

ROX, M. Mais Esports. **League of Legends chega a 180 milhões de jogadores mensais**. 2021. Disponível em: https://maisesports.com.br/league-of-legends-180-milhoes-jogadores-mensais/. Acesso em: 18 mai. 2022.

RUNAS. **Quantas pessoas jogam League of Legends?** 2021. Disponível em: https://www.runas.lol/blog/quantas-pessoas-jogam-a-league-of-legends/ Acesso em: 18 mai. 2022.

VIEIRA, P. R. C. **Análise Multivariada com o uso do SPSS.** Rio de Janeiro: Ciência Moderna, 2011.

APÊNDICE A – Script utilizado

```
#CARREGAR PACOTES
if(!require(pacman)) install.packages("pacman")
if(!require(lmtest)) install.packages("lmtest")
if(!require(lmtest)) install.packages("readxl")
if(!require(lmtest)) install.packages("dplyr")
if(!require(lmtest)) install.packages("ResourceSelection")
library(ResourceSelection)
library(pacman)
library(lmtest)
library(readxl)
library(dplyr)
library(pacman)
library(car)
pacman::p_load(dplyr, psych, car, MASS, DescTools, QuantPsyc, ggplot2)

#CARREGAR DADOS / ANALISAR
dados <- read_excel("C:/Users/Italo/Downloads/dados.xlsx")
glimpse(dados) #ANALISAR COLUNAS
View(dados) #VERDADOS
summary(dados) #ANALISAR RESUMO / ##AMOSTRA ATENDE MIN PARA REGRESSÃO

#AJUSTAR COLUNA PARA FACTOR
dados$default <- as.factor(dados$default)
class(dados$default)

#AJUSTAR REFERENCIA DA COLUNA DEFAULT / ANALISAR AJUSTE
dados$default <- relevel(dados$default, ref = "PERDEU")
levels(dados$default)
dados$default <- as.factor(dados$default)
glimpse(dados) #ANALISAR COLUNAS
summary(dados) # RESUMO AJUSTADO

#REGRESSÃO PROPOSTAS

modelo1 <- glm(default ~ abate + morte + ass,
               family = binomial(link = 'logit'), data = dados)

modelo2 <- glm(default ~ abate + ass,
               family = binomial(link = 'logit'), data = dados)

modelo3 <- glm(default ~ abate + morte,
               family = binomial(link = 'logit'), data = dados)

modelo4 <- glm(default ~ morte + ass,
               family = binomial(link = 'logit'), data = dados)
```

```
modelo5 <- glm(default ~ morte,
               family = binomial(link = 'logit'), data = dados)

modelo6 <- glm(default ~ ass,
               family = binomial(link = 'logit'), data = dados)

modelo7 <- glm(default ~ abate,
               family = binomial(link = 'logit'), data = dados)

#GRAFICOS VERIFICAR OUTLIERS
par(mfrow= c(2,4))
plot(modelo1, which = 5) #OK SEM OUTLIERS
plot(modelo2, which = 5) #OK SEM OUTLIERS
plot(modelo3, which = 5) #OK SEM OUTLIERS
plot(modelo4, which = 5) #OK SEM OUTLIERS
plot(modelo5, which = 5) #OK SEM OUTLIERS
plot(modelo6, which = 5) #OK SEM OUTLIERS
plot(modelo7, which = 5) #OK SEM OUTLIERS
##ANALISE ATENDE A AUSENCIA DE OUTLIERS

#GRAFICOS VERIFICAR INDEPEDÊNCIA DOS RESÍDUOS
par(mfrow= c(2,4))
plot(modelo1, which = 2) #OK NORMAL / OK SEM OUTLIERS
plot(modelo2, which = 2) #OK NÃO NORMAL / OK SEM OUTLIERS
plot(modelo3, which = 2) #OK NÃO NORMAL / OK SEM OUTLIERS
plot(modelo4, which = 2) #OK NORMAL / OK SEM OUTLIERS
plot(modelo5, which = 2) #OK NÃO NORMAL / OK SEM OUTLIERS
plot(modelo6, which = 2) #OK NÃO NORMAL / OK SEM OUTLIERS
plot(modelo7, which = 2) #OK NÃO NORMAL / OK SEM OUTLIERS
##ANALISE ATENDE EM ALGUNS CASOS A INDEPEDÊNCIA DOS RESÍDUOS

#VERIFICAR MULTICOLINEARIDADE
pairs.panels(dados) #SEM CORRELAÇÃO SIGN. (Multicolinearidade quando r > 0.8)
vif(modelo1) # OK SEM CORRELAÇÃO SIGN. (Multicolinearidade quando VIF > 10)
vif(modelo2) # OK SEM CORRELAÇÃO SIGN. (Multicolinearidade quando VIF > 10)
vif(modelo3) # OK SEM CORRELAÇÃO SIGN. (Multicolinearidade quando VIF > 10)
vif(modelo4) # OK SEM CORRELAÇÃO SEGN. (Multicolinearidade quando VIF > 10)
##ANALISE ATENDE, NÃO CONTEM INDICAÇÕES DE MULTICOLINEARIDADE

#VERIFICAR SE VI ESTÃO LINEARMENTE RELACIONADAS AO LOG DAS PROB.
  #CALCULO
intlogabate <- dados$abate * log(dados$abate)
intlogmorte <- dados$morte * log(dados$morte)
intlogass <- dados$ass * log(dados$ass)
dados$intlogabate <- intlogabate
dados$intlogmorte <- intlogmorte
dados$intlogass <- intlogass
```

```
 #MODELO COM INT LOG
modelo1int <- glm(default ~ abate + morte + ass +intlogabate + intlogmorte + intlogass,
                  family = binomial(link = 'logit'), data = dados)

modelo2int <- glm(default ~ abate + ass + intlogabate + intlogass,
                  family = binomial(link = 'logit'), data = dados)

modelo3int <- glm(default ~ abate + morte + intlogabate + intlogmorte,
                  family = binomial(link = 'logit'), data = dados)

modelo4int <- glm(default ~ morte + ass + intlogmorte + intlogass,
                  family = binomial(link = 'logit'), data = dados)

modelo5int <- glm(default ~ morte + intlogmorte,
                  family = binomial(link = 'logit'), data = dados)

modelo6int <- glm(default ~ ass + intlogass,
                  family = binomial(link = 'logit'), data = dados)

modelo7int <- glm(default ~ abate + intlogabate,
                  family = binomial(link = 'logit'), data = dados)

 #RESULTADO DOS MODELOS COM INT LOG
summary(modelo1int) #NÃO ATENDE?
summary(modelo2int) #NÃO ATENDE
summary(modelo3int) #NÃO ATENDE?
summary(modelo4int) #NÃO ATENDE
summary(modelo5int) #NÃO ATENDE
summary(modelo6int) #NÃO ATENDE
summary(modelo7int) #ATENDE

 #ANALISE GRAF. 2 DE VERIFICAR SE VI ESTÃO LINEARMENTE RELACIONADAS AO LOG DAS PROB.
#MODELO1
par(mfrow= c(2,2))
logito1abate <- modelo1$linear.predictors
dados$logito1abate <- logito1abate
ggplot(dados, aes(logito1abate, abate)) +
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_classic()

logito1morte <- modelo1$linear.predictors
dados$logito1morte <- logito1morte
ggplot(dados, aes(logito1morte, morte)) +
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_classic()
```

```
logito1ass <- modelo1$linear.predictors
dados$logito1ass <- logito1ass
ggplot(dados, aes(logito1ass, ass)) +
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_classic()

#MODELO2
logito2abate <- modelo2$linear.predictors
dados$logito2abate <- logito2abate
ggplot(dados, aes(logito2abate, abate)) +
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_classic()

logito2ass <- modelo2$linear.predictors
dados$logito2ass <- logito2ass
ggplot(dados, aes(logito2ass, ass)) +
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_classic()

#MODELO3
logito3abate <- modelo3$linear.predictors
dados$logito3abate <- logito3abate
ggplot(dados, aes(logito3abate, abate)) +
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_classic()

logito3morte <- modelo3$linear.predictors
dados$logito3morte <- logito3morte
ggplot(dados, aes(logito3morte, morte)) +
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_classic()


#MODELO4
logito4morte <- modelo4$linear.predictors
dados$logito4morte <- logito4morte
ggplot(dados, aes(logito4morte, morte)) +
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_classic()

logito4ass <- modelo4$linear.predictors
dados$logito4ass <- logito4ass
ggplot(dados, aes(logito4ass, ass)) +
  geom_point(size = 0.5, alpha = 0.5) +
```

```
  geom_smooth(method = "loess") +
  theme_classic()
##MODELO 1 A 4 ATENDE 5 E 6 NÃO ATENDE E 7 ATENDE.

#RESULTADO DOS MODELOS / TESTE DE WALD
summary(modelo1) #OK ACEITO / TESTE DE WALD ACEITO
summary(modelo2) #OK ACEITO / TESTE DE WALD ACEITO
summary(modelo3) #OK ACEITO / TESTE DE WALD ACEITO
summary(modelo4) #OK ACEITO / TESTE DE WALD ACEITO
summary(modelo5) #OK ACEITO / TESTE DE WALD ACEITO
summary(modelo6) #OK ACEITO / TESTE DE WALD ACEITO
summary(modelo7) #OK ACEITO / TESTE DE WALD ACEITO

#TESTE RAZÃO DE VEROSSIMILHANÇA
anova(modelo1, test="Chisq") #OK ACEITO
anova(modelo2, test="Chisq") #OK ACEITO
anova(modelo3, test="Chisq") #OK ACEITO
anova(modelo4, test="Chisq") #OK ACEITO
anova(modelo5, test="Chisq") #OK ACEITO
anova(modelo6, test="Chisq") #OK ACEITO
anova(modelo7, test="Chisq") #OK ACEITO

#RAZÕES DE CHANCES PARA CADA MODELO
OR1=exp(modelo1$coefficients)
OR1
OR2=exp(modelo2$coefficients)
OR2
OR3=exp(modelo3$coefficients)
OR3
OR4=exp(modelo4$coefficients)
OR4
OR5=exp(modelo5$coefficients)
OR5
OR6=exp(modelo6$coefficients)
OR6
OR7=exp(modelo7$coefficients)
OR7

#RAZÕES DE CHANCES PARA CADA MODELO COM INTERVALO
exp(cbind(OR = coef(modelo1), confint(modelo1)))
exp(cbind(OR = coef(modelo2), confint(modelo2)))
exp(cbind(OR = coef(modelo3), confint(modelo3)))
exp(cbind(OR = coef(modelo4), confint(modelo4)))
exp(cbind(OR = coef(modelo5), confint(modelo5)))
exp(cbind(OR = coef(modelo6), confint(modelo6)))
exp(cbind(OR = coef(modelo7), confint(modelo7)))

#PSEUDO R² DE Nagelkerke
PseudoR2(modelo1, which = "Nagelkerke")
PseudoR2(modelo2, which = "Nagelkerke")
```

```
PseudoR2(modelo3, which = "Nagelkerke")
PseudoR2(modelo4, which = "Nagelkerke")
PseudoR2(modelo5, which = "Nagelkerke")
PseudoR2(modelo6, which = "Nagelkerke")
PseudoR2(modelo7, which = "Nagelkerke")

# AIC e BIC
AIC(modelo1, modelo2, modelo3, modelo4, modelo5, modelo6, modelo7)
BIC(modelo1, modelo2, modelo3, modelo4, modelo5, modelo6, modelo7)

#Tabela de classificação
ClassLog(modelo1, dados$default)
ClassLog(modelo2, dados$default)
ClassLog(modelo3, dados$default)
ClassLog(modelo4, dados$default)
ClassLog(modelo5, dados$default)
ClassLog(modelo6, dados$default)
ClassLog(modelo7, dados$default)

#Teste Hosmer e Lemeshow
h1=hoslem.test(dados$default2,fitted(modelo1),g=10)
h2=hoslem.test(dados$default2,fitted(modelo2),g=10)
h3=hoslem.test(dados$default2,fitted(modelo3),g=10)
h4=hoslem.test(dados$default2,fitted(modelo4),g=10)
h5=hoslem.test(dados$default2,fitted(modelo5),g=10)
h6=hoslem.test(dados$default2,fitted(modelo6),g=10)
h7=hoslem.test(dados$default2,fitted(modelo7),g=10)
h1
h2
h3
h4
h5
h6
h7
cbind(h1$expected, h1$observed)
cbind(h2$expected, h2$observed)
cbind(h3$expected, h3$observed)
cbind(h4$expected, h4$observed)
cbind(h5$expected, h5$observed)
cbind(h6$expected, h6$observed)
```