UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE MATEMÁTICA

MARIA CAROLINA PICOLI DURANTE

REGRESSÃO MÚLTIPLA E LOGÍSTICA: EXPLORANDO SUAS RELAÇÕES E APLICAÇÕES

SÃO CARLOS – SP
2024

MARIA CAROLINA PICOLI DURANTE

REGRESSÃO MÚLTIPLA E LOGÍSTICA: EXPLORANDO SUAS RELAÇÕES E APLICAÇÕES

Monografia apresentada ao Curso de Licenciatura em Matemática da Universidade Federal de São Carlos.

Orientador: Prof. Dr. Wladimir Seixas

SÃO CARLOS – SP
2024

AGRADECIMENTOS

Primeiramente, agradeço ao meu orientador, Wladimir Seixas pela orientação dedicada, apoio constante e valiosas sugestões ao longo deste processo. Seu conhecimento e entusiasmo não apenas orientaram meu trabalho, mas também instigaram meu interesse em aprofundar-me ainda mais no tema.

À minha mãe, irmãs e avós que sempre me apoiaram nos estudos e foram fundamentais para que tornasse realidade de sair de uma cidade tão pequena como Divinolândia - SP e vim viver novos desafios em São Carlos.

Ao meu companheiro Rafael, que esteve sempre ao meu lado, oferecendo muito amor, incentivo e compreensão durante toda essa jornada.

Às minhas amigas, Alice e Carol que mesmo de longe, sempre estiveram presentes, prontas para ouvir e compartilhar suas experiências durante nossas graduações.

Um agradecimento especial à minha psicóloga, Aline, cujo papel foi fundamental na etapa final desta graduação, auxiliando-me a compreender que limitações são parte do percurso e que é necessário calcular as rotas para alcançar a linha de chegada.

À UFSCar, expresso minha gratidão por proporcionar não apenas conhecimento, mas também amizades e experiências únicas de crescimento pessoal e profissional.

Por último, mas não menos importante, agradeço a todos que, direta ou indiretamente, estiveram envolvidos não apenas neste trabalho, mas em toda a jornada da graduação.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

No contexto deste Trabalho de Conclusão de Curso, exploraremos um conjunto robusto de técnicas estatísticas conhecidas como regressão, que oferecem uma abordagem sistemática para modelar e compreender relações entre variáveis.

Este trabalho está dividido em três partes. A primeira parte abordará a construção da regressão linear, utilizando o método dos mínimos quadrados. Este método, reconhecido por sua eficácia na minimização das discrepâncias entre os dados observados e os valores previstos, será o alicerce fundamental para nossa análise. Em seguida, aprofundaremos as principais técnicas da regressão linear simples, que contribuem para interpretar e extrair informações significativas relacionadas a uma única variável independente.

Na segunda parte, abordaremos a Regressão Múltipla, na qual a complexidade é ampliada ao considerarmos múltiplas variáveis independentes. Neste estágio da pesquisa, não apenas ampliamos nosso entendimento da modelagem, mas também nos permitimos analisar relações de vários fatores, fundamentais para uma compreensão mais abrangente dos fenômenos estudados.

Na terceira e última parte, tratamos da construção das principais técnicas utilizadas na regressão logística binária, estendendo nosso escopo para além da linearidade. Esta técnica está intimamente interligada tanto à regressão simples quanto à múltipla, entretanto, diferentes das regressões anteriores que sua variável dependente era contínua, a sua variável dependente será binária. Esta técnica proporcionará uma compreensão essencial para lidar com problemas de classificação, frequentemente encontrados em estudos de predição e categorização.

O objetivo principal deste trabalho é o de apresentar os conceitos e métodos fundamentais para as análises de regressão, contribuindo para o desenvolvimento de habilidades analíticas e interpretativas, essenciais para a pesquisa e tomada de decisões em diversas áreas do conhecimento.

2 REGRESSÃO LINEAR

Neste capítulo, abordaremos a regressão linear, destacando conceitos fundamentais como: diagrama de dispersão, método dos mínimos quadrados e adequação do modelo ajustado.

O diagrama de dispersão é uma ferramenta visual de grande importância para compreender as relações entre variáveis, enquanto o método dos mínimos quadrados representa uma abordagem matemática essencial na determinação da reta de melhor ajuste em um conjunto de dados, através da minimização dos erros residuais. Esses elementos, em conjunto, constituem a base para a compreensão e aplicação da regressão linear, uma técnica estatística notável utilizada para modelar e prever relações entre variáveis.

2.1 MÉTODO DOS MÍNIMOS QUADRADOS

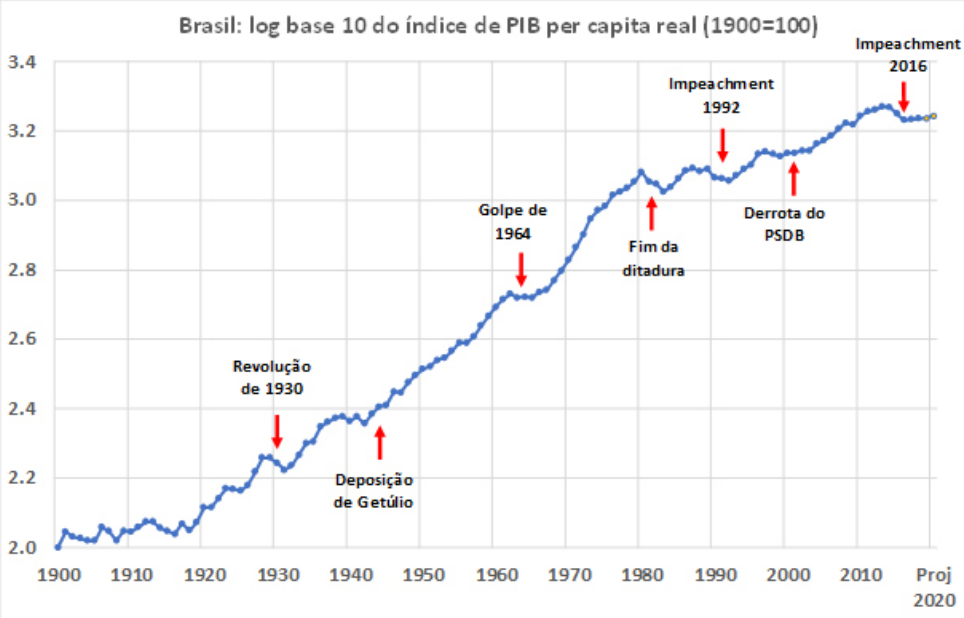
O Método dos Mínimos Quadrados é uma técnica matemática utilizada para determinar a melhor aproximação de uma reta ou curva que se ajuste a um conjunto de pontos (dados) no plano. Nesse sentido, o objetivo é minimizar a soma dos quadrados das diferenças entre os valores observados e os valores teóricos previstos pela curva ajustada. Isso é realizado ajustando os parâmetros da reta ou curva de forma a reduzir ao máximo a discrepância entre os dados reais e os valores teóricos calculados. Em outras palavras, considerando os pontos X_i com valores observados Y_i, onde $i = 1, \ldots, m$, o método irá determinar a função $Y = g(X)$ (reta ou curva teórica) que melhor se ajusta aos pares de valores (X_i, Y_i), com $i = 1, \ldots, m$, de forma que a soma dos quadrados dos erros $\varepsilon_i = g(X_i) - Y_i$, com $i = 1, \ldots, m$, seja o menor valor possível.

2.1.1 Diagrama de dispersão

O diagrama de dispersão é um gráfico que ilustra a relação entre duas variáveis, sendo uma considerada a variável independente (X) e a outra a variável dependente (Y). O diagrama de dispersão exibe os m pontos (X_i, Y_i) no plano cartesiano. Esse gráfico auxilia na visualização para determinar se existe alguma tendência geral, comportamento ou relação entre as variáveis (X, Y). Se houver um comportamento de proporcionalidade, onde o aumento da variável independente acarreta em um aumento ou diminuição proporcional na variável dependente, será possível supor a adequação de uma reta passando entre os pontos. Nesse caso, diz-se que existe uma relação linear entre as variáveis.

Exemplo 2.1. Considere a evolução do log do índice de PIB real per capita desde 1900. Definindo o valor de 1990 como 100 (log decimal igual a 2) temos:

Figura 1 – PIB real per capita versus ano.



Fonte: BARBOSA, N. Evolução do PIB per capita e situação política[1].

Um ajuste de uma reta por entre os pontos do gráfico indicariam uma taxa de crescimento no período considerado. O mesmo levantamento de dados para diferentes países poderia mostrar diferentes taxas de crescimento comparativas entre os países considerados.

2.1.2 Ajuste linear

Consideremos um conjunto genérico de dados no qual desejamos analisar o comportamento do fenômeno. Vamos supor que a distribuição dos pontos (X_i, Y_i), $i = 1, \ldots, m$ no diagrama de dispersão apresentem uma grande proximidade gráfica a uma reta, isto é, estão quase alinhados em relação a uma reta "imaginária" que passa por entre eles. O objetivo será encontrar a função

$$Y = f(X) = a_0 + a_1 X$$

que se adapte melhor a esses dados.

Para compreender como determinar a reta que melhor se ajusta aos pontos de um conjunto de dados específico, vamos empregar o método dos mínimos quadrados. Inicialmente, vemos que os erros presentes em cada ponto é dado por $e(X_i) = Y_i - f(X_i)$. O método dos mínimos

[1] <https://https://blogdoibre.fgv.br/posts/evolucao-do-pib-capita-e-situacao-politica>. Acesso em: 20 ago. 2023.

quadrados consiste em minimizar a soma dos quadrados dos desvios ou erros, ou seja,

$$\min \sum_{i=1}^{m} e(X_i)^2 = \min \sum_{i=1}^{m} (Y_i - f(X_i))^2.$$

Quando nos referimos ao ajuste linear, a função a ser ajustada será $f(X) = a_0 + a_1 X$ e o erro dado por

$$\begin{aligned} E(a_0, a_1) &= \sum_{i=1}^{m} (Y_i - f(X_i))^2 \\ &= \sum_{i=1}^{m} (Y_i - a_0 - a_1 X_i)^2. \end{aligned}$$

O erro é, portanto, uma função dos parâmetros a_0 e a_1.

Com o intuito de estabelecer a função $f(X) = a_0 + a_1 X$ que melhor se ajusta (isto é, fornece o menor erro quadrático) em relação aos valores observados, é necessário determinar os parâmetros a_0 e a_1 que minimizam o erro $E(a_0, a_1)$. Nesse sentido, as derivadas parciais da função E em relação às variáveis a_0 e a_1 devem ser igualadas a zero, localizando os pontos críticos de E. Uma vez que a função de erro E é crescente e ilimitada superiormente, o ponto crítico determinado será o ponto de mínimo da função. Sendo assim,

$$\frac{\partial E}{\partial a_0} = -2 \sum_{i=1}^{m} (Y_i - a_0 - a_1 X_i) = 0 \tag{2.1}$$

ou

$$\sum_{i=1}^{m} Y_i - a_0 \sum_{i=1}^{m} 1 - a_1 \sum_{i=1}^{m} X_i = 0$$

Segue que

$$a_0 m + a_1 \sum_{i=1}^{m} X_i = \sum_{i=1}^{m} Y_i. \tag{2.2}$$

Analogamente,

$$\frac{\partial E}{\partial a_1} = -2 \sum_{i=1}^{m} [X_i (Y_i - a_0 - a_1 X_i) = 0 \tag{2.3}$$

ou

$$\sum_{i=1}^{m} X_i Y_i - a_0 \sum_{i=1}^{m} X_i - a_1 \sum_{i=1}^{m} X_i^2 = 0$$

Segue que

$$a_0 \sum_{i=1}^{m} X_i + a_1 \sum_{i=1}^{m} X_i^2 = \sum_{i=1}^{m} X_i Y_i. \tag{2.4}$$

Portanto, os parâmetros a_0 e a_1 que resultam na minimização do erro são soluções do sistema de equações lineares (2.2) e (2.4) conhecido como sistema de equações normais. Reescrevendo na forma matricial:

$$\begin{bmatrix} m & \sum_{i=1}^{m} X_i \\ \sum_{i=1}^{m} X_i & \sum_{i=1}^{m} X_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} Y_i \\ \sum_{i=1}^{m} X_i Y_i \end{bmatrix}.$$

Aplicando o Método de Cramer ao sistema linear de 2 equações à 2 incógnitas resulta:

$$a_0 = \frac{\sum_{i=1}^{m} X_i^2 \sum_{i=1}^{m} Y_i - \sum_{i=1}^{m} X_i \sum_{i=1}^{m} X_i Y_i}{m \sum_{i=1}^{m} X_i^2 - \left(\sum_{i=1}^{m} X_i \right)^2}. \tag{2.5}$$

e

$$a_1 = \frac{m \sum_{i=1}^{m} X_i Y_i - \sum_{i=1}^{m} X_i \sum_{i=1}^{m} Y_i}{m \sum_{i=1}^{m} X_i^2 - \left(\sum_{i=1}^{m} X_i \right)^2}. \tag{2.6}$$

Para maiores detalhes ver Ruggiero e Lopes (1996) e Salvador e Arenales (2012).

Exemplo 2.2. Um professor de matemática decide conduzir uma pesquisa sobre como o tempo dedicado ao estudo influencia o desempenho acadêmico de seus alunos. Visando entender graficamente a relação entre essas duas variáveis (tempo versus desempenho).

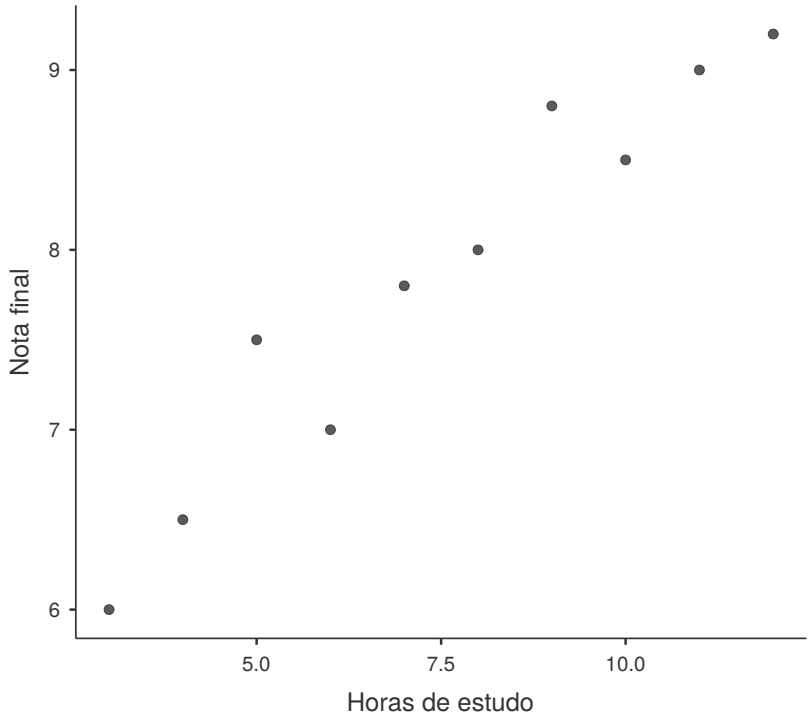
Tabela 1 – Horas semanais dedicadas ao estudo e as respectivas notas finais.

Horas de Estudo (X)	**Nota Final (Y)**
5	7,5
10	8,5
7	7,8
3	6,0
8	8,0
12	9,2
6	7,0
9	8,8
4	6,5
11	9,0

Fonte: Elaborada pela autora.

Pergunta-se: qual é a função matemática que melhor aproxima os dados? Inicialmente verificamos o diagrama de dispersão. Ver Figura 2.

Figura 2 – Diagrama de dispersão.



Fonte: JAMOVI (2023).

Observamos um certo “alinhamento” dos pontos (dados) indicando um possível comportamento linear. Fazemos então um ajuste linear segundo o método dos mínimos quadrados. Nesse sentido, utilizamos a Tabela 2.

Tabela 2 – Horas semanais dedicadas ao estudo e as respectivas notas finais, e seus cálculos.

Aluno	**Horas de Estudo** (X_i)	**Nota Final** (Y_i)	X_i^2	Y_i^2	X_iY_i
1	5	7,5	25	56,25	37,5
2	10	8,5	100	72,25	85
3	7	7,8	49	60,84	54,6
4	3	6	9	36	18
5	8	8	64	64	64
6	12	9,2	144	84,64	110,4
7	6	7	36	49	42
8	9	8,8	81	77,44	79,2
9	4	6,5	16	42,25	26
10	11	9	121	81	99
Soma	**75**	**78,3**	**645**	**623,67**	**615,7**

Fonte: Elaborada pela autora.

Substituindo os valores da Tabela 2 nas expressões (2.5) e (2.6) obtemos

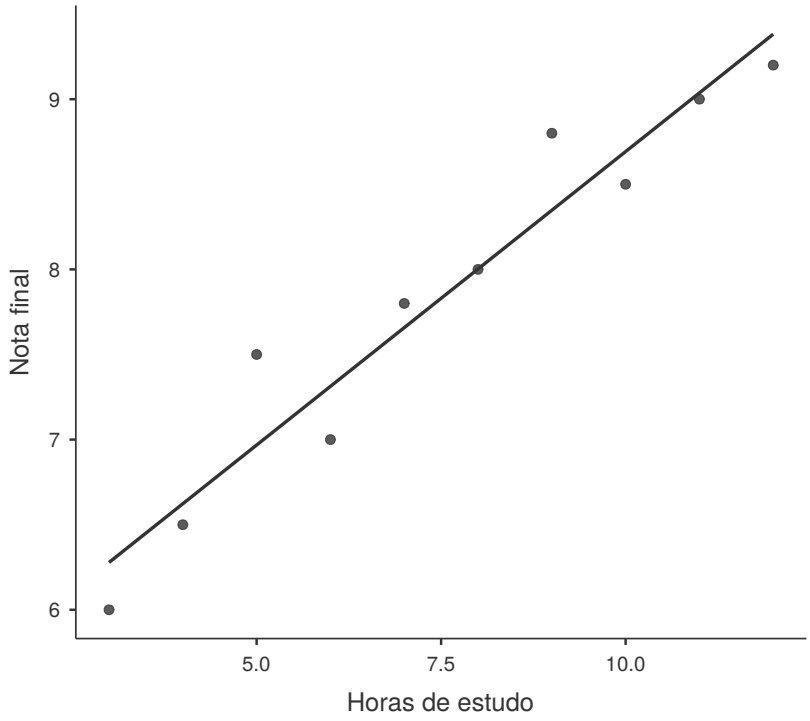
$$a_0 = \frac{645 \times 78,3 - 75 \times 615,7}{10 \times 645 - 75^2} \approx 5,2436364$$

e

$$a_1 = \frac{10 \times 615,7 - 75 \times 78,3}{10 \times 645 - 75^2} \approx 0,3448485$$

respectivamente. A equação da reta do ajuste é dada por $y = 5,2436364 + 0,3448485x$. A Figura 3 mostra a reta ajustada para o conjunto de dados.

Figura 3 – Diagrama de dispersão com reta ajustada.



Fonte: JAMOVI (2023).

2.2 REGRESSÃO LINEAR SIMPLES

Nesta seção utilizamos como referência o texto de Levine, Stephan e Szabat (2016).

A regressão é um método estatístico empregado para explorar e construir um modelo que descreva a relação entre uma variável dependente (ou resposta) e uma ou mais variáveis independentes (ou preditoras). O propósito fundamental da análise de regressão é compreender de que maneira as alterações nas variáveis independentes estão relacionadas às mudanças na variável dependente. Em termos mais simples, a regressão busca estabelecer uma fórmula matemática que melhor representa a conexão entre essas variáveis.

O modelo de regressão simples supõe que a relação entre as variáveis pode ser aproximada por uma função linear, isto é , uma reta. A equação geral para o modelo de regressão simples é dada por

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{2.7}$$

sendo X a variável independente, Y a variável dependente, β_0 o intercepto da reta de regressão em $X = 0$, β_1 a declividade da reta e ε o erro da parte não explicada no modelo.

Nesta seção abordaremos diversos conceitos e técnicas necessárias para compreender e aprimorar o modelo estatístico da regressão.

2.2.1 Medidas de variação na regressão e na correlação

Soma Total dos Quadrados: é a medida de variação dos valores de Y_i, em torno da média aritmética[2] $\overline{Y}$. Em outras palavras, essa medida reflete o quão espalhados os valores individuais de Y_i estão em relação à média $\overline{Y}$, revelando a distância dos pontos em relação à média,

$$\begin{aligned} STQ &= \sum_{i=1}^{m}(Y_i-\overline{Y})^2 \\ &= \sum_{i=1}^{m}(Y_i^2-2\overline{Y}Y_i+\overline{Y}^2) \\ &= \sum_{i=1}^{m}Y_i^2-\overline{Y}\left(2\sum_{i=1}^{m}Y_i-\sum_{i=1}^{m}\overline{Y}\right) \\ &= \sum_{i=1}^{m}Y_i^2-\overline{Y}\left(2\sum_{i=1}^{m}Y_i-m\overline{Y}\right) \\ &= \sum_{i=1}^{m}Y_i^2-\overline{Y}\left(2\sum_{i=1}^{m}Y_i-\sum_{i=1}^{m}Y_i\right) \end{aligned}$$

Logo,

$$STQ = \sum_{i=1}^{m}Y_i^2-\frac{1}{m}\left(\sum_{i=1}^{m}Y_i\right)^2. \tag{2.8}$$

Soma dos Quadrados dos Resíduos representa a parte das variações em Y que não é explicada pela regressão. Isso é determinado comparando cada valor real, denotado por Y_i, com o valor previsto pelo modelo, $\widehat{Y}_i$. Em outras palavras, essa medida quantifica a diferença entre o que realmente aconteceu e o que o nosso modelo de previsão sugeriu que aconteceria. Isso nos ajuda a entender o quanto nosso modelo não consegue explicar e prever com

[2] A média, frequentemente referida como "média aritmética" na linguagem comum, é calculada somando-se todos os valores de um conjunto e, em seguida, dividindo-se pelo número total de valores presentes no conjunto. Este cálculo proporciona uma medida representativa do valor típico ou central do conjunto de dados em questão.

precisão.

$$\begin{aligned} SQR &= \sum_{i=1}^{m} (Y_i - \widehat{Y}_i)^2 \\ &= \sum_{i=1}^{m} (Y_i^2 - 2Y_i\widehat{Y}_i + \widehat{Y}_i^2) \\ &= \sum_{i=1}^{m} \left[Y_i^2 - Y_i\widehat{Y}_i - \left(Y_i - \widehat{Y}_i\right)\widehat{Y}_i \right] \end{aligned}$$

Substituindo $\widehat{Y}_i = a_0 + a_1 X_i$ segue que

$$\begin{aligned} SQR &= \sum_{i=1}^{m} \left[Y_i^2 - a_0 Y_i - a_1 X_i Y_i - (Y_i - a_0 - a_1 X_i)(a_0 + a_1 X_i) \right] \\ &= \sum_{i=1}^{m} Y_i^2 - a_0 \sum_{i=1}^{m} Y_i - a_i \sum_{i=1}^{m} X_i Y_i \\ &\qquad - a_0 \sum_{i=1}^{m} (Y_i - a_0 - a_1 X_i) - a_1 \sum_{i=1}^{m} X_i (Y_i - a_0 - a_1 X_i) \end{aligned}$$

Lembrando as condições (2.1) e (2.3) segue que as duas últimas parcelas da equação anterior se anulam. Logo,

$$SQR = \sum_{i=1}^{m} Y_i^2 - a_0 \sum_{i=1}^{m} Y_i - a_i \sum_{i=1}^{m} X_i Y_i. \tag{2.9}$$

Soma dos quadrados devida à regressão. Este é um conceito que reflete a diferença entre o valor médio de Y (a média das respostas) e o valor previsto $\widehat{Y}$ (calculado usando a relação de regressão). Isso nos ajuda a avaliar a variação nos resultados explicada pelas informações disponíveis. Podemos pensar como a distância entre os valores que nosso modelo prevê e a média geral dos valores. Em essência, essa medida nos mostra o quanto a relação de regressão está contribuindo para entender a variação nos dados.

$$\begin{aligned} SQReg &= \sum_{i=1}^{m} (\widehat{Y}_i - \overline{Y})^2 \\ &= \sum_{i=1}^{m} \left[(Y_i - \overline{Y}) - (Y_i - \widehat{Y}_i) \right]^2. \\ &= \sum_{i=1}^{m} \left[(Y_i - \overline{Y})^2 + (Y_i - \widehat{Y}_i)^2 - 2(Y_i - \overline{Y})(Y_i - \widehat{Y}_i) \right] \\ &= \sum_{i=1}^{m} (Y_i - \overline{Y})^2 + \sum_{i=1}^{m} (Y_i - \widehat{Y}_i)^2 - 2\sum_{i=1}^{m} (Y_i - \overline{Y})(Y_i - \widehat{Y}_i) \end{aligned}$$

Considerando as definições de STQ, SQR segue que

$$SQReg = STQ + SQR - 2\sum_{i=1}^{m} Y_i(Y_i - \widehat{Y}_i) + 2\overline{Y}\sum_{i=1}^{m}(Y_i - \widehat{Y}_i).$$

Aplicando a condição (2.1) tem-se

$$SQReg = STQ + SQR - 2\sum_{i=1}^{m} Y_i(Y_i - \widehat{Y}_i).$$

Substituindo $\widehat{Y}_i = a_0 + a_1 X_i$ segue que

$$SQReg = STQ + SQR - 2\sum_{i=1}^{m} Y_i^2 + 2a_0\sum_{i=1}^{m} Y_i + 2a_1\sum_{i=1}^{m} X_i Y_i.$$

Por fim, substituindo as expressões (2.8) e (2.9)

$$\begin{aligned} SQReg = &\sum_{i=1}^{m} Y_i^2 - \frac{1}{m}\left(\sum_{i=1}^{m} Y_i\right)^2 + \sum_{i=1}^{m} Y_i^2 - a_0\sum_{i=1}^{m} Y_i - a_i\sum_{i=1}^{m} X_i Y_i \\ &- 2\sum_{i=1}^{m} Y_i^2 + 2a_0\sum_{i=1}^{m} Y_i + 2a_1\sum_{i=1}^{m} X_i Y_i. \end{aligned}$$

Logo,

$$SQReg = a_0\sum_{i=1}^{m} Y_i + a_1\sum_{i=1}^{m} X_i Y_i - \frac{1}{m}\left(\sum_{i=1}^{m} Y_i\right)^2.$$

É fácil verificar que

$$STQ = SQR + SQReg.$$

Portanto, a soma total dos quadrados será igual a soma dos quadrados dos resíduos mais a soma dos quadrados devido a regressão.

2.2.2 Coeficiente de Determinação

O coeficiente de determinação denotado por r_{XY}^2 pode ser descrito como a razão entre a soma dos quadrados devido à regressão e a soma total dos quadrados, isto é,

$$r_{XY}^2 = \frac{\text{Soma dos Quadrados devida à regressão}}{\text{Soma total dos quadrados}} = \frac{\sum_{i=1}^{m}(\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{m}(Y_i - \overline{Y})^2}$$

Basicamente, o valor de r^2_{XY} avalia a contribuição da regressão para explicar as variações na variável dependente em relação ao modelo de regressão. O valor do coeficiente de determinação varia entre 0 a 1: quanto mais próximo de 1, mais eficaz é o ajuste das variáveis independentes à amostra; quanto mais próximo de 0, menos eficaz é esse ajuste. Em resumo, o coeficiente r^2_{XY} reconhece o grau de ajuste do modelo de regressão aos dados e a influência das variáveis independentes na variável dependente.

Para obter uma avaliação mais precisa do coeficiente de determinação, consideramos o r^2_{XY} ajustado, que leva em consideração o número de variáveis independentes no modelo. Isso é dado por

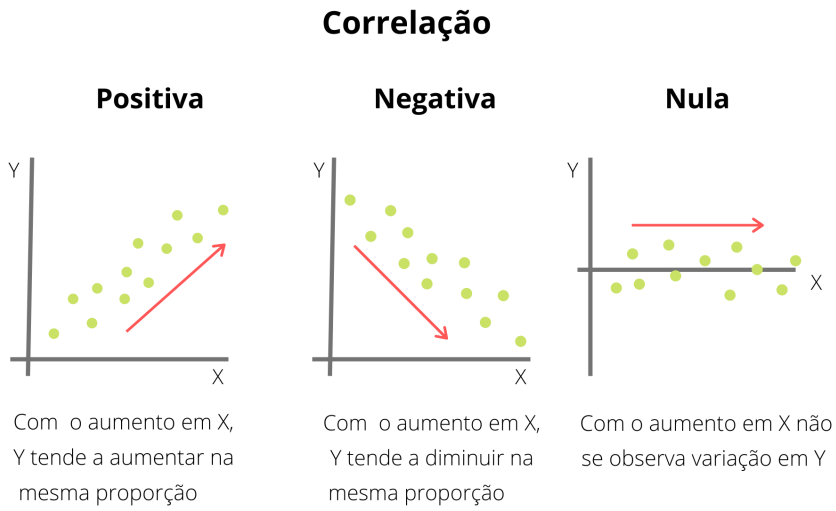
$$r^2_{\text{ajust}} = 1 - \left[(1 - r^2_{XY})\frac{m-1}{m-2}\right].$$

2.2.3 Coeficiente de Correlação

A análise de correlação tem como objetivo quantificar o grau de associação entre duas variáveis e entender como elas se relacionam. Essa conexão é frequentemente avaliada por meio do coeficiente de correlação r que varia de -1 (correlação negativa perfeita) a +1 (correlação positiva perfeita). O valor de r_{XY} reflete a força e a direção da relação entre as variáveis, permitindo discernir se elas tendem a se mover juntas, se afastar juntas ou não têm uma ligação linear significativa.

A Figura 4 ilustra os tipos de associação entre as variáveis.

Figura 4 – A correlação visualizada no diagrama de dispersão.



Fonte: Figura 4.2 (ESQUERRE et al., 2023).

A fórmula para calcular o coeficiente de correlação pode ser derivada a partir da equação do coeficiente de determinação. Isso significa que:

$$r_{XY} = [\text{sinal de } a_1]\sqrt{r_{XY}^2}.$$

Apesar da distinção entre regressão, que prevê, e correlação, que relaciona, o coeficiente de correlação em regressão linear é usado para avaliar a força e direção da relação linear entre variáveis independentes e dependentes. Este coeficiente não apenas ajuda a entender a relação entre as variáveis, mas também desempenha um papel significativo na interpretação dos resultados da regressão.

2.3 TESTE DE SIGNIFICÂNCIA NA REGRESSÃO LINEAR SIMPLES

O modelo de regressão linear simples é dado pela expressão (2.7). O método dos mínimos quadrados irá fornecer as estimativas para a determinação dos parâmetros β_0 e β_1 para o modelo de regressão linear simples. Na equação de regressão linear simples, a média de Y é modelada como $E(Y) = a_0 + a_1X$. Se a_1 é zero, não há relação linear entre X e Y. Se a_1 não é zero, há uma relação linear. Para avaliar a significância dessa relação, realizamos um teste para a_1, com duas abordagens, ambas precisando de uma estimativa de σ^2, que representa a variância[3] do termo de erro ε no modelo de regressão.

2.3.1 Erro Padrão da Estimativa

Ao aplicar o método dos mínimos quadrados para derivar a equação da regressão linear simples, calculamos previsões específicas para um conjunto de dados. É crucial observar que, embora esse método minimize a variação e ajuste uma reta aos dados, a equação da regressão linear não é uma previsão perfeita, a menos que todos os pontos estejam precisamente alinhados com a reta de regressão. Reconhecemos que os valores do conjunto de dados não coincidem exatamente com suas médias aritméticas. Para avaliar adequadamente os valores reais de Y, em comparação com as previsões $\widehat{Y}$, e refletir a variabilidade em torno da linha de regressão, utilizamos o **erro padrão da estimativa** denotado por s_{YX}^2. Essa medida de variabilidade é denotada como erro quadrático médio obtida pela divisão da soma dos quadrados dos resíduos

[3] A variância é uma medida que quantifica o quão dispersos ou espalhados estão os valores de um conjunto de dados em relação à média desse conjunto. Ela é calculada encontrando a média dos quadrados das diferenças entre cada valor individual e a média do conjunto de dados. A variância é expressa em unidades ao quadrado (por exemplo, se os dados representam medidas em metros, a variância seria em metros ao quadrado).

$$\sigma^2 = \frac{\sum_{i=1}^{m}(X_i - \overline{X})}{m}$$

onde X_i é cada ponto de dados, $\overline{X}$ é a média dos dados e m é o número de observações.

$(SQRes)$ pelo grau de liberdade $(m-2)$, isto é,

$$s_{YX}^2 = SMRes = \frac{SQRes}{m-2}.$$

Obtemos dessa forma o quadrado médio dos resíduos definindo um estimador não viesado da variância σ^2.

Para obter uma estimativa de σ^2, calculamos sua raiz quadrada. O resultado é denotado por s e é dado por

$$s_{YX} = \sqrt{\frac{\sum_{i=1}^{m}(Y_i - \widehat{Y}_i)^2}{m-2}}$$

sendo m o número total de observações no conjunto de dados representando o tamanho da amostra, p o número de parâmetros estimados no modelo de regressão e $(Y_i - \widehat{Y}_i)^2$ se refere aos quadrados dos resíduos. Cada resíduo é a diferença entre o valor observado (Y_i) e o valor previsto ou estimado $\widehat{Y}_i$ para cada ponto de dados (X_i, Y_i). s é conhecido como o erro padrão da estimativa.

2.3.2 Teste da hipótese de ausência de relação linear

O teste t é frequentemente utilizado para avaliar a significância estatística dos coeficientes em uma regressão linear simples, com foco especial no coeficiente a_1, que representa a inclinação da linha de regressão.

A hipótese do teste é

$$\begin{cases} H_0: & a_1 = 0 \\ H_a: & a_1 \neq 0. \end{cases}$$

A hipótese H_0 nos diz que não existe relação linear entre X e Y, isto é, não há correlação entre X e Y.

A estatística de teste t para a_1 é calculada dividindo o valor estimado de a_1 pelo seu desvio padrão estimado s_{a_1}, isto é,

$$t = \frac{a_1}{s_{a_1}} \qquad \text{com} \qquad s_{a_1} = \frac{s}{\sqrt{\sum_{i=1}^{m}(X_i - \overline{X})^2}}.$$

Esta estatística é a versão padrão para o coeficiente angular a_1 determinado no método de mínimos quadrados. Temos assim a seguinte regra de rejeição:

Critério do Valor-p: Rejeitamos H_0 se o valor-p associado ao teste for menor que o nível de

significância escolhido.

Critério do Valor Crítico: Rejeitamos H_0 se a estatística de teste t estiver fora do intervalo crítico definido pelos valores críticos de $t_{\frac{\alpha}{2}}$ e $-t_{\frac{\alpha}{2}}$, onde $t_{\frac{\alpha}{2}}$ é determinado pelos graus de liberdade $n-2$ e α é o nível de significância.

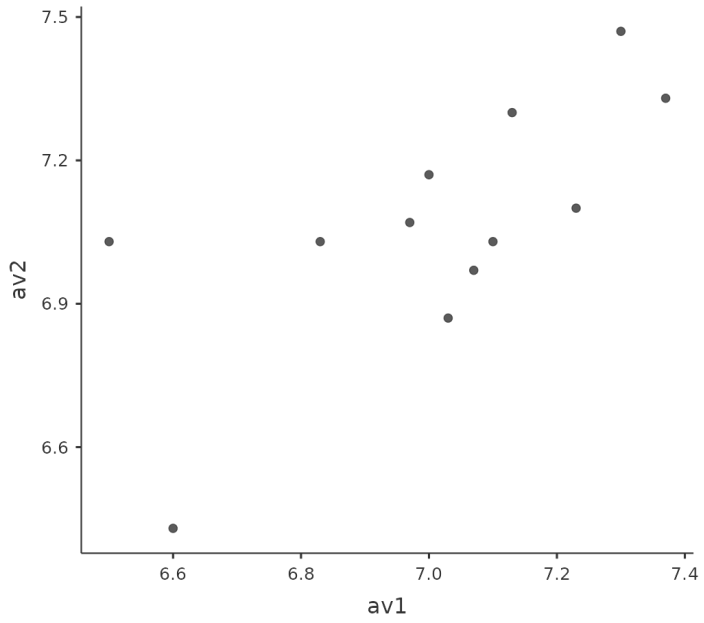
Exemplo 2.3. *Adaptado do Exemplo 10.7 de Moore (2000, p. 426).* Doze estudantes fizeram duas avaliações de uma determinada disciplina. As notas obtidas nas duas avaliações são mostradas na Tabela 3. Podemos nos perguntar se as notas obtidas na primeira avaliação podem predizer as notas na segunda avaliação, isto é, um bom rendimento na primeira prova irá levar a um aumento da nota na segunda avaliação?

Tabela 3 – Notas obtidas de doze estudantes em duas avaliações.

Estudante	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Avaliação 1 (av1)	7,03	7,00	7,10	6,83	7,13	7,30	6,60	6,50	7,23	7,07	6,97	7,37
Avaliação 2 (av2)	6,87	7,17	7,03	7,03	7,30	7,47	6,43	7,03	7,10	6,97	7,07	7,33

O diagrama de dispersão entre as variáveis (av1) e (av2) é mostrado na Figura 5.

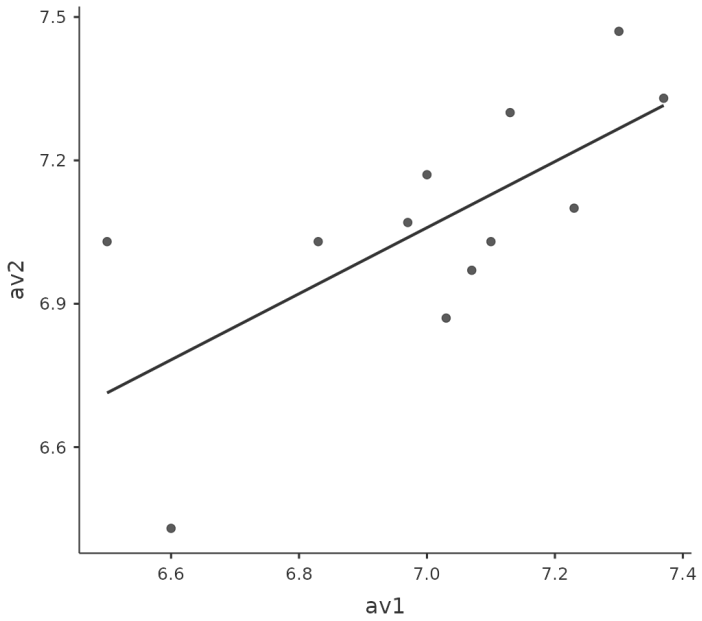
Figura 5 – Gráfico de dispersão entre as avaliações 1 e 2.



Fonte: (JAMOVI, 2023).

Visualmente é aceitável que há uma relação linear entre (av1) e (av2). Em seguida, determinamos a regressão linear entre as duas avaliações. Isto é mostrado na Figura 6.

Figura 6 – A correlação visualizada no diagrama de dispersão.



Fonte: (JAMOVI, 2023).

Visualmente verificamos que os estudantes 7 e 8 são pontos incomuns (os dois primeiros pontos mais distantes da reta) em um possível modelo de regressão linear por estarem distantes da reta de regressão.

Em seguida determinamos o erro padrão da estimativa e o teste da hipótese de ausência de relação linear, com mostra a Figura 7.

Figura 7 – Análise de regressão para os dados da Tabela tab:exemplo2.3.

Medidas de Ajustamento do Modelo

Modelo	R	R^2
1	0.69	0.48

Coeficientes do Modelo - av2

Preditor	Estimativas	Erro-padrão	t	p
Intercepto	2.22	1.61	1.38	0.199
av1	0.69	0.23	3.01	0.013

Fonte: (JAMOVI, 2023).

Com resposta obtemos o modelo de regressão

$(av2) = 2,22 + 0,69(av1)$.

Recorrendo aos valores tabelados para o teste t verificamos que: com grau de liberdade igual à 10 e p-valor de 0,005, o valor crítico de t é de 3,169. O valor t encontrado foi de 3,01, abaixo do valor crítico e valor p igual a 0,013. Portanto, há uma forte evidência de que as notas da segunda avaliação aumentam linearmente com as notas obtidas na primeira avaliação.

2.4 AVALIAÇÃO DA ADEQUAÇÃO DO MODELO AJUSTADO

Os valores de erros estimados (e_i) ou resíduos são definidos como a diferença entre os valores de Y_i observados e os valores previstos $\widehat{Y_i}$ de variável dependente para valores dados de X_i.

O resíduo padronizado é o quociente entre o resíduo e a estimativa do seu desvio padrão determinado para cada observação. Os resíduos de Student são resíduos padronizados ajustados para a distância do valor de X médio, ou seja,

$$\text{Resíduo de Student} = RS_i = \frac{e_i}{S_{YX}\sqrt{1-h_i}}$$

onde

$$h_i = \frac{1}{m} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{m}(X_i - \overline{X})^2}$$

Os resíduos de Student permitem considerar a magnitude dos resíduos em unidades que refletem as variações padronizadas em torno da linha de regressão.

2.4.1 A estimativa do intervalo de confiança para a média aritmética de Y_i

A estimativa do intervalo de confiança para a média aritmética de Y_i na análise de regressão é uma ferramenta estatística fundamental para avaliar a incerteza em torno da média populacional de Y_i, especialmente quando fazemos previsões para um valor específico de X_i. Essa abordagem fornece uma faixa de valores prováveis para a média de Y_i com base na equação de regressão linear e é crucial para uma interpretação robusta dos resultados.

A fórmula do intervalo de confiança, $\widehat{Y_i} \pm t_{\alpha/2} S_{YX}\sqrt{h_i}$, incorpora vários elementos essenciais:

a) $\widehat{Y_i}$: Representa a previsão para Y em um valor específico X_i, utilizando a equação de regressão linear.

b) $t_{\alpha/2}$: O valor crítico associado a uma probabilidade de cauda superior de $\alpha/2$ na distribuição t com $m-2$ graus de liberdade. Este valor é crucial para ajustar o intervalo de acordo com o nível de confiança desejado $(1-\alpha)$.

c) S_{YX}: O erro-padrão da estimativa irá refletir a dispersão dos pontos de dados em torno da linha de regressão. Quanto maior o valor de S_{YX}, maior será a amplitude do intervalo, indicando maior incerteza.

d) h_i é um termo que considera a variabilidade dos dados em torno do valor específico de X_i. Incorpora a variabilidade média $\frac{1}{n}$ e a influência do valor X_i na variabilidade de Y.

O intervalo resultante $\widehat{Y}_i \pm t_{\alpha/2} S_{YX} \sqrt{h_i}$ fornece uma faixa dentro da qual a verdadeira média populacional de Y_i para um dado X_i é estimada com um nível de confiança de $(1-\alpha)$. Se aumentarmos o nível de confiança, a amplitude do intervalo cresce, refletindo uma maior incerteza na estimativa.

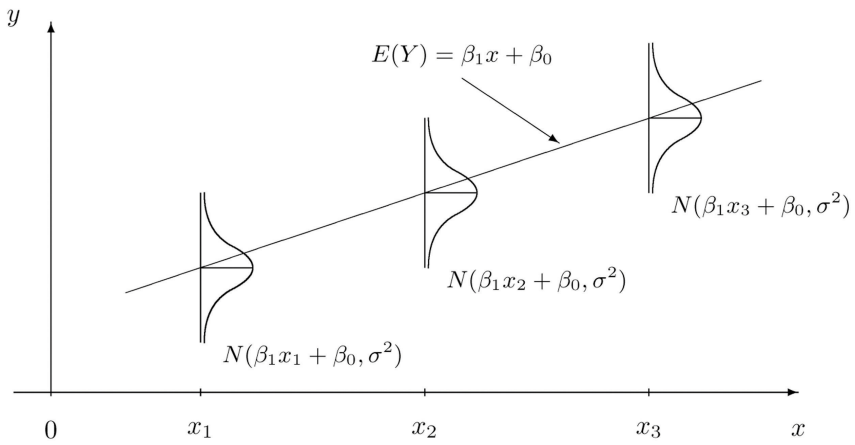
É crucial notar que a largura do intervalo é influenciada por diferentes fatores. Maior flutuação em torno da linha de regressão, indicada por um S_{YX} maior, resulta em um intervalo mais amplo. Entretanto, um aumento no tamanho da amostra tende a reduzir a largura do intervalo, proporcionando maior confiança na estimativa. Além disso, a posição dos valores de X_i também desempenha um papel importante, pois, estimativas para valores de X_i próximos à média aritmética resultam em intervalos mais estreitos, enquanto estimativas para valores distantes da média ampliam o intervalo.

Visualmente, esse efeito de banda de confiança pode ser observado em um gráfico, onde diferentes níveis de confiança geram intervalos de larguras variadas, destacando a dinâmica complexa envolvida na predição e na incerteza associada à regressão linear.

2.5 PRESSUPOSTOS DA REGRESSÃO E CORRELAÇÃO

Normalidade: exige que os valores de Y apresentem uma distribuição normal para cada valor de X. Isso significa que, para diferentes níveis de X, os valores de Y devem seguir uma distribuição que se assemelha a uma curva normal.

Figura 8 – Pressuposto da normalidade.



Fonte: Figura 10.5 (Introductory Statistics, 2012).

Homocedasticidade: exige que as variações em torno da linha de regressão se mantenham uniformes para todos os valores de X. Isso significa que as flutuações em Y devem ocorrer de maneira consistente, independentemente de X ser baixo ou alto. A homocedasticidade é essencial para uma estimativa precisa dos coeficientes da regressão e para garantir a validade das conclusões baseadas no modelo.

Independência dos erros: exige que os resíduos (a diferença entre os valores observados de Y e os valores previstos) não tenham relação entre si e sejam independentes para cada valor de X. Isso significa que os erros não devem ser afetados por fatores específicos ou pela ordem dos dados. Essa suposição é essencial para garantir que as conclusões tiradas da análise de regressão sejam válidas e não estejam sujeitas a viés ou influências externas.

Linearidade: estabelece que a relação entre as variáveis deve ser aproximadamente linear, o que significa que os pontos de dados devem formar uma tendência que se aproximem de uma linha reta quando localizados em um gráfico.

3 REGRESSÃO LINEAR MÚLTIPLA

A regressão linear múltipla é uma técnica estatística que busca modelar a relação entre uma variável dependente e duas ou mais variáveis independentes de maneira simultânea. Essa abordagem permite capturar a influência conjunta dessas variáveis independentes na variável dependente. A formulação matemática da regressão linear múltipla é expressa por meio da seguinte equação:

$$Y = a_0 + a_1X_1 + a_2X_2 + \ldots + a_kX_k + \varepsilon$$

em que:

- a_0: intercepto de Y.
- a_i: inclinação de Y em relação à variável X_i quando as variáveis $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k$ são mantidas constantes para $i = 1, \ldots, k$.
- ε: erro. Este termo irá corresponder à variabilidade de Y que não pode ser explicada quando as variáveis independentes variam linearmente.

3.1 MÉTODO DOS MÍNIMOS QUADRADOS NA REGRESSÃO MÚLTIPLA

A utilização do Método dos Mínimos Quadrados na regressão múltipla é semelhante ao seu emprego na regressão simples, onde o objetivo é encontrar a representação mais precisa de um plano (de dimensão k) que se ajusta, segundo um critério, a um conjunto de m pontos $(X_{1i}, \ldots, X_{ki})$, $i = 1, \ldots, m$. O critério e principal objetivo é minimizar a soma dos quadrados dos resíduos, ou seja, as discrepâncias entre os valores observados e previstos para diversas variáveis independentes.

A representação dos mínimos quadrados é expressa da seguinte forma:

$$\min \sum_{i=1}^{m} (Y_i - \widehat{Y})^2$$

onde:

- Y_i: valor observado da variável dependente para a i-ésima observação.
- $\widehat{Y}$: valor previsto (ou ajustado) da variável dependente para a i-ésima observação.

Para construir o método iremos considerar o modelo de forma semelhante ao realizado anteriormente para a regressão linear simples. Assim, admita que para m observações:

$$
\begin{aligned}
Y_1 &= a_1 + a_2X_{21} + a_3X_{31} + \ldots + a_kX_{k1} + \varepsilon_1 \\
Y_2 &= a_1 + a_2X_{22} + a_3X_{32} + \ldots + a_kX_{k2} + \varepsilon_2 \\
&\vdots \\
Y_m &= a_1 + a_2X_{2m} + a_3X_{3m} + \ldots + a_kX_{km} + \varepsilon_m
\end{aligned}
$$

Este conjunto de equações pode ser reescrito em notação matricial como

$$
\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \ldots & X_{k1} \\ 1 & X_{22} & X_{32} & \ldots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2m} & X_{3m} & \ldots & X_{km} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{bmatrix}
$$

Denotando as matrizes por

$$
Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{21} & X_{31} & \ldots & X_{k1} \\ 1 & X_{22} & X_{32} & \ldots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2m} & X_{3m} & \ldots & X_{km} \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} \quad e = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{bmatrix},
$$

Podemos escrever na forma $Y = X\mathbf{a} + e$.

Apesar de estar formulada como uma equação matricial, a sua estrutura é comparável à equação de regressão simples sem intercepto. O estimador de mínimos quadrados para o vetor será, de maneira bastante próxima, análogo ao utilizado na regressão simples, isto é,

$$
\widehat{a} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})
$$

onde

- $\mathbf{X}'$: matriz transposta de $\mathbf{X}$. A transposta de uma matriz é obtida trocando suas linhas por colunas e vice-versa.
- $\mathbf{X}'\mathbf{X}$: produto escalar da transposta de $\mathbf{X}$ com $\mathbf{X}$. esse produto resulta em uma matriz simétrica chamada matriz de produtos cruzados.
- $(\mathbf{X}'\mathbf{X})^{-1}$: matriz inversa da matriz de produtos cruzados.
- $\mathbf{X}'\mathbf{Y}$: produto escalar da transposta de $\mathbf{X}$ com $\mathbf{Y}$. O resultado é um vetor coluna.
- $\widehat{a}$: vetor de coeficientes estimados.

Observa-se que o produto $\mathbf{X}'\mathbf{Y}$ é análogo a $\sum_{i=1}^{m} x_i y_i$ na regressão simples, enquanto o produto $\mathbf{X}'\mathbf{X}$ é análogo a $\sum_{i=1}^{m} x_i^2$. Como não há divisão de matrizes, a multiplicação pela matriz inversa desempenha o papel do elemento inverso multiplicativo.

A condição essencial para que o estimador $\widehat{a}$ seja viável é garantir que a matriz $X'X$ seja invertível. Isso ocorre quando cada coluna da matriz X é linearmente independente, ou seja, nenhuma variável é uma combinação linear das demais. Dessa forma, evita-se situações onde uma variável (X_2, por exemplo) é precisamente o dobro de outra (X_3), ou casos onde X é representado como $2X_1 + 3X_2$.

3.1.1 Coeficiente de determinação múltiplo

O coeficiente de determinação múltiplo é semelhante ao utilizado na regressão linear simples, onde a soma total dos quadrados pode ser decomposta em duas componentes: a soma dos quadrados devido à regressão *SQReg* e a soma dos quadrados devido ao erro *SQRes*. A relação entre essas grandezas é expressa por:

$$SQTotal = SQReg + SQRes$$

onde

- *SQT* é a soma dos quadrados total calculada como $\sum_{i=1}^{m} (Y_i - \bar{Y}_i)^2$.
- *SQReg* é a soma dos quadrados devido à regressão determinada por $\sum_{i=1}^{m} \left(\widehat{Y}_i - \bar{Y}_j\right)^2$.
- *SQRes* é a soma dos quadrados devido ao resíduos obtida por $\sum_{i=1}^{m} \left(Y_i - \widehat{Y}_i\right)^2$.

O coeficiente de determinação múltiplo, R^2 é calculado como:

$$R^2 = \frac{SQReg}{SQTotal}.$$

O coeficiente de determinação múltiplo R^2 indica a proporção da variabilidade na variável dependente explicada pelo modelo, variando entre 0 e 1. À medida que incorporamos mais variáveis, a precisão preditiva da equação melhora, reduzindo erros de previsão e diminuindo a soma dos quadrados devido ao erro (*SQRes*). Isso aumenta a *SQReg*, contribuindo para o aumento de R^2. Quando novas variáveis são adicionadas, R^2 aumenta, mesmo que algumas delas não sejam estatisticamente significativas. O R^2 ajustado compensa automaticamente o acréscimo no número de variáveis independentes. Assim, um aumento em R^2 nem sempre indica uma

melhoria significativa, podendo ser influenciado pela inclusão de variáveis com contribuição limitada.

Em análises de regressão múltipla, ajustar o R^2 para o número de variáveis independentes evita superestimar o impacto ao adicionar uma variável sobre a variabilidade explicada pela equação de regressão estimada. Usando m para o número de observações e p para o número de variáveis independentes, o coeficiente de determinação múltiplo ajustado é calculado como:

$$R^2_{\text{ajustado}} = 1 - \left(1 - R^2\right) \frac{m-1}{m-p-1}$$

3.1.2 Pressupostos da regressão linear múltipla

Veremos agora quais são os pressupostos para regressão linear múltipla. São eles:

Média do Termo de Erro (ε). O termo de erro (ε) é uma variável aleatória[4] com média zero ($E(\varepsilon) = 0$). Em outras palavras, para um conjunto específico de valores das variáveis independentes ($X_1, X_2, \ldots, X_k$), a média dos valores observados de Y é determinada pela equação de regressão múltipla.

Variância constante do termo de erro (ε). A variância do termo de erro (ε), representada por σ^2, é constante para todos os valores das variáveis independentes ($X_1, X_2, \ldots, X_k$). Isso implica que a dispersão dos valores de Y ao longo da linha de regressão permanece constante.

Independência dos termos de erro (ε). Os valores do termo de erro (ε) são independentes entre si. Em outras palavras, o valor do erro para um conjunto específico de valores das variáveis independentes não está relacionado ao valor do erro para qualquer outro conjunto de valores.

Distribuição normal do termo de erro (ε). O termo de erro (ε) segue uma distribuição normal. Isso significa que, dados os coeficientes $a_0, a_1, a_2, \ldots, a_k$ para um conjunto de valores de ($X_1, X_2, \ldots, X_k$), a variável dependente Y segue uma distribuição normal[5]

Essas suposições são cruciais para garantir uma análise estatística apropriada da relação

[4] Uma variável aleatória é uma função que associa um número real a cada resultado possível de um experimento probabilístico.

[5] A função de densidade de probabilidade (PDF) da distribuição normal é dada por

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

onde onde: x é a variável aleatória, μ é a média da distribuição e σ é o desvio padrão da distribuição. O Teorema Central do Limite afirma que a soma (ou média) de um grande número de variáveis aleatórias independentes e identicamente distribuídas, independentemente da forma da distribuição original, se aproxima de uma distribuição normal.(MOORE, 2000, p. 211).

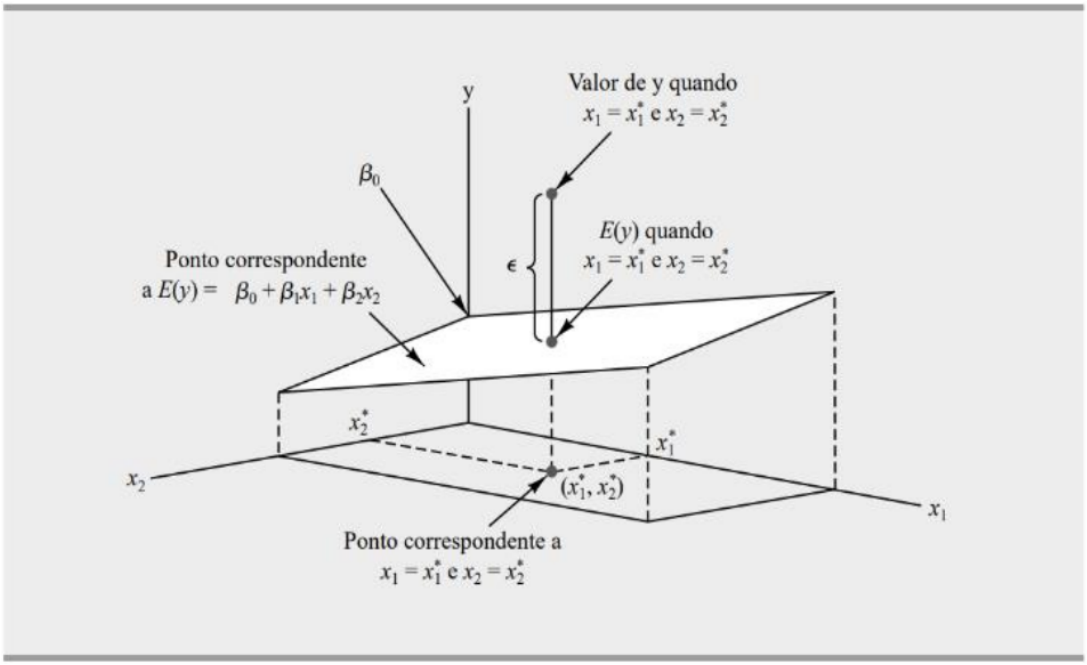
expressa pela equação de regressão múltipla, considerando a média, variância e distribuição dos erros associados ao modelo.

Para obtermos uma melhor compreensão da relação do modelo de regressão múltipla, iremos nos restringir a equação de regressão múltipla com duas variáveis independentes, isto é,

$$E(y) = a_0 + a_1 x_{i1} + a_2 x_{i2}. \tag{3.1}$$

A equação (3.1) descreve um plano em um espaço tridimensional. A Figura 9 fornece um exemplo gráfico desse cenário.

Figura 9 – Gráfico da equação de regressão para a análise de regressão múltipla com duas variáveis independentes.



Fonte: Figura 15.5 (ANDERSON et al., 2021, p. 617).

No contexto da análise de regressão, o termo "variável resposta" é frequentemente empregado em lugar de "variável dependente". Além disso, dado que a equação de regressão múltipla gera um plano ou superfície, o gráfico associado é denominado superfície de resposta.

3.2 TESTE DE SIGNIFICÂNCIA

Nesta seção, abordamos a realização de testes de significância em uma análise de regressão múltipla. Ao contrário da regressão linear simples, os testes t e F geralmente convergem para a mesma conclusão. Em resumo, a rejeição da hipótese nula implica que o coeficiente (a_1) é diferente de zero. No entanto, na regressão múltipla, esses testes servem a propósitos distintos.

3.2.1 Teste F

O teste F verifica se há uma relação globalmente significativa entre a variável dependente e todas as variáveis independentes, sendo conhecido como teste de significância global.

3.2.1.1 Teste para significância geral

Expressão das Hipóteses Nula e Alternativa:

$$\begin{cases} H_0 : a_1 = a_2 = a_3 = \ldots = a_k \\ H_a : \text{Pelo menos um dos parâmetros não é igual a zero} \end{cases}$$

Estatística do teste:

$$F = \frac{\text{Soma dos Quadrados devido à Regressão}}{\text{Soma dos Quadrados dos Resíduos}/(n-p-1)}$$

sendo

- *SQReg* é a Soma dos Quadrados devido à Regressão.
- *SQRes* é a Soma dos Quadrados dos Resíduos.
- p é o número de parâmetros no modelo (número de variáveis independentes, excluindo a constante).
- n é o número total de observações.

Regra de Rejeição:

- **Critério do Valor-p:** Rejeitar H_0 se o valor-p for menor que o nível de significância escolhido.
- **Critério do Valor Crítico:** Rejeitar H_0 se F for maior que o valor crítico, onde F é baseado em uma distribuição F com p graus de liberdade no numerador e $n-p-1$ no denominador.

3.2.1.2 Teste t

Se o teste F mostra que a relação da regressão múltipla é significativa, o teste t serva para avaliar a importância individual de cada variável independente. Realizamos um teste t separado para cada variável no modelo, chamando essas análises de testes t individuais de significância.

– Para qualquer parâmetro a_i:

$$\begin{cases} H_0 : a_i = 0 \\ H_a : a_i \neq 0 \end{cases}$$

– Estatística do Teste:

$$t = \frac{b_i}{sb_i}$$

– Regra de Rejeição:

  – Critério do valor$-p$: rejeitar H_0 se valor$-p \leqslant \alpha$.
  – Critério do valor crítico: rejeitar H_0 se $t \geqslant -t_{\frac{\alpha}{2}}$ ou se $t \geqslant t_{\frac{\alpha}{2}}$ onde $t_{\frac{\alpha}{2}}$ e baseado na distribuição t com $n-p-1$ graus de liberdade.

Exemplo 3.1. *Adaptado do Exemplo de Downing e Clark (2002, p. 263-270).* Um professor coletou dados sobre a venda de livros de matemática em quinze cidades durante um período específico. Ele deseja discutir com seus alunos a relação que a variável dependente (vendas do livro) possui com duas variáveis independentes: $X1$ (preço do livro) e $X2$ (renda da população).

A Tabela 4 mostra os dados obtidos.

Tabela 4 – Dados de vendas de livros por preço e renda em 15 cidades.

cidade	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
venda - Y	166	180	73	81	229	182	233	102	190	150	221	137	173	150	92
preço - X1	10	9	10	14	8	15	6	10	7	10	11	15	8	12	10
renda - X2	20	21	12	16	24	24	23	15	20	19	25	21	19	20	14

Fonte: Adaptada pela autora de Downing e Clark (2002, p. 264).

A relação entre as variáveis é modelada pela fórmula da regressão múltipla, onde buscamos estimar os parâmetros mais apropriados para compreender a relação das variáveis independentes com a variável dependente.

Através dos resultados obtidos no software (JAMOVI, 2023) (Ver Figura 10, temos que:

Figura 10 – Resultados da regressão múltipla

Medidas de Ajustamento do Modelo

				Teste ao Modelo Global			
Modelo	R	R^2	R^2 Ajustado	F	gl1	gl2	p
1	0.998	0.996	0.995	1391.898	2	12	< .001

Teste ANOVA omnibus

	Soma de Quadrados	gl	Quadrado médio	F	p
PREÇO - X1	6184.849	1	6184.849	452.482	< .001
RENDA - X2	31346.598	1	31346.598	2293.311	< .001
Resíduos	164.025	12	13.669		

Nota. Soma de Quadrados de tipo 3

Coeficientes do Modelo - VENDA - Y

Preditor	Estimativas	Erro-padrão	t	p
Intercepto	−2.765	6.391	−0.433	0.673
PREÇO - X1	−7.738	0.364	−21.272	< .001
RENDA - X2	12.286	0.257	47.889	< .001

Fonte: Resposta do programa (JAMOVI, 2023).

A relação entre as variáveis é modelada pela fórmula da regressão múltipla, onde buscamos estimar os parâmetros mais apropriados para compreender a relação das variáveis independentes com a variável dependente. Obtemos, dessa forma, os seguintes estimadores para expressão:

$$y = -2,76 - 7,74X1 + 12,29X2. \tag{3.2}$$

Através do resultado obtido em (3.2), podemos fazer algumas interpretações:

– O coeficiente $X1$ é negativo. Assim, quanto maior for o preço de um livro, menor será o valor de venda, ou seja, teremos uma diminuição média de 7.74 unidades nas vendas para cada aumento unitário no preço do livro.

– O R^2 indica que a proporção de 99,5% de variabilidade total com relação a variável dependente explicada pelas variáveis independentes.

– A uma relação globalmente significativa entre a variável dependente e todas as variáveis

independente, já que através do teste F concluímos que o valor de p é pois o valor de p<0,05.

– Analisando individualmente cada variável, obtemos através do teste t que as variáveis independentes são significadas no modelo. Entretanto, apenas o intercepto que não, onde poderíamos interpretar como sendo o custo inicial de cada livro.

4 REGRESSÃO LOGÍSTICA

Neste capítulo abordaremos o método da regressão logística e faremos uso de Fávero e Belfiore (2017). O objetivo é determinar o modelo mais adequado para descrever a relação entre a variável resultado (dependente) e as variáveis independentes (preditoras). Ao contrário da regressão linear, em que a variável de dependente é contínua, na regressão logística, a variável de resultado é binária.

Temos diferentes tipos de regressão logística, por exemplo:

Regressão Logística Binária: Envolve situações em que a variável dependente tem apenas duas possibilidades de resposta, como por exemplo, "sim" ou "não". Esta técnica é utilizada para estimar a probabilidade dessas alternativas ocorrerem com base em variáveis explicativas, tais como o tempo dedicado aos estudos e a idade, por exemplo.

Regressão Logística Multinomial: Envolve situações em que a variável dependente possui mais que duas possibilidades de respostas, por exemplo: "gato", "cachorro" ou "pássaro". Nesse contexto, cada resposta é tratada como uma categoria distinta. Esta técnica é utilizada para prever a probabilidade de cada categoria ocorrer para um indivíduo específico com base em variáveis explicativas.

Regressão Logística Ordinal: Envolve situações em que as variáveis dependentes, além de serem categorizadas, possuem uma ordem específica, ou seja, são classificadas em diferentes níveis e graus de intensidade. Esta técnica é empregada, por exemplo, em pesquisas de satisfação ao cliente, avaliação de produtos, educação e qualidade de vida. Ela é aplicada sempre que se busca compreender não apenas se algo acontecerá ou não, mas também em que medida ou intensidade isso ocorrerá.

Regressão Logística Ponderada: O modelo tem como objetivo ajustar a importância relativa de diversas observações. Nesse caso, irá classificar em diferentes pesos. Isto indica que algumas observações podem ser mais essenciais ou têm mais impacto do que as outras na estimativa dos parâmetros e previsões no modelo.

Regressão Logística Regularizada: A técnica visa simplificar e aprimorar a eficiência do modelo de regressão logística em situações altamente complexas, selecionando apenas variáveis independentes significativas para explicar a variabilidade na variável dependente. O método é utilizado, por exemplo, em classificação de e-mail, previsão de compras online e análise de risco de crédito, em que informações irrelevantes podem afetar a precisão do modelo.

Apesar de diversos tipos de técnicas da regressão logística, o foco do nosso estudo será apresentar as principais ferramentas utilizadas na regressão logística binária.

4.1 CONSTRUÇÃO DA REGRESSÃO LOGÍSTICA BINÁRIA

A regressão logística binária é uma técnica estatística utilizada para avaliar a probabilidade de ocorrência de um evento específico. Nesse contexto, o evento é representado pela variável Y, que assume valores qualitativos dicotômicos, geralmente 1 para indicar a presença do evento e 0 para indicar a sua ausência. A análise é conduzida considerando a influência de variáveis explicativas, que podem ser métricas ou variáveis *dummy*. O modelo de regressão logística binária expressa a relação entre as variáveis explicativas e a probabilidade condicional do evento ocorrer.

Podemos expressar o modelo de regressão logística da seguinte maneira :

$$Z_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$$

sendo

- Z_i é conhecido como logito,
- a é a constante,
- β_j (para $j = 1, \ldots, k$ são os parâmetros estimados de cada variável explicativa,
- X_{ji} são as variáveis explicativas (métricas ou *dummies*),
- i representa cada observação da amostra (com $i = 1, \ldots, n$, onde n é o tamanho da amostra).

É importante esclarecer que Z não representa a variável dependente Y. O objetivo, neste momento, é estabelecer a expressão da probabilidade $P(Y = 1)$ para cada observação, em termos do logito Z_i e, portanto, dos parâmetros estimados para cada variável explicativa.

Para realizar essa transição, é necessário introduzir o conceito de *odds* (chance) de ocorrência de um evento. Os *odds* são definidos como a razão entre a probabilidade de o evento acontecer ($P(Y = 1)$) e a probabilidade de não acontecer ($P(Y = 0)$). A expressão para os *odds* é dada por:

$$odds = \frac{P(Y = 1)}{P(Y = 0)} = \frac{P_i}{1 - P_i}$$

Assim, o objetivo é relacionar essa expressão de *odds* com o logito Z_i e, consequentemente, com os parâmetros estimados das variáveis explicativas. Essa relação é fundamental para compreender como as variáveis explicativas influenciam a probabilidade de ocorrência do evento de interesse em um contexto de regressão logística binária.

Dessa forma, podemos definir o logito de Z como o logaritmo natural da chance, de modo que

$$\ln(odds_{Y=1}) = Z_i$$

substituindo o valor de *odds*, temos :

$$\ln\left(\frac{P_i}{1-P_i}\right) = Z_i$$

Para que possamos definir a expressão da probabilidade de ocorrência em estudo da função logito, temos a seguinte equação :

$$\frac{P_i}{1-P_i} = e^{Z_i} \iff P_i = (1-P_i)e^{Z_i} \iff P_i(1+e^{Z_i}) = e^{Z_i}.$$

Assim, temos as seguintes resultados:

– Probabilidade de ocorrência do evento:

$$P_i = \frac{e^{Z_i}}{1+e^{Z_i}} = \frac{1}{1+e^{-Z_i}}.$$

– Probabilidade de ocorrência do não evento:

$$1-P_i = 1 - \frac{e^{Z_i}}{1+e^{Z_i}} = \frac{1}{1+e^{Z_i}}.$$

Para visualizarmos graficamente o comportamento da probabilidade de ocorrência de um evento P_i em função do logito Z_i, $p = f(Z)$, é necessário que Z assuma valores no intervalo de $-\infty$ a $+\infty$ pois, desta forma, podemos calcular a probabilidade para qualquer valor possível de Z_i.

Para o esboço do gráfico temos os seguintes resultados:

a) $f(Z_i)$ é sempre positiva, isto é $f(Z_i) > 0$ para todo $Z_i \in \mathbb{R}$.

b) O limite da função da probabilidade (HOSMER; LEMESHOW; STURDIVANT, 2013) de ocorrer um evento nos intervalos $-\infty$ a $+\infty$, temos os seguintes resultados:

$$\lim_{Z_i \to -\infty} f(Z_i) = \lim_{Z_i \to -\infty} \frac{1}{1+e^{-Z_i}} = 0.$$

e

$$\lim_{Z_i \to +\infty} f(Z_i) = \lim_{Z_i \to +\infty} \frac{1}{1+e^{-Z_i}} = 1.$$

c) A derivada de $f(Z-i)$ é dada por

$$f'(Z_i) = \frac{e^{-Z_i}}{(1+e^{-Z_i})^2}.$$

Segue que $f'(Z_i) > 0$ para todo $Z_{\in}\mathbb{R}$. Logo, f é crescente.

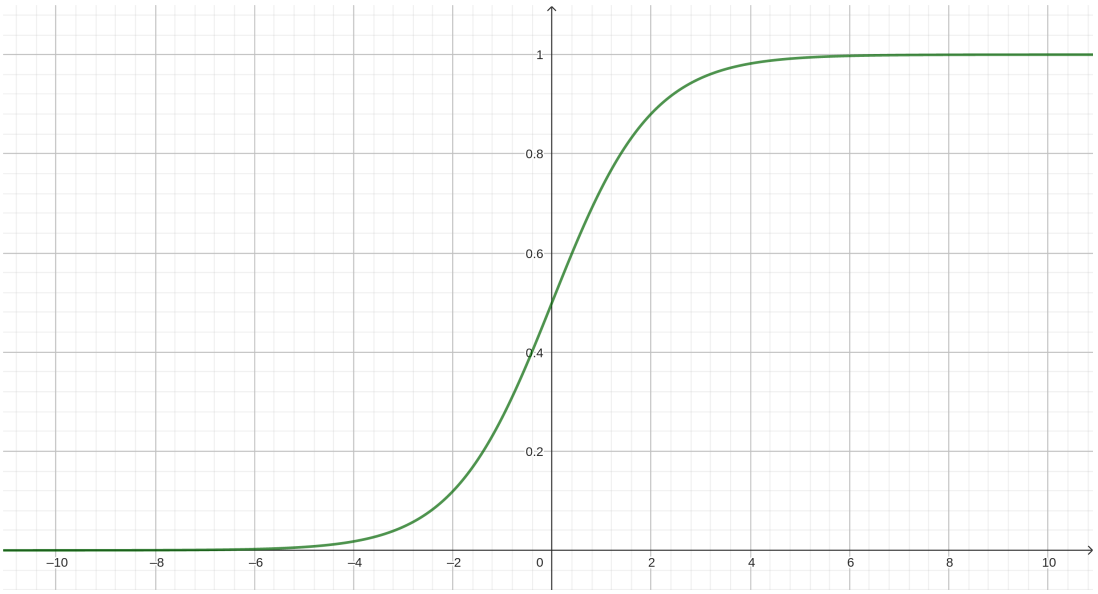
d) A derivada segunda de $f(Z-i)$ é dada por

$$f''(Z_i) = \frac{e^{-Z_i}}{(1+e^{-Z_i})^4}\left(e^{-2Z_i}-1\right).$$

Assim, $Z_i = 0$ será ponto de inflexão e f tem concavidade para cima quando $Z_i < 0$ e concavidade para baixo para $Z_i > 0$.

Concluímos que, independente dos valores de Z_i, as probabilidades resultantes permaneceram dentro do intervalo 0 e 1. Ver Figura 11.

Figura 11 – Função $P = f(Z)$.



Fonte: Elaborada pela autora com uso do software Geogebra.

As expressões obtidas anteriormente permitem representar, de forma dicotômica uma observação i, através da expressão geral da probabilidade estimada de ocorrência do evento da seguinte forma:

$$P_i = \frac{1}{1+e^{-(\alpha+\beta_1 X_{1i}+\beta_2 X_{2i}+\ldots+\beta_k X_{ki})}}.$$

A regressão logística binária estima a probabilidade de ocorrência do evento em estudo

para cada observação, ao contrário da regressão linear, que estima valores previstos da variável dependente.

4.2 ESTIMAÇÃO DO MODELO DA REGRESSÃO LOGÍSTICA POR MÁXIMO VEROSSIMILHANÇA

Na regressão linear simples, utilizamos o método dos mínimos quadrados para estimar os parâmetros, minimizando a soma dos desvios quadráticos entre os valores observados de Y e os valores previstos $\widehat{Y}$, supondo uma distribuição normal dos erros. Na regressão logística, dado que a variável dependente é dicotômica, esse procedimento não é apropriado. Na regressão logística, a variável dependente reflete a ocorrência ou não do evento de interesse em uma observação i. De forma análoga, ao modelo de distribuição de Bernoulli, utilizamos o método da máxima verossimilhança para estimar parâmetros desconhecidos que maximizam a probabilidade de obter o conjunto observado de dados.

Durante a estimação dos parâmetros do modelo logístico, a variável dependente é tratada como binária, indicando a ocorrência (1) ou não (0) do evento de interesse em cada observação. O método da máxima verossimilhança busca encontrar os valores dos parâmetros que tornam mais provável a observação dos resultados reais, considerando a natureza dicotômica da variável dependente.

A probabilidade de ocorrência de Y é dada por:

$$L = \prod_{i=1}^{n} \left[P_i^{Y_i} (1 - P_i)^{1-Y_i} \right].$$

sendo que seus valores são oriundos das expressões da probabilidade de um evento ocorrer e da probabilidade de um evento não ocorrer, isto é,

$$L = \prod_{i=1}^{n} \left[\left(\frac{e^{Z_i}}{1 + e^{Z_i}} \right)^{Y_i} \left(\frac{1}{1 + e^{Z_i}} \right)^{1-Y_i} \right]$$

Em termos práticos, iremos trabalhar com o logaritmo da função de verossimilhança para simplificação dos cálculos e tornar o processo de maximização mais eficiente.

$$LL = \sum_{i=1}^{n} \left[Y_i \ln \left(\frac{e^{Z_i}}{1 + e^{Z_i}} \right) + (1 - Y_i) \ln \left(\frac{1}{1 + e^{Z_i}} \right) \right]$$

A maximização da função de verossimilhança na regressão logística envolve derivar a função logarítmica da verossimilhança em relação aos parâmetros e, em seguida, usar métodos de otimização para encontrar os valores dos parâmetros que maximizam essa função.

4.3 TESTE DE SIGNIFICÂNCIA DO MODELO E DOS PARÂMETROS DA REGRESSÃO LOGÍSTICA

Após a estimação por máxima verossimilhança dos parâmetros na equação de probabilidade do evento, é importante avaliarmos a significância de todos os parâmetros estimados em um determinado nível de confiança.

O teste qui-quadrado (χ^2) proporciona uma avaliação da importância global do modelo de regressão logística, formulando as hipóteses nula e alternativa da seguinte maneira:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \\ H_1 : \text{pelo menos um } \beta_j \neq 0 \end{cases} .$$

O teste qui-quadrado é particularmente apropriado para modelos estimados pelo método da verossimilhança, permitindo uma análise inicial da validade do modelo proposto. Se todos os parâmetros estimados β_j com $j = 1, \ldots, k$ forem estatisticamente iguais a zero, isso sugere que nenhuma das variáveis analisadas têm um efeito significativo na probabilidade do evento ocorrer. Caso algum parâmetro não seja estatisticamente significativo, é necessário reavaliar o modelo, mantendo apenas os parâmetros significativos.

A estatística χ^2 é calculada pela expressão:

$$\chi^2 = -2 \ln \left(\frac{L_0}{L_1} \right)$$

onde L_0 representa a verossimilhança máxima possível para um modelo conhecido como nulo, que inclui apenas a constante, sem nenhuma variável explicativa, e $L_{\text{máximo}}$ é a verossimilhança do modelo ajustado com as variáveis explicativas. Para determinar a rejeição ou não da hipótese nula, é essencial calcular o valor crítico do teste qui-quadrado, obtido a partir da tabela de distribuição do qui-quadrado.

Se o valor da estatística χ^2 for maior que o valor crítico, podemos rejeitar a hipótese nula de que todos os parâmetros β_j são estatisticamente iguais a zero, indicando que pelo menos uma variável X é estatisticamente significativa para explicar a probabilidade de ocorrência do evento.

O teste χ^2 avalia a importância conjunta das variáveis explicativas, sem identificar individualmente quais são estatisticamente significativas para influenciar a probabilidade do evento.

4.4 TESTE DE WALD

Após identificar, no teste de significância, que pelo menos um parâmetro X é diferente de zero, é necessário realizar uma análise mais detalhada do modelo de regressão logística binária usando a estatística z de Wald. Este teste visa determinar a significância estatística de cada

parâmetro considerado no modelo.

As hipóteses nula e alternativa do teste para α e para os parâmetros β_i são, respectivamente :

$$\begin{cases} H_0 : \alpha = 0 \\ H_1 : \alpha \neq 0 \end{cases}$$

e

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

As fórmulas para calcular as estatísticas z de Wald para cada parâmetro α e β_i são dadas, respectivamente, por:

$$z_\alpha = \frac{\alpha}{s.e.(\alpha)}$$

e

$$z_{\beta_i} = \frac{\beta_i}{s.e.(\beta_i)}$$

onde $s.e.$ é o erro padrão de cada parâmetro da análise. Dado a complexidade dos cálculos envolvidos, não faremos esse estudo, porém recomendamos ao leitor a leitura de (ENGLE, 1984).

Após obter as estatísticas z de Wald, os valores críticos são obtidos da tabela de distribuição normal padrão. Para um nível de significância de 5%, os valores críticos são -1,96 para a cauda inferior (probabilidade de 0,025 na cauda inferior em uma distribuição bicaudal) e 1,96 para a cauda superior (probabilidade de 0,025 na cauda superior em uma distribuição bicaudal). Se as estatísticas z de Wald estiverem no intervalo de -1,96 a 1,96, não rejeitamos a hipótese nula, indicando que o parâmetro β_i não é estatisticamente significativo.

4.5 TESTE DA RAZÃO DE VEROSSIMILHANÇA

Em situações em que a variável β_j não demonstra significância estatística no modelo, torna-se necessário considerar a exclusão dessa variável no modelo final. Essa decisão visa obter um modelo mais simplificado e eficiente, focado nas variáveis verdadeiramente relevantes. A exclusão de variáveis não significativas pode facilitar a interpretação do modelo e reduzir a complexidade sem prejudicar sua capacidade preditiva.

Para determinar se o novo modelo estimado (modelo final) mantém a qualidade do ajuste em comparação com o modelo completo estimado, que inclui todas as variáveis explicativas, é empregado o teste de razão de verossimilhança (*likelihood-ratio test*). Este teste avalia a

adequação do ajuste do modelo completo em comparação com o ajuste do modelo final. A expressão para o teste de razão de verossimilhança é a seguinte:

$$\chi^2_{1\ g.l.} = -2(LL_{\text{modelo final}} - LL_{\text{modelo completo}}).$$

Para verificação se a qualidade do teste utilizaremos a tabela da distribuição do qui-quadrado para verificarmos se a exclusão de uma variável

4.6 PSEUDO R^2 DE MCFADDEN

Sabemos que nas regressões lineares as variáveis dependentes são contínuas e estimadas por mínimos quadrados ordinários. O coeficiente de ajuste R^2 é utilizado para avaliar a qualidade do ajuste do modelo aos dados. No entanto, ao aplicar a regressão logística, em que a variável dependente é qualitativa, a interpretação direta do percentual de variância explicado torna-se inadequada. Utiliza-se então, o coeficiente de pseudo R^2_{MF} é definido como

$$R^2_{MF} = 1 - \frac{\ln(L)}{\ln(L_0)}.$$

O pseudo R^2_{MF} de McFadden compara o logaritmo da verossimilhança do modelo completo (L) com o logaritmo da verossimilhança do modelo apenas com intercepto L_0, ou o modelo nulo.

O objetivo desse coeficiente é comparar dois ou mais modelos distintos, ou seja, avaliar diferentes modelos e determinar quão bem cada um se ajusta aos dados em comparação com um modelo nulo. Destaca-se que um dos critérios predominantes para a escolha do modelo é o critério de maior pseudo R^2 de McFadden. Este coeficiente é particularmente útil para decisões de seleção de modelo, proporcionando uma medida relativa da qualidade de ajuste em relação ao modelo

4.7 INTERVALO DE CONFIANÇA PARA OS PARÂMETROS DA REGRESSÃO LOGÍSTICA BINÁRIA

Os intervalos de confiança dos coeficientes da expressão

$$P_i = \frac{1}{1 + e^{-(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki})}}.$$

para os parâmetros α e β_j, $j = 1, \ldots, k$ no nível de confiança de 95%, podem ser escritos, respectivamente, na forma $\alpha \pm 1{,}96[s.e.(\alpha)]$ e $\beta_j \pm [s.e.(b_j)]$. O número 1,96 representa o *z-score* associado a um nível de confiança de 95%, com uma margem de significância de 5%. Utilizando esse valor, é viável calcular os coeficientes estimados dos parâmetros na

expressão de probabilidade do evento em questão. Além disso, podemos determinar os erros padrão correspondentes, as estatísticas z de Wald e estabelecer intervalos de confiança, todos considerando um nível de significância de 5%.

5 CONSIDERAÇÕES FINAIS

Este Trabalho de Conclusão de Curso teve como objetivo entender como podemos modelar as relações entre diferentes variáveis. A motivação do estudo surgiu de um estímulo pessoal de querer entender e interpretar as técnicas utilizadas nas áreas de análise de crédito, *machine learning* e prevenção a fraudes.

Assim, nosso estudo foi desenvolver as principais técnicas e construções das regressões estudadas: simples, múltipla e logística binária. Com este propósito foi possível compreender como cada tipo de regressão se conecta e amplifica o entendimento como um todo, onde foram exibidos exemplos simples e práticos.

Ao longo da análise, as técnicas de regressão linear e logística apresentadas vão além das convencionalmente fornecidas. Isso se justifica pela necessidade de abordagens mais refinadas em situações complexas, onde a tomada de decisões exige *insights* mais profundos e precisos.

Entre uma ampla gama de áreas que utilizam a análise de regressão, destacamos um breve contexto da utilização para a análise de crédito.

A análise de crédito é um processo realizado por instituições financeiras, como bancos e empresas de cartão de crédito, para avaliar a capacidade de um indivíduo ou empresa de cumprir suas obrigações financeiras, especialmente em relação a empréstimos ou linhas de crédito. Dessa forma, as técnicas de regressões são fundamentais para avaliar variáveis relevantes para determinar a probabilidade de um indivíduo ou empresa cumprir suas obrigações de pagamento. Com isso, temos que a regressão múltipla irá estimar os coeficientes associados a cada variável independente (o histórico de pagamento, renda, idade, quantidade de dívidas, entre outras), indicando a contribuição relativa de cada uma para a probabilidade de inadimplência, por exemplo. Já a regressão logística é fundamental para calcular a probabilidade de um evento ocorrer (como a inadimplência) com base nas variáveis independentes analisadas na regressão múltipla. Assim, fará uso da função logística para transformar a soma ponderada das variáveis explicativas em uma probabilidade, garantindo que o resultado esteja entre 0 e 1. Dessa maneira, irá distinguir entre clientes de alto risco e baixo risco de inadimplência. Por meio dessas técnicas, é possível encontrar um equilíbrio entre oferecer oportunidades de crédito e proteger contra possíveis inadimplências (VIEIRA, 2009).

Por fim, nosso estudo revelou algumas limitações, uma vez que o escopo da formação em Matemática não proporciona uma exploração mais profunda das técnicas estatísticas. No entanto, esse cenário abriu perspectivas para futuros aprofundamentos mais detalhado e esclarecedores, especialmente no contexto de possíveis especializações profissionais.

REFERÊNCIAS

ANDERSON, D. R. et al. **Estatística Aplicada a Administração e Economia**. 5. ed. São Paulo: Cengage Learning, 2021. Citado na página 30.

DOWNING, D.; CLARK, J. **Estatística Aplicada**. 2. ed. São Paulo: Saraiva, 2002. Citado na página 32.

ENGLE, R. F. W. Handbook of econometrics. In: ____. Amsterdam: North Holland, 1984. cap. Likelihood ratio, and Lagrange multiplier tests in econometrics, p. 796–801. Citado na página 41.

ESQUERRE, K. P. O. et al. **Uma introdução gentil à Ciência de Dados**. [s.n.], 2023. Disponível em: <https://bookdown.org/cienciadedadosnaep/ebook_ciencia_de_dados/_book/fig_cap4/Fig_correlacao.png>. Acesso em: ago. 2023. Citado na página 18.

FáVERO, L. P.; BELFIORE, P. **Manual de Análise de Dados, Estatística e Modelagem Multivariada com Excel, SPSS e Stata**. 1rd. ed. [S.l.]: Elsevier, 2017. Citado na página 35.

HOSMER, D.; LEMESHOW, S.; STURDIVANT, R. **Applied Logistic Regression**. 3. ed. [S.l.]: Wiley, 2013. Citado na página 37.

Introductory Statistics. **Introductory Statistics**. Saylor Academy, 2012. Disponível em: <<https://saylordotorg.github.io/text_introductory-statistics/s14-03-modelling-linear-relationships.html>>. Citado na página 25.

JAMOVI. **The jamovi project, version 2.3**. 2023. Programa estatístico. Disponível em: <https://www.jamovi.org>. Acesso em: jan. 2024. Citado 6 vezes nas páginas 12, 14, 21, 22, 32 e 33.

LEVINE, D. M.; STEPHAN, D. F.; SZABAT, K. A. **Estatística - Teoria e Aplicações usando MS Excel em Português**. 7. ed. Rio de Janeiro: LTC, 2016. Citado na página 14.

MOORE, D. S. **A estatística básica e sua prática**. Rio de Janeiro: Livros Técnicos e Científicos, 2000. Citado 2 vezes nas páginas 21 e 29.

RUGGIERO, M.; LOPES, V. L. **Cálculo Numérico: aspectos teóricos e computacionais**. São Paulo: McGraw-Hill, 1996. Citado na página 11.

SALVADOR, J. A.; ARENALES, S. H. V. **Modelagem matemática de problemas ambientais**. São Carlos: EdUFSCar, 2012. Citado na página 11.

VIEIRA, L. **Risco de crédito e a aplicação da modelagem regressão logística**. Dissertação (Mestrado) — Universidade Estadual Paulista (Unesp), Instituto de Geociências e Ciências Exatas, Rio Claro, 2009. Citado na página 44.