

Parte II

*Modelos de Regressão: Linear Simples e
Múltipla, Ridge, LASSO e Elastic Net*

*Modelos de Neurônio Único para Regressão
e Classificação*

Capítulo 12

Modelos de Regressão

Linear Simples e Múltipla

e

Regressão usando Redes

Neurais

Como todos sabem, o tempo máximo permitido em uma conexão é de nove horas. A determinação deste número é fruto de um estudo detalhado, dos novos mapas cerebrais. Nós chegamos à conclusão, de que não seria saudável permitir que um tempo maior do que esse fosse colocado à disposição, tendo em vista que habitamos um corpo físico cheio de necessidades. É sempre bom lembrar, que não importa o quanto a Rede Neural nos dê a sensação de realidade, nada substitui o contato direto, como este aqui, agora, onde posso olhar pessoalmente para cada um de vocês.

*Miller Aloísio de Britto
Rede Neural: Corra e Não Olhe para Trás*

Antes de iniciar a abordagem de Redes Neurais Artificiais para realizar tarefas de Regressão e Classificação, será apresentada uma revisão sobre modelos de regressão linear simples e múltipla.

A teoria estatística para construção de Modelos de Regressão, tem início em 1885 com Sir Francis Galton, o qual definiu inicialmente o termo “regressão” na análise estatística. Veja os trabalhos: (i) *Family Likeness in Stature*, dos autores Francis Galton e J. D. Hamilton Dickson no *Proceedings of the Royal Society of London Vol. 40 (1886), pp. 42-73 (32 pages)* Publicado pela: Royal Society; e (ii) *Regression towards Mediocrity in Hereditary Stature*, publicado no *Journal of the Anthropological Institute of Great Britain and Ireland, (Vol. 15, pp. 246-263)*.

Vale destacar que modelos estatísticos são construídos baseados em premissas e uma vez que o modelo foi determinado, são necessários diversos testes para verificação da adequabilidade dos parâmetros calculados. Modelos estatísticos são construídos com base em uma teoria matemática muito bem estabelecida e, conforme Salsburg (2009) “só podem ser totalmente compreendidos em termos de fórmulas e símbolos matemáticos”. Os resultados estatísticos dependem de verificações de hipóteses subjacentes nas diversas estimativas que são realizadas, e esse processo de “auditoria” torna muito explícitos todos os passos dos procedimentos adotados, não possibilitando modelos do tipo black-box (caixa-preta), comuns em machine learning. Pelo contrário, teoria estatística tem sido desenvolvida e usada para entender o comportamento de algoritmos black-box de machine learning.

Ou seja, sempre deve-se levar em consideração que um termo, ou definição estatística, é um termo, ou definição matemática, para o qual são consideradas muitas vezes, algumas pressuposições e hipóteses que precisam ser atendidas.

Existem várias técnicas em análise de regressão, iniciando com modelos de regressão linear simples, naturalmente evoluindo para a explicação de uma variável a partir de diversas outras, a métodos mais complexos como regressão Ridge, Lasso e Elastic Net. Dessa forma, este capítulo se concentrará na análise matemática fundamental, fornecendo uma base sólida para entender técnicas mais avançadas.

A análise de regressão tem inúmeras aplicações em diferentes campos que precisam entender relações entre variáveis. Em economia, é usada para descrever tendências e correlações de mercado e medir o impacto de mudanças causadas por adoção de determinadas políticas (Wooldridge, 2016; Tibshirani, 1996). O próprio termo correlação, foi criado por Galton em 1888 trabalhando com dados antropométricos. Na medicina, ajuda a entender a relação entre fatores de risco e resultados de saúde (Harrell, 2015). Na engenharia, a análise de regressão auxilia no controle de qualidade e na melhoria dos processos de fabricação (Montgomery *et al.*, 2012). Iniciamos, com uma das técnicas mais primordiais para o entendimento de uma variável sendo explicada por uma outra variável, o Modelo de Regressão Linear Simples.

3.1 Modelo Estatístico de Regressão Linear

A análise de regressão é um método para estimar as relações entre variáveis (Montgomery *et al.*, 2012). Permite entender como uma variável dependente muda quando qualquer uma das variáveis independentes é alterada (Freedman, 2009). Este processo estima o valor

médio da variável dependente quando as variáveis independentes são mantidas constantes. O objetivo é a estimativa de uma função das variáveis independentes, conhecida como função de regressão (Hastie, Tibshirani e Friedman, 2009). De forma geral, a análise de regressão é uma técnica estatística para investigar e modelar a relação entre variáveis.

Em certas situações, a análise de regressão pode inferir relações causais entre variáveis independentes e dependentes. Por exemplo, Angrist e Pischke (2008) discutem o uso da análise de regressão em econometria para inferir efeitos causais. No entanto, é importante observar que a correlação não implica necessariamente causalidade, portanto, é necessário cautela (Pearl, 2009).

3.1.1 Regressão Linear Simples

De acordo com Neter *et al.* (1983) a Análise de Regressão serve a três propósitos principais: (1) descrição, (2) controle e (3) predição. Dessa forma, vamos considerar a construção de um modelo de predição que consiste em explicar uma variável utilizando uma ou mais outras variáveis. Para iniciar a apresentação dessa modelagem que envolve o desenvolvimento de um modelo linear, utilizaremos valores observados de uma variável X (denominada de variável independente ou explicativa) que serão usados para predizer uma variável Y (chamada de variável dependente, resposta ou explicada). A variável que será explicada, Y , constitui-se dos valores observados que correspondem aos respectivos valores de X . Valores conhecidos de X e Y são descritos de forma genérica como x e y , respectivamente. Na prática, nunca existirá uma relação matemática exata (por exemplo, linear) entre duas variáveis, de tal forma que temos um erro associado a cada valor gerado pelo modelo, ou seja

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad (1)$$

descreve o **Modelo de Regressão Linear Simples (MRLS)**. Note que aplicando o operador esperança a (1), obtemos (2). A especificação do modelo precisa da estimativa de β_0 e β_1 , para isso necessitamos dos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$. Em termos de concepção do modelo, sua descrição geral é dada por

$$E[y|x] = \beta_0 + \beta_1 x, \quad (2)$$

que é denominada de **Função de Regressão Populacional (FRP)**. Além disso, consideramos que

$$\text{Var}[y|x] = \sigma^2. \quad (3)$$

De maneira geral, temos que (2) descreve que a média condicional da variável dependente Y é linear com relação a X , sendo β_0 o intercepto e β_1 a inclinação da reta que será ajustada aos dados observados. A função de regressão em (1) é linear no parâmetro β_1 , se β_1 aparece somente tendo expoente 1 e não estiver multiplicado ou dividido por nenhum outro parâmetro (por exemplo, $\beta_1\beta_2$, $\frac{\beta_1}{\beta_2}$, etc.). Como estamos iniciando a construção de modelos de regressão, os resultados gerados serão obtidos a partir de um modelo que é

tanto linear nos parâmetros, como na variável, tal como mostrado em (2). Em (3) é considerado que a variância de Y dado X é constante. De acordo com Gujarati (2011): “geometricamente, a curva de regressão populacional é simplesmente o lugar geométrico das médias condicionais da variável dependente para os valores fixos da(s) variável(is) explicativa(s). Mais simplesmente, é a curva que conecta as médias das subpopulações de Y correspondentes aos valores dados do regressor X ”. Graficamente, a Figura-1 mostra essa situação onde $\beta_0 > 0$ e $\beta_1 > 0$.

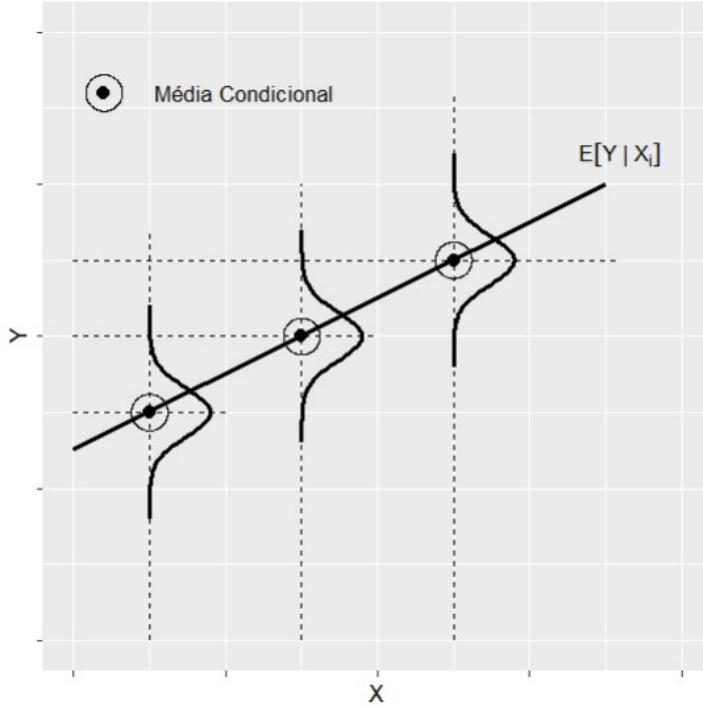


Figura-1 Modelo de Regressão Linear Simples. Fonte: Adaptado de Wooldridge (2000, pg. 26).

Como não temos acesso a todos os valores da população, para estimar os parâmetros β_0 e β_1 , precisamos de uma amostra obtida da população. Considerando que $\{(x_i, y_i); i = 1, 2, \dots, n\}$ denota uma amostra aleatória de tamanho n da população, temos

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \quad (4)$$

Das equações (1), (2) e (4) podemos também escrever que $\varepsilon_i = y_i - E[y_i]$. Ou seja, o que estamos fazendo é ajustando um modelo para explicar y , a partir dos valores coletados de x , tal que $y|x \sim N(\beta_0 + \beta_1 x; \sigma^2)$, atente como é apresentada a distribuição de cada valor de y na Figura-1. Isso implica que o modelo gerado está restrito a amplitude dos valores da variável independente (ou **preditora**) que foram coletados e inseridos no nosso conjunto de dados amostrais. Essa limitação, do modelo predizer valores de y considerando o intervalo disponível de x , em muitas áreas é uma justificativa para fazer-se uma distinção entre os conceitos de **predizer** e **prever**. Aqui, vamos utilizar o MRLS para a tarefa de **predição** e o termo **previsão** será reservado para casos em que um modelo retorna um valor a frente no tempo de uma variável de interesse (Veja o Quadro Conceitual-1). Dessa forma, na equação (4) temos que ε_i , o termo de **erro (resíduo ou desvio)** para a observação i , é uma variável

aleatória não observável que assume valores positivos ou negativos.

Quadro Conceitual - 1

É comum, usarem de forma intercambiável as palavras previsão e predição. Porém, dependendo da área, temos conceitos distintos associados a cada uma delas. Por exemplo, em Silver (*The Signal and the Noise: Why So Many Predictions Fail-but Some Don't*, 2012) encontramos que:

“A ideia do homem como mestre de seu destino estava ganhando força. As palavras predizer e prever são amplamente usadas indistintamente hoje em dia, mas na época de Shakespeare, elas significavam coisas diferentes. Uma predição (prediction) era o que o adivinho dizia; uma previsão (forecast) era algo mais parecido com a ideia de Cassius. O termo previsão (forecast) veio das raízes germânicas do inglês, diferente de predizer (predict), que é do latim. A previsão refletia a nova mundanidade protestante em vez da sobrenaturalidade do Sacro Império Romano. Fazer uma previsão tipicamente implicava planejar sob condições de incerteza. Sugeria ter prudência, sabedoria e diligência, mais como a maneira como agora usamos a palavra previsão.”

Também escrevendo que:

“Uma **predição** é uma declaração definitiva e específica sobre quando e onde ocorrerá um terremoto” e “Enquanto uma **previsão** é uma afirmação probabilística, geralmente em uma escala de tempo mais longa: há 60 por cento de chance de um terremoto”.

Considere a Figura-2 em que temos as representações de uma variável y em termos de uma outra variável x e também sua evolução no tempo t . Para muitos autores, a predição consiste em ajustar um modelo para explicar y , porém somente no intervalo em que x está definida. Na previsão, queremos saber os valores futuros de y , tomando como base uma série histórica de seus valores.

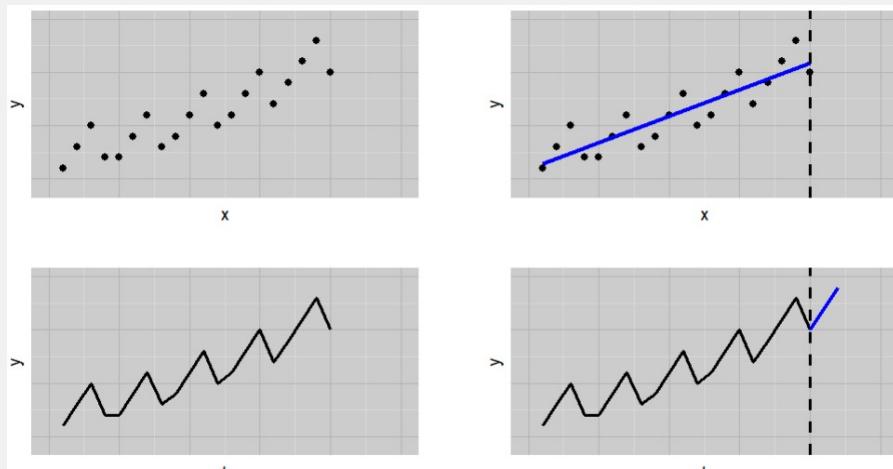


Figura-2 Predição e Previsão.

Esses dois tipos de modelos geram intervalos de predição e previsão. No caso das predições temos intervalos hiperbólicos, e no caso das previsões temos os intervalos no formato de “cones” de previsão. Portanto, a construção de modelos preditivos está associada com a precisão de estimativas considerando os intervalos dos dados observados, enquanto a previsão leva em conta a incerteza sobre os valores futuros da variável (Kuhn e Johnson, 2013; Hyndman e Athanasopoulos, 2018 e Box et al., 2015).

Sobre os erros, consideramos duas suposições importantes: (i) assumimos que $E[\varepsilon_i | x_i] = 0$ e (ii) que estes erros são independentes, ou seja, o valor de erro inerente a uma observação não fornece nenhuma informação sobre o erro associado a outra observação.

O MRLS estará plenamente especificado quando os coeficientes β_0 e β_1 estiverem determinados e as premissas do MRLS forem atendidas. As premissas do modelo de regressão são as seguintes (Gujarati, 2011):

1. O modelo de regressão é linear nos parâmetros.
2. Os valores dos regressores, os X , são fixados em amostras repetidas.
3. Para dados X , o valor médio do termo de erro ε_i é zero.
4. Para dados X , a variância de ε_i é constante e homocedástica.
5. Para dados X , não há autocorrelação entre os termos de erro.
6. Se os X forem estocásticos, o termo de erro e os X (estocásticos) são independentes, ou pelo menos não correlacionados.
7. O número de observações deve ser maior que o número de regressores.
8. Deve haver suficiente variabilidade nos valores assumidos pelos regressores.
9. O modelo de regressão está especificado da forma correta.
10. Não há relação linear exata (ou seja, multicolinearidade) entre os regressores.
11. O termo estocástico (de erro) ε_i se distribui normalmente.

A partir da equação (4) podemos considerar o seguinte resultado

$$\begin{aligned} \sum_{i=1}^n y_i &= \sum_{i=1}^n (\beta_0 + \beta_1 x_i + \varepsilon_i) = n\beta_0 + \beta_1 \sum_{i=1}^n x_i + \sum_{i=1}^n \varepsilon_i, \\ \frac{\sum_{i=1}^n y_i}{n} &= \frac{n}{n} \beta_0 + \beta_1 \frac{\sum_{i=1}^n x_i}{n} + \frac{\sum_{i=1}^n \varepsilon_i}{n}, \\ \bar{y} &= \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}. \end{aligned} \tag{4A}$$

Isso permite escrever as diferenças entre cada valor de y e sua média como sendo

$$\begin{aligned} &\left\{ \begin{array}{l} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \\ \bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon} \end{array} \right. + \\ &\quad \underline{\left(\begin{array}{l} \\ \times (-1) \end{array} \right)} \\ (y_i - \bar{y}) &= \beta_1 (x_i - \bar{x}) + \varepsilon_i - \bar{\varepsilon}. \end{aligned} \tag{4B}$$

Para determinarmos os estimadores de β_0 e β_1 , vamos considerar duas técnicas de estimação, o Método dos Minímos Quadrados Ordinários (MQO) e o Método da Máxima Verossimilhança (MV). O motivo para apresentar a estimação considerando duas técnicas, é de destacar o MV, mostrar que ele gera o mesmo resultado e é uma poderosa ferramenta de estimação de modelos econométricos.

Sendo assim, precisamos determinar a reta de regressão dada por

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{5}$$

A equação (5) é chamada de **Função de Regressão Amostral (FRA)**. Sendo que: (i) a

notação \hat{y} é lida como “y chapéu” e refere-se aos valores de y preditos; (ii) $\hat{\beta}_0$ é o intercepto, ou seja, corresponde ao valor predito de y quando $x=0$; (iii) e $\hat{\beta}_1$ é a inclinação da reta de regressão. Portanto, dada a amostra $\{(x_i, y_i); i=1,2,\dots,n\}$, os valores ajustados para y quando $x=x_i$ serão obtidos como

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \quad (6)$$

3.1.1.1 Estimação de β_0 e β_1 pelo Método dos Minímos Quadrados Ordinários (MQO)

O método dos mínimos quadrados ordinários considera a soma dos quadrados dos desvios (diferenças, erros, resíduos) entre as observações e os valores estimados pelo modelo, na direção vertical no diagrama de dispersão (scatterplot), para obter as estimativas de β_0 e β_1 . Para a aplicação do MQO não é necessária nenhuma suposição com relação a forma da distribuição de probabilidade de ε_i . Como o MQO é utilizado em um processo de inferência estatística, assumimos que os erros ε_i ($i=1,2,\dots,n$) possuem as seguintes características:

$$E[\varepsilon_i] = 0, \quad (7)$$

$$Var[\varepsilon_i] = \sigma^2, \quad (8)$$

$$Cov[\varepsilon_i; \varepsilon_j] = 0, \quad \forall i \neq j (i, j = 1, 2, \dots, n). \quad (9)$$

Ou seja, os erros têm média zero, variância constante e são não correlacionados.

Dessa forma, considerando a FRA temos que determinar os estimadores do intercepto e da inclinação tal que

$$\sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \quad (10)$$

seja a menor possível, onde $\hat{\varepsilon}_i^2$ são os resíduos estimados elevados ao quadrado. Para isso, precisamos das derivadas parciais do somatório dos erros ao quadrado com relação a $\hat{\beta}_0$ e $\hat{\beta}_1$, o que produz

$$\frac{\partial \sum_{i=1}^n \hat{\varepsilon}_i^2}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = -2 \sum_{i=1}^n \hat{\varepsilon}_i \quad (11)$$

e

$$\frac{\partial \sum_{i=1}^n \hat{\varepsilon}_i^2}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = -2 \sum_{i=1}^n \hat{\varepsilon}_i x_i. \quad (12)$$

Fazendo $\frac{\partial \sum_{i=1}^n \hat{\varepsilon}_i^2}{\partial \hat{\beta}_0} = 0$, temos que

$$\begin{aligned}
\frac{\partial \sum_{i=1}^n \hat{\varepsilon}_i^2}{\partial \hat{\beta}_0} &= -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = -2 \sum_{i=1}^n \hat{\varepsilon}_i = 0, \\
-2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0, \\
\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= \sum_{i=1}^n y_i - \sum_{i=1}^n \hat{\beta}_0 - \sum_{i=1}^n \hat{\beta}_1 x_i = \sum_{i=1}^n y_i - n \hat{\beta}_0 - \sum_{i=1}^n \hat{\beta}_1 x_i = 0, \\
\hat{\beta}_0 &= \frac{\sum_{i=1}^n y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^n x_i}{n},
\end{aligned}$$

o que resulta no estimador de $\hat{\beta}_0$ sendo dado por

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \quad (13)$$

No caso em que $\frac{\partial \sum_{i=1}^n \hat{\varepsilon}_i^2}{\partial \hat{\beta}_1} = 0$, temos

$$\begin{aligned}
\frac{\partial \sum_{i=1}^n \hat{\varepsilon}_i^2}{\partial \hat{\beta}_1} &= -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = -2 \sum_{i=1}^n \hat{\varepsilon}_i x_i = 0, \\
-2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i &= 0, \\
\sum_{i=1}^n y_i x_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 &= 0, \\
\sum_{i=1}^n y_i x_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 &= 0, \\
\sum_{i=1}^n y_i x_i - \left(\frac{\sum_{i=1}^n y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^n x_i}{n} \right) \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 &= 0, \\
\hat{\beta}_1 \left[\frac{\left(\sum_{i=1}^n x_i \right)^2}{n} - \sum_{i=1}^n x_i^2 \right] &= \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n} - \sum_{i=1}^n y_i x_i,
\end{aligned}$$

$$\hat{\beta}_1 = \frac{\frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n} - n \frac{\sum_{i=1}^n y_i x_i}{n}}{\frac{\left(\sum_{i=1}^n x_i\right)^2}{n} - n \frac{\sum_{i=1}^n x_i^2}{n}},$$

o que resulta no estimador de $\hat{\beta}_1$ sendo dado por

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}. \quad (14)$$

Fazendo alguma álgebra, podemos reescrever a equação (14). Não apenas para obter uma forma diferente de expressar $\hat{\beta}_1$, mas pela característica que essa nova equação possui em termos de processamento computacional.

Vamos iniciar considerando o numerador de (14), de modo que podemos reescrevê-lo como

$$\begin{aligned} n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i &= n \sum_{i=1}^n y_i x_i - n \bar{y} n \bar{x} = n \sum_{i=1}^n y_i x_i - n \sum_{i=1}^n \bar{y} \bar{x}, \\ n \sum_{i=1}^n (y_i x_i - \bar{y} \bar{x}) &= n \sum_{i=1}^n (y_i x_i - \bar{y} \bar{x} - \bar{y} \bar{x} + \bar{y} \bar{x} + \bar{y} \bar{x} - \bar{y} \bar{x}), \\ n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i &= n \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}). \end{aligned} \quad (15B)$$

Agora, considerando o denominador, temos

$$\begin{aligned} n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2 &= n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right) \left(\sum_{i=1}^n x_i\right) = n \sum_{i=1}^n x_i^2 - n \bar{x} \sum_{i=1}^n x_i, \\ n \sum_{i=1}^n (x_i^2 - \bar{x} x_i) &= n \sum_{i=1}^n (x_i^2 - \bar{x} x_i - \bar{x} x_i + \bar{x}^2 + \bar{x} x_i - \bar{x}^2), \\ n \sum_{i=1}^n (x_i^2 - 2 \bar{x} x_i + \bar{x}^2 + \bar{x} x_i - \bar{x}^2) &= n \sum_{i=1}^n [(x_i^2 - 2 \bar{x} x_i + \bar{x}^2) + (\bar{x} x_i - \bar{x}^2)], \\ n \sum_{i=1}^n [(x_i - \bar{x})^2 + (\bar{x} x_i - \bar{x}^2)] &= n \sum_{i=1}^n (x_i - \bar{x})^2 + n \bar{x} \sum_{i=1}^n x_i - n^2 \bar{x}^2, \\ n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2 &= n \sum_{i=1}^n (x_i - \bar{x})^2. \\ n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2 &= n \sum_{i=1}^n (x_i - \bar{x})^2. \end{aligned} \quad (16B)$$

Com as equações (15) e (16) podemos reescrever o estimador de $\hat{\beta}_1$ dado em (14) como

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (17B)$$

Da obtenção dos estimadores para os parâmetros β_0 e β_1 , podemos extrair duas igualdades importantes,

$$\sum_{i=1}^n y_i - n\hat{\beta}_0 - \sum_{i=1}^n \hat{\beta}_1 x_i = 0 \rightarrow \sum_{i=1}^n y_i = n\hat{\beta}_0 + \sum_{i=1}^n \hat{\beta}_1 x_i, \quad (NN1)$$

e

$$\sum_{i=1}^n y_i x_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0 \rightarrow \sum_{i=1}^n y_i x_i = \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2. \quad (NN2)$$

Essas equações (NN1) e (NN2) são denominadas de **equações normais** da regressão.

Da equação (NN1) podemos inferir que

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^n \varepsilon_i = 0. \quad (NN3)$$

Da equação (NN2) podemos inferir que

$$\sum_{i=1}^n x_i \varepsilon_i = \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^n y_i x_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0. \quad (NN4)$$

Holder (1985) no livro *Multiple Regression in Hydrology*, salienta que a primeira expressão para $\hat{\beta}_1$ na equação (14) é a geralmente recomendada para realização do cálculo desse estimador porque preserva a precisão. No entanto, este ponto só é válido desde que a precisão total possa ser mantida durante todo o cálculo. Se houver um grande número de observações de dados e y e x forem valores relativamente grandes, isso pode levar a

$n \sum_{i=1}^n y_i x_i$ e $\sum_{i=1}^n y_i \sum_{i=1}^n x_i$ ambos serem grandes e semelhantes; portanto, sua diferença pode ser seriamente afetada pelos erros de arredondamento gerados ao calcular usando qualquer uma dessas expressões. Em tais circunstâncias, a segunda opção de expressão para calcular $\hat{\beta}_1$, apresentada na equação (17B), é mais satisfatória. Problemas de arredondamento geralmente surgem quando um computador digital é usado para cálculo e, nessas circunstâncias, há a recomendação para uso de expressão alternativa, mostrada em (18B)..

Erros de arredondamento têm impacto direto sobre projetos de sistemas computadorizados. No trabalho *What Every Computer Scientist Should Know About Floating-Point Arithmetic*, David Goldberg (1991) descreve que: “*Comprimir infinitamente muitos números reais em um número finito de bits requer uma representação aproximada. Embora haja infinitamente muitos inteiros, na maioria dos programas o resultado de cálculos inteiros pode ser armazenado em 32 bits. Em contraste, dado qualquer número fixo de bits, a maioria dos cálculos com números reais produzirá quantidades que não podem ser*

exatamente representadas usando tantos bits. Portanto, o resultado de um cálculo de ponto flutuante deve frequentemente ser arredondado para se ajustar de volta à sua representação finita. Esse erro de arredondamento é a característica do cálculo de ponto flutuante”.

Como a soma de quadrados e produtos cruzados aparecem frequentemente nos cálculos de modelos de regressão, os termos definidos a seguir são utilizados para representar alguns dos estimadores e estatísticas construídos. Sendo assim, considerando

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n},$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n},$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}.$$

Podemos reescrever a equação (14) como

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}. \quad (18B)$$

Em (18B) temos que o coeficiente da regressão $\hat{\beta}_1$, é escrito como a razão entre a covariância amostral $Cov[x, y]$ e $Var[x]$, a variância amostral. Pensando em termos das unidades e escalas de x e y , é como se a covariância entre x e y fosse corrigida para a escala de x . O uso de (18B) simplifica os cálculos e fornece um caminho direto para encontrar o coeficiente de regressão. Quando expressamos $\hat{\beta}_1$ na forma S_{xy}/S_{xx} , o coeficiente $\hat{\beta}_1$ indica a mudança esperada em y para cada unidade de mudança em x .

3.1.1.2 Estimação de β_1 e β_2 pelo Método de Máxima Verossimilhança (MV)

Considere que, no modelo de duas variáveis $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ os y_i se distribuem de maneira independente, com média $\beta_0 + \beta_1 x_i$ e variância σ^2 . Em consequência disso, a função de densidade de probabilidade conjunta de y_1, y_2, \dots, y_n dadas a média e a variância anteriores, pode ser escrita como

$$f(y_1, y_2, \dots, y_n | \beta_0 + \beta_1 x_i, \sigma^2). \quad (15)$$

Mas tendo em vista a independência dos y 's, essa função de densidade de probabilidade conjunta pode ser expressa como um produto de n funções de densidades individuais, conforme mostrado a seguir

$$f(y_1, y_2, \dots, y_n | \beta_0 + \beta_1 x_i, \sigma^2) = \quad (16)$$

$$f(y_1 | \beta_0 + \beta_1 x_i, \sigma^2) \times f(y_2 | \beta_0 + \beta_1 x_i, \sigma^2) \times \dots \times f(y_n | \beta_0 + \beta_1 x_i, \sigma^2),$$

onde

$$f(y_i | x_i; \beta_0, \beta_1, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{1}{2} \frac{(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sigma^2} \right]. \quad (17)$$

que é a função densidade de uma distribuição Normal com média e variância dadas. Considerando essa informação, podemos escrever a expressão (13) como

$$f(y_1, y_2, \dots, y_n | \beta_0 + \beta_1 x_i, \sigma^2) = \frac{1}{\sigma^n (\sqrt{2\pi})^n} \exp \left[-\frac{1}{2} \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sigma^2} \right]. \quad (18)$$

Se y_1, y_2, \dots, y_n forem conhecidos ou dados mas β_0 , β_1 e σ^2 não forem, a função acima é chamada de função de verossimilhança, denotada por $L(\beta_0, \beta_1, \sigma^2)$, e expressa como

$$L(\beta_0, \beta_1, \sigma^2) = \frac{1}{\sigma^n (\sqrt{2\pi})^n} \exp \left[-\frac{1}{2} \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sigma^2} \right]. \quad (19)$$

O método da máxima verossimilhança, consiste na estimativa dos parâmetros desconhecidos de maneira que a probabilidade de observar o dado y seja a maior (ou máxima possível). Portanto, precisamos encontrar o máximo da função $L(\beta_0, \beta_1, \sigma^2)$.

Para derivar, é mais fácil expressar $L(\beta_0, \beta_1, \sigma^2)$ na forma logarítmica, tal como mostrado a seguir.

$$\begin{aligned} \ln L(\beta_0, \beta_1, \sigma^2) &= \ln \frac{1}{\sigma^n (\sqrt{2\pi})^n} \exp \left[-\frac{1}{2} \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sigma^2} \right], \\ \ln L(\beta_0, \beta_1, \sigma^2) &= -n \ln \sigma - \frac{n}{2} \ln 2\pi - \frac{1}{2} \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sigma^2}. \end{aligned}$$

Fazendo $l(\beta_0, \beta_1, \sigma^2) = \ln L(\beta_0, \beta_1, \sigma^2)$ e derivando parcialmente em relação a β_0 , β_1 e σ^2 temos

$$\frac{\partial l(\beta_0, \beta_1, \sigma^2)}{\partial \hat{\beta}_0} = -\frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1), \quad (20)$$

$$\frac{\partial l(\beta_0, \beta_1, \sigma^2)}{\partial \hat{\beta}_1} = -\frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i), \quad (21)$$

$$\frac{\partial l(\beta_0, \beta_1, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \quad (22)$$

Igualando essas primeiras derivadas a zero (condição de primeira ordem para indentificar os candidatos a pontos de mínimo) obtemos

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \quad (23)$$

$$\sum_{i=1}^n (y_i x_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) = 0, \quad (24)$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2. \quad (25)$$

As equações (23) e (24) são análogas às obtidas na estimação usando MQO e após simplificação chegamos nos mesmos resultados que foram obtidos no MQO. Conforme mencionado anteiormente uma das premissas do modelo de regressão linear é que para dados X , a variância de ε_i é constante e homocedástica. É importante que sejam utilizadas técnicas de coleta de dados adequadas e que haja uma melhora a medida que novas amostras são obtidas e as bases de dados são construídas, isso ajuda a diminuir a variabilidade dos dados coletados. Além disso, precisam ser consideradas: (i) estratégias para lidar com dados discrepantes (outliers); (ii) correta identificação das distribuições dos regressores no modelo e (iii) consequências das transformações realizadas nos dados.

3.1.1.3 Valor Esperado, Variância e Covariância dos Estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$

Uma vez obtidos os estimadores de β_0 e β_1 , podemos considerar a equação (4B) e reescrever as expressões para $\hat{\beta}_0$ e $\hat{\beta}_1$ em função do erro de estimação ε_i . Isso é particularmente útil porque permite calcular o valor esperado e a variância de cada um desses estimadores de maneira mais direta e permitindo desenvolver os resultados sobre inferências estatísticas associadas ao intercepto, inclinação da reta de regressão e a variância do termo de erro. Facilitando também a determinação da covariância entre $\hat{\beta}_0$ e $\hat{\beta}_1$. Começamos mostrando que $\hat{\beta}_0$ e $\hat{\beta}_1$ são estimadores não viesados, ou seja, o valor esperado é igual ao parâmetro populacional. No estimador de β_1 , dado por (17B), podemos substituir a diferença entre y_i e \bar{y} conforme mostrada em (4B) o que resulta em

$$\hat{\beta}_1 = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x}) [\beta_1 (x_i - \bar{x}) + \varepsilon_i - \bar{\varepsilon}],$$

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} - \frac{\sum_{i=1}^n (x_i - \bar{x}) \bar{\varepsilon}}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

porém, a segunda razão de somatórios, dessa última igualdade apresentada, é igual a zero. Conforme facilmente podemos verificar

$$\frac{\sum_{i=1}^n (x_i - \bar{x}) \bar{\varepsilon}}{\sum_{i=1}^n (x_i - \bar{x})^2} = \bar{\varepsilon} \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \bar{\varepsilon} \frac{\sum_{i=1}^n x_i - n\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} = \bar{\varepsilon} \frac{\overbrace{\sum_{i=1}^n x_i - n\bar{x}}^{=n\bar{x}}}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0.$$

Sendo assim,

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (\text{B1N})$$

Valor Esperado de $\hat{\beta}_1$

Podemos determinar o valor esperado de $\hat{\beta}_1$ usando (B1N), tal que

$$E[\hat{\beta}_1] = E\left[\beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) E[\varepsilon_i]}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1.$$

O que permite afirmar que $\hat{\beta}_1$, dado por (B1N), é um estimador não viesado de β_1 .

Variância de $\hat{\beta}_1$

Podemos determinar a variância de $\hat{\beta}_1$ da seguinte forma

$$\begin{aligned} Var[\hat{\beta}_1] &= Var\left[\beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \stackrel{\text{independência}}{=} Var\left[\frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] = \frac{\sum_{i=1}^n Var[(x_i - \bar{x}) \varepsilon_i]}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2}, \\ &\frac{\sum_{i=1}^n (x_i - \bar{x})^2 Var[\varepsilon_i]}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2}, \\ Var[\hat{\beta}_1] &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}. \end{aligned} \quad (\text{VB1})$$

Valor Esperado de $\hat{\beta}_0$

Para $\hat{\beta}_0$, dado por (13), substituindo \bar{y} conforme mostrado em (4A), resulta em

$$\hat{\beta}_0 = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon} - \hat{\beta}_1 \bar{x} = \beta_0 + \bar{\varepsilon} - (\hat{\beta}_1 - \beta_1) \bar{x}.$$

Usando (B1N), temos

$$\hat{\beta}_0 = \beta_0 + \bar{\varepsilon} - \frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \bar{x} = \beta_0 + \bar{\varepsilon} - \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i.$$

Podemos considerar reescrever $\bar{\varepsilon}$ da seguinte forma

$$\bar{\varepsilon} = \frac{\sum_{i=1}^n \varepsilon_i}{n} \rightarrow \bar{\varepsilon} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \frac{\sum_{i=1}^n \varepsilon_i}{n} \rightarrow \bar{\varepsilon} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \frac{\sum_{i=1}^n \varepsilon_i}{n}.$$

Substituindo esse $\bar{\varepsilon}$ na expressão imediatamente anterior de $\hat{\beta}_0$, resulta em

$$\begin{aligned} \hat{\beta}_0 &= \beta_0 + \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \frac{\sum_{i=1}^n \varepsilon_i}{n} - \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i, \\ \hat{\beta}_0 &= \beta_0 + \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n \varepsilon_i}{n \bar{x}} - \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i \right], \\ \hat{\beta}_0 &= \beta_0 + \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \left\{ \frac{\sum_{i=1}^n \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] \varepsilon_i}{n \bar{x}} - \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i \right\}, \end{aligned}$$

Sendo assim,

$$\hat{\beta}_0 = \beta_0 + \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n \bar{x}} - (x_i - \bar{x}) \right] \varepsilon_i. \quad (\text{B2N})$$

Podemos determinar o valor esperado de $\hat{\beta}_0$ usando (B2N), tal que

$$E[\hat{\beta}_0] = E \left\{ \beta_0 + \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n \bar{x}} - (x_i - \bar{x}) \right] \varepsilon_i \right\} = \beta_0.$$

O que permite afirmar que $\hat{\beta}_0$, dado por (B2N), é um estimador não viesado de β_0 .

Uma outra forma, mais simples, para determinar o valor esperado (esperança) de $\hat{\beta}_0$, pode ser usando a expressão (13). De tal forma que tomando a esperança dos dois lados da igualdade, resulta em

$$\begin{aligned} E[\hat{\beta}_0] &= E[\bar{y} - \hat{\beta}_1 \bar{x}], \\ &= E\left[\frac{\sum_{i=1}^n y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^n x_i}{n}\right] = E\left[\frac{\sum_{i=1}^n (\beta_0 + \beta_1 x_i + \varepsilon_i)}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^n x_i}{n}\right], \\ &= \frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 x_i) - E[\hat{\beta}_1] \frac{\sum_{i=1}^n x_i}{n} = \frac{1}{n} \left(n\beta_0 + \beta_1 \sum_{i=1}^n x_i \right) - \beta_1 \frac{\sum_{i=1}^n x_i}{n} = \beta_0. \end{aligned}$$

Optou-se por apresentar a primeira forma de determinação da esperança de $\hat{\beta}_0$ dada por (B2N) porque é uma expressão que depende de ε_i . E isso é interessante no desenvolvimento de novos resultados.

Variância de $\hat{\beta}_0$

Considerando a independência entre $\hat{\beta}_0$ e os termos de erro, além da independência também dos termos de erro entre si e com a variável explicativa, podemos determinar a variância de $\hat{\beta}_0$ da seguinte forma

$$\begin{aligned} Var[\hat{\beta}_0] &= Var\left\{\beta_0 + \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n\bar{x}} - (x_i - \bar{x}) \right] \varepsilon_i\right\}, \\ &= \left[\frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^2 Var\left\{ \sum_{i=1}^n \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n\bar{x}} - (x_i - \bar{x}) \right] \varepsilon_i \right\} \\ &= \left[\frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^2 \sum_{i=1}^n \left\{ Var\left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n\bar{x}} - (x_i - \bar{x}) \right] \varepsilon_i \right\} \\ &= \left[\frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^2 \sum_{i=1}^n \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n\bar{x}} - (x_i - \bar{x}) \right]^2 Var[\varepsilon_i] \end{aligned}$$

$$\begin{aligned}
&= \left[\frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^2 \sum_{i=1}^n \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n\bar{x}} - (x_i - \bar{x}) \right]^2 \sigma^2 \\
&= \frac{\bar{x}^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \sum_{i=1}^n \left\{ \frac{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2}{n^2 \bar{x}^2} - 2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n\bar{x}} (x_i - \bar{x}) + (x_i - \bar{x})^2 \right\} \sigma^2 \\
&= \frac{\bar{x}^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \left\{ \frac{1}{n^2 \bar{x}^2} n \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2 - 2 \sum_{i=1}^n \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n\bar{x}} (x_i - \bar{x}) \right] + \sum_{i=1}^n (x_i - \bar{x})^2 \right\} \sigma^2 \\
&= \frac{\bar{x}^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \left\{ \frac{1}{n^2 \bar{x}^2} n \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2 - 2 \sum_{i=1}^n \underbrace{\left[(x_i - \bar{x}) \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n\bar{x}} \right]}_{=0} + \sum_{i=1}^n (x_i - \bar{x})^2 \right\} \sigma^2,
\end{aligned}$$

que resulta em

$$Var[\hat{\beta}_0] = \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \sigma^2. \quad (\text{VB0})$$

Também poderíamos usar a expressão (13) para determinar a variância do estimador do intercepto ($\hat{\beta}_0$), nesse caso teríamos

$$\begin{aligned}
Var[\hat{\beta}_0] &= Var[\bar{y} - \hat{\beta}_1 \bar{x}], \\
&= Var\left[\frac{\sum_{i=1}^n y_i}{n}\right] + Var[\hat{\beta}_1 \bar{x}] + 2Cov[\bar{y}, -\hat{\beta}_1 \bar{x}] \\
&= \frac{1}{n^2} Var[y_1 + y_2 + \dots + y_n] + \bar{x}^2 Var[\hat{\beta}_1] - \bar{x} 2Cov[\bar{y}, \hat{\beta}_1],
\end{aligned}$$

considerando o pressuposto de que os valores de y são independentes, ficamos com

$$= \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \bar{x}^2 - 2\bar{x} Cov[\bar{y}, \hat{\beta}_1].$$

Nesse ponto, precisamos avaliar a covariância entre \bar{y} e $\hat{\beta}_1$ antes de prosseguir com o desenvolvimento da expressão anterior. Sendo assim, temos

$$\begin{aligned}
 Cov[\bar{y}, \hat{\beta}_1] &= Cov\left[\frac{\sum_{i=1}^n y_i}{n}, \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \\
 &\stackrel{Cov[aX, bY] = abCov[X, Y]}{=} \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} Cov\left[\sum_{i=1}^n y_i, \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})\right] \\
 &\stackrel{\sum_{i=1}^n \bar{y}(x_i - \bar{x}) = 0}{=} \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} Cov\left[\sum_{i=1}^n y_i, \sum_{i=1}^n y_i(x_i - \bar{x})\right] \\
 &= \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} Cov\left[\sum_{i=1}^n y_i, \sum_{i=1}^n y_i(x_i - \bar{x})\right] \\
 &= \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n \sum_{j=1}^n (x_i - \bar{x}) Cov[y_i, y_j].
 \end{aligned}$$

A covariância entre y_i e y_j pode ser escrita como

$$Cov[y_i, y_j] = E[y_i - E[y_i], y_j - E[y_j]],$$

e como sabemos que $\varepsilon_i = y_i - E[y_i]$, temos que $Cov[y_i, y_j]$ assume os seguintes valores

$$Cov[y_i, y_j] = E[\varepsilon_i, \varepsilon_j] = \begin{cases} \sigma^2, & \text{se } i = j \\ 0, & \text{se } i \neq j. \end{cases}$$

Usando esse resultado em $Cov[\bar{y}, \hat{\beta}_1]$ produz

$$Cov[\bar{y}, \hat{\beta}_1] = \begin{cases} \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2 \sum_{i=1}^n \sum_{j=1}^n (x_i - \bar{x}) = 0 \\ \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n \sum_{j=1}^n (x_i - \bar{x}) \cdot 0 = 0. \end{cases}$$

Finalmente, concluímos que

$$Var[\hat{\beta}_0] = \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \sigma^2. \quad (\text{VB02})$$

Verificamos que (VB02) corresponde ao mesmo resultado apresentado em (VB0).

Covariância entre $\hat{\beta}_0$ e $\hat{\beta}_1$

Uma forma simplificada para determinar a covariância entre $\hat{\beta}_0$ e $\hat{\beta}_1$ é através do emprego diretamente da definição de covariância e de suas propriedades, o que permite escrever

$$\begin{aligned} Cov[\hat{\beta}_0, \hat{\beta}_1] &= Cov[\bar{y} - \bar{x}\hat{\beta}_1, \hat{\beta}_1] = Cov[\bar{y}, \hat{\beta}_1] - Cov[\bar{x}\hat{\beta}_1, \hat{\beta}_1] \\ &= Cov[\bar{y}, \hat{\beta}_1] - \bar{x}Cov[\hat{\beta}_1, \hat{\beta}_1] \\ &= -\bar{x}Var[\hat{\beta}_1], \\ Cov[\hat{\beta}_0, \hat{\beta}_1] &= -\frac{\bar{x}\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}. \end{aligned} \quad (\text{COVB1})$$

A outra maneira que podemos utilizar para determinar $Cov[\hat{\beta}_0, \hat{\beta}_1]$ é usando os resultados (B1N) e (B2N).

$$\begin{aligned} Cov[\hat{\beta}_0, \hat{\beta}_1] &= E\left\{(\hat{\beta}_0 - E[\hat{\beta}_0])(\hat{\beta}_1 - E[\hat{\beta}_1])\right\} = E\left[(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1)\right], \\ &= E\left\{\frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n\bar{x}} - (x_i - \bar{x}) \right] \varepsilon_i \cdot \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right\}, \\ &= \frac{\bar{x}}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} E\left\{\sum_{i=1}^n \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n\bar{x}} - (x_i - \bar{x}) \right] \varepsilon_i \cdot \sum_{i=1}^n (x_i - \bar{x})\varepsilon_i \right\}, \end{aligned}$$

as parcelas $(x_i - \bar{x})\varepsilon_i$ do último somatório podem ser separadas em duas partes com índices que coincidem com o ε_i do somatório anterior e naqueles em que diferem ($i \neq j$). Sendo assim, temos

$$\begin{aligned} & \frac{\bar{x}}{\left[\sum_{i=1}^n(x_i-\bar{x})^2\right]^2} E\left\{\sum_{i=1}^n\left[\frac{\sum_{i=1}^n(x_i-\bar{x})^2}{n\bar{x}}-(x_i-\bar{x})\right](x_i-\bar{x})\varepsilon_i^2+\right. \\ & \quad \left.+\sum_{i \neq j}\left[\frac{\sum_{i=1}^n(x_i-\bar{x})^2}{n\bar{x}}-(x_i-\bar{x})\right](x_j-\bar{x})\varepsilon_i\varepsilon_j\right], \end{aligned}$$

considerando o pressuposto de independência entre os termos de erro, ficamos com

$$\begin{aligned} & \frac{\bar{x}}{\left[\sum_{i=1}^n(x_i-\bar{x})^2\right]^2} E\left\{\sum_{i=1}^n\left[\frac{\sum_{i=1}^n(x_i-\bar{x})^2}{n\bar{x}}-(x_i-\bar{x})\right](x_i-\bar{x})\varepsilon_i^2\right\}, \\ & \frac{\bar{x}}{\left[\sum_{i=1}^n(x_i-\bar{x})^2\right]^2} E\left\{\sum_{i=1}^n\left[\frac{\sum_{i=1}^n(x_i-\bar{x})^2}{n\bar{x}}(x_i-\bar{x})-(x_i-\bar{x})^2\right]\varepsilon_i^2\right\}, \\ & \frac{\bar{x}}{\left[\sum_{i=1}^n(x_i-\bar{x})^2\right]^2} E\left[-\sum_{i=1}^n(x_i-\bar{x})^2\varepsilon_i^2\right]=-\frac{\bar{x}}{\left[\sum_{i=1}^n(x_i-\bar{x})^2\right]^2} \sum_{i=1}^n(x_i-\bar{x})^2Var[\varepsilon_i], \end{aligned}$$

portanto,

$$Cov[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\bar{x}}{\left[\sum_{i=1}^n(x_i-\bar{x})^2\right]^2} \sigma^2. \quad (\text{COVB2})$$

Considerando os resultados obtidos em (VB1), (VB0) e (COVB1) temos

$$Var[\hat{\beta}_0] = \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n(x_i-\bar{x})^2} \right] \sigma^2, \quad Var[\hat{\beta}_1] = \frac{\sigma^2}{\sum_{i=1}^n(x_i-\bar{x})^2} \quad \text{e} \quad Cov[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\bar{x}}{\left[\sum_{i=1}^n(x_i-\bar{x})^2\right]^2} \sigma^2.$$

Dessas equações verificamos que as magnitudes das variabilidades de $\hat{\beta}_0$, $\hat{\beta}_1$ e a covariância entre eles diminui a medida que a dispersão dos valores x , em torno de sua média, aumenta. No caso de $\hat{\beta}_0$, aumentando o tamanho da amostra, n , contribui também para a diminuição de sua variância. Tanto para as variâncias de $\hat{\beta}_0$ e $\hat{\beta}_1$, como para a covariância, verifica-se que suas magnitudes (em valores absolutos) aumentam quando a variância do termo de erro, σ_ε^2 , aumenta. Também é necessário ficar claro que $Var[\hat{\beta}_0]$,

$Var[\hat{\beta}_1]$ e $Cov[\hat{\beta}_0, \hat{\beta}_1]$ poderiam ser escritas como $\sigma_{\hat{\beta}_0}^2$, $\sigma_{\hat{\beta}_1}^2$ e $\sigma_{\hat{\beta}_0, \hat{\beta}_1}$, respectivamente.

Ou seja, principalmente as variâncias de $\hat{\beta}_0$ e $\hat{\beta}_1$, são importantes para realizar as inferências referentes à inclinação e ao intercepto do MRLS. Como as variâncias dos estimadores de mínimos quadrados de β_0 e β_1 dependem da variância do termo de erro (ε) no modelo de regressão, e essa variância do erro é desconhecida, torna-se necessário estimá-la. A seguir, discutiremos o procedimento para estimar a variância do termo de erro no contexto do modelo de regressão linear simples.

3.1.1.4 Inferência na Análise de Regressão Linear Simples

A seguir, são apresentadas inferências sobre os parâmetros β_0 e β_1 do MRLS, considerando tanto a estimativa de intervalo desses parâmetros, quanto os testes de significância associados. Após a estimação do modelo, adotamos a abordagem da Análise de Variância (ANOVA) para testar a significância da regressão. Em seguida, discutimos a estimativa de intervalo da média da variável dependente Y , ou seja $E[Y]$, para um dado X , e os intervalos de predição para uma nova observação Y , dado X . Os resíduos são analisados para verificar se os pressupostos de autocorrelação serial, normalidade, linearidade e homocedasticidade são atendidos.

Uma vez que os dados referentes à variável dependente e à variável independente são coletados e organizados, uma das primeiras providências é construir um gráfico de dispersão (**scatter plot**). Esse tipo de gráfico permite fazer uma primeira avaliação sobre o tipo de relacionamento entre Y e X . Na Figura-2 são apresentados alguns tipos de relacionamentos lineares, visualizados a partir dos gráficos de dispersão.

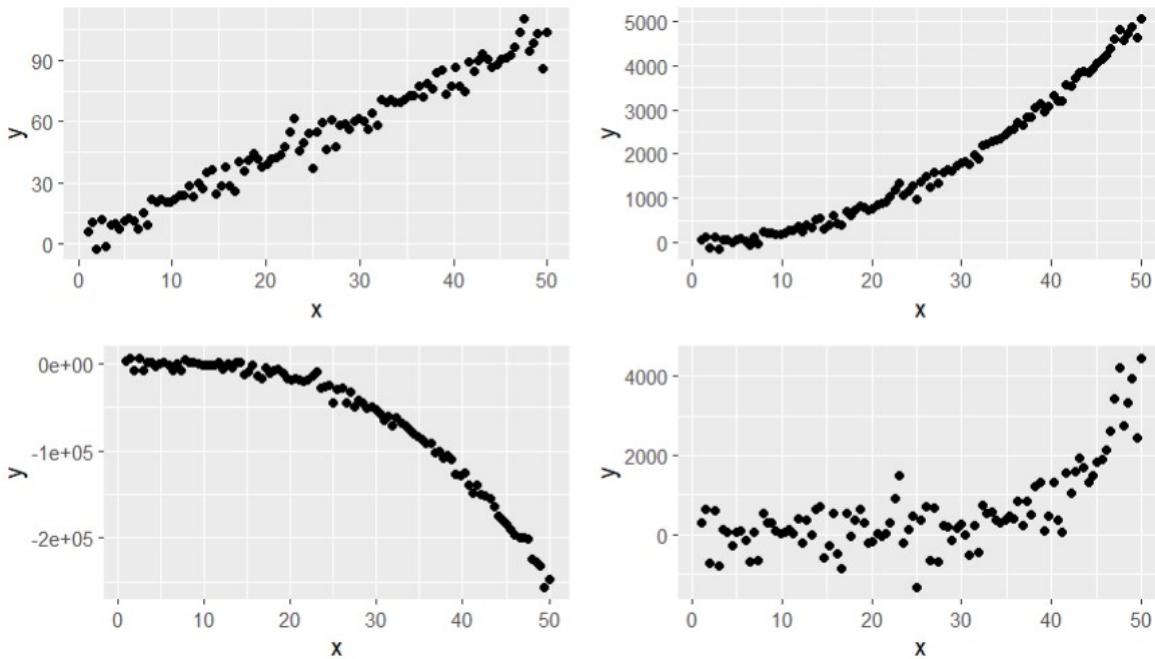


Figura-2 Gráficos de dispersão, servem de apoio para visualizar possíveis relacionamentos funcionais.

Cálculo do Erro Quadrado Médio (Mean Squared Error - MSE)

Quando construímos um modelo de regressão linear, estimamos os parâmetros da regressão de tal forma que, uma inclinação particular da reta de regressão consiga explicar os dados

que coletamos. Uma medida para avaliar o ajuste do modelo aos dados é obtida a partir da soma dos quadrados das diferenças entre y_i e \hat{y}_i considerando a Soma dos Quadrados dos Erros (Sum od Squared Errors - SSE)

$$SSE = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \quad (10)$$

Dessa forma, a medida de ajuste utilizada depende dos dois parâmetros estimados, $\hat{\beta}_0$ e $\hat{\beta}_1$, e é denominada de Mean Squared Error (MSE), sendo definida como

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}. \quad (26)$$

Utilizaremos o MSE como o estimador não viesado de σ^2 . Portanto, temos que y_i é a variável de resposta observada e $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, o valor ajustado com base no modelo estimado. Enquanto y_i representa um valor da base de dados e \hat{y}_i um valor calculado, ambos estão disponíveis para a realização das análises. A variabilidade σ_i , associada ao erro verdadeiro cometido no ajuste do modelo, não é diretamente observável e, portanto, precisa ser estimada. A diferença $y_i - \hat{y}_i$ é a versão empírica do erro ε_i . Essa diferença, conhecida como resíduo da regressão, desempenha um papel crucial no diagnóstico do modelo de regressão. Também podemos escrever (26) das seguintes formas

$$MSE = \frac{SSE}{n-2} = \frac{SS_{yy} - \frac{(SS_{xy})^2}{SS_{xx}}}{n-2} = \frac{n-1}{n-2} \left\{ s_y^2 - \frac{(\text{Cov}[x, y])^2}{s_x^2} \right\}. \quad (26)$$

As diferenças mais importantes que envolvem os dados observados e o modelo estimado são apresentadas na Figura-3.

Análise de Variância Aplicada à Regressão

As representações das variáveis dependente e independente utilizadas na Figura-3, foram feitas usando letras maiúsculas (X e Y) para descrever de forma geral essas variáveis. Vamos voltar a considerar os casos particulares de respostas individuais observadas, y_i , que apresentam média \bar{y} . Utilizamos a Análise de Variância para testar a significância da regressão, de tal forma que particionamos os desvios entre y_i e a média \bar{y} , tal como mostrado na Figura-3. A diferença $y_i - \bar{y}$, denominada de variabilidade total, pode ser separada em uma componente devido ao desvio entre o valor observado y_i e o respectivo valor ajustado $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ e outra parte que considera o desvio entre o valor ajustado \hat{y}_i e a média \bar{y} . Ou seja, particionamos o desvio total $(y_i - \bar{y})$ em uma soma do desvio em torno da reta de regressão $(y_i - \hat{y}_i)$ e o desvio do valor ajustado em torno da média $(\hat{y}_i - \bar{y})$ da variável dependente.

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}).$$

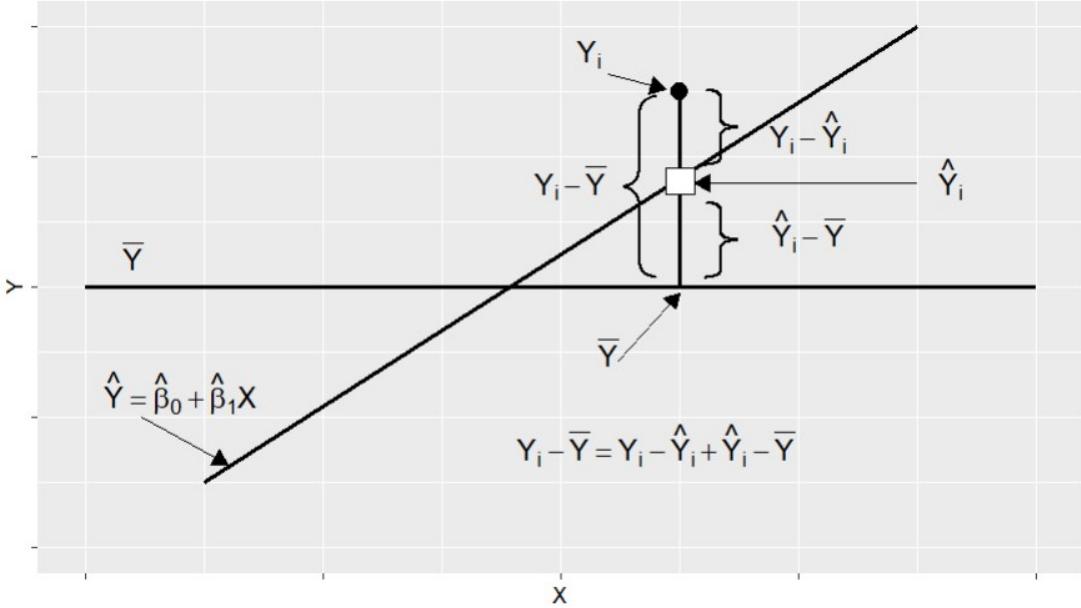


Figura-3 Partição dos desvios $Y_i - \bar{Y}$.

Dessa equação, notamos que a igualdade é mantida quando tomamos o quadrado de cada lado da equação e somamos considerando todos os n valores observados e ajustados pelo modelo. Dessa forma, temos que

$$(y_i - \bar{y})^2 = [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 = (y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}),$$

e aplicando o somatório

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}), \\ &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^n (y_i \hat{y}_i - y_i \bar{y} - \hat{y}_i \hat{y}_i + \hat{y}_i \bar{y}), \\ &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^n [\hat{y}_i (y_i - \hat{y}_i) - \bar{y} (y_i - \hat{y}_i)], \\ &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^n \hat{y}_i (y_i - \hat{y}_i) + 2 \bar{y} \sum_{i=1}^n (y_i - \hat{y}_i), \end{aligned}$$

porém,

$$2 \bar{y} \sum_{i=1}^n (y_i - \hat{y}_i) = 2 \bar{y} \left(\underbrace{\sum_{i=1}^n y_i - n \hat{\beta}_0 - \sum_{i=1}^n \hat{\beta}_1 x_i}_{=0 \text{ [Devido a equação normal]}} \right) = 0,$$

e

$$\sum_{i=1}^n \hat{y}_i (y_i - \hat{y}_i) = \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i) \varepsilon_i = \underbrace{\hat{\beta}_0 \sum_{i=1}^n \varepsilon_i + \hat{\beta}_1 \sum_{i=1}^n x_i \varepsilon_i}_{\text{Consequências das equações normais}} = 0.$$

Portanto,

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2. \quad (\text{SSTO})$$

Sendo,

$$\sum_{i=1}^n (y_i - \bar{y})^2, \quad \text{denominada de soma de quadrados total} \quad [\text{SST}], \quad (s1)$$

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad \text{denominada de soma de quadrados dos erros} \quad [\text{SSE}], \quad (s2)$$

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \quad \text{denominada de soma de quadrados da regressão} \quad [\text{SSR}]. \quad (s3)$$

Conforme mostrado na Figura-3, podemos observar que SST está associada a variação total das observações de y com relação a sua média, SSE mede a variação entre os valores observados e aqueles gerados pelo modelo de regressão e SSR mede o quanto bem o modelo estimado ajusta-se aos dados observados. Essas medidas são utilizadas para construir uma estatística que permite testar a hipótese $H_0 : \beta_1 = 0$ contra $H_a : \beta_1 \neq 0$, e são organizadas em uma tabela chamada de Tabela ANOVA, onde ANOVA é um acrônimo para ANalysis Of VAriance, a Figura-4 apresenta essa tabela para o caso do MRLS.

Tabela-1 Análise de Variância para o MRLS.

ANOVA

| Fonte da Variação | Soma de Quadrados | G.L. | Média Quadrada | Estatística F |
|-------------------|--|---------|---------------------------|-----------------------|
| Regressão | $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ | 1 | $MSR = \frac{SSR}{1}$ | $F = \frac{MSR}{MSE}$ |
| Erro | $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ | $n - 2$ | $MSE = \frac{SSE}{n - 2}$ | |
| Total | $SST = \sum_{i=1}^n (y_i - \bar{y})^2$ | $n - 1$ | | |

Na Tabela-1 da ANOVA para o MRLS a coluna G.L. refere-se ao número de graus de liberdade consideradas em cada uma das medidas associadas ao ajuste do modelo. Nessa tabela, a Estatística F é utilizada para no teste da associação linear entre as variáveis Y e X , tendo distribuição F com 1 e $n - 2$ graus de liberdade no numerador e denominador, respectivamente. O teste é delineado estabelecendo-se um nível de significância α e em seguida realizando os seguintes passos:

(i) Determinação das hipóteses

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_a : \beta_1 \neq 0 \end{cases}$$

(ii) Cálculo da estatística de teste $F^* = \frac{MSR}{MSE}$, que de acordo com o Teorema de Cochran

(1934) possui distribuição $F^* \sim \frac{\chi_1^2}{1} \div \frac{\chi_{n-2}^2}{n-2} \rightarrow F^* \sim F_{1;n-2}$;

- (iii) Determinação da região de rejeição, considerando α e uma distribuição $F_{1;n-2;\alpha}$;
- (iv) A partir da estatística de teste, determinar o valor probabilístico do teste (p-valor) como $P(F > F^*)$;
- (v) Conclusão do teste será

$$\begin{aligned} \text{Se } |F^*| \leq F_{1;n-2;\alpha}, & \text{ não rejeitar } H_0, \\ \text{Se } |F^*| > F_{1;n-2;\alpha}, & \text{ rejeitar } H_0. \end{aligned}$$

Coeficiente de Determinação (R^2) e Coeficiente de Correlação (r)

Na construção do MRLS temos duas medidas que desempenham papel importante para avaliar o grau de relacionamento linear entre as variáveis y e x . Elas são denominadas de coeficiente de determinação (denotado por R^2) e coeficiente de correlação (r). O R^2 é definido como

$$R^2 = \frac{SSR}{SST}. \quad (\text{R1})$$

Nessa razão que define o R^2 , o denominador não tem influência do comportamento da variável x , uma vez que é construído baseado nas diferenças de cada y_i com a média de y e o numerador, através do modelo estimado, incorpora a variabilidade em cada y_i associada aos respectivos valores de x . O coeficiente de determinação R^2 representa a fração da variação de cada y_i em relação à média de y que pode ser explicada pela relação linear entre x e y . Em outras palavras, indica a fração da variabilidade total de y que pode ser atribuída à regressão linear em x . Como $SST = SSR + SSE$, também podemos escrever que

$$R^2 = 1 - \frac{SSE}{SST}. \quad (\text{R2})$$

Uma vez que $SSE \leq SST$, o coeficiente de determinação está situado no intervalo entre 0 e 1 ($0 \leq R^2 \leq 1$). Quando o valor do R^2 está próximo de 1, temos que quase toda variação em y com relação a \bar{y} pode ser explicada pela relação linear entre x e y . No caso em que obtemos R^2 esté perto de 0, implica que pouco da variação de y com relação a \bar{y} pode ser explicada pelo relacionamento linear entre x e y . O R^2 então expressa a força do relacionamento linear na forma $y = \beta_0 + \beta_1 x + \varepsilon$, o que permite avaliar a utilidade de x como uma variável para explicar y nesse tipo de modelo.

Existem outras formas de representação do R^2 , a seguir são apresentadas algumas opções.

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \hat{\beta}_1^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{S_{xy}^2}{S_{xx} S_{yy}}. \quad (\text{R3})$$

A razão $\left[\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \right] / \sum_{i=1}^n (y_i - \bar{y})^2$ usada para expressar o R^2 pode ser provada

considerando uma forma alternativa de escrever a SSR , que sabemos ser igual a $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$. Uma vez que também sabemos que $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ e $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ (o que implica $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$), então ficamos com $\hat{y}_i - \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}) = \hat{\beta}_1 (x_i - \bar{x})$.

A partir do coeficiente de determinação que ainda pode ser escrito como

$$R^2 = \frac{\left[\sum_{i=1}^n (x_i - \bar{x})(\hat{y}_i - \bar{y}) \right]^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2},$$

definimos o coeficiente de correlação (amostral) r como sendo

$$r = \pm \sqrt{R^2}, \quad (R4)$$

ou

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (R5)$$

O sinal positivo para r indica uma inclinação positiva da reta de regressão estimada e um sinal negativo indica inclinação negativa. Os valores de r estão situados no intervalo de -1 a 1 ($-1 \leq r \leq 1$). Para que r seja zero, a covariância entre x e y precisa ser zero, o que significa que não existe relacionamento linear entre estas variáveis, de tal forma que dizemos que x e y são não correlacionadas. Se o r é igual a 1, implicada que x e y têm um relacionamento linear positivo perfeito. Se o r é igual a -1 (menos um), implicada que x e y têm um relacionamento linear negativo perfeito.

Quando $Cov[x, y]$ é zero, o R^2 também é zero, indicando completa falta de ajuste de um modelo do tipo $y = \beta_0 + \beta_1 x + \varepsilon$. Nessa situação temos que $\sum_{i=1}^n \hat{\varepsilon}_i^2$ é igual a S_{yy} . Ou seja, tanto o R^2 quanto o r são iguais a zero, alcançando respectivamente seus valores mínimos. Para que tenhamos $R^2 = 1$, precisamos necessariamente que $\sum_{i=1}^n \hat{\varepsilon}_i^2 = 0$. Podemos reescrever a soma de erros quadrados da seguinte forma

$$\begin{aligned} 1 - \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} &= \hat{\beta}_1^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \rightarrow \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \hat{\beta}_1^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \\ \sum_{i=1}^n \hat{\varepsilon}_i^2 &= \sum_{i=1}^n (y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (y_i - \bar{y})^2 - \frac{\left[\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) \right]^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \end{aligned}$$

$$\sum_{i=1}^n \hat{\varepsilon}_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}. \quad (\text{R5})$$

Sendo que

$$R^2 = 0 \rightarrow \sum_{i=1}^n \hat{\varepsilon}_i^2 = S_{yy}$$

$$R^2 = 1 \rightarrow \sum_{i=1}^n \hat{\varepsilon}_i^2 = 0.$$

Para qualquer valor de R^2 entre 0 e 1, implica que r será menor que $|1|$. O coeficiente de correlação r somente assumirá seus valores limites quando $\sum_{i=1}^n \hat{\varepsilon}_i^2 = 0$. Dessa forma, temos

$$\begin{aligned} \sum_{i=1}^n \hat{\varepsilon}_i^2 &= S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 0, \\ \frac{S_{xy}^2}{S_{xx}} &= S_{yy} \rightarrow \frac{S_{xy}^2}{S_{xx} S_{yy}} = 1 \rightarrow \left[\frac{\sum_{i=1}^n (x_i - \bar{x})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \right]^2 = [r]^2 = 1. \end{aligned}$$

Então, temos que os valores limites que r pode assumir são -1 e 1.

Preicisamos considerar alguns aspectos sobre os usos e limitações associados ao R^2 e ao r . (i) o primeiro consiste no uso dos modelos de regressão linear para tarefas de predição (veja Quadro Conceitual - 1), ou seja, a relação linear imposta pelo modelo pode ser apropriada apenas para a faixa limitada aos valores de x ; (ii) Apesar do R^2 ser utilizado como uma medida de qualidade de ajuste, pode não fornecer a precisão necessária para uma aplicação específica; (iii) independentemente do valor de R^2 , o gráfico de dispersão dos pares de dados deve sempre ser inspecionado para ver se um modelo de regressão linear simples é garantido. Valores altos e baixos de R^2 podem ser associados a uma forte relação não linear entre x e y ; (iv) no caso em que o analista pode controlar os valores da variável x , a magnitude de R^2 depende dessas escolhas e isso pode influenciar sua interpretação; (v) a interpretação usual do coeficiente de correlação r como um estimador da correlação populacional $\rho = \text{Cov}[x, y]/(\sigma_x \sigma_y)$ é apropriada apenas quando x e y são variáveis aleatórias, o que não é o caso na regressão linear simples porque x é assumido como observado sem erro [Leemis, 2023]; (vi) é importante tomar cuidado com o uso do modelo de regressão linear para atribuir causalidade. Por exemplo, imagine a relação entre o número de pessoas usando guarda-chuvas (x) e a quantidade de chuva registrada em um dia (y). É provável que exista uma forte correlação entre essas variáveis, permitindo construir um modelo de regressão entre a quantidade de chuva (y) e o número de guarda-chuvas (x). Entretanto, a correlação entre o número de guarda-chuvas e a quantidade de chuva não implica causalidade. Ou seja, o fato de ambas as variáveis estarem relacionadas não significa que uma é a causa da outra. O número de guarda-chuvas sendo usados é consequência da chuva, não sua causa. Assim, mesmo que exista uma relação linear

aparente, se manipulássemos o número de guarda-chuvas sendo usados (distribuindo mais guarda-chuvas aleatoriamente, por exemplo), isso não teria impacto na quantidade de chuva que cai.

Inferências Considerando β_1

No MRLS o parâmetro β_1 corresponde a inclinação da reta do modelo, e representa a mudança na média da variável dependente y , quando a variável independente x é aumentada de 1 unidade. As inferências a respeito de β_1 , consideram que sua distribuição amostral seja Normal quando tratamos do modelo descrito em (4). Assim, a partir da suposição prévia de que $y \sim N(\beta_0 + \beta_1 x; \sigma^2)$, e de que os termos de erro $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ são variáveis aleatórias independentes distribuídas normalmente, $\varepsilon \sim N(0, \sigma^2)$, temos

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \text{ ou } \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

A prova de que a distribuição amostral de $\hat{\beta}_1$ é Normal, segue do fato de que $\hat{\beta}_1$ é uma combinação linear de y_i , e a combinação linear de variáveis aleatórias independentes que têm distribuição Normal é normalmente distribuída (Neter, 1983; Casella e Berger, 2018, Pág. 471; Feller, 1991; Pitman, 1993).

Sabendo que $\hat{\beta}_1$ tem distribuição Normal, implica que a padronização $(\hat{\beta}_1 - \beta_1)/\sigma_{\hat{\beta}_1}$, gera uma variável que possui distribuição Normal Padrão. Como não conhecemos $\sigma_{\hat{\beta}_1}$ ($\sqrt{Var[\hat{\beta}_1]}$), precisamos considerar o uso de $s_{\hat{\beta}_1}$ em seu lugar. Isso leva à necessidade de saber a distribuição de $(\hat{\beta}_1 - \beta_1)/s_{\hat{\beta}_1}$, sendo que $s_{\hat{\beta}_1}$ é obtido como

$$s_{\hat{\beta}_1} = \frac{MSE}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{MSE}{\frac{\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}{n}}.$$

Também sabemos que a razão $SSE/\sigma_{\hat{\beta}_1}$, é uma estatística que possui distribuição χ^2 com $n - 2$ graus de liberdade (G.L.) e é independente de $\hat{\beta}_0$ e $\hat{\beta}_1$. Para determinar a distribuição de $(\hat{\beta}_1 - \beta_1)/s_{\hat{\beta}_1}$, tomamos

$$\frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \text{ e dividimos numerador e denominador por } \sigma_{\hat{\beta}_1},$$

o que resulta em

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{s_{\hat{\beta}_1}}{\sigma_{\hat{\beta}_1}}} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}}{\sqrt{\frac{s_{\hat{\beta}_1}^2}{\sigma_{\hat{\beta}_1}^2}}} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}}{\sqrt{\frac{(n-2)}{(n-2)} \frac{s_{\hat{\beta}_1}^2}{\sigma_{\hat{\beta}_1}^2}}} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}}{\sqrt{\frac{1}{(n-2)} \underbrace{\frac{(n-2)s_{\hat{\beta}_1}^2}{\sigma_{\hat{\beta}_1}^2}}_Q}} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}}{\sqrt{\frac{Q}{(n-2)}}},$$

sendo que o numerador, $(\hat{\beta}_1 - \beta_1)/\sigma_{\hat{\beta}_1}$ tem distribuição Normal Padrão (e será denotado por Z) e no denominador, $Q \sim \chi^2_{n-2}$ (ou seja, Q tem distribuição Qui-quadrado com $n-2$ graus de liberdade). Denominando essa razão como t , temos que

$$t = \frac{Z}{\sqrt{\frac{Q}{(n-2)}}} \text{ tem distribuição } t\text{-de-Student } t_{n-2},$$

ou, de forma simplificada:

$$\frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \sim t_{n-2}.$$

Uma vez que

$$s_{\hat{\beta}_1}^2 = \frac{MSE}{SS_{xx}} \rightarrow s_{\hat{\beta}_1} = \frac{\sqrt{MSE}}{\sqrt{SS_{xx}}} \rightarrow s_{\hat{\beta}_1} = \frac{\sqrt{MSE}}{\sqrt{(n-1)s_x^2}},$$

com esse resultado e considerando um nível de significância α podemos escrever o intervalo de confiança (IC) para β_1 .

Intervalo de Confiança para β_1

Uma vez que $(\hat{\beta}_1 - \beta_1)/s_{\hat{\beta}_1}$ tem distribuição t de Student com $n-2$ graus de liberdade, a probabilidade de que o valor gerado por esta razão esteja dentro de um intervalo contendo um limite inferior e um limite superior, considerando um nível de confiança de $1-\alpha$, pode ser escrita como

$$P\left(t_{\frac{\alpha}{2}, n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \leq t_{1-\frac{\alpha}{2}, n-2}\right) = 1-\alpha. \quad (\text{IC1})$$

Porém, uma vez que a distribuição t de Student é simétrica, sabemos que

$$t_{\frac{\alpha}{2}, n-2} = -t_{1-\frac{\alpha}{2}, n-2}. \quad (\text{IC2})$$

Usando o R, podemos verificar esse resultado, por exemplo, escrevendo

```
> alfa <- 0.05
> n <- 30
> qt(alfa/2, n - 2)
[1] -2.048407
> -qt(1 - alfa/2, n - 2)
[1] -2.048407
```

Isolando β_1 em (IC1) e considerando o resultado (IC2) temos

$$P\left(\hat{\beta}_1 - t_{1-\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{1-\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1}\right) = 1 - \alpha. \quad (\text{IC3})$$

Então, o intervalo de confiança de β_1 é dado por

$$\hat{\beta}_1 \mp t_{1-\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1} \equiv \hat{\beta}_1 \pm t_{\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1} = \hat{\beta}_1 \pm t_{\frac{\alpha}{2}, n-2} \sqrt{\frac{MSE}{SS_{xx}}}. \quad (\text{IC4})$$

Teste de Hipóteses para β_1

De forma geral, o teste para β_1 poderia ser realizado considerando valores específicos de interesse na alegação inicial, β_{10} , e as seguintes opções alternativas

$$\begin{cases} H_0 : \beta_1 = \beta_{10} & H_0 : \beta_1 \leq \beta_{10} \\ H_a : \beta_1 \neq \beta_{10} & H_a : \beta_1 > \beta_{10} \\ & H_a : \beta_1 < \beta_{10} \end{cases}$$

O que geraria uma estatística de teste $t^* = (\hat{\beta}_1 - \beta_{10}) / s_{\hat{\beta}_1}$, sendo β_{10} o valor alegado na hipótese nula.

No MRLS, interessa saber a existência da associação linear entre x e y , e apesar do Teste de Hipóteses ter as opções alternativas de $\beta_1 < 0$, $\beta_1 > 0$ e $\beta_1 \neq 0$, vamos considerar o Teste de Hipóteses para β_1 como

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_a &: \beta_1 \neq 0 \end{aligned}$$

Para fazer o teste, a estatística de teste usada é dada por

$$t^* = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}.$$

Uma vez estabelecido o nível de significância α , a decisão do teste pode ser baseada no valor probabilístico do teste (p-valor), tal que

$$\begin{aligned} \text{Região de Rejeição: } |t| &\geq t_{1-\frac{\alpha}{2}, n-2}, \\ \text{p-valor: } 2.P(|t^*|), \end{aligned}$$

concluindo que

$$\text{Se } |t^*| \leq t_{1-\frac{\alpha}{2}, n-2}, \text{ não rejeitar } H_0,$$

$$\text{Se } |t^*| > t_{1-\frac{\alpha}{2}, n-2}, \text{ rejeitar } H_0.$$

Inferências Considerando β_0

As obtenções do Intervalo de Confiança (IC) e do Teste de Hipóteses (TH) para β_0 são realizadas de forma similar ao que foi realizado para β_1 . Lembrando que β_0 corresponde ao valor esperado de y quando x é igual a zero ($E[y|x=0]$), e de que a reta de regressão

sempre passa pelo ponto ($E[x], E[y]$). Para mostrar essa última característica partimos dos resultados

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

e

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Reescrevendo \hat{y} temos

$$\hat{y} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x.$$

Subtraindo \hat{y} dessa última expressão resulta em

$$\hat{y} - \hat{y} = 0 = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x - \hat{y}.$$

Isso implica que, se o ponto (\bar{x}, \bar{y}) estiver na reta de regressão ($x = \bar{x}$ e $\hat{y} = \bar{y}$), essa igualdade é verdadeira. Ou seja,

$$\hat{y} - \hat{y} = 0 = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} - \bar{y}.$$

Conforme determinamos anteriormente, o estimador de β_0 ($\hat{\beta}_0$), é igual a diferença entre \bar{y} e $\hat{\beta}_1 \bar{x}$. De tal forma que, quanto mais próximo $\hat{\beta}_1$ estiver de zero, mais $\hat{\beta}_0$ se aproxima de \bar{y} .

As inferências a respeito de β_0 , consideram que sua distribuição amostral seja Normal quando tratamos do modelo descrito em (4). Assim, a partir da suposição prévia de que $y \sim N(\beta_0 + \beta_1 x; \sigma^2)$, e de que os termos de erro $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ são variáveis aleatórias independentes distribuídas normalmente, $\varepsilon \sim N(0, \sigma^2)$, temos

$$\hat{\beta}_0 \sim N\left[\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \sigma^2\right] \text{ ou } \hat{\beta}_0 \sim N\left[\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \sigma^2\right].$$

Da mesma forma que fizemos para β_1 , podemos mostrar que:

$$\frac{\hat{\beta}_0 - \beta_0}{s_{\hat{\beta}_0}} \sim t_{n-2},$$

com esse resultado e considerando um nível de significância α podemos escrever o intervalo de confiança (IC) para β_0 .

Intervalo de Confiança para β_1

Uma vez que $(\hat{\beta}_1 - \beta_1)/s_{\hat{\beta}_1}$ tem distribuição t de Student com $n-2$ graus de liberdade, a probabilidade de que o valor gerado por esta razão esteja dentro de um intervalo contendo um limite inferior e um limite superior, considerando um nível de confiança de $1-\alpha$, pode ser escrita como

$$P\left(t_{\frac{\alpha}{2}, n-2} \leq \frac{\hat{\beta}_0 - \beta_0}{S_{\hat{\beta}_0}} \leq t_{1-\frac{\alpha}{2}, n-2}\right) = 1 - \alpha. \quad (\text{ICB0})$$

Isolando β_0 em (ICB0) temos

$$P\left(\hat{\beta}_0 - t_{\frac{\alpha}{2}, n-2} S_{\hat{\beta}_0} \leq \beta_0 \leq \hat{\beta}_0 + t_{\frac{\alpha}{2}, n-2} S_{\hat{\beta}_0}\right) = 1 - \alpha. \quad (\text{IC3})$$

Então, o intervalo de confiança de β_0 é dado por

$$\hat{\beta}_0 \pm t_{\frac{\alpha}{2}, n-2} S_{\hat{\beta}_0} = \hat{\beta}_0 \pm t_{\frac{\alpha}{2}, n-2} \sqrt{MSE \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}. \quad (\text{IC4})$$

Teste de Hipóteses para β_0

Da mesma forma como fizemos no caso da inclinação do MRLS, o teste para β_0 pode ser realizado considerando valores específicos de interesse na alegação inicial (Hipótese Nula), β_{00} , e as seguintes opções alternativas

$$\begin{cases} H_0 : \beta_0 = \beta_{00} \\ H_a : \beta_0 \neq \beta_{00} \end{cases} \quad \begin{cases} H_0 : \beta_0 \leq \beta_{00} \\ H_a : \beta_0 > \beta_{00} \end{cases} \quad \begin{cases} H_0 : \beta_0 \geq \beta_{00} \\ H_a : \beta_0 < \beta_{00} \end{cases}$$

O que geraria uma estatística de teste $t^* = (\hat{\beta}_0 - \beta_{00}) / S_{\hat{\beta}_0}$, sendo β_{00} o valor alegado na hipótese nula.

No MRLS, interessa saber a existência da associação linear entre x e y , e apesar do Teste de Hipóteses ter as opções alternativas de $\beta_0 < 0$, $\beta_0 > 0$ e $\beta_0 \neq 0$, vamos considerar o Teste de Hipóteses para β_0 como

$$\begin{aligned} H_0 &: \beta_0 = 0 \\ H_a &: \beta_0 \neq 0 \end{aligned}$$

Para fazer o teste, a estatística de teste usada é dada por

$$t^* = \frac{\hat{\beta}_0}{S_{\hat{\beta}_0}}.$$

Uma vez estabelecido o nível de significância α , a decisão do teste pode ser baseada no valor probabilístico do teste (p-valor), tal que

$$\text{Região de Rejeição: } |t| \geq t_{1-\frac{\alpha}{2}, n-2},$$

$$\text{p-valor: } 2.P(|t| \geq |t^*|),$$

concluindo que

Se $|t^*| \leq t_{1-\frac{\alpha}{2};n-2}$, não rejeitar H_0 ,
 Se $|t^*| > t_{1-\frac{\alpha}{2};n-2}$, rejeitar H_0 .

Inferências Considerando σ^2

O ajuste do modelo ao conjunto de dados observados gera uma diferença entre cada y_i e o respectivo $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Essa diferença, $y_i - \hat{y}_i = \varepsilon_i$, desempenha um papel importante no diagnóstico do modelo de regressão. Como mencionado anteriormente, a variabilidade σ_i , associada ao erro ε_i , não é diretamente observável e, portanto, precisa ser estimada. O interesse é obter um estimador que seja não viesado e escrito em termos da soma dos quadrados das diferenças $SSE = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$.

Assim, o estimador utilizado é dado por

$$\hat{\sigma}^2 = s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2} = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-2}. \quad (S1)$$

Uma vez que precisamos de um estimador não viesado para a variância do erro (σ^2), consideramos o denominador de (S1) sendo $n-2$. Isso porque o MRLS apresenta dois parâmetros (β_0 e β_1) a serem estimados. Em um modelo de regressão linear múltipla, é fundamental ajustar corretamente o número de graus de liberdade ao estimar a variância dos erros, representada por σ^2 . Para obter um estimador não viesado dessa variância, o denominador deve ser $n-p$, onde n representa o número total de observações e p o número de parâmetros estimados. Esse ajuste é necessário para corrigir o viés associado ao uso de estimativas amostrais, como destacado por Wooldridge (2013) e Montgomery, Peck e Vining (2021). A escolha do denominador $n-p$, em vez de apenas n , é amplamente justificada na literatura como uma medida para compensar a perda de graus de liberdade decorrente da estimação dos parâmetros, conforme discutido por Gujarati e Porter (2009).

Podemos mostrar que o MSE é um estimador não viesado de σ^2 , ou seja, $E[MSE] = \sigma^2$.

Para isso consideramos

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{y}_i)^2 &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \stackrel{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}{=} \sum_{i=1}^n (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)^2, \\ \sum_{i=1}^n [(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})]^2 &= \sum_{i=1}^n [(y_i - \bar{y})^2 - 2\hat{\beta}_1 (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_1^2 (x_i - \bar{x})^2] \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2 \sum_{i=1}^n \hat{\beta}_1 (x_i - \bar{x})(y_i - \bar{y}) + \sum_{i=1}^n \hat{\beta}_1^2 (x_i - \bar{x})^2, \end{aligned}$$

lambendo que $\hat{\beta}_1$ pode ser escrito como S_{xy}/S_{xx} , temos

$$\sum_{i=1}^n (y_i - \bar{y})^2 - 2 \sum_{i=1}^n \hat{\beta}_1^2 (x_i - \bar{x})^2 + \sum_{i=1}^n \hat{\beta}_1^2 (x_i - \bar{x})^2 = \sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n \hat{\beta}_1^2 (x_i - \bar{x})^2,$$

é interessante levar em conta a possibilidade de escrever o somatório $\sum_{i=1}^n (y_i - \bar{y})^2$ em termos de ε_i , uma vez que sabemos que $E[\varepsilon_i] = 0$. Dessa forma, usando a expressão (4B) temos

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n [\beta_1(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})]^2 = \sum_{i=1}^n \beta_1^2 (x_i - \bar{x})^2 + 2 \sum_{i=1}^n \beta_1 (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) + \sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2.$$

Continuando, temos

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \beta_1^2 (x_i - \bar{x})^2 + 2 \sum_{i=1}^n \beta_1 (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) + \sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2 - \sum_{i=1}^n \hat{\beta}_1^2 (x_i - \bar{x})^2.$$

Tomando a esperança dos dois lados da igualdade, resulta em

$$\begin{aligned} E\left[\sum_{i=1}^n (y_i - \hat{y}_i)^2\right] &= \\ &= E\left[\underbrace{\sum_{i=1}^n \beta_1^2 (x_i - \bar{x})^2}_{(1)}\right] + 2E\left[\underbrace{\sum_{i=1}^n \beta_1 (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}_{(2)}\right] + E\left[\underbrace{\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2}_{(3)}\right] - E\left[\underbrace{\sum_{i=1}^n \hat{\beta}_1^2 (x_i - \bar{x})^2}_{(4)}\right]. \end{aligned}$$

Vamos obter, separadamente, o resultado de cada uma das esperanças do lado direito dessa igualdade.

$$\begin{aligned} (1) \quad E\left[\beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2\right] &= \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2. \\ (2) \quad 2E\left[\beta_1 \sum_{i=1}^n (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})\right] &= 2\beta_1 E\left[\sum_{i=1}^n (x_i \varepsilon_i - \bar{\varepsilon} x_i - \bar{x} \varepsilon_i + \bar{x} \bar{\varepsilon})\right] \\ &= 2\beta_1 E\left[\sum_{i=1}^n x_i \varepsilon_i - \bar{\varepsilon} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n \varepsilon_i + \sum_{i=1}^n \bar{x} \bar{\varepsilon}\right] = 2\beta_1 E\left[\sum_{i=1}^n x_i \varepsilon_i - n \bar{\varepsilon} \bar{x} - n \bar{\varepsilon} \bar{x} + n \bar{\varepsilon} \bar{x}\right] \\ &= 2\beta_1 E\left[\sum_{i=1}^n x_i \varepsilon_i - n \bar{\varepsilon} \bar{x} - n \bar{\varepsilon} \bar{x} + n \bar{\varepsilon} \bar{x}\right] = 2\beta_1 E[-n \bar{\varepsilon} \bar{x}] = -2n \bar{x} \beta_1 E[\bar{\varepsilon}] = 0. \end{aligned}$$

$$\begin{aligned} (3) \quad E\left[\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2\right] &= E\left[\sum_{i=1}^n (\varepsilon_i^2 - 2\bar{\varepsilon}\varepsilon_i + \bar{\varepsilon}^2)\right] = E\left[\sum_{i=1}^n \varepsilon_i^2 - 2\bar{\varepsilon} \sum_{i=1}^n \varepsilon_i + \sum_{i=1}^n \bar{\varepsilon}^2\right], \\ E\left[\sum_{i=1}^n \varepsilon_i^2 - 2n \bar{\varepsilon}^2 + n \bar{\varepsilon}^2\right] &= E\left[\sum_{i=1}^n \varepsilon_i^2 - n \bar{\varepsilon}^2\right] = E\left[\sum_{i=1}^n \varepsilon_i^2\right] - n E[\bar{\varepsilon}^2], \end{aligned}$$

atente que na esperança de $\bar{\varepsilon}^2$ precisamos considerar todos os produtos do tipo $\varepsilon_i \varepsilon_j$ e $\varepsilon_i \varepsilon_j$ ($i \neq j$). Como consideramos que os termos de erro são não correlacionados, teremos

$$\begin{aligned}
E \left[\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2 \right] &= \sum_{i=1}^n E[\varepsilon_i^2] - nE \left[\frac{(\varepsilon_1 + \varepsilon_2 + \dots + \varepsilon_n)^2}{n^2} \right] \\
&= n\sigma^2 - \frac{1}{n} E \left[\sum_{i=1}^n \varepsilon_i^2 + \sum_{i \neq j} \varepsilon_i \varepsilon_j \right] = n\sigma^2 - \frac{n\sigma^2}{n} - 0 = n\sigma^2 - \sigma^2, \\
E \left[\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2 \right] &= (n-1)\sigma^2. \\
(4) \quad E \left[\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \right] &= E \left\{ \left[\frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^2 \sum_{i=1}^n (x_i - \bar{x})^2 \right\} \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} E \left\{ \left[\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) \right]^2 \right\},
\end{aligned}$$

nesse ponto, vamos utilizar a seguinte forma de escrever a variância para uma variável aleatória $Var[X] = E[X^2] - E^2[X]$, mais propriamente, no nosso caso vamos reescrever essa expressão em termos de S_{xy} , ou seja, $Var[S_{xy}] = E[S_{xy}^2] - E^2[S_{xy}]$. Sendo assim, temos

$$E \left[\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) \right] = E \left[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right] = \sum_{i=1}^n (x_i - \bar{x}) E[y_i - \bar{y}],$$

lembrando novamente da equação (4B)

$$E \left[\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) \right] = \sum_{i=1}^n (x_i - \bar{x}) E[\beta_1(x_i - \bar{x}) + \varepsilon_i - \bar{\varepsilon}] = \beta_1 \sum_{i=1}^n (x_i - \bar{x})^2.$$

Quanto a variância de S_{xy} temos

$$\begin{aligned}
Var \left[\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) \right] &= Var \left[\sum_{i=1}^n (y_i x_i - y_i \bar{x} - \bar{y} x_i + \bar{y} \bar{x}) \right] \\
&= Var \left[\sum_{i=1}^n y_i x_i - \sum_{i=1}^n \bar{x} y_i - \bar{y} \sum_{i=1}^n x_i + n \bar{y} \bar{x} \right] = Var \left[\sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \bar{x} - n \bar{y} \bar{x} + n \bar{y} \bar{x} \right] \\
&= Var \left[\sum_{i=1}^n (x_i - \bar{x}) y_i \right] = \left[\sum_{i=1}^n (x_i - \bar{x}) \cdot \sum_{i=1}^n (x_i - \bar{x}) \right] Var[y_i] = \sum_{i=1}^n (x_i - \bar{x})^2 Var[y_i] \\
&= \sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2 = \sigma^2 \sum_{i=1}^n (x_i - \bar{x})^2.
\end{aligned}$$

Voltando à nossa quarta esperança

$$E \left[\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \right] = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \left\{ \sigma^2 \sum_{i=1}^n (x_i - \bar{x})^2 + \left[\beta_1 \sum_{i=1}^n (x_i - \bar{x})^2 \right]^2 \right\}$$

$$= \sigma^2 + \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2.$$

Portanto,

$$\begin{aligned}
E\left[\sum_{i=1}^n (y_i - \hat{y}_i)^2\right] &= \\
&= \underbrace{\beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}_{(1)} + \underbrace{0}_{(2)} + \underbrace{(n-1)\sigma^2}_{(3)} - \underbrace{\left[\sigma^2 + \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2\right]}_{(4)} \\
&= (n-1)\sigma^2 - \sigma^2 = (n-2)\sigma^2, \\
E\left[\sum_{i=1}^n (y_i - \hat{y}_i)^2\right] &= (n-2)\sigma^2, \\
\frac{E\left[\sum_{i=1}^n (y_i - \hat{y}_i)^2\right]}{n-2} &= E\left[\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}\right] = \sigma^2, \\
E[MSE] &= \sigma^2. \tag{MSE}
\end{aligned}$$

Ou seja, o MSE é um estimador não viésado de σ^2 .

Intervalo de Confiança para σ^2

Considerando os pressupostos do MRLS podemos mostrar que

$$\frac{(n-2)MSE}{\sigma_\varepsilon^2} \sim \chi^2_{n-2}.$$

Ou seja, a razão da soma dos quadrados dos erros ($SSE = (n-2)MSE$), pela variância (σ_ε^2), segue uma distribuição Qui-quadrado com $n-2$ graus de liberdade. Escrevemos σ_ε^2 para deixar claro que essa variância é obtida a partir dos erros ε_i oriundos das diferenças $y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. Assim, para um nível de confiança de $100(1-\alpha)\%$, os limites inferior e superior para σ_ε^2 são dados por:

$$\frac{(n-2)MSE}{\chi^2_{\frac{\alpha}{2};n-2}} \leq \sigma_\varepsilon^2 \leq \frac{(n-2)MSE}{\chi^2_{1-\frac{\alpha}{2};n-2}},$$

sendo $\chi^2_{\alpha/2;n-2}$ e $\chi^2_{1-\alpha/2;n-2}$ os quantis da distribuição Qui-quadrado com $n-2$ graus de liberdade.

3.1.1.5 Aplicações da Análise de Regressão Linear Simples: Estimação Intervalar para $E[Y_p]$ e Intervalo de Predição para $Y_{p(novo)}$

Antes de iniciar um tratamento mais técnico sobre o assunto, vamos considerar o aspecto intuitivo associado às incertezas que envolvem a predição de uma variável de interesse. Para isso, consideraremos a situação em que uma agência de pesquisa automotiva esteja

avaliando os preços de carros elétricos de acordo com a autonomia. Nesse caso a variável dependente (\hat{Y}) é o preço, que será estimado com base em um valor particular (x_p) de autonomia (X). De acordo com Faraway (2009), temos dois tipos de predições que podem ser realizadas para um dado valor : (i) suponha que um carro elétrico seja lançado no mercado com a característica x_p . O preço predito será dado por $\hat{\beta}_0 + \hat{\beta}_1 x_p$, mas ao avaliar a variância dessa predição precisamos incluir a variância de ε ; (ii) outro interesse seria saber, a qual preço seria vendido, em média, um carro elétrico com a autonomia x_p ? Novamente a predição é obtida por $\hat{\beta}_0 + \hat{\beta}_1 x_p$, mas agora, somente as variâncias $\hat{\beta}_0$ e de $\hat{\beta}_1$ precisam ser levadas em conta. Na maioria das vezes, a primeira forma de predição, denominada de **predição para uma nova observação**, é mais aplicada. Enquanto o segundo caso, chamado de **predição da resposta média**, é menos comumente utilizado. Isso porque, no nosso exemplo, é como se estivéssemos interessados no preço médio populacional dos carros elétricos a partir de uma autonomia específica.

Sendo assim, a construção de modelos de regressão tem como objetivos principais as seguintes aplicações: (i) predição de um valor individual de Y , digamos Y_p [um valor particular de Y] correspondente a um dado X , que chamaremos de x_p , e para a qual também construiremos um intervalo de confiança; (ii) para uma determinada distribuição de um valor Y , predizer o valor médio condicionado de Y correspondente a um X selecionado, por exemplo x_p [um valor particular de X], que pertence a linha de regressão populacional (FRP). Para isso, construímos um intervalo de confiança para um determinado valor de Y correspondente a um valor escolhido de X . De forma geral, descrevemos esses objetivos como estimativa de intervalo de predição (IP) para uma nova observação $Y_{p(novo)}$ e estimativa intervalar para $E[Y_p]$. Comparando esses dois tipos diferentes de intervalos de confiança, temos que o primeiro é um intervalo de confiança maior, o que reflete a menor precisão resultante da estimativa de um único valor predito de Y em vez do valor médio (Draper e Smith, 1998).

Estimação Intervalar para $E[Y = y_h]$

Digamos que em uma análise, para um valor particular da variável preditora, $X = x_p$, haja interesse em conhecer um intervalo de confiança para o valor esperado da variável resposta, $E[Y_p]$. Esse intervalo de confiança para a média do valor Y , fornece um intervalo para um valor estimado Y , dado um X , com um nível de confiança desejável de $1 - \alpha$ (Neter *et al.*, 2004). Quando toda a linha de regressão é de interesse, uma região de confiança pode fornecer resultados simultâneos sobre estimativas de Y para vários valores da variável preditora X ; ou seja, para um conjunto de valores do regressor, $100(1 - \alpha)\%$ dos valores de resposta correspondentes estarão neste intervalo (Weisberg, 2005).

Nesse contexto, o modelo de regressão linear simples é usado para predizer o valor esperado condicional de Y . Portanto, denotaremos o valor de interesse X por x_p , uma constante fixa observada sem erro intrínseco dentro do escopo do modelo de regressão linear simples e o valor aleatório associado Y denotado por Y_p , que tem um valor esperado

condicional $E[Y_p]$. Esta notação compacta para o valor esperado condicional será adotada em vez da mais precisa $E[Y|X=x_p]$ (Neter *et al.*, 2004).

Para obter o intervalo de confiança para $E[Y_p]$, primeiro verificamos que sendo Y_p um valor aleatório da variável independente Y , para um dado valor específico e não aleatório x_p , tem o valor esperado dado por

$$E[Y_p] = E[\beta_0 + \beta_1 x_p + \varepsilon_p] = \beta_0 + \beta_1 x_p.$$

Como

$$\hat{Y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p,$$

temos que

$$E[\hat{Y}_p] = E[\hat{\beta}_0 + \hat{\beta}_1 x_p] = E[\hat{\beta}_0] + E[\hat{\beta}_1] x_p = \beta_0 + \beta_1 x_p. \quad (\text{EYp})$$

Ou seja, \hat{Y}_p é um estimador não viesado de $E[Y_p]$.

A variância de \hat{Y}_p , obtemos da seguinte forma

$$\text{Var}[\hat{Y}_p] = \text{Var}[\hat{\beta}_0 + \hat{\beta}_1 x_p],$$

como sabemos que $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, então

$$\begin{aligned} \text{Var}[\hat{Y}_p] &= \text{Var}[\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_p] = \text{Var}[\bar{y} + (x_p - \bar{x}) \hat{\beta}_1] \\ &= \text{Var}[\bar{y}] + (x_p - \bar{x})^2 \text{Var}[\hat{\beta}_1] = \text{Var}\left[\frac{\sum_{i=1}^n y_i}{n}\right] + (x_p - \bar{x})^2 \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \text{Var}\left[\frac{y_1 + y_2 + \dots + y_n}{n}\right] + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2 = \frac{n}{n^2} \sigma^2 + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2 \\ \text{Var}[\hat{Y}_p] &= \sigma^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]. \end{aligned} \quad (\text{VARYp})$$

Também poderíamos denotar $\text{Var}[\hat{Y}_p]$ como $\sigma^2[\hat{Y}_p]$.

Uma vez que \hat{Y}_p é uma combinação linear de Y_1, Y_2, \dots, Y_n e os Y_i 's são normalmente distribuídos, além das suposições iniciais sobre os termos de erro, temos que

$$\hat{Y}_p \sim N\left(E[\hat{Y}_p]; \text{Var}[\hat{Y}_p]\right), \quad (\text{NORMYp})$$

ou

$$\hat{Y}_p \sim N\left(\beta_0 + \beta_1 x_p; \sigma^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]\right). \quad (\text{NORMYp2})$$

A padronização de \hat{Y}_p , $\frac{\hat{Y}_p - E[\hat{Y}_p]}{\sqrt{Var[\hat{Y}_p]}}$, terá distribuição $N(0,1)$. Porém, como a variância de \hat{Y}_p não é conhecida na maior parte das vezes, iremos utilizar o MSE de tal forma que teremos sua variância amostral sendo dada por

$$s^2[\hat{Y}_p] = MSE\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]. \quad (\text{VAMYp})$$

Isso implica que, no MRLS com termos de erro independentes e normalmente distribuídos, teremos

$$\frac{\hat{Y}_p - E[\hat{Y}_p]}{s[\hat{Y}_p]} \sim t_{n-2}. \quad (\text{tYp})$$

Consequentemente, para essa distribuição, considerando um nível de significância α , além do que $E[\hat{Y}_p] = E[Y_p]$, teremos que um intervalo de confiança de $100(1-\alpha)\%$ para $E[Y_p]$ é dado por

$$\hat{Y}_p \pm t_{\frac{\alpha}{2}; n-2} s[\hat{Y}_p], \quad (\text{ICYp})$$

ou

$$\hat{\beta}_0 + \hat{\beta}_1 x_p - t_{\frac{\alpha}{2}, n-2} \sqrt{MSE\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]} < E[Y_p] < \hat{\beta}_0 + \hat{\beta}_1 x_p + t_{\frac{\alpha}{2}, n-2} \sqrt{MSE\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}.$$

Note que esse intervalo de confiança (IC) é uma função de x_p , de tal forma que terá seu comprimento mínimo em $x_p = \bar{x}$. Além disso, a medida que a diferença $|x_p - \bar{x}|$ torna-se maior, o intervalo também aumenta. Por isso, é sugerido que x_p seja um valor situado entre o menor e o maior valores de X , não necessariamente sendo igual a qualquer valor observado de X . Para usar o MRLS para dados fora da amostra (out of sample) é necessário uma análise mais cuidadosa de sua capacidade de generalização e extração.

Intervalo de Predição para $Y_{p(novo)}$

Quando construímos um modelo de regressão linear simples, precisamos dos pares de valores observados de X e Y , respectivamente dados por $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$. O modelo presta-se, principalmente, para que a partir de um valor particular de interesse x_p consigamos gerar informação sobre o valor de Y , a variável a ser explicada. Uma vez que construímos o MRLS, e formos questionados, ou de alguma maneira, precisarmos de uma estimativa para Y , necessariamente devemos ser informados sobre o valor particular de interesse de X . Tendo esse x_p , primeiro verificamos se ele está entre o menor e o maior valor de X que foram observados. Ou seja, $\min(x_1, x_2, \dots, x_n) \leq x_p \leq \max(x_1, x_2, \dots, x_n)$. Essa recomendação é devido a característica dos modelos de regressão terem intervalos de confiança e predição que ficam mais amplos a medida que a diferença $|x_p - \bar{x}|$ aumenta.

Sendo assim, uma vez tendo o x_p , também teremos um valor que não necessariamente é igual a um dos valores observados de Y . Como Y é uma variável aleatória, é importante conhecer esse $Y_{p(novo)}$ com algum grau de certeza.

Para construir o intervalo de predição para $Y_{p(novo)}$, usamos a mesma estratégia do caso anterior, ou seja, comparando as esperanças entre $Y_{p(novo)}$ e \hat{Y}_p .

Retomando, porquê dessa estratégia. Lembre, nosso objetivo é construir um intervalo de predição, e não confiança, como definimos anteriormente. O intervalo de confiança mais fundamental é o IC para a média populacional, porque como sabemos, pelo Teorema Limite Central (TLC) a média amostral tem distribuição Normal. Ou seja, construímos um IC para um parâmetro populacional. Porém, quando temos uma nova observação individual, também queremos um intervalo que contemple a variabilidade associada a esse dado. Então, consideraremos um novo Y pertencente ao conjunto de observações, tal como se tivéssemos um novo par de valores observados X e Y , $(X_{n+1} = x_p, Y_{n+1} = Y_{p(novo)})$, no conjunto de pares $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$. Esse X_{n+1} particular, corresponde ao x_p . Uma vez que temos nosso modelo estimado, o valor da variável dependente estimado a partir de x_p é, como sabemos

$$\hat{Y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p.$$

Obviamente, podemos considerar um novo erro de predição ($\hat{\epsilon}_{novo}$) associado, dado por

$$\hat{\epsilon}_{novo} = Y_{p(novo)} - \hat{Y}_p.$$

Sendo

$$E[\hat{\epsilon}_{novo}] = E[Y_{p(novo)} - \hat{Y}_p] = 0.$$

Além disso, para a variância de $\hat{\epsilon}_{novo}$ temos

$$Var[\hat{\epsilon}_{novo}] = Var[Y_{p(novo)} - \hat{Y}_p] = Var[Y_{p(novo)}] + Var[\hat{Y}_p],$$

isso porque os Y 's são independentes. E considerando (VARYp), e que os Y 's também são identicamente distribuídos gera

$$Var[\hat{\epsilon}_{novo}] = \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right],$$

$$Var[\hat{\epsilon}_{novo}] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]. \quad (\text{var_eps_pred})$$

Além disso, uma vez que os erros de predição estimados, $\hat{\epsilon}_{pred}$, são independentes e combinações lineares de variáveis aleatórias normalmente distribuídas, temos

$$\hat{\epsilon}_{pred} \sim N \left(0; \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \right).$$

Novamente, como a variância de $\hat{Y}_p(\sigma^2)$ não é conhecida na maior parte das vezes, iremos utilizar o MSE, de tal forma que teremos o desvio padrão amostral do erro de predição sendo dado por

$$s_{pred} = \sqrt{MSE \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}. \quad (\text{spred})$$

Agora, considerando que

$$\frac{\hat{\epsilon}_{pred}}{s_{pred}} = \frac{Y_{p(novo)} - \hat{Y}_p}{s_{pred}} \sim t_{n-2},$$

podemos então derivar o intervalo de predição $100(1-\alpha)\%$ para $Y_{p(novo)}$ como sendo

$$\hat{Y}_p \pm t_{\frac{\alpha}{2}; n-2} s_{pred},$$

ou

$$\hat{\beta}_0 + \hat{\beta}_1 x_p \pm t_{\frac{\alpha}{2}; n-2} \sqrt{MSE \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}. \quad (\text{IC_Ypnovo})$$

Ou seja, a partir do conjunto de observações coletadas, utilizando o IP da expressão (IC_Ypnovo), podemos estar $100(1-\alpha)\%$ confiantes de que uma nova observação $Y_{p(novo)}$ da variável de interesse, considerando um valor particular x_p , estará nesse intervalo.

3.1.1.6 Pressupostos do Modelo de Regressão

Tal como mencionado anteriormente, para que um modelo de regressão esteja bem especificado é necessário que sejam realizadas análises que levem em conta se os pressupostos da construção do modelo estão atendidos. Considerando, tanto o modelo de regressão linear simples, como o modelo de regressão linear múltipla, os principais pressupostos que precisam ser verificados nesses modelos são:

- (i) Normalidade dos resíduos;
- (ii) Homocedasticidade dos resíduos;
- (iii) Ausência de outliers;
- (iv) Linearidade nos parâmetros e nas variáveis;
- (v) Ausência de autocorrelação serial nos resíduos;
- (vi) Multicolinearidade entre as variáveis independentes [verificado no MRLM].

A construção de modelos de regressão contempla um conjunto de ferramentas bastante amplo da teoria estatística. Existem diversos livros que são dedicados exclusivamente à análise de regressão. A seguir vamos descrever de maneira bastante sucinta como avaliar cada um dos pressupostos mencionados acima.

■ Normalidade dos resíduos

Existem diversas causas que podem acarretar a falta de normalidade dos resíduos gerados pelo modelo de regressão. Entre as mais comuns, estão a presença de outliers e a mal especificação do modelo. Uma vez que os resíduos do modelo foram obtidos, é comum construírem-se gráficos para avaliar o comportamento distribucional dos resíduos. Apesar dos apoios gráficos, o diagnóstico de normalidade é realizado através de testes de normalidade em que os mais comuns são:

- Teste de Kolmogorov-Smirnov (teste mais amplamente utilizado);
- Teste de Shapiro-Wilk (mais indicado para conjuntos de dados com menos de 50 observações);
- Teste de Jarque-Bera (é um teste sensível a grandes tamanhos de amostra, mesmo quando temos pequenos desvios da normalidade podem resultar na rejeição de hipótese nula);
- Teste de Anderson-Darling (é um teste menos sensível a presença de outliers do que outros teste, mas quando os dados proveem de uma distribuição Normal, ele pode rejeitar mais do que outros testes);
- Teste de Lilliefors (considera a correção de Lilliefors, devido ao fato de que na maior parte das vezes não são conhecidos a verdadeira média e desvio padrão da população).

Como os testes de Kolmogorov-Smirnov e Shapiro-Wilk estão entre os testes de normalidade mais amplamente utilizados, vamos detalhar a construção de cada um desses testes. Para isso, será utilizada uma abordagem procedural com foco na implementação das funções que definem esses testes,

Estatísticas e comandos do R para realizar testes de normalidade dos resíduos

● *Teste de Kolmogorov-Smirnov*

O comando do R que gera a estatística do Teste de Kolmogorov-Smirnov é o `ks.test`. A estatística do teste, o estimador, usado nesse comando é dada por:

$$D = \max(D^+, D^-). \quad (\text{ks1})$$

Sendo

$$D^+ = \max_x \left[\frac{k}{n} - F(x) \right], \quad k = 1, 2, \dots, n \quad (\text{ks2})$$

e

$$D^- = \max_x \left[F(x) - \frac{k-1}{n} \right], \quad k = 1, 2, \dots, n. \quad (\text{ks3})$$

A obtenção da estatística D do teste de normalidade de Kolmogorov é obtida da seguinte forma:

- ordene, de forma crescente, os dados que interessam para fazer o teste de normalidade. Por exemplo, os dados podem ser os resíduos gerados pelo modelo de regressão;
- você pode, ou não, padronizar os dados (aqui, vamos fazer o teste para os resíduos padronizados);
- considerando cada valor padronizado como um quantil, determine as áreas geradas/acumuladas por esses valores sob uma distribuição Normal. Essas áreas correspondem aos $F(x)$ das duas expressões acima;
- calcule todos os n valores das diferenças $k/n - F(x)$ e atribua ao D^+ o maior valor encontrado;
- calcule todos os n valores das diferenças $F(x) - (k-1)/n$ e atribua ao D^- o maior valor encontrado;
- a estatística D do teste de Kolmogorov, obtida a partir do comando `ks.test` do R, é o maior valor entre D^+ e D^- .

Essa é uma descrição muito sucinta de como obter a estatística D . Porém, sabemos que quando realizamos um teste estatístico, é fundamental determinar o p-valor. O comando `ks.test` do R retorna dois valores importantes, a estatística D e p-valor. A obtenção do primeiro foi descrita acima e para obter o mesmo p-valor, precisamos saber a expressão matemática que foi implementada. Para isso, recorremos ao `help` do R sobre o `ks.test`, de onde extraímos as seguintes referências: Birnbaum e Tingey (1951); Conover (1971); Durbin (1973); Feller (1948); Marsaglia *et al.* (2003); Schröer (1991); Schröer e Trenkler (1995) e Viehmann (2021). Entre esses, o R informa que no caso de um teste bicaudal para uma amostra, p-valores exatos são obtidos conforme descrito em Marsaglia *et al.* (2003) (mas sem usar a aproximação opcional na cauda direita, o que pode ser lento para p-valores pequenos). Recorrendo ao trabalho de Marsaglia *et al.* (2003), temos o desenvolvimento para obter uma função $K(n, D)$ que possibilita a determinação do p-valor no `ks.test` do R. Disponibilizam essa função escrita com a linguagem C, de tal forma que se convertermos para a linguagem R obtemos o p-valor igual ao que é apresentado na saída do `ks.test`.

EXEMPLO 12.1 Modelo de Regressão Linear Simples e Teste de Normalidade de Kolmogorov-Smirnov dos Resíduos

Vamos considerar a construção de um MRLS para explicar uma variável y usando uma variável x . Os dados observados de cada uma das variáveis são apresentados a seguir. O

interesse é somente obter o modelo do tipo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, encontrar os resíduos e testar se não rejeitamos a hipótese de serem provenientes de uma distribuição Normal. Obter o MRLS no R é muito simples. Para isso, usamos a função `lm` na qual indicamos a variável dependente (v. d.) y e a variável independente (v. i.) x . Sendo assim, o primeiro trecho de comandos é dado por

```
dev.off(dev.list()["RStudioGD"])
rm(list=ls())
cat("\f") # Limpa a área de plots
# Limpa o environment
# Limpa o console

x <- c(1,2,3,4,5)
y <- c(1,1,2,2,4)

dados <- data.frame(x, y)

rls <- lm(y ~ x, data = dados); rls
```

Essa sequência de comandos gera como resultado:

| Coefficients: | | | | | |
|----------------|----------|------------|----------|----------|-----------|
| | Estimate | Std. Error | t value | Pr(> t) | |
| (Intercept) | -0.1000 | 0.6351 | -0.157 | 0.8849 | |
| x | 0.7000 | 0.1915 | 3.656 | 0.0354 * | |
| --- | | | | | |
| Signif. codes: | 0 ‘***’ | 0.001 ‘**’ | 0.01 ‘*’ | 0.05 ‘.’ | 0.1 ‘ ’ 1 |

Sendo o modelo

$$\hat{y} = -0,1 + 0,7x,$$

com coeficientes que são estatisticamente significantes, considerando um nível de significância de 0,01.

Esse modelo ajustado corresponde exatamente ao ajuste de uma reta de regressão ao conjunto de valores observados. A Figura-R1 mostra a reta de regressão ajustada aos dados e o correspondente intervalo de confiança. Vamos discutir mais detalhadamente sobre a implementação dessa reta aos dados em um exemplo mais adiante. Nesse momento, vamos focar nos pressupostos da construção do MRLS. Os comandos para gerar a reta ajustada aos dados estão a seguir.

```
# Scatter Plot e Reta de Regressão
scatter_e_regressao <- grafico +
  stat_smooth(method = "lm",
              formula = y ~ x,
              geom = "smooth") +
  stat_regrline_equation(label.x = 1.2, label.y = 4) +
  stat_cor(label.x = 1.2, label.y = 3.5)

scatter_e_regressao
```

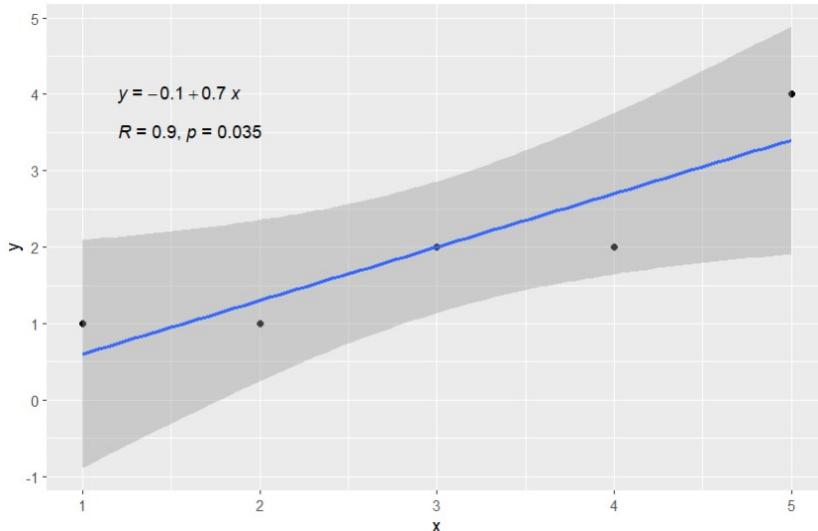


Figura-R1 Reta de regressão e intervalo de confiança.

A partir desse modelo podemos obter os resíduos escrevendo:
`residuos_r1s <- resid(r1s); residuos_r1s`

Agora ordenamos os valores do vetor de resíduos, explicitamos o tamanho da amostra (n) e padronizamos os valores (lembre que padronizar significa, para cada um dos valores, subtrair a média e dividir pelo desvio-padrão), ou seja

```
residuos_ordenados <- sort(residuos_r1s)
n <- length(residuos_ordenados)
z_residuos_ordenados <- scale(residuos_ordenados)
```

Uma vez que temos os resíduos padronizados, aplicamos o comando `ks.test`.
`ks.test(z_residuos_ordenados, "pnorm")`

Que gera como resultado
`Exact one-sample Kolmogorov-Smirnov test`

```
data: z_residuos_ordenados
D = 0.1772, p-value = 0.9895
alternative hypothesis: two-sided
```

Para obtermos a mesma estatística D do teste de Kolmogorov-Smirnov fazemos

```
rank <- 1:n
F_X <- pnorm(z_residuos_ordenados)
D_mais <- max(rank/n - F_X)
D_menos <- max(F_X - (rank - 1)/n)

D <- max(D_mais, D_menos); D
```

O que gera
`[1] 0.1771997`

E para obter o p-valor, usamos a informação do `help` do R que menciona ter utilizado o trabalho de Marsaglia *et al.* (2003) que disponibiliza uma rotina escrita em C. Convertendo essa rotina para a linguagem R temos
`# Função para multiplicar duas matrizes`

```

mMultiply <- function(A, B, m) {
  C <- matrix(0, nrow = m, ncol = m) # Matriz resultado
  for (i in 1:m) {
    for (j in 1:m) {
      C[i, j] <- sum(A[i, ] * B[, j])    }  }
  return(C)}

# Função para calcular a potência de uma matriz
mPower <- function(A, eA, m, n) {
  if (n == 1) {
    return(list(v = A, ev = eA))  }
  half <- mPower(A, eA, m, floor(n / 2))
  v <- half$v
  ev <- half$ev
  B <- mMultiply(v, v, m)
  eB <- 2 * ev
  if (n %% 2 == 0) {
    v <- B
    ev <- eB
  } else {
    v <- mMultiply(A, B, m)
    ev <- eA + eB
  }
  if (v[floor(m / 2) + 1, floor(m / 2) + 1] > 1e140) {
    v <- v * 1e-140
    ev <- ev + 140
  }
  return(list(v = v, ev = ev))
}

# Função principal K(n, d)
K <- function(n, d) {
  s <- d^2 * n
  # Aproximação rápida (linha omitida se for necessário alta precisão)
  if (s > 7.24 || (s > 3.76 && n > 99)) {
    return(1 - 2 * exp(-(2.000071 + 0.331 / sqrt(n) + 1.409 / n) * s))
  }
  k <- floor(n * d) + 1
  m <- 2 * k - 1
  h <- k - n * d
  H <- matrix(0, nrow = m, ncol = m)
  for (i in 1:m) {
    for (j in 1:m) {
      if ((i - j + 1) >= 0) {
        H[i, j] <- 1      }    }
  }
  for (i in 1:m) {
    H[i, 1] <- H[i, 1] - h^(i)
    H[m, i] <- H[m, i] - h^(m - i + 1)
  }
  if ((2 * h - 1) > 0) {
    H[m, 1] <- H[m, 1] + (2 * h - 1)^m
  }
  for (i in 1:m) {
    for (j in 1:m) {
      if ((i - j + 1) > 0) {
        for (g in 1:(i - j + 1)) {
          H[i, j] <- H[i, j] / g      }    }
    }
  }
  eH <- 0
  power_result <- mPower(H, eH, m, n)
  Q <- power_result$v
  eQ <- power_result$ev
  s <- Q[k, k]
  for (i in 1:n) {
    s <- s * i / n
  }
}

```

```

if (s < 1e-140) {
  s <- s * 1e140
  eQ <- eQ - 140      }  }
s <- s * 10^eQ
return(s)
}

```

que aplicada aos nossos dados

$$1-K(n, D)$$

$$1-K(5, 0.1771997)$$

gera
[1] 0.9894705

● Teste de Shapiro-Wilk

O teste de Shapiro-Wilk também é um dos testes mais amplamente recomendado para avaliar a normalidade, especialmente para tamanhos amostrais pequenos (< 50 observações). A verificação da suposição de normalidade é baseada na estatística W , que varia de 0 (zero) a 1 (um), sendo que um valor W próximo de 1 (um) indica que os dados ($n \geq 3$) são provenientes de uma distribuição Normal. Por outro lado, quanto mais próximo o valor W estiver de 0 (zero), menor será a probabilidade dos dados terem distribuição Normal. O teste de Shapiro-Wilk pode ser realizado no R utilizando a função `shapiro.test`. A estatística W é dada por:

$$W = \frac{\left[\sum_{i=1}^n a_i x_{(i)} \right]^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (\text{sw1})$$

onde n corresponde ao tamanho da amostra, x_i é um valor particular de x , $x_{(i)}$ indica um valor ordenado particular de x , \bar{x} é a média amostral de x e a_i é um coeficiente do teste de Shapiro-Wilk que representa a melhor estimativa linear do desvio padrão de x_i , assumindo normalidade. Os coeficientes a_i podem ser obtidos a partir de Tabelas, ou serem calculados através das expressões que serão apresentadas a seguir. De acordo com o `help` do R para o comando `shapiro.test`, é mencionado que a implementação segue o algoritmo de Royston, sendo indicados três trabalhos: (i) *An extension of Shapiro and Wilk's W test for normality to large samples* (1982), (ii) *Algorithm AS 181: The W test for Normality* (1982) e (iii) *Remark AS R94: A remark on Algorithm AS 181: The W test for normality*. Além desses também acrescento o (iv) *Approximating the Shapiro-Wilk W-test for non-normality. Statistics and computing*, 2(3), 117-119. Todos esses trabalhos são importantes para o delineamento e implementação do teste de Shapiro-Wilk que serão apresentados a seguir. Para calcular a estatística W usando o algoritmo de Royston, é considerada a aproximação assintótica dos coeficientes a_i pelas estatísticas c_i , que são definidas como

$$c_i = (\mathbf{m}^T \mathbf{m})^{-\frac{1}{2}} m_i, \quad (\text{sw2})$$

sendo que $\mathbf{m} = [m_i] = (m_1, m_2, \dots, m_n)$ corresponde aos valores esperados das estatísticas de ordem das variáveis aleatórias independentes e identicamente distribuídas x_i , amostradas de

uma distribuição Normal padrão. Temos que m_i é obtido a partir da fórmula aproximada de Blom (1958) para quantis como

$$m_i = \Phi^{-1} \left(\frac{i - 0.375}{n + 0.25} \right), \quad (\text{sw3})$$

sendo Φ a função de distribuição acumulada da Normal.

O desenvolvimento do procedimento é realizado a partir de aproximações para determinar os valores de a_i , de tal forma que

$$a_n = c_n + 0,221157y - 0,147981y^2 - 2,071190y^3 + 4,434685y^4 - 2,706056y^5 \quad (\text{sw4})$$

e

$$a_{n-1} = c_{n-1} + 0,042981y - 0,293762y^2 - 1,752461y^3 + 5,682633y^4 - 3,582663y^5, \quad (\text{sw5})$$

sendo $y = \frac{1}{\sqrt{n}}$. Na sequência calculamos a_i como

$$a_i = \frac{m_i}{\sqrt{\delta}}, \quad 2 < i < n-1 \quad (\text{sw6})$$

onde δ é obtido usando

$$\delta = \begin{cases} \frac{\mathbf{m}^T \mathbf{m} - 2m_n^2}{1 - a_n^2}, & \text{se } n \leq 5 \\ \frac{\mathbf{m}^T \mathbf{m} - 2m_n^2 - 2m_{n-1}^2}{1 - 2a_n^2 - 2a_{n-1}^2}, & \text{se } n > 5. \end{cases} \quad (\text{sw7})$$

Royston (1992) realizou a aproximação da função de distribuição da estatística W de Shapiro-Wilk como uma função da distribuição Normal padrão cumulativa usando o seguinte conjunto de transformações, que dependem do tamanho da amostra:

$$\phi_w(W) \approx \phi_z \left(\frac{w - \mu_w}{\sigma_w} \right), \quad (\text{sw8})$$

onde

$$w = \begin{cases} -\ln[-2,273 + 0,459n - \ln(1-W)], & 4 \leq n \leq 11 \\ \ln(1-W), & 12 \leq n \leq 2000 \end{cases} \quad (\text{sw9})$$

$$\begin{aligned} \mu_w = & \\ & \begin{cases} 0,5440 - 0,39978n + 0,025054n^2 - 0,0006714n^3, & 4 \leq n \leq 11 \\ -1,5861 - 0,31082x - 0,083751x^2 + 0,0038915x^3, & 12 \leq n \leq 2000 \end{cases} \\ \text{com } x = \ln(n) \end{aligned} \quad (\text{sw10})$$

e

$$\sigma_w = \begin{cases} e^{1,3822 - 0,77857n + 0,062767n^2 - 0,0020322n^3}, & 4 \leq n \leq 11 \\ e^{-0,4803 - 0,082676x + 0,0030302x^2}, & 12 \leq n \leq 2000 \end{cases} \quad (\text{sw11})$$

com $x = \ln(n)$.

O p-valor do teste de normalidade de Shapiro-Wilk usando o algoritmo de Royston é simplesmente determinado como:

$$p = 1 - \phi_Z \left(\frac{w - \mu_w}{\sigma_w} \right). \quad (\text{sw12})$$

O teste de Shapiro-Wilk é utilizado para verificar se um conjunto de dados segue uma distribuição Normal. A hipótese nula (H_0) deste teste assume que os dados são provenientes de uma população com distribuição Normal. Se o p-valor do teste for menor que o nível de significância (α) escolhido, rejeitamos H_0 e concluímos que os dados não são normalmente distribuídos. Caso contrário, se o p-valor for maior ou igual a α , não rejeitamos H_0 e não há evidências suficientes para afirmar que os dados não seguem uma distribuição Normal. Em outras palavras:

p-valor $< \alpha$: Os dados não têm distribuição Normal.

p-valor $\geq \alpha$: Não podemos afirmar que os dados não têm distribuição Normal.

É importante lembrar que um teste estatístico não prova que uma hipótese é verdadeira, mas sim fornece evidências para rejeitá-la ou não. Um p-valor alto não significa que os dados são definitivamente normais, mas sim que não encontramos evidências suficientes para concluir o contrário.

EXEMPLO 12.2 Modelo de Regressão Linear Simples e Teste de Normalidade de Shapiro-Wilk dos Resíduos

Considerando os mesmos dados do Exemplo 12.1, construímos o MRLS e realizamos o teste de normalidade de Shapiro-Wilk com algoritmo de Royston conforme é especificado na descrição da função `shapiro.test` do R e comparamos os resultados.

Primeira forma: usando os valores de a_i provenientes de uma Tabela.

```
x <- c(1,2,3,4,5); y <- c(1,1,2,2,4); dados <- data.frame(x, y)

rls <- lm(y ~ x, data = dados); rls

residuos_rls <- resid(rls); residuos_rls

residuos_ordenados <- sort(residuos_rls)
n <- length(residuos_ordenados)
z_residuos_ordenados <- scale(residuos_ordenados)
shapiro.test(z_residuos_ordenados)

soma_dif_quad <- sum((z_residuos_ordenados - mean(z_residuos_ordenados))^2)

a_i <- c(0.6646, 0.2413, 0, -0.2413, -0.6646)
soma_ai_xi <- sum(a_i*z_residuos_ordenados)

w <- (soma_ai_xi^2)/soma_dif_quad; w
```

Nessa sequência de comandos anterior, note que os coeficientes a_i foram atribuídos a partir de uma Tabela conforme mostrado na figura a seguir.

| <u>n</u> | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 1 | 0.7071 | 0.7071 | 0.6872 | 0.6646 | 0.6431 | 0.6233 | 0.6052 | 0.5888 | 0.5739 |
| 2 | - | 0.0000 | 0.1677 | 0.2413 | 0.2806 | 0.3031 | 0.3164 | 0.3244 | 0.3291 |
| 3 | - | - | - | 0.0000 | 0.0875 | 0.1401 | 0.1743 | 0.1976 | 0.2141 |
| 4 | - | - | - | - | - | 0.0000 | 0.0561 | 0.0947 | 0.1224 |
| 5 | - | - | - | - | - | - | - | 0.0000 | 0.0399 |

Figura-4 Coeficientes a_i para o teste de Shapiro-Wilk para normalidade. (Fonte: Shapiro e Wilk, 1965)

Usando o comando `shapiro.test` o resultado é:
Shapiro-wilk normality test

```
data: z_residuos_ordenados
W = 0.97005, p-value = 0.8756
```

Fazendo os cálculos a partir da expressão para a estatística W temos:

```
W
[1] 0.9698743
```

Ou seja, houve uma diferença por questões de arredondamentos das funções empregadas, mas os valores estão próximos.

Segunda forma: usando o algoritmo de Royston.

```
x <- c(1,2,3,4,5); y <- c(1,1,2,2,4); dados <- data.frame(x, y)
rls <- lm(y ~ x, data = dados); rls
residuos_rls <- resid(rls); residuos_rls
residuos_ordenados <- sort(residuos_rls)
n <- length(residuos_ordenados)
z_residuos_ordenados <- scale(residuos_ordenados)
shapiro.test(z_residuos_ordenados)

# Passo-1 [Soma das diferenças quadradas]
soma_dif_quad <- sum((z_residuos_ordenados - mean(z_residuos_ordenados))^2)

# Passo-2 [Valores esperados das estatísticas de ordem]
m <- qnorm((1:n - 0.375) / (n + 0.25)) # Fórmula aproximada de Blom para quantis

# Passo-3 [Cálculo dos coeficientes ci's]
c_i <- ((m %*% m)^(-0.5)) %*% m

# Passo-4 [Cálculo do y]
y <- 1/sqrt(n)

# Passo-5 [Cálculo dos coeficientes an = a5 e a1]
a_5 <- c_i[5] + 0.221157*y - 0.147981*(y^2) - 2.071190*(y^3) +
4.434685*(y^4) - 2.706056*(y^5)
a_1 <- -a_5

# Passo-6 [Cálculo dos coeficientes an-1 = a4 e a2]
a_4 <- c_i[4] + 0.042981*y - 0.293762*(y^2) - 1.752461*(y^3) +
5.682633*(y^4) - 3.582663*(y^5)
a_2 <- -a_4
```

```

# Passo-7 [Cálculo do delta]
delta_w <- ((m%%m) - 2*m[n])/(1 - (a_5)^2)

# Passo-8 [Cálculo dos coeficiente a3 (2 < i < n - 1)]
a_3 <- m[3]/(delta_w^0.5)

a_i <- c(a_1, a_2, a_3, a_4, a_5)
soma_ai_xi <- sum(a_i*z_residuos_ordenados)

w <- (soma_ai_xi^2)/soma_dif_quad; w

# Passo-9 [Cálculo da estatística w]
w <- -log((0.459*n) - 2.273 - log(1 - w))

# Passo-10 [Cálculo de mi]
mi_w <- 0.544 - 0.39978*n + 0.025054*(n^2) - 0.0006714*(n^3)

# Passo-11 [Cálculo do sigma]
sigma_w <- exp(1.3822 - 0.77857*n + 0.062767*(n^2) - 0.0020322*(n^3))

# Passo-12 [Estatística phi_w]
phi_w = (w - mi_w)/sigma_w

# Passo-13 [Determinação do p-valor]
p_valor_sw <- 1 - pnorm(phi_w)

```

Os resultados dessa segunda sequência de comandos são:

```

w
[1] 0.9703044
p_valor_sw
[1] 0.8772036

```

Também observamos diferenças, devido a arredondamentos das funções utilizadas no ambiente do software R. Mas, novamente os valores estão iguais até a segunda casa decimal.

O pressuposto de normalidade dos resíduos é considerado uma suposição fraca. Sendo que alguns autores (Wooldridge, 2020; Greene, 2018; Neter *et al.*, 2013; Fox, 2015) destacam que para amostras grandes, a maioria das inferências permanece válida devido à utilização de métodos estatísticos assintóticos na construção dos modelos.

■ Homocedasticidade dos resíduos

Em Estatística, uma sequência ou vetor de variáveis aleatórias é classificado como heterocedástico quando as variâncias dessas variáveis não são constantes ao longo das observações, ou seja, variam de acordo com algum fator ou índice (Gujarati e Porter, 2009; Wooldridge, 2020; Yan e Su, 2009).

O pressuposto da homocedasticidade implica que todos os erros aleatórios (resíduos) gerados pelo modelo, tenham a mesma variância constante. Ou seja, considerando o MRLS dado por $y = \beta_0 + \beta_1 x + \varepsilon$, a hipótese nula associada a suposição de variância constante é escrita como

$$H_0 : \text{Var}[\varepsilon|x] = \sigma^2. \quad (\text{bp1})$$

Uma vez que também é assumido que ε tem valor esperado condicional igual a zero, temos que $Var[\varepsilon|x] = E[\varepsilon^2|x] - E^2[\varepsilon|x] = E[\varepsilon^2|x]$, o que indica que a média dos erros quadrados não deve variar com os valores de x e isso implica que a hipótese nula de homocedasticidade pode ser expressa de forma equivalente como

$$H_0 : E[\varepsilon^2|x] = E[\varepsilon^2] = \sigma^2. \quad (\text{bp2})$$

A heterocedasticidade é a violação desse pressuposto.

Novamente a presença de outliers pode ser um fonte de heterocedasticidade dos resíduos, além do que mudanças na magnitude da variável dependente devido a certos valores da variável dependente (ou variáveis dependentes, no caso do Modelo de Regressão Linear Múltipla - MRLM) podem influenciar a variabilidade dos resíduos. Por isso, é necessário verificar inicialmente os dados observados para avaliar as amplitudes dos valores que podem causar diferenças que levem a situação de heterocedasticidade. Porque se isso acontece, os estimadores deixam de ser os melhores estimadores lineares não viesados (BLUE, como será visto logo mais no MRLM), gerando previsões que serão ineficientes (Yan e Su, 2009).

Aqui, o diagnóstico da homocedasticidade será realizado utilizando o teste de Breusch-Pagan (1979).

O teste de Breusch-Pagan verifica a heterocedasticidade, comparando a variância dos resíduos de um modelo de regressão aos valores preditos do modelo. O teste calcula uma estatística Qui-quadrado com base nessa comparação, e o p-valor resultante é usado para determinar se há ou não evidência de heterocedasticidade.

Para fazer esse teste no R é muito simples, basta utilizar a função `bptest` que está na library chamada `lmtest`. De acordo com o `help` do R, a implementação desse função usa como referências os trabalhos: (i) T.S. Breusch & A.R. Pagan (1979), *A Simple Test for Heteroscedasticity and Random Coefficient Variation*. *Econometrica* 47, 1287–1294; (ii) R. Koenker (1981), *A Note on Studentizing a Test for Heteroscedasticity*. *Journal of Econometrics* 17, 107–112; e (iii) W. Krämer & H. Sonnberger (1986), *The Linear Regression Model under Test*. Heidelberg: Physica.

No teste de Breusch e Pagan (1979), é realizada uma regressão dos resíduos quadrados em todas as variáveis explicativas e testado se essa regressão tem poder explicativo, ou seja, estatisticamente significante. Estamos a princípio considerando somente uma variável independente e essa regressão é simplesmente denotada como

$$\hat{\varepsilon}^2 = \alpha_0 + \alpha_1 x + \nu, \quad (\text{bp3})$$

sendo ν um termo de erro com média zero dado o x . Como veremos no MRLM teremos mais de uma variável explanatória e a regressão para o ε^2 poderá ser escrita como

$$\hat{\varepsilon}^2 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_k x_k + \nu. \quad (\text{bp4})$$

Testar a homocedasticidade corresponde verificar a hipótese

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_k = 0. \quad (\text{bp2})$$

A estatística de teste utilizada é dada por

$$LM = nR_{\hat{\varepsilon}^2}^2. \quad (\text{bp2})$$

Essa estatística de teste é assintoticamente distribuída com χ_k^2 sob a hipótese nula de homocedasticidade.

O nome LM para a estatística de teste de Breusch e Pagan é devido a abordagem ser baseada no teste do multiplicador de Lagrange (LM) de Aitchinson e Silvey (1960), também conhecido como teste de pontuação eficiente de Rao (1973a, 1973b).

Em termos de implementação computacional esse teste pode ser constituído das seguintes etapas:

- faça a regressão de y contra x ;
- determine os resíduos;
- construa um vetor dos resíduos ao quadrado;
- faça uma nova regressão, em que a variável dependente constitui-se do vetor de resíduos ao quadrado [do item anterior] e a variável independente continua sendo x ;
- dessa nova regressão extraia o R^2 ;
- a estatística do teste BP é o produto de n (tamanho da amostra) pelo R^2 ;
- determine o p-valor considerando que o produto $n.R^2$ tem distribuição Qui-quadrado com n graus de liberdade.

EXEMPLO 12.3 Modelo de Regressão Linear Simples e Teste de Heterocedasticidade de Breusch-Pagan

Ainda levando em conta os dados do Exemplo-12.1, realizar o Teste de Breusch-Pagan para os resíduos do MRLS.

```
x <- c(1,2,3,4,5); y <- c(1,1,2,2,4); dados <- data.frame(x, y)
rls <- lm(y ~ x, data = dados); rls
residuos_rls <- resid(rls); residuos_rls

# Heterocedasticidade dos Resíduos [Breusch-Pagan]
# H0: Resíduos são Homocedásticos [Ou seja, os resíduos são distribuídos
# com igual varânci]

library(lmtest)
bptest(rls) # bptest está na library "lmtest"
```

O resultado desse comando é
studentized Breusch-Pagan test

```
data: rls
BP = 1.9827, df = 1, p-value = 0.1591
```

E agora conduzindo o teste conforme as etapas descritas previamente, temos a sequência de comandos mostrada a seguir.

```
n <- length(residuos_rls)
residuos_quad <- residuos_rls^2

dados_bp <- data.frame(x, residuos_quad)
bp_rls <- lm(residuos_quad ~ x, data = dados_bp); bp_rls
summary(bp_rls)

r_quadrado_bp <- summary(bp_rls)$r.squared
r_quadrado_bp
```

```

estat_bp <- n*r_quadrado_bp
estat_bp

p_value_bp <- 1 - pchisq(estat_bp,1)
#p_value_bp <- pchisq(estat_bp,1, lower.tail = F)
p_value_bp

```

O que gera
 estat_bp
 [1] **1.982652**

e
 p_value_bp
 [1] **0.1591113**

Por esses resultados ($BP \approx 1,983$ e $p\text{-valor} \approx 0,159$), considerando um nível de significância de 0,05, não rejeitamos a hipótese de que os resíduos são homocedásticos.

Quadro Conceitual - 2

Com relação as palavras homocedástico e heterocedástico. O significado do termo “cedástico”, no que diz respeito à estatística, pode ser obtido do Dicionário Brasileiro de Estatística [Seguido de um Vocabulário Inglês-Português, 2a. Edição Revista e Aumentada], da Fundação IBGE [Instituto Brasileiro de Estatística], publicado no Rio de Janeiro em 1970. Tendo como autor, Milton da Silva Rodrigues [Professor Emérito de Estatística da Universidade de São Paulo - USP]. Na página 27 desse Dicionário [disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv81219.pdf>] temos a definição do termo CEDASTICIA.

CEDASTICIA – É a propriedade que têm as distribuições condicionadas de uma d.f. a dois atributos de apresentarem dispersões iguais ou diferentes. O termo e o conceito foram dados por K. PEARSON, em *On the general theory of skew correlation...*, in DCRM, 1905.

Figura-ZZ Descrição de Cedasticia no Dicionário do IBGE de 1970.

Na língua inglesa, temos dois trabalhos seminais que tratam do conceito de heterocedasticidade. O primeiro, foi publicado em 1982 pelo Engle [Prêmio Nobel de Economia de 2003] com título: *Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation* [*Econometrica*, 50, pp. 987–1008]. O outro é de 1986 e foi publicado pelo Bollerslev: *Generalized Autoregressive Conditional Heteroskedasticity* [*Journal of Econometrics*, 31, pp. 307–327].

Note como o termo heterocedástico é escrito nos dois trabalhos.

Em Março de 1985, McCulloch publicou na *Econometrica* [Vol. 53, No. 2] um breve trabalho intitulado: *Miscellanea on Heteros*edasticity*. Onde é descrito que: “*Nossa palavra é uma cunhagem moderna, derivada das duas raízes gregas hetero-(έτερο-), que significa "outro" ou "diferente", e skedannumi (σκεδάννυμι), que significa "espalhar". A letra em questão é, portanto, a transliteração do grego kappa (κ). Em palavras científicas que os estudiosos retiraram diretamente do grego para o inglês, a letra kappa é sempre transliterada como k. Exemplos são skeptic (σκεπτικός) e skeleton (σκελετός).*”

Depois de algumas considerações, McCulloch conclui que:

Heteroskedasticity é, portanto, a grafia correta em inglês.

■ Ausência de outliers

Precisa estar muito claro que quando construímos um modelo a partir da técnica de regressão, não basta simplesmente estimar o modelo, para que a análise esteja completa é necessário realizar o estudo dos resíduos e verificar se os pressupostos da análise de regressão são todos satisfeitos. Essa análise dos resíduos também possibilita a identificação de valores discrepantes (outliers) que estão presentes em praticamente todas as pesquisas que são realizadas utilizando dados reais. Outliers em modelos de regressão linear podem gerar diversos problemas que prejudicam tanto a precisão quanto a interpretabilidade do modelo. A presença de outliers em modelos de regressão linear pode acarretar diversos problemas, afetando tanto a precisão quanto a interpretabilidade do modelo. Os principais problemas a serem destacados são:

- (i) O método de minimização de erros quadrados utilizado no modelo de regressão é afetado devido a presença de valores discrepantes, uma vez que esses valores geram maiores diferenças e maior peso na soma realizada. Assim, os coeficientes estimados podem apresentar tendenciosidade, e esse viés pode influenciar de forma desproporcional a estimação da reta de regressão, distorcendo os coeficientes estimados.
- (ii) Se os outliers não forem representativos do padrão geral dos dados, eles podem levar a previsões inadequadas para os casos regulares, reduzindo a precisão das previsões.
- (iii) Outliers podem aumentar a variabilidade do modelo, resultando em erros-padrão maiores. Isso pode diminuir a significância estatística de variáveis preditoras importantes.
- (iv) Outliers podem inflar medidas de erro, distorcendo métricas de avaliação:, como o erro médio absoluto (MAE) e o erro quadrático médio (MSE), dando a impressão de que o modelo tem desempenho ruim.
- (v) Em casos extremos, outliers podem influenciar a relação entre variáveis preditoras, exacerbando problemas de multicolinearidade e tornando os coeficientes suscetíveis a variações.
- (vi) Os outliers podem violar pressupostos fundamentais da regressão linear, como: (a) linearidade entre as variáveis preditoras e a variável resposta; (b) homocedasticidade (variância constante dos erros); (c) normalidade dos resíduos e (d) independência dos erros.
- (vii) Outliers podem se tornar pontos influentes, ou seja, pontos que afetam consideravelmente o ajuste do modelo. Esses pontos podem levar ao aumento da sensibilidade do modelo a se ajustar a padrões errôneos.

Como a identificação visual é difícil em conjuntos de dados que contêm grande número de observações, então é fundamental que sejam utilizadas medidas numéricas para realizar esta tarefa. Existem várias medidas utilizadas para identificação de outliers, as mais comuns, também chamadas de medidas clássicas são: Leverage, DFFit, DFBeta, Distância de Cook, Distância de Mahalanobis e Covratio.

Vamos explorar o uso da distância de Cook.

● *Distância de Cook*

A Distância de Cook também é considerada um método clássico para identificar outliers e é definida, no caso geral do MRLM (em que temos p variáveis explicativas), como

$$CD_i = \frac{\sum_{j=1}^n [\hat{y}_j - \hat{y}_{j(i)}]^2}{p\hat{\sigma}^2}, \quad (\text{cd1})$$

onde

\hat{y}_j é o valor da j -ésima resposta ajustada quando todas as observações são incluídas;

$\hat{y}_{j(i)}$ é o valor da j -ésima resposta ajustada, onde o ajuste não inclui a observação i ;

p é o número de coeficientes estimados no modelo de regressão;

$\hat{\sigma}^2$ é a variância estimada (MSE), calculada como

$$\hat{\sigma}^2 = \frac{SSE}{n - p}. \quad (\text{cd2})$$

Como no caso do MRLS estimamos dois coeficientes (β_0 e β_1), então temos p igual a 2.

Dessa forma, a Distância de Cook para a i -ésima observação é baseada nas diferenças entre as respostas preditas a partir do modelo construído considerando todos os dados e as respostas preditas a partir do modelo desconsiderando a i -ésima observação. Por isso escrevemos $\hat{y}_{j(i)}$, o que significa que estimamos y deixando de lado a observação i . O menor valor que a Distância de Cook pode assumir é zero. No entanto, com relação ao valor da Distância de Cook que caracteriza uma observação como valor discrepante (outlier), quando comparada com as outras observações da amostra, não há regras bem definidas que estabeleçam esses limiares (thresholds) ou pontos de corte (cut offs). O que comumente é aceito, é que distâncias de Cook maiores que 1 (um), caracterizam observações que são outliers. Sendo também consideradas as opções: (i) $4/(n - k - 1)$, onde n é o tamanho da amostra e k é o número de variáveis independentes e (ii) $4/n$ também é bastante utilizada, inclusive sendo aplicada nos comandos `ols_plot_resid_lev` e `ols_plot_cooksd_chart` da library `olsrr` do R.

No exemplo a seguir são calculadas as Distâncias de Cook utilizando as expressões apresentadas e também usando comandos do pacote `olsrr` do R.

EXEMPLO 12.4 Modelo de Regressão Linear Simples e Identificação de Outliers utilizando Distância de Cook

Mantendo nossas variáveis dependente e independente iguais ao que foi apresentado no Exemplo-12.1, constituindo uma amostra com cinco valores para y e cinco valores para x , o que torna mais fácil realizar as contas e comparar com os resultados gerados pelos comandos do R, vamos calcular as Distâncias de Cook para cada uma das observações. Primeiro usando comandos da library `olsrr` do R, depois usando as expressões necessárias.

Vale comentar que os comandos do R gerados para aplicar as fórmulas da Distância de Cook estão escritas para cada observação, para deixar bem claro o cálculo do $\hat{y}_{j(i)}$ que é o valor da j -ésima resposta ajustada, onde o ajuste não inclui a observação i .

Primeira forma: usando a library `olsrr`.

```
x <- c(1,2,3,4,5); y <- c(1,1,2,2,4); dados <- data.frame(x, y)
```

```
r1s <- lm(y ~ x, data = dados); r1s
y_hat <- fitted(r1s)
residuos_r1s <- resid(r1s); residuos_r1s
sum_residuos_r1s_quad <- sum(residuos_r1s^2)
```

```
library(car)
```

```

residualPlots(rls)

# Verificação de outliers
qqPlot(rls,id.n=2)

# Plota a Leverage (Alavancagem) e outliers
library(olsrr)
ols_plot_resid_lev(rls)

# Plota a Distância de Cook
ols_plot_cooksd_chart(rls)

```

Na Figura-C1 temos o gráfico gerado pelo comando `ols_plot_resid_lev`. Corresponde a um gráfico de resíduos vs alavancagem, usado para identificar casos influentes, ou seja, valores extremos que podem influenciar os resultados da regressão quando incluídos ou excluídos da análise. Essa alavancagem indica se os valores de X para a i -ésima observação são ou não discrepantes, porque pode ser demonstrado que a alavancagem é uma medida da distância entre os valores de X para a i -ésima observação e as médias dos valores de X para todas as n observações (Veja o Neter, 1983, na página 402).

Na Figura-C1 (b) destacamos seis áreas que são usadas para o diagnóstico de outliers. Na área 3 é onde encontramos a maior parte dos valores observados, na área 4 temos os pontos potenciais que podem influenciar na estimação do MRLS, nas áreas 1 e 5 temos os outliers e nas áreas 2 e 6 é onde estariam os pontos mais influentes e problemáticos para a construção do modelo. Atente que a linha vertical está indicando o leverage threshold de 0,8 e o outlier threshold igual a 2 que é o default do comando `ols_plot_resid_lev`.

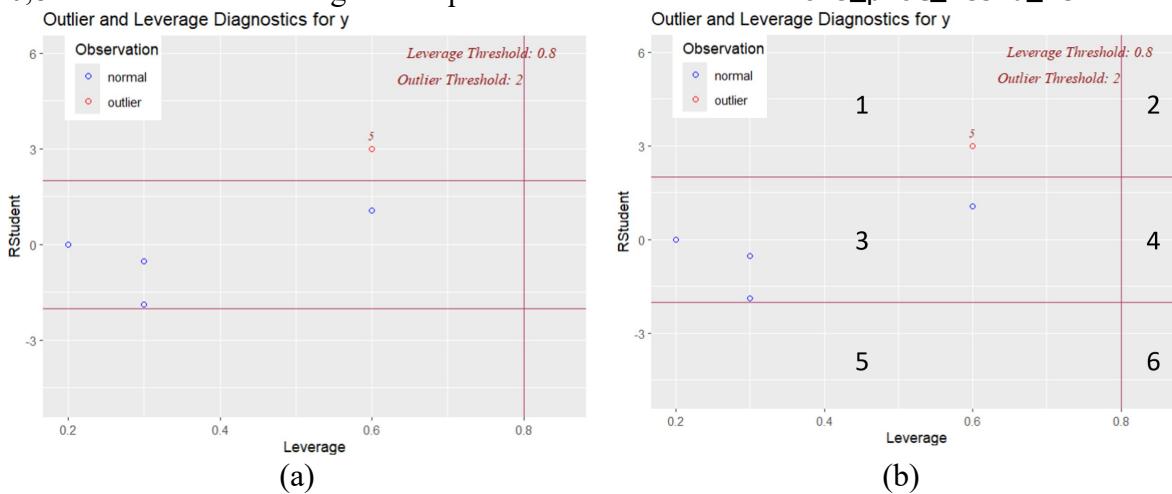


Figura-C1 Gráfico de diagnóstico de outliers.

De acordo com o `help` do R para o comando `ols_plot_cooksd_chart`, temos as informações apresentadas a seguir.

A distância de Cook foi introduzida pelo estatístico americano Ralph Dennis Cook em 1977. Ela é usada para identificar pontos de dados influentes. Depende tanto do resíduo quanto da alavancagem (leverage), ou seja, leva em conta tanto o valor x quanto o valor y da observação.

Etapas para calcular a distância de Cook: (i) exclua as observações uma de cada vez; (ii) reajuste o modelo de regressão nas observações restantes e (iii) examine o quanto todos os valores ajustados mudam quando a i -ésima observação é excluída.

Uma observação com um grande CD de Cook indica que o ponto influencia fortemente os

valores ajustados. Existem vários métodos/fórmulas para calcular o limite usado para detectar ou classificar observações como outliers e os listamos abaixo.

Tipo 1: $4 / n$

Tipo 2: $4 / (n - k - 1)$

Tipo 3: ~ 1

Tipo 4: $1 / (n - k - 1)$

Tipo 5: 3 * média (vetor de valores de distância de Cook)

onde n e k representam

n : Número de observações;

k : Número de preditores.

Utilizando o comando `ols_plot_cooksd_chart` podemos extrair os valores de Distâncias de Cook para cada ponto/observação escrivendo `ols_plot_cooksd_chart[1]`. De tal forma que obtemos como resultados:

```
$data
  obs      cd  color fct_color txt
1   1 8.181818e-01 outlier  outlier   1
2   2 7.513915e-02 normal   normal    NA
3   3 1.313134e-33 normal   normal    NA
4   4 4.090909e-01 normal   normal    NA
5   5 1.840909e+00 outlier  outlier   5
```

Onde cd é a Distância de Cook e na coluna `fct_color` temos a indicação se o ponto é caracterizado como outlier. A Figura-C2 apresenta um gráfico construído a partir das Distâncias de Cook calculadas. Observe que a linha horizontal está indicando o threshold de $0,8 (= 4/n = 4/5)$. Sendo que as observações 1 e 5 foram as que excederam DC igual a 0,8.

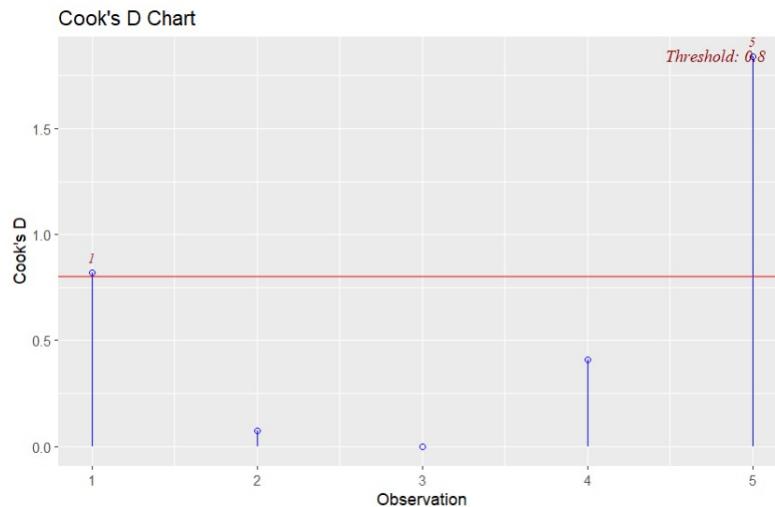


Figura-C2 Gráfico com Distâncias de Cook.

Os comandos a seguir implementam as fórmulas para determinação da Distância de Cook.

Segunda forma: implementando o cálculo da Distância de Cook.

```
# Distâncias de Cook
n <- length(residuos_rls)
p <- length(rls$coefficients)
```

```

sigma_hat <- sum_residuos_rls_quad/(n - p)
new_data <- c(1,2,3,4,5)

# DC_1
x <- c(2,3,4,5); y <- c(1,2,2,4); rls_DC_1 <- lm(y ~ x); #rls_DC_1
y_hat_DC_1 <- predict(rls_DC_1, newdata = data.frame(x = new_data))
DC_1 <- (1/(p*sigma_hat))*(sum((y_hat - y_hat_DC_1)^2)); DC_1

# DC_2
x <- c(1,3,4,5); y <- c(1,2,2,4); rls_DC_2 <- lm(y ~ x); #rls_DC_2
y_hat_DC_2 <- predict(rls_DC_2, newdata = data.frame(x = new_data))
DC_2 <- (1/(p*sigma_hat))*(sum((y_hat - y_hat_DC_2)^2)); DC_2

# DC_3
x <- c(1,2,4,5); y <- c(1,1,2,4); rls_DC_3 <- lm(y ~ x); #rls_DC_3
y_hat_DC_3 <- predict(rls_DC_3, newdata = data.frame(x = new_data))
DC_3 <- (1/(p*sigma_hat))*(sum((y_hat - y_hat_DC_3)^2)); DC_3

# DC_4
x <- c(1,2,3,5); y <- c(1,1,2,4); rls_DC_4 <- lm(y ~ x); #rls_DC_4
y_hat_DC_4 <- predict(rls_DC_4, newdata = data.frame(x = new_data))
DC_4 <- (1/(p*sigma_hat))*(sum((y_hat - y_hat_DC_4)^2)); DC_4

# DC_5
x <- c(1,2,3,4); y <- c(1,1,2,2); rls_DC_5 <- lm(y ~ x); #rls_DC_5
y_hat_DC_5 <- predict(rls_DC_5, newdata = data.frame(x = new_data))
DC_5 <- (1/(p*sigma_hat))*(sum((y_hat - y_hat_DC_5)^2)); DC_5

DC_1; DC_2; DC_3; DC_4; DC_5

```

O que gera

```

[1] 0.8181818
[1] 0.07513915
[1] 1.69762e-30
[1] 0.4090909
[1] 1.840909

```

Nessa implementação temos os mesmos resultados, ou seja, as observações 1 e 5 foram as que tiveram DC superior a 0,8.

Note que a DC para a observação 3 é bastante pequena, podendo ser considerada zero. A diferença de valores printados na Primeira forma (1.313134e-33) e na Segunda forma (1.69762e-30), provavelmente acontecem devido a arredondamentos e ponto flutuante distintos empregados nas duas fórmulas.

■ Linearidade nos parâmetros e nas variáveis

Nos modelos de regressão linear que estamos tratando, o pressuposto de linearidade estabelece *a priori* que a relação funcional entre a variável dependente (y) e a variável independente (x) é linear. Apesar de podermos considerar os casos de linearidade em termos das variáveis e dos parâmetros do modelo, aqui o objetivo é construir modelos de regressão que sejam lineares nos parâmetros. Conforme destaca Gujarati (2011): “... a expressão ‘linear’ significará sempre uma regressão linear nos parâmetros: os β ’s (isto é, os parâmetros são elevados apenas à primeira potência).”

Como mencionado no início do capítulo, estamos tratando do MRLS que é linear tanto nos parâmetros, como na variável.

Uma forma bastante empregada para verificar a linearidade do MRLS é construir um gráfico de dispersão entre os valores preditos pelo modelo e os resíduos gerados. A seguir é

adicionada uma linha ajustada usando uma função LOESS (Locally Estimated Scatterplot Smoothing, sendo no R designada como Local Polynomial Regression Fitting), ou seja, de forma geral essa função ajusta uma superfície polinomial local determinada por um ou mais preditores numéricos, usando ajuste local. No nosso caso de MRLS, essa linha ajustada permite visualizar a relação entre as variáveis de interesse. Sendo que, se a linha LOESS oscilar em torno da linha horizontal de valor zero para os resíduos, podemos assumir o relacionamento linear entre as variáveis.

Do ponto de vista da aplicação prática, isso é muito fácil de realizar usando os recursos do R. A seguir, na Figura-L1 são apresentados dois gráficos, o primeiro, Figura-L1 (a), é construído com a função `plot` do R e usamos a opção `which = 1` que constrói o gráfico de resíduos versus valores ajustados. No segundo é usada a função `geom_smooth` com as opções `method = "lm"` e `method = "loess"`.

EXEMPLO 12.5 Modelo de Regressão Linear Simples e Diagnóstico de Linearidade

Continuando com os dados apresentados no Exemplo-12.1,vamos verificar a linearidade no modelo.

```
x <- c(1,2,3,4,5); y <- c(1,1,2,2,4); dados <- data.frame(x, y)

rls <- lm(y ~ x, data = dados); rls
y_hat <- fitted(rls)
residuos_rls <- resid(rls); residuos_rls

# Linearidade
y_hat <- rls$fitted.values # valores ajustados

plot(rls, which = 1) # Verificamos se os erros se distribuem em torno
# da linha zero

# Também pode ser verificado usando o ggplot
# Verificamos se a linha de LOESS está em torno da linha zero
library(ggplot2)
ggplot(mapping = aes(x = y_hat, y = residuos_rls)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color="red") +
  geom_smooth(method = "loess", se = FALSE)
```

Para atender ao pressuposto de linearidade, idealmente, o gráfico de resíduos versus valores ajustados não pode apresentar nenhum padrão funcional, por exemplo, ter a forma de uma parábola. Ou seja, a linha ajustada pela LOESS deve ser aproximadamente horizontal em zero. Caso seja observado algum padrão, isso implica em problema de especificação do modelo. Conforme mostrado na Figura-C1, apesar da oscilação dos resíduos em torno do zero, não podemos descartar que uma função linear (nas variáveis) pode ser empregada para explicar y em termos de x .

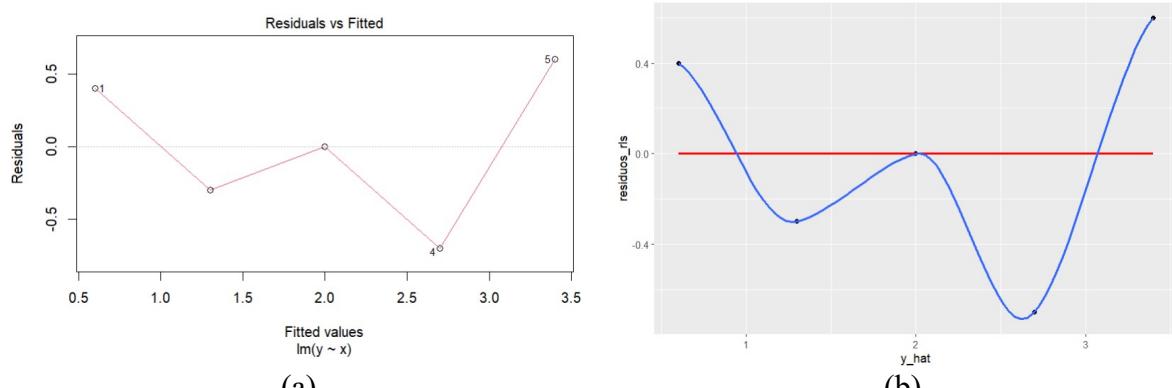


Figura-C1 Diagnóstico de linearidade: (a) resíduos vs valores ajustados e (b) usando LOESS

■ Ausência de autocorrelação serial nos resíduos

Um outro pressuposto da construção de modelos de regressão está relacionado à correlação entre os resíduos, também denominada de autocorrelação, que constitui-se de um fenômeno estatístico, em que os termos de erro de um modelo de regressão são correlacionados entre si. O problema associado a essa correlação é que ela pode levar à violação da suposição de independência dos termos de erro, que é fundamental para a análise de regressão de mínimos quadrados ordinários (MQO). Isso pode acarretar a obtenção de estimativas ineficientes que irão distorcer os erros gerados pelo modelo, levando a testes de hipóteses não confiáveis.

Um dos testes mais amplamente utilizado para verificar autocorrelação dos resíduos é o teste de Durbin-Watson. Porém, quando consultamos as tabelas que fornecem os valores necessários para proceder a determinação da estatística do teste de Durbin-Watson, verificamos que são construídas iniciando com tamanho de amostra igual a seis (muitas vezes as tabelas consideram tamanhos de amostra no intervalo de 6 a 2000). A alternativa, para verificar a autocorrelação dos resíduos para amostras pequenas, é utilizar o teste de Breusch-Godfrey. Sendo assim, a seguir estes dois testes de autocorrelação são apresentados brevemente.

● *Teste de Durbin-Watson*

Uma das pressuposições da construção do modelo de regressão é que a correlação entre os resíduos gerados a partir das variáveis independentes (v. i.) seja igual a zero. O principal problema da autocorrelação é que os estimadores obtidos pelo método dos mínimos quadrados deixam de ser eficientes, ou seja, não apresentam variância mínima. Para verificar esta suposição utilizamos a estatística de Durbin-Watson em que as hipóteses testadas são:

$$\begin{aligned} H_0 &: \rho = 0 \\ H_a &: \rho \neq 0. \end{aligned} \tag{DW1}$$

A estatística de teste é dada por:

$$d = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=2}^n e_i^2}. \quad (\text{DW2})$$

Onde $e_i = y_i - \hat{y}_i$ e y_i e \hat{y}_i são, respectivamente, os valores observados e preditos para a variável resposta (v. d.) para um caso, ou indivíduo i . A medida d torna-se tanto menor quanto maior a correlação serial. Limites inferiores e superiores da estatística d , comumente denotados por d_U e d_L , respectivamente, são tabulados para diferentes quantidades de variáveis explanatórias n . Assim, dados d , d_U e d_L , temos as seguintes regras de decisão para o teste de autocorrelação:

- Se $d < d_L$, rejeitar $H_0 : \rho = 0$.
- Se $d_L \leq d \leq d_U$, o teste é inconclusivo.
- Se $4 - d_L < d < 4$, rejeitar $H_0 : \rho = 0$.
- Se $4 - d_U \leq d \leq 4 - d_L$, o teste é inconclusivo.
- Se $d_U < d < 4 - d_U$, não rejeitar $H_0 : \rho = 0$.

Sendo ρ denotando a autocorrelação populacional.

De maneira geral, estas regras de decisão são apresentadas na forma de um esquema como o mostrado a seguir.

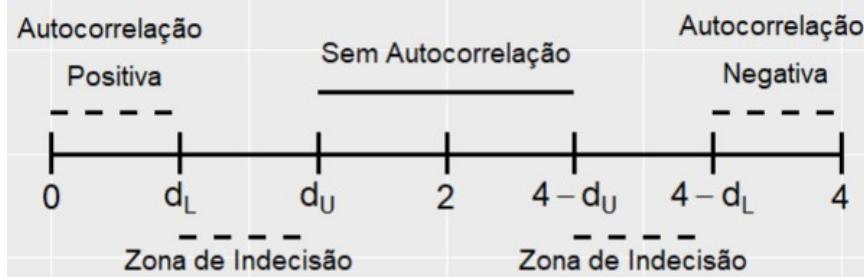


Figura-DW1 Análise da estatística d de Durbin-Watson .

O R fornece a estatística de Durbin-Watson, por exemplo, usando o comando `dwtest`. Porém, a conclusão do teste precisa dos valores de d_U e d_L que são tabelados.

| Valores Críticos para a Estatística de Durbin-Watson (d) | | | | | | | | | | |
|--|---------|-------|---------|-------|---------|-------|---------|-------|---------|-------|
| Nível de Significância $\alpha = 0,05$ | | | | | | | | | | |
| n | $k = 1$ | | $k = 2$ | | $k = 3$ | | $k = 4$ | | $k = 5$ | |
| | d_L | d_U |
| 6 | 0.61 | 1.40 | | | | | | | | |
| 7 | 0.70 | 1.36 | 0.47 | 1.90 | | | | | | |
| 8 | 0.76 | 1.33 | 0.56 | 1.78 | 0.37 | 2.29 | | | | |
| 9 | 0.82 | 1.32 | 0.63 | 1.70 | 0.46 | 2.13 | 0.30 | 2.59 | | |
| 10 | 0.88 | 1.32 | 0.70 | 1.64 | 0.53 | 2.02 | 0.38 | 2.41 | 0.24 | 2.82 |

Figura-DW2 Parte da Tabela de Durbin-Watson para $\alpha = 0,05$.

● *Teste de Breusch-Godfrey*

Vamos utilizar o teste de Breusch-Godfrey para avaliar autocorrelação serial dos resíduos da regressão. Esse teste é desenvolvido nos trabalhos de Breusch (1978) e Godfrey (1978a, 1978b). Para realizar esse teste no R pode ser utilizado o comando `bgtest` que está na library `lmtest`.

Considerando nosso MRLS dado por $y = \beta_0 + \beta_1 x + \varepsilon$, supomos que o termo de erro ε pode ser expresso como um processo autoregressivo de ordem p descrito como

$$\varepsilon_i = \rho_1 \varepsilon_{i-1} + \rho_2 \varepsilon_{i-2} + \dots + \rho_p \varepsilon_{i-p} + \nu_i, \quad (\text{BP1})$$

onde ν_i é um termo de erro com média zero e variância finita constante. A hipótese nula do teste de correlação serial assume que não há autocorrelação dos resíduos até a ordem p . A rejeição dessa hipótese indica a presença de correlação serial, o que pode violar uma das premissas do modelo de regressão linear e exigir ajustes (Gujarati e Porter, 2009; Box *et al.*, 2015; Wooldridge, 2020). A hipótese nula é escrita como

$$H_0 : \rho_1 = \rho_2 = \dots = \rho_p = 0. \quad (\text{BP2})$$

Para a obtenção da estatística de teste realizamos uma regressão auxiliar dos resíduos contra as variáveis independentes (no MRLS, temos somente x) e os resíduos defasados, ou seja

$$\hat{\varepsilon}_i = \alpha_0 + \alpha_1 x_i + \rho_1 \hat{\varepsilon}_{i-1} + \rho_2 \hat{\varepsilon}_{i-2} + \dots + \rho_p \hat{\varepsilon}_{i-p} + \nu_i. \quad (\text{BP3})$$

A estatística do teste de Breusch-Godfrey, denotada por LM (para ficar igual à saída do comando `bgtest` do R) é calculada a partir do valor R-quadrado da regressão auxiliar, multiplicado pelo número de observações. Assintoticamente essa estatística tem distribuição Qui-quadrado com graus de liberdade iguais ao número de defasagens usadas na regressão auxiliar.

$$LM = nR^2 \sim \chi_p^2. \quad (\text{BP3})$$

A regra de decisão implica que se a estatística do teste exceder o valor crítico da distribuição Qui-quadrado, rejeita-se a hipótese nula de nenhuma correlação serial.

EXEMPLO 12.5 Modelo de Regressão Linear Simples e Verificação de Autocorrelação Serial dos Resíduos

Vamos aplicar o teste de autocorrelação serial de Breusch-Godfrey aos dados do Exemplo 12-1. Primeiro usando o comando `bgtest` que está na library `lmtest` do R e depois seguindo as equações que levam à estatística do teste e encontrando o p-valor conforme a respectiva distribuição do teste.

Primeira forma: usando a library lmtest.
`library(lmtest)`

```
x <- c(1,2,3,4,5); y <- c(1,1,2,2,4); dados <- data.frame(x, y)
bgtest(y ~ x, order = 1, data = dados)
```

O que gera como saída

```
Breusch-Godfrey test for serial correlation of order up to 1
```

```
data: y ~ x
LM test = 3.8531, df = 1, p-value = 0.04966
```

Segunda forma: aplicando as equações do teste de Breusch-Godfrey.

```
r1s <- lm(y ~ x, data = dados); r1s
y_hat <- fitted(r1s)
residuos_r1s <- resid(r1s); residuos_r1s
```

```
# Para obter o lag do vetor de resíduos
library(zoo)
```

```
residuos_zoo <- zoo(residuos_r1s)
residuos_lag <- lag(residuos_zoo, -1, na.pad = TRUE)
```

```
# Cria o data frame com os dados para a regressão
dados_bg <- data.frame(x, y, residuos_r1s, residuos_lag)
dados_bg[is.na(dados_bg)] <- 0
dados_bg
```

```
# Estatística de Breusch-Godfrey
erro_hat <- lm(residuos_r1s ~ x + residuos_lag, data = dados_bg)
r_quad <- summary(erro_hat)$r.squared
n <- length(residuos_r1s)
p <- 1
```

```
bg_estat <- n*r_quad
```

```
p_valor_bg <- pchisq(bg_estat, 1, lower.tail = F)
```

Obtemos os seguintes resultados

```
bg_estat
[1] 3.853066
```

e

```
p_valor_bg
[1] 0.04965516
```

que correspondem exatamente aos mesmos resultados da aplicação do `bptest`. Considerando um nível de significância de 4%, não rejeitamos a hipótese de que não há autocorrelação dos resíduos até a ordem 1.

A seguir é apresentado um exemplo do uso prático do MRLS em que utilizamos todas as expressões construídas anteriormente para estimação dos parâmetros (coeficientes) da regressão, verificar as suposições do modelo, construir gráficos de apoio e testes necessários para assegurar a qualidade do modelo. Nesse exemplo o MRLS é utilizado para determinar uma das quantidades mais fundamentais da teoria de finanças, o **beta** de um ativo financeiro.

EXEMPLO 12.6 Modelo de Regressão Linear Simples

Todos os comandos do R utilizados na construção desse Exemplo estão disponíveis na seção de Apêndices. Nesses comandos, são implementadas as expressões matemáticas do MRLS e os resultados são comparados com os que são gerados pelos comandos/funções disponíveis nos pacotes no R.

De acordo com Bodie *et al.* (2015), o Modelo de Precificação de Ativos de Capital (em inglês, Capital Asset Pricing Model - CAPM), constitui-se em uma ferramenta de

análise central na economia financeira moderna. Em 1952, Markowitz apresenta as bases da teoria de gestão moderna de carteiras. O CAPM é desenvolvido e divulgado em uma sequência de trabalhos de Sharpe (1964), Lintner (1965) e Mossin (1966), permitindo avaliar a relação entre o risco e o retorno de um ativo financeiro e seu retorno esperado. No CAPM, a medida usada para medir o desempenho de um ativo financeiro é o **beta**. Quando consultamos o trabalho do Sharpe (*Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. Journal of Finance*, 19, 425-442) extraímos o seguinte trecho: “*Mas parte da dispersão é devido a uma relação subjacente com o retorno da combinação, mostrada por B_{ig} , a inclinação da linha de regressão*”. Como o Sharpe usou uma máquina de escrever para escrever (datilografar) o texto, o que ele realmente quis dizer com B era a letra grega β ou beta.

No glossário do Banco Central do Brasil (<https://www.bcb.gov.br/meubc/glossario>) temos o beta sendo descrito como a: “*estimativa do nível de oscilação que se deve esperar de um fundo (ou ativo qualquer) como resposta a variações do mercado de ações. Uma ação com beta igual a 2, por exemplo, indica que quando o Ibovespa sobe 1%, a ação tende a subir 2%. Quanto maior o beta, maior o risco do papel.*”

O beta é uma medida do risco de mercado ou risco sistemático e é utilizado por investidores para entender o risco de mercado inerente a uma ação. O beta também pode ajudar a criar um portfólio diversificado combinando algumas ações com betas superiores a 1 e algumas ações betas inferiores a 1.

Como o beta de um ativo (ação, título) mede sua volatilidade em relação ao mercado geral (ou uma carteira mais ampla), é natural considerar que envolva medidas de variabilidade. Dessa forma, o beta é escrito como

$$\beta_i = \frac{\text{Cov}[R_i, R_M]}{\text{Var}[R_M]}, \quad (\text{EMB1})$$

onde,

R_i = Retorno do ativo i .

R_M = Retorno da proxy do mercado.

Ou seja, o coeficiente beta de um ativo i é determinado pela razão entre a covariância dos retornos do ativo i e os retornos de uma proxy do mercado, e a variância dos retornos da carteira de mercado. Entenda por proxy a variável que representará o comportamento do mercado, que é formado por um conjunto de ativos. Esta razão apresentada na expressão (EMBR1) corresponde exatamente a mesma para o cálculo do coeficiente angular da reta de regressão. Sendo assim, vamos construir um MRLS em que um dos resultados de interesse será o beta do ativo escolhido.

O ativo escolhido será o EMBR3. A Figura-EX12.6 apresenta o histórico do beta desse ativo no período de 1998 a 2024. Nessa figura a linha tracejada corresponde ao valor médio dos betas no período analisado, sendo de aproximadamente 0,53. E a linha contínua corresponde ao beta igual a 1.

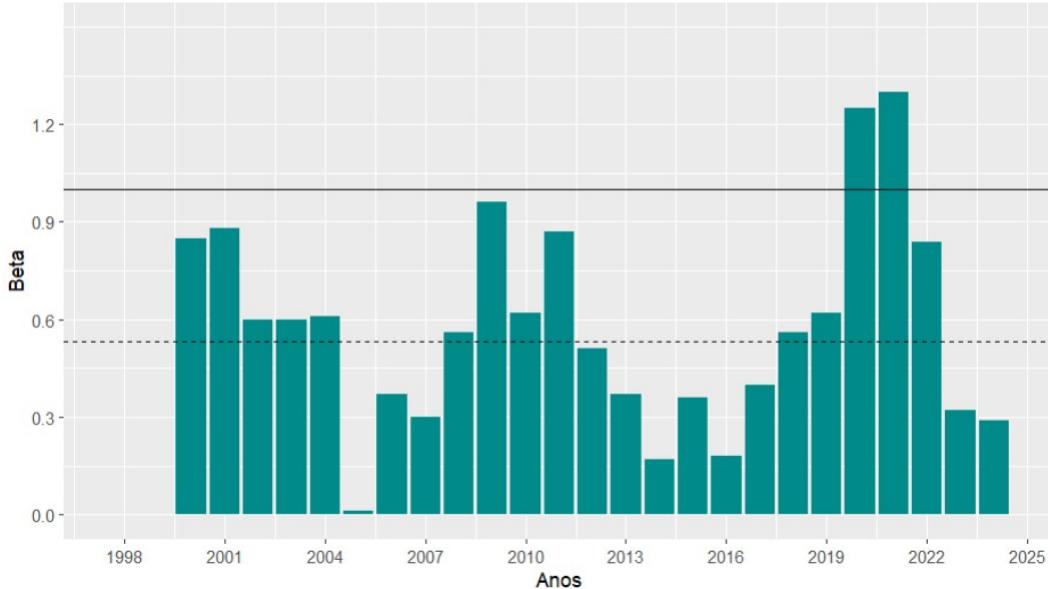


Figura-EMB1 Evolução do beta do ativo EMBR3 no período 1998-2024.

Vamos construir o modelo de regressão para o ativo EMB3, em que a variável dependente é o retorno do ativo (R_A) e a variável independente é a taxa de retorno da carteira de mercado (R_M). Para representar o mercado, é comum utilizar um índice de bolsa como proxy para o índice de mercado, nesse caso utilizaremos o IBOVESPA.

$$R_A = \alpha + \beta R_M + \varepsilon. \quad (\text{EMB2})$$

O beta representa a influência do risco sistemático sobre a taxa de retorno da empresa. Vamos determinar o beta da empresa EMBR3 para o ano de 2021. As etapas da análise de regressão são apresentadas as seguir.

Obteração dos dados

Os dados da empresa EMBR3 podem ser obtidos, por exemplo, do yahoo finance, utilizando o comando `getSymbols` da library `quantmod`.

```
EMBR <- getSymbols("EMBR3.SA", src = "yahoo", from = "2021-01-01", to = "2021-12-31", auto.assign = FALSE)
```

Entre os dados obtidos temos os preços de fechamento para cada dia de 2021.

Da mesma forma carregamos os dados do IBOVESPA.

```
IBOV <- getSymbols("^BVSP", src = "yahoo", from = "2021-01-01", to = "2021-12-31", auto.assign = FALSE)
```

Criação do vetores de retornos do ativo e do mercado

A partir dos preços de fechamento e dos valores do IBOVESPA para dia de negociação em 2021, calculas o retorno do ativo e o retorno de mercado (taxa de variação do IBOVESPA). Lembrando que o retorno em um instante t é calculado usando os valores de preço dos instantes t e $t - 1$, conforme mostrado a seguir.

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}. \quad (\text{EMB3})$$

Scatter plot dos retornos

A seguir é apresentado o scatter plot dos retornos. Esse gráfico possibilita verificar o tipo de relacionamento linear (ou não) entre as variáveis.

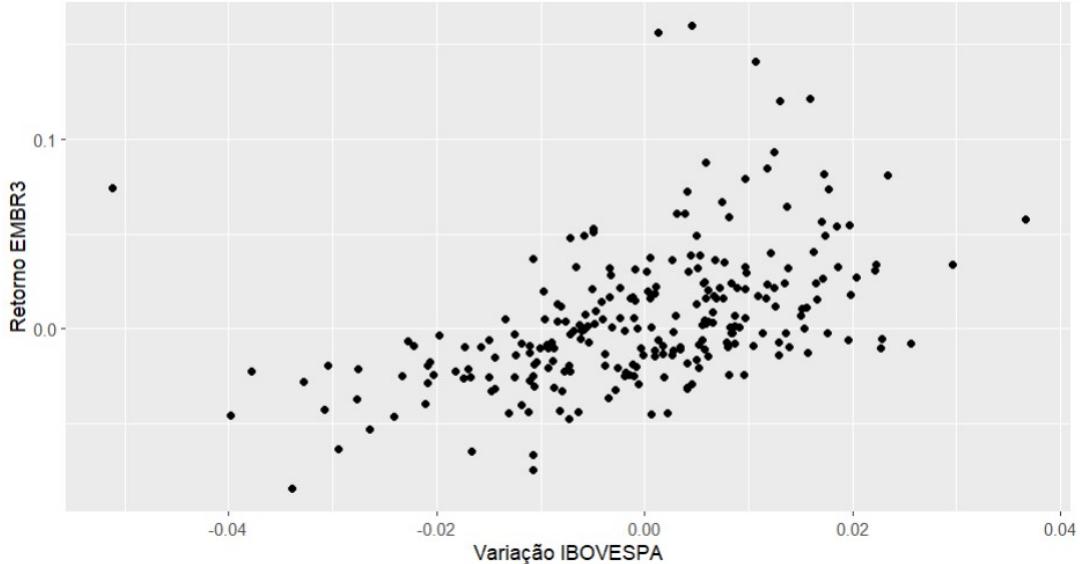


Figura-EMB2 Gráfico de dispersão entre o retorno do ativo e o retorno de mercado.

Estimacão do modelo de regressão

Na figura a seguir podemos observar a reta de regressão ajustada aos dados e o coeficiente angular sendo exatamente igual ao beta da empresa EMBR3 que aparece nos históricos financeiros. Essa reta de regressão foi obtida utilizando-se o comando
`stat_smooth(method = "lm", formula = y ~ x, geom = "smooth")`
Esse comando já gera um intervalo de confiança para $E[y]$.

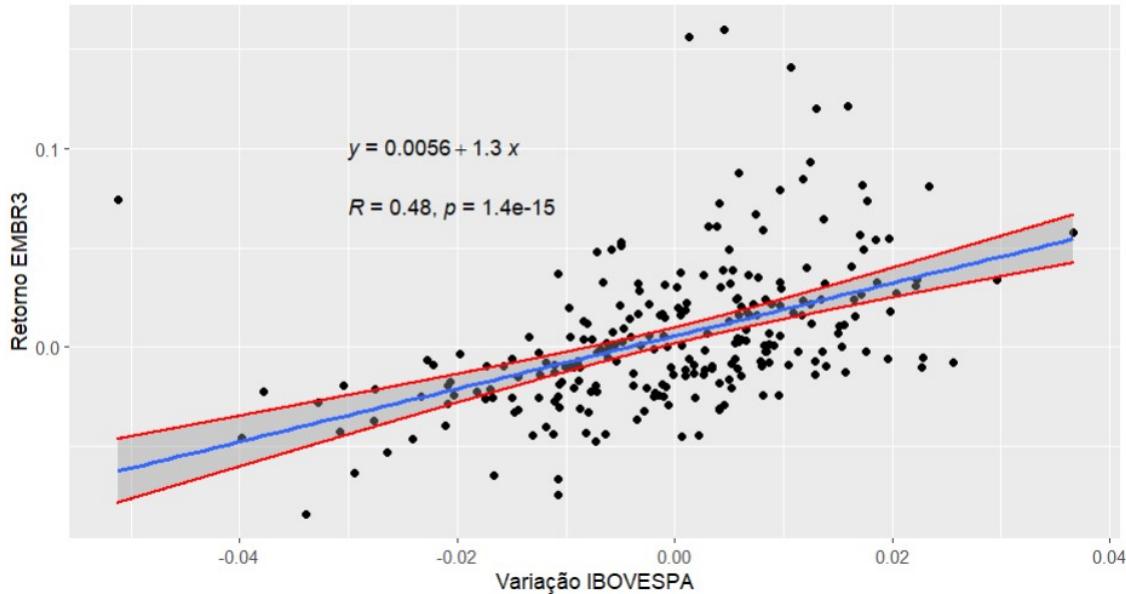


Figura-EMB3 Reta de regressão ajustada para R_A e R_M com intervalo de confiança para $E[R_A]$.

O sumário da regressão é apresentado a seguir. Note que o p-valor associado ao coeficiente angular da reta de regressão é bastante pequeno, permitindo rejeitar a hipótese de que seja o beta estatisticamente igual a zero.

```

Call:
lm(formula = EMBR3.SA.Close ~ BVSP.Close, data = IBOV_EMBR)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.06581 -0.01959 -0.00315  0.01253  0.14871 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.005626  0.002053   2.741  0.00658 **  
BVSP.Close  1.331904  0.155762   8.551 1.36e-15 *** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03218 on 244 degrees of freedom
Multiple R-squared:  0.2306, Adjusted R-squared:  0.2274 
F-statistic: 73.12 on 1 and 244 DF,  p-value: 1.359e-15

```

Intervalo de Predição para $Y_{p(novo)}$

Na próxima figura é inserido o intervalo de predição. Observe que o intervalo de predição é mais amplo que o intervalo para $E[y]$, como pode ser observado das expressões matemáticas para esses dois intervalos. Essa maior amplitude do IP é devido à incerteza adicional de novos dados observados que introduz um erro de predição e quanto mais longe o ponto de predição estiver dos dados utilizados para ajustar o modelo, maior será a incerteza e, consequentemente, mais amplo será o intervalo de predição.

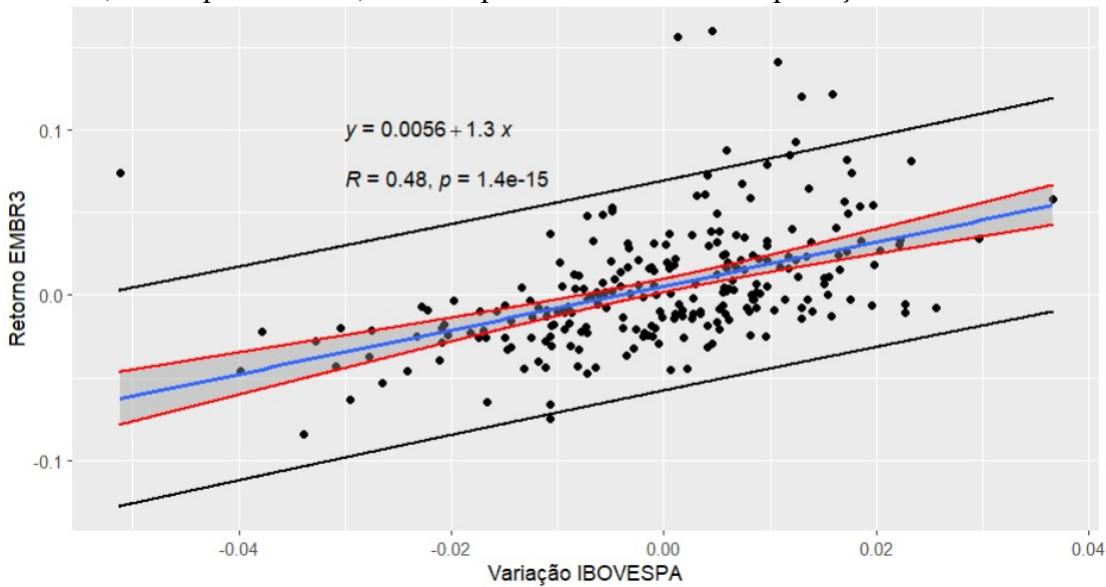


Figura-EMB4 Reta de regressão ajustada para R_A e R_M com intervalo de predição.

Verificação da autorrelação entre os resíduos

Utilizamos o teste de Durbin-Watson para avaliar a autocorrelação dos resíduos. O resultado é apresentado a seguir.

Durbin-Watson test

```

data: regressao
DW = 2.1336, p-value = 0.8565
alternative hypothesis: true autocorrelation is greater than 0

```

Concluindo que a hipótese nula (H_0 : resíduos não são correlacionados) não é rejeitada.

Verificação da homocedasticidade dos resíduos

Faendo o Teste de Breusch-Pagan em que a hipótese nula é dada por H_0 : Resíduos são Homocedásticos [Ou seja, os resíduos são distribuídos com igual varância], temos o seguinte resultado:

```
bptest(regressao) # Está na library "lmtest"
studentized Breusch-Pagan test

data: regressao
BP = 0.011568, df = 1, p-value = 0.9143
```

Ou seja, não rejeitamos (com um nível de confiança de 5%, por exemplo) a hipótese de que os resíduos gerados pelo modelo de regressão são homocedásticos.

Teste de normalidade dos resíduos

Aplicando os Testes de Kolmogorov-Smirnov, Shapiro-Wilk, Anderson-Darling e de Lilliefors temos os resultados:

```
# KOLMOGOROV-SMIRNOV> ks.test(residuos, "pnorm")
Asymptotic one-sample Kolmogorov-Smirnov test

data: residuos
D = 0.47376, p-value < 2.2e-16
alternative hypothesis: two-sided

# SHAPIRO-WILK> shapiro.test(residuos)
Shapiro-Wilk normality test

data: residuos
W = 0.89143, p-value = 2.691e-12

# ANDERSON-DARLING> ad.test(residuos)
Anderson-Darling normality test

data: residuos
A = 5.0146, p-value = 1.971e-12

# LILLIEFORST> lillie.test(residuos) # Também precisa está na library "nortest"
Lilliefors (Kolmogorov-Smirnov) normality test

data: residuos
D = 0.10293, p-value = 1.241e-06
```

Ou seja, em todos os testes de normalidade, rejeitamos a hipótese de que os resíduos se distribuem com distribuição Normal. Porém, como mencionado anteriormente, não normalidade dos erros não é um grande problema com amostras de tamanho grande. Nesse caso, o tamanho da amostra é de 246 observações. Para mais informações sobre não normalidade de resíduos, pode-se consultar as referências Wooldridge (2019) e Lumley *et al.* (2002).

Linearidade do modelo

A linha ajustada pela LOESS deve ser aproximadamente horizontal em zero. Na Figura-EMB5 verificamos que a linha LOESS que ajusta aos dados não apresenta um comportamento funcional bem definido. Temos que a linha LOESS aproxima-se da linha horizontal próximo de zero, e dessa forma concluímos que a linearidade é satisfeita.

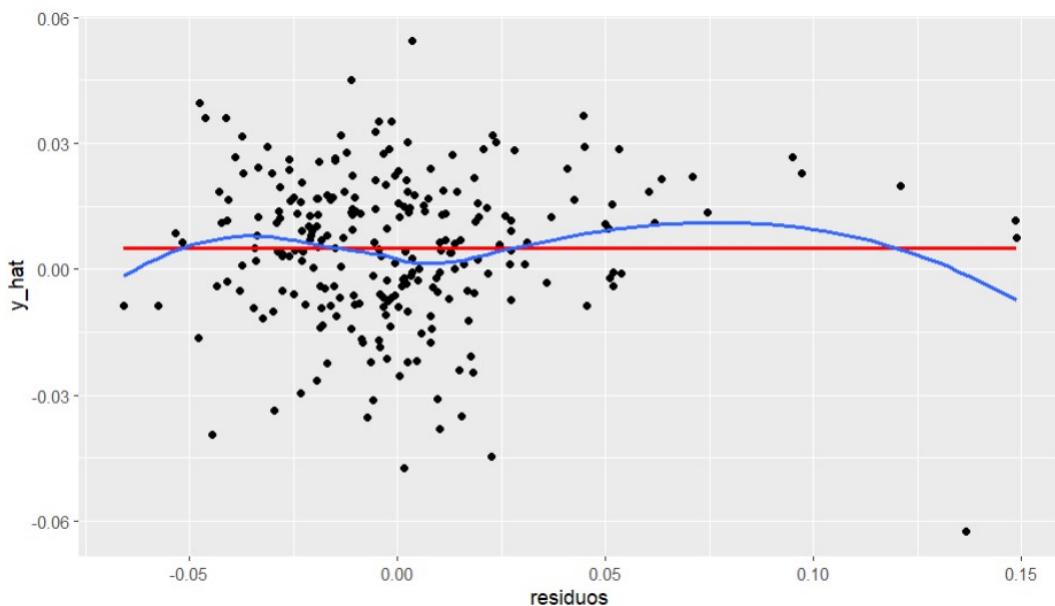


Figura-EMB5 Linha ajustada LOESS.

Identificação de outliers

A Figura-EMB6 e a Figura-EMB7 mostram o gráfico Outlier and Leverage Diagnostics e a distância de Cook para identificar outliers, verificamos a presença de vários outliers, sendo um desses valores excendendo bastante o limiar da distância de Cook.

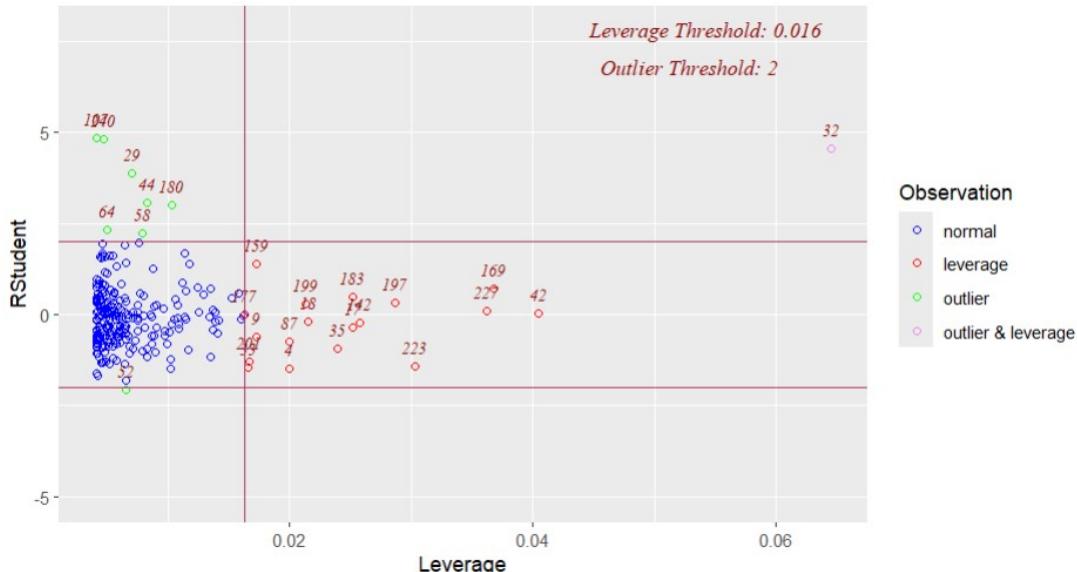


Figura-EMB6 Gráfico Outlier and Leverage Diagnostics.

Esse outlier, denotado como sendo a observação 32, corresponde exatamente ao valor da data 22 de Fevereiro de 2021. Esse ponto refere-se ao fato de que agência de classificação de risco Standard & Poor's (S&P) rebaixou nesta data (22/02/2021) o rating da empresa de BB+ para BB com perspectiva negativa, diante da expectativa de demanda mais fraca por causa da pandemia de coronavírus.

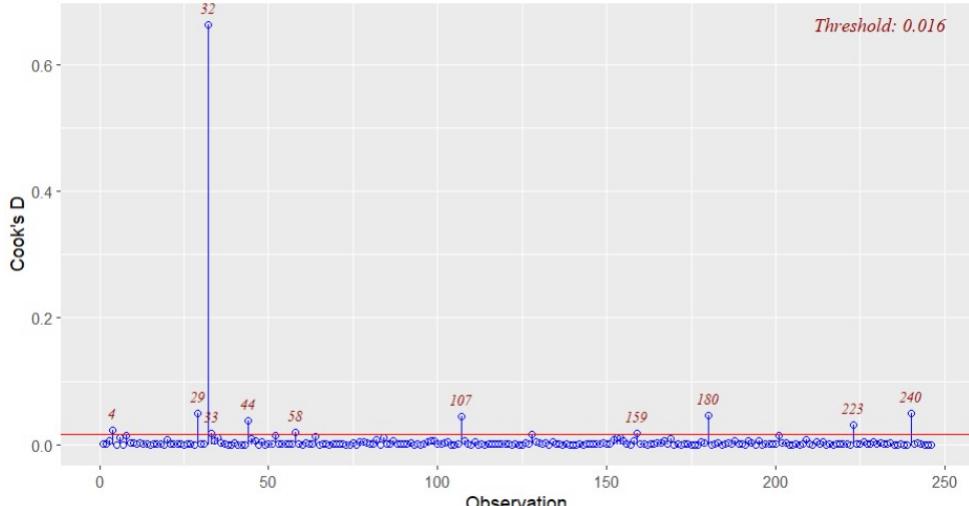


Figura-EMB7 Distâncias de Cook para EMBR3 em 2021.

3.2 Regressão Linear Múltipla

O modelo de regressão linear múltipla generaliza o modelo de regressão linear simples que considera uma única variável independente, de tal forma que são acrescentadas outras variáveis para predizer a variável dependente. Um modelo de regressão linear múltipla construído para explicar uma variável dependente Y a partir de k variáveis independentes, denominadas, X_1, X_2, \dots, X_k e um termo de erro ε , é denominado de **Modelo de Regressão Linear Múltipla (MRLM)**. Este modelo pode ser escrito como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon. \quad (26)$$

Da mesma forma que no modelo de regressão linear simples algumas suposições devem ser feitas sobre o termo de erro aleatório:

- O erro do modelo é uma variável aleatória com média zero;
- Os erros possuem distribuição Normal;
- A variância dos erros é constante;
- Os erros associados com quaisquer duas observações são independentes;
- O número de observações deve ser superior ao número de variáveis;
- Multicolinearidade entre as variáveis independentes (não pode existir combinação linear entre as variáveis independentes).

Nesse caso, tanto o método dos mínimos quadrados, quanto o método da máxima verossimilhança podem ser utilizados para obter os estimadores dos parâmetros do MRLM. A equação de regressão linear múltipla estimada é dada por

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k, \quad (27)$$

sendo que:

x_1, x_2, \dots, x_k são os preditores,

$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$, são os estimadores dos parâmetros $\beta_0, \beta_1, \beta_2, \dots, \beta_k$, e \hat{y} é o valor estimado da variável independente.

Os modelos de regressão linear múltipla também podem ser escritos em termos de potências, interações e mistura de potências e interações entre as variáveis preditoras. Por exemplo, considerando potência

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^3 + \beta_3 x_3^2 + \beta_4 x_4 + \varepsilon, \quad (28)$$

interação entre as variáveis preditoras

$$y = \beta_0 + \beta_1 x_1 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \varepsilon, \quad (29)$$

ou usando tanto potência como interações entre elas

$$y = \beta_0 + \beta_1 x_1 + \beta_{12} x_1 x_2 + \beta_{33} x_3^2 + \beta_{24} x_2 x_4^2 + \varepsilon. \quad (30)$$

Os modelos representados em (28), (29) e (30) continuam sendo modelos lineares em termos dos parâmetros (coeficientes que precisam ser estimados).

Considerando uma amostra $\{(x_{i1}, x_{i2}, \dots, x_{ik}; y_i); i = 1, 2, \dots, n\}$, os valores ajustados para y quando as k variáveis independentes (preditores) assumem os valores $x_{i1}, x_{i2}, \dots, x_{ik}$ serão obtidos pelo **Modelo de Regressão Linear Múltipla Amostral (MRLMA)**.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i. \quad (31)$$

Assumimos que os resíduos ε_i ($i = 1, 2, \dots, n$) possuem as seguintes características:

$$E[\varepsilon_i] = 0, \quad (32)$$

$$Var[\varepsilon_i] = \sigma^2, \quad (33)$$

$$Cov[\varepsilon_i; \varepsilon_j] = 0, \quad \forall i \neq j \quad (34)$$

O MRLMA é representado pelo seguinte sistema de equações

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_k x_{1k} + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_k x_{2k} + \varepsilon_2$$

$$\vdots \quad \vdots$$

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_k x_{nk} + \varepsilon_n$$

Podemos representar esse sistema de equações utilizando notação matricial, em que cada matriz é dada por

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Dessa forma, o MRLM, considerando todos os n valores que a variável independente pode assumir simultaneamente, é matricialmente representado como

$$\underbrace{\mathbf{y}}_{(n \times 1)} = \underbrace{\mathbf{X}}_{(n \times r)} \cdot \underbrace{\boldsymbol{\beta}}_{(r \times 1)} + \underbrace{\boldsymbol{\varepsilon}}_{(n \times 1)}, \quad (35)$$

sendo que $r = k + 1$ representa o número de parâmetros da regressão e k o número de regressores do modelo. Também podemos reescrever matricialmente a média e a variância do vetor de erros como

$$E[\boldsymbol{\varepsilon}] = \underbrace{\mathbf{0}}_{(n \times 1)}, \quad (36)$$

$$Var[\boldsymbol{\varepsilon}] = \underbrace{\sigma^2 \mathbf{I}}_{(n \times n)}, \text{ sendo } \mathbf{I}, \text{ a matriz identidade.} \quad (37)$$

Se adicionarmos a suposição de normalidade aos erros (resíduos da regressão), temos que

$$\boldsymbol{\varepsilon} \sim N\left(\underbrace{\mathbf{0}}_{(n \times 1)}, \underbrace{\sigma^2 \mathbf{I}}_{(n \times n)}\right). \quad (38)$$

Os estimadores dos coeficientes que permitem a especificação do MRLM serão obtidos pelo método de mínimos quadrados, de tal forma que sejam escolhidos os estimadores que minimizem a soma dos quadrados dos resíduos. Ou seja, será ajustado o modelo de regressão em que $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$, minimizando a soma das diferenças quadradas entre cada observação i de y da amostra e o respectivo \hat{y}_i gerado pelo modelo. A soma a ser minimizada é uma função dos coeficientes β do modelo, denotada por $S(\hat{\beta})$.

$$\min_{\hat{\beta}} S(\hat{\beta}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \varepsilon_i^2, \quad (27)$$

e,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_k x_{ik}, \quad (27)$$

gerado para cada valor $1 \leq i \leq n$ da amostra de tamanho n . A Figura-2 apresenta essa ideia, para o caso de uma variável dependente (Y) sendo explicada por duas variáveis preditoras (X_1 e X_2).

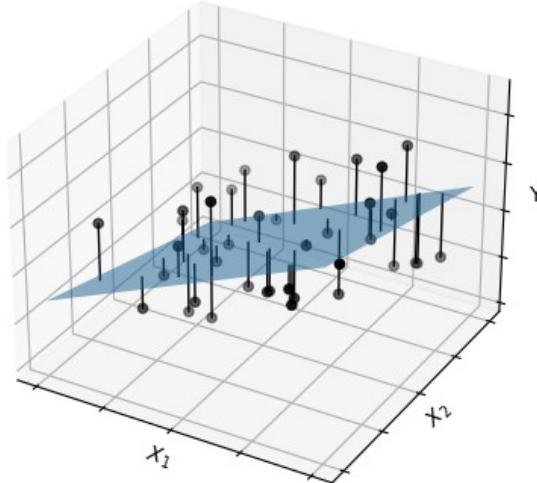


Figura-2 Modelo de Regressão Linear Múltipla de (Y) sendo explicada por duas variáveis (X_1 e X_2).

Podemos utilizar a representação matricial para obter os estimadores dos parâmetros da regressão. Sendo $\boldsymbol{\varepsilon} = (\varepsilon_i) \in \mathbb{R}^n$ e $\hat{\mathbf{y}} = (\hat{y}_i) = \mathbf{X}\hat{\boldsymbol{\beta}} \in \mathbb{R}^n$, temos $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}}$.

| Representação Escalar do MRLM | Representação Matricial do MRLM |
|---|--|
| Para os valores estimados de y temos $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_{ij}.$ | Para os valores estimados de y temos $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}.$ |
| Os erros de estimação são obtidos como $\hat{\boldsymbol{\epsilon}}_i = y_i - \hat{y}_i.$ | Os erros de estimação são obtidos como $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}}.$ |
| Lembrando que $\ \mathbf{y} - \hat{\mathbf{y}}\ ^2 = \ \mathbf{y}\ ^2 + \ \hat{\mathbf{y}}\ ^2 - 2\mathbf{y}^T \hat{\mathbf{y}}$ podemos escrever a soma dos quadrados dos resíduos como | |
| $S(\hat{\boldsymbol{\beta}}) = \ \mathbf{y}\ ^2 + \ \mathbf{X}\hat{\boldsymbol{\beta}}\ ^2 - 2\mathbf{y}^T \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y}^T \mathbf{y} + \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{X}\hat{\boldsymbol{\beta}} - 2\mathbf{y}^T \mathbf{X}\hat{\boldsymbol{\beta}}.$ | |
| O último termo da matriz acima corresponde ao produto de matrizes com as dimensões $(1 \times n)$, $(n \times r)$ e $(r \times 1)$ o que gera como resultado uma matriz (1×1) , que é sempre simétrica. Ou seja, podemos reescrever esse último termo também como $-2\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{y}$. | |
| Dessa forma, ficamos com | |
| $S(\hat{\boldsymbol{\beta}}) = \mathbf{y}^T \mathbf{y} + \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{X}\hat{\boldsymbol{\beta}} - 2\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{y}.$ | |
| Derivando com relação a $\hat{\boldsymbol{\beta}}$ produz | |
| $\frac{\partial S(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} = 0 + 2\mathbf{X}^T \mathbf{X}\hat{\boldsymbol{\beta}} - 2\mathbf{X}^T \mathbf{y}.$ | |
| Igualando essa derivada a zero, resulta em | |
| $\mathbf{X}^T \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y}.$ | |
| Assim, podemos agora isolar $\hat{\boldsymbol{\beta}}$ da seguinte forma | |
| $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad \begin{cases} \mathbf{X}^T = \text{transposta de } \mathbf{X} \\ (\mathbf{X}^T \mathbf{X})^{-1} = \text{inversa de } \mathbf{X}^T \mathbf{X} \end{cases}$ | |
| considerando que $\mathbf{X}^T \mathbf{X}$ seja não singular (ou seja, possua inversa), a matriz \mathbf{X} deve ter posto completo. Como \mathbf{X} é uma matriz $(n \times r)$, sendo $r = k+1$, isso requer em particular que $n \geq k$, ou seja, o número de parâmetros é menor ou igual ao número de observações. Na prática, quase sempre exigiremos que k seja consideravelmente menor que n (Heiji <i>et al.</i> , 2004). | |
| $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \tag{28}$ | |

A pergunta que surge é: *O resultado apresentado na equação (28) para $\hat{\boldsymbol{\beta}}$, gera a soma dos quadrados mínima para os erros do MRLM?*

Para responder essa pergunta consideramos o método dos mínimos quadrados, que

seleciona $\hat{\beta}$ de tal forma a minimizar a soma dos quadrados das diferença $y - \hat{y}$. Podemos escrever esse problema de diversas formas, por exemplo

$$\min_{\hat{\beta}} S(\hat{\beta}) = \min_{\hat{\beta}} \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij} \right)^2 = \min_{\hat{\beta}} \|y - X\hat{\beta}\|^2 = \min_{\hat{\beta}} \|\hat{\epsilon}\|^2 = \min_{\hat{\beta}} (y - X\hat{\beta})^T (y - X\hat{\beta}).$$

Escrevendo $\hat{\beta}$, o candidato a mínimo de $S(\hat{\beta})$, usando o resultado (28) temos

$$\begin{aligned}\hat{\epsilon} &= y - \hat{y} = y - X\hat{\beta}, \\ \hat{\epsilon} &= y - X(X^T X)^{-1} X^T y, \\ \hat{\epsilon} &= [I - X(X^T X)^{-1} X^T] y, \text{ sendo } I \text{ a Matriz Identidade.}\end{aligned}$$

A matriz $X(X^T X)^{-1} X^T$ é chamada de matriz chapéu (hat matrix), denotada por H e a matriz $[I - X(X^T X)^{-1} X^T]$ possui as seguintes propriedades:

i. Simetria

$$[I - X(X^T X)^{-1} X^T]^T = I^T - [X(X^T X)^{-1} X^T]^T = I - [X(X^T X)^{-1}] X^T,$$

usando a propriedade de que $(MN)^T = N^T M^T$ ficamos com

$$\begin{aligned}I - (X^T)^T [X(X^T X)^{-1}]^T &= I - X[(X^T X)^{-1}]^T X^T = I - X[(X^T X)^T]^{-1} X^T, \\ [I - X(X^T X)^{-1} X^T]^T &= I - X(X^T X)^{-1} X^T.\end{aligned}$$

Também infere-se que

$$H^T = H.$$

ii. Idempotência

$$\begin{aligned}[I - X(X^T X)^{-1} X^T][I - X(X^T X)^{-1} X^T] &= I - 2X(X^T X)^{-1} X^T + X[(X^T X)^{-1} X^T X](X^T X)^{-1} X^T, \\ [I - X(X^T X)^{-1} X^T][I - X(X^T X)^{-1} X^T] &= [I - X(X^T X)^{-1} X^T]\end{aligned}$$

Também infere-se que

$$\begin{aligned}X(X^T X)^{-1} X^T [X(X^T X)^{-1} X^T] &= X[(X^T X)^{-1} X^T X](X^T X)^{-1} X^T = X(X^T X)^{-1} X^T, \\ H^2 &= H.\end{aligned}$$

iii. $H[I - H] = 0$

$$\begin{aligned}X(X^T X)^{-1} X^T [I - X(X^T X)^{-1} X^T] &= X(X^T X)^{-1} X^T - [X(X^T X)^{-1} X^T][X(X^T X)^{-1} X^T], \\ X(X^T X)^{-1} X^T - X[(X^T X)^{-1} X^T X](X^T X)^{-1} X^T &= 0.\end{aligned}$$

iv. $X^T[I - H] = 0$

$$\mathbf{X}^T \left[\mathbf{I} - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] = \mathbf{X}^T - \left[\mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \right] \mathbf{X}^T = \mathbf{X}^T - \mathbf{X}^T = \mathbf{0}.$$

Dessa forma, conseguimos verificar que

$$\mathbf{X}^T \boldsymbol{\varepsilon} = \mathbf{X}^T (\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{X}^T \left[\mathbf{I} - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \mathbf{y} = \mathbf{0},$$

e consequentemente

$$\hat{\mathbf{y}}^T \hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \hat{\boldsymbol{\varepsilon}} = \mathbf{0}..$$

Também podemos escrever $\hat{\mathbf{y}}$ como

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}, \\ \hat{\mathbf{y}} &= \mathbf{H} \mathbf{y}. \end{aligned}$$

A partir dos resultados anteriores podemos escrever matricialmente a soma dos quadrados dos resíduos como

$$\begin{aligned} \hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} &= \mathbf{y}^T \left[\mathbf{I} - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \left[\mathbf{I} - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \mathbf{y} = \mathbf{y}^T \left[\mathbf{I} - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \mathbf{y}, \\ \hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} &= \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \left[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \right] \\ \hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} &= \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \hat{\boldsymbol{\beta}}. \end{aligned} \tag{29}$$

Considere que o ponto de mínimo da soma de quadrados dos resíduos ocorra em um valor particular de $\boldsymbol{\beta}$, denotado por $\boldsymbol{\beta}^*$. Isso significa que esse $\boldsymbol{\beta}^*$ deve gerar o menor valor dessa soma, tal que $S(\boldsymbol{\beta}^*)$ é mínima. Escrevendo o erro (ou resíduo) em termos de $\boldsymbol{\beta}^*$ temos $\boldsymbol{\varepsilon}^* = \mathbf{y} - \mathbf{X} \boldsymbol{\beta}^*$. Para representar esse erro também considerando $\hat{\boldsymbol{\beta}}$, podemos escrever esse erro como $\boldsymbol{\varepsilon}^* = \mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}} + \mathbf{X} \hat{\boldsymbol{\beta}} - \mathbf{X} \boldsymbol{\beta}^* = (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}) + \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$. Dessa forma,

$$\begin{aligned} (\hat{\boldsymbol{\varepsilon}}^*)^T \hat{\boldsymbol{\varepsilon}}^* &= \left\{ (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})^T + \left[\mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \right]^T \right\} \left[(\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}) + \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \right], \\ (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})^T (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}) &+ (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \left[\mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \right]^T (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}) + \left[\mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \right]^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*), \end{aligned}$$

usando as propriedades de que $\mathbf{M}^T \mathbf{N} = \mathbf{N}^T \mathbf{M}$ e $(\mathbf{MN})^T = \mathbf{N}^T \mathbf{M}^T$ para $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{n \times n}$ temos

$$(\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})^T (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}) + (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*),$$

o que produz

$$S(\boldsymbol{\beta}^*) = (\hat{\boldsymbol{\varepsilon}}^*)^T \hat{\boldsymbol{\varepsilon}}^* = (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})^T (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}) + 2 (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*),$$

porém, como $(\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})^T \mathbf{X} = \hat{\boldsymbol{\varepsilon}}^T \mathbf{X} = \mathbf{0}^T$ resulta em

$$S(\boldsymbol{\beta}^*) = (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})^T (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*). \tag{29}$$

Como a matriz \mathbf{X} tem posto completo, $\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \neq \mathbf{0}$ se $\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*$, portanto a soma dos quadrados mínima é única e ocorre em $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ (Johnson e Wichern, 1998).

3.1.2.1 Não Tendenciosidade do Estimador de MQO para o MRLM

O estimador dos parâmetros da regressão determinados em (28) corresponde ao estimador que gera a menor soma dos quadrados dos erros. Podemos verificar se esse estimador é não viesado (não tendencioso), reescrevendo (28) em termos de $\boldsymbol{\epsilon}$ e tomando a esperança. Dessa forma, temos

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon}), \\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\epsilon}, \\ \hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\epsilon},\end{aligned}$$

tomando a esperança, temos

$$\begin{aligned}E[\hat{\boldsymbol{\beta}}] &= E[\boldsymbol{\beta} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\epsilon}] = E[\boldsymbol{\beta}] + E[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\epsilon}], \\ E[\hat{\boldsymbol{\beta}}] &= E[\boldsymbol{\beta}] + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E[\boldsymbol{\epsilon}], \\ E[\hat{\boldsymbol{\beta}}] &= \boldsymbol{\beta}.\end{aligned}\tag{30}$$

Portanto, $\hat{\boldsymbol{\beta}}$ é não viesado.

3.1.2.2 Matriz de Variância-Covariância do Estimador de MQO para o MRLM

Para determinar a matriz de variância-covariância (ou simplesmente, covariância) de $\hat{\boldsymbol{\beta}}$, vamos considerar a situação análoga do cálculo da variância de uma constante escalar (m) multiplicando uma variável aleatória (y), onde temos que $Var[my] = m^2 Var[y]$. No caso matricial $Var[\mathbf{My}] = \mathbf{M} Var[\mathbf{y}] \mathbf{M}^T$, então

$$\begin{aligned}Var[\hat{\boldsymbol{\beta}}] &= Var[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\epsilon}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Var[\boldsymbol{\epsilon}] (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T, \\ Var[\hat{\boldsymbol{\beta}}] &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\sigma^2 \mathbf{I}) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}, \\ Var[\hat{\boldsymbol{\beta}}] &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}.\end{aligned}\tag{31}$$

A expressão (31) também poderia ter sido escrita como $Cov[\hat{\boldsymbol{\beta}}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$, uma vez que $Var[\hat{\boldsymbol{\beta}}] = Cov[\hat{\beta}_i, \hat{\beta}_j]_{0 \leq i, j \leq k}$.

Outros resultados importantes podem ser obtidos para $E[\hat{\boldsymbol{\epsilon}}]$, $Cov[\hat{\boldsymbol{\epsilon}}]$ e $Cov[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\epsilon}}]$, respectivamente como:

i. $E[\hat{\boldsymbol{\epsilon}}]$

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}} &= \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \mathbf{y}, \\
E[\hat{\boldsymbol{\epsilon}}] &= E \left\{ \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \mathbf{y} \right\} = \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] E[\mathbf{y}], \\
\left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \mathbf{X} \boldsymbol{\beta} &= \mathbf{X} \boldsymbol{\beta} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X} \boldsymbol{\beta} - \mathbf{X} \boldsymbol{\beta}, \\
E[\hat{\boldsymbol{\epsilon}}] &= \mathbf{0}.
\end{aligned}$$

ii. $Cov[\hat{\boldsymbol{\epsilon}}]$

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}} &= \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \mathbf{y}, \\
Var[\hat{\boldsymbol{\epsilon}}] &= Var \left\{ \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \mathbf{y} \right\} = \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] Var[\mathbf{y}] \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right]^T, \\
\left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \sigma^2 \mathbf{I} \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right]^T &= \sigma^2 \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \left\{ \mathbf{I}^T - \left[\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right]^T \right\}, \\
\sigma^2 \left\{ \mathbf{I} \mathbf{I}^T - \mathbf{I} \left[\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right]^T - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{I}^T + \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \left[\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right]^T \right\}, \\
\text{como sabemos que } \mathbf{I} \mathbf{I}^T &= \mathbf{I}, \mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T, \mathbf{H}^T = \mathbf{H} \text{ e } \mathbf{H}^2 = \mathbf{H}, \text{ então} \\
\sigma^2 \left\{ \mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{I}^T + \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{I}^T \right\}, \\
Var[\hat{\boldsymbol{\epsilon}}] &= Cov[\hat{\boldsymbol{\epsilon}}] = \sigma^2 \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right].
\end{aligned}$$

iii. $Cov[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\epsilon}}]$

$$Cov[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\epsilon}}] = E \left\{ \left[\hat{\boldsymbol{\beta}} - E[\hat{\boldsymbol{\beta}}] \right] \left[\hat{\boldsymbol{\epsilon}} - E[\hat{\boldsymbol{\epsilon}}] \right]^T \right\} = E \left\{ \left[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right] \left[\hat{\boldsymbol{\epsilon}} - \mathbf{0} \right]^T \right\} = E \left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \hat{\boldsymbol{\epsilon}}^T \right].$$

Antes de prosseguir, vamos considerar o seguinte resultado sobre $\hat{\boldsymbol{\epsilon}}$ em que

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}} &= \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \mathbf{y} = \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] (\mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon}), \\
&= \left[\mathbf{X} \boldsymbol{\beta} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} \right] + \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \boldsymbol{\epsilon}, \\
&= \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \boldsymbol{\epsilon}.
\end{aligned}$$

Para a transposta de $\hat{\boldsymbol{\epsilon}}$ temos

$$\hat{\boldsymbol{\epsilon}}^T = \boldsymbol{\epsilon}^T \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right].$$

Sabendo ainda que $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\epsilon}$, temos

$$\begin{aligned}
Cov[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\epsilon}}] &= E \left\{ \left[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\epsilon} \right] \hat{\boldsymbol{\epsilon}}^T \right\} = E \left\{ (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\epsilon} \boldsymbol{\epsilon}^T \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \right\}, \\
(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E \left[\boldsymbol{\epsilon} \boldsymbol{\epsilon}^T \right] \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right],
\end{aligned}$$

$$\sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \underbrace{\left\{ \mathbf{X}^T \left[\mathbf{I} - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \right\}}_0 = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \cdot \mathbf{0},$$

$$Cov[\hat{\beta}, \hat{\epsilon}] = \mathbf{0}.$$

Mostramos que o estimador $\hat{\beta}$ produz a menor soma dos quadrados dos resíduos e é não viesado, mesmo assim: *O resultado apresentado na equação (28) para $\hat{\beta}$, é um estimador ótimo obtido pelo método MQO, ou seja, gera variância mínima?*

Faltou deixar explícita a Matriz Var-Cov.

Precisa mostrar que o MSE é estimador para Σ^2

NÃO ESQUECER O R^2 AJUSTADO.

3.1.2.3 O Estimador $\hat{\beta}$ é BLUE

Mostramos que $\hat{\beta}$ é o estimador que gera a menor soma dos quadrados dos resíduos e que também é um estimador não tendencioso, além disso, o Teorema de Gauss-Markov constitui-se em um resultado poderoso que garante que o estimador $\hat{\beta}$ é de mínima variância.

Teorema de Gauss-Markov

Considere o modelo de regressão linear $\mathbf{y} = \mathbf{X}\beta + \epsilon$, onde $E[\epsilon] = \mathbf{0}$, $Cov[\epsilon] = \sigma^2 \mathbf{I}$ e \mathbf{X} tem posto completo $k + 1$. O melhor estimador não viesado de variância mínima (e única) $\hat{\beta}$ de β existe, é dado por $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ e tem variância $Var[\hat{\beta}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$.

Generalizamos a prova para toda combinação linear de $\hat{\beta}$ na forma \mathbf{Gy} . Suponha que tenhamos um estimador alternativo $\tilde{\beta}$, tal que $\tilde{\beta} = \mathbf{Gy}$ e $E[\tilde{\beta}] = \beta$. Dessa forma,

$$E[\tilde{\beta}] = E[\mathbf{Gy}] = \mathbf{GE[y]} = \mathbf{GX}\beta = \beta,$$

implica, para todo β , que $\mathbf{GX} = \mathbf{I}$. Considerando os dois estimadores não viesados para β sendo $\hat{\beta}$ e $\tilde{\beta}$, sabemos pelo método de MQO que $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$. Para o modelo de regressão linear $\mathbf{y} = \mathbf{X}\beta + \epsilon$, temos

$$E[\hat{\beta}] = \beta \quad \text{e} \quad Var[\hat{\beta}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1},$$

fazendo $\mathbf{G} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + \mathbf{M}$, então

$$E[\tilde{\beta}] = \beta \quad \text{e} \quad Var[\tilde{\beta}] = Var[\mathbf{Gy}] = \mathbf{G}Var[\mathbf{y}]\mathbf{G}^T.$$

Determinando a variância, temos

$$\begin{aligned}
Var[\tilde{\beta}] &= \mathbf{G}\sigma^2\mathbf{G}^T = \sigma^2\mathbf{G}\mathbf{G}^T = \sigma^2 \left[(\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T + \mathbf{M} \right] \left[(\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T + \mathbf{M} \right]^T, \\
&\quad \sigma^2 \left[(\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T + \mathbf{M} \right] \left[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} + \mathbf{M}^T \right], \\
&\quad \sigma^2 \left[(\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T\mathbf{X})^{-1} + (\mathbf{X}^T\mathbf{X})^{-1} \underbrace{\mathbf{X}^T \mathbf{M}^T}_0 + \underbrace{\mathbf{M} \mathbf{X}}_0 (\mathbf{X}^T\mathbf{X})^{-1} + \mathbf{M} \mathbf{M}^T \right],
\end{aligned}$$

sabemos que $\mathbf{G}\mathbf{X} = \mathbf{I}$, isso implica que

$$\mathbf{G}\mathbf{X} = (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} + \mathbf{M}\mathbf{X} \rightarrow \mathbf{I} = \mathbf{I} + \mathbf{M}\mathbf{X} \rightarrow \mathbf{M}\mathbf{X} = \mathbf{0}.$$

Então,

$$Var[\tilde{\beta}] = \underbrace{\sigma^2 (\mathbf{X}^T\mathbf{X})^{-1}}_{Var[\hat{\beta}]} + \sigma^2 \mathbf{M} \mathbf{M}^T.$$

Assim, temos que $Var[\tilde{\beta}] - Var[\hat{\beta}] = \sigma^2 \mathbf{M} \mathbf{M}^T$, a qual é positiva definida e somente será zero se $\mathbf{M} = \mathbf{0}$, ou seja, se tivermos $\mathbf{G} = (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T$. Portanto a variância mínima é única e ocorre em $\tilde{\beta} = \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$. O estimador $\hat{\beta}$ é chamado de o melhor (variância mínima) estimador linear não viesado (**Best Linear Unbiased Estimator - BLUE**) (Johnson e Wichern, 1998).

LIVRO MAURI [GESTÃO DE ESTOQUES]

3.4 Coeficiente de Determinação

O coeficiente de determinação é uma medida da qualidade do ajuste do modelo de regressão linear. Ele pode ser calculado para a regressão linear simples e para a regressão linear múltipla.

O coeficiente de determinação é dado por:

Coeficiente de determinação

$$R^2 = \frac{\text{Soma dos quadrados da regressão}}{\text{Soma de quadrado total}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}, \quad 0 \leq R^2 \leq 1 \quad (4)$$

O coeficiente de determinação representa a proporção da soma dos quadrados dos desvios de y em torno da média \bar{y} que pode ser atribuída a equação de regressão linear múltipla estimada. Desta forma, o coeficiente de determinação representa a proporção da variabilidade da variável dependente explicada pelo modelo. Portanto, $R^2=0$ implica em uma completa falta de ajuste do modelo aos dados, enquanto que $R^2=1$ implica em um ajuste perfeito, com o modelo ajustado passando por todos os pontos. Em geral, quanto maior o R^2 melhor será o ajuste do modelo aos dados da amostra.

Na regressão linear múltipla deve-se calcular o coeficiente de determinação ajustado pelo número de variáveis no modelo e pelo tamanho da amostra. Esse coeficiente é denominado coeficiente de determinação ajustado, é denotado por R_a^2 e é dado pela seguinte expressão:

Coeficiente de determinação ajustado

$$R_a^2 = 1 - \frac{(n-1)}{n-(p+1)}(1-R^2), \quad R_a^2 \leq R^2 \quad (5)$$

sendo que

n = número de observações na amostra;

p = número de variáveis independentes.

Os coeficientes R^2 e R_a^2 têm interpretações similares, porém, o R_a^2 leva em conta tanto o

tamanho da amostra (n) quanto o número de parâmetros no modelo (p). O R^2_a sempre será menor que o R^2 .

3.5 Teste de Hipótese Global

Os testes apresentados na seção 3.3 testam individualmente se cada variável independente deve permanecer no modelo.

Pode-se realizar um teste de hipótese com o objetivo de verificar se pelo menos uma variável independente explica, por meio de um modelo de regressão linear, o comportamento da variável dependente. Este teste é apresentado a seguir:

Hipótese de interesse

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

H_a : pelo menos um dos parâmetros $\beta_1, \beta_2, \dots, \beta_p$ é diferente de zero

Para testar a hipótese apresentada pode-se utilizar a seguinte estatística de teste:

$$F = \left(\frac{R^2}{1-R^2} \right) \left[\frac{n-(p+1)}{p} \right]$$

em que:

n = número de observações na amostra;

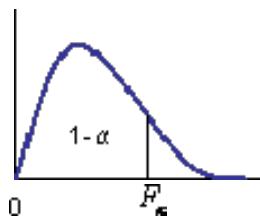
p = número parâmetros no modelo (excluindo β_0);

R^2 = coeficiente de determinação.

Esta estatística de teste possui distribuição F de Snedecor com p e n-p-1 graus de liberdade. A distribuição F de Snedecor também é conhecida como distribuição F de Fisher-Snedecor ou distribuição F de Fisher.

Seja F uma variável aleatória com distribuição F de Snedecor com p e n-p-1 graus de liberdade. O valor crítico F_α é um valor da distribuição F, tal que $P(F < F_\alpha) = 1 - \alpha$. A distribuição F de Snedecor é apresentada na Figura 3.3.

Figura 3.3 Distribuição F de Snedecor



Rejeita-se a hipótese de interesse quando $F > F_\alpha$. Sendo que a probabilidade α é o nível de significância para o teste de hipótese e F_α , é obtido em uma distribuição F de Snedecor com p e $n-p-1$ graus de liberdade. A tabela com a distribuição F de Snedecor é apresentada no Apêndice.

A hipótese de interesse, também, pode ser testada por meio da análise de variância apresentada na próxima seção.

3.6 Análise de Variância

A hipótese de que pelo menos uma variável independente explica, por meio de um modelo de regressão linear, o comportamento da variável dependente pode ser testada por meio da ANOVA.

Na Tabela 3.1 é apresentada a análise de variância (ANOVA) para um modelo de regressão linear múltipla. O valor da estatística de teste (F), apresentado na quinta coluna da tabela, deve ser comparado com o valor crítico, F_α , obtido por meio da distribuição F de Snedecor com p graus de liberdade no numerador e $n-p-1$ graus de liberdade no denominador. Sendo que α é o nível de significância para o teste de hipótese. Rejeita-se a hipótese de interesse quando $F > F_\alpha$.

Tabela 3.1 Tabela ANOVA para regressão múltipla.

| Fonte de variação | Soma de quadrados | Graus de liberdade | Quadrado médio | Razão F |
|-------------------|--|--------------------|--|--|
| Modelo | $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ | p | $\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{p}$ | $F = \frac{\text{Quadrado médio do modelo}}{\text{Quadrado médio dos resíduos}}$ |

| | | | |
|----------|------------------------------------|-------------|--|
| Resíduos | $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ | $n - p - 1$ | $\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n-p-1}$ |
| Total | $\sum_{i=1}^n (y_i - \bar{y})^2$ | $n - 1$ | |

Pressupostos da Análise de Regressão

Os principais pressupostos da análise de regressão que serão avaliados seguir são:

- Normalidade dos resíduos;
- Homocedasticidade dos resíduos;
- Ausência de autocorrelação serial dos resíduos;
- Multicolinearidade entre as variáveis independentes;
- Linearidade dos coeficientes;

Multicolinearidade

O problema da multicolinearidade existe quando dois ou mais regressores do modelo estão fortemente correlacionados ou são linearmente dependentes. Na construção de modelos de regressão, as variáveis regressoras (independentes) são assumidas como sendo independentes uma da outra. No entanto, dependendo do comportamento das variáveis independentes pode acontecer que haja algum grau de similaridade na informação fornecida por algumas variáveis. Uma maneira para verificar a multicolinearidade de maneira mais direta é construindo uma matriz de correlações para todas as variáveis do modelo. Uma forte relação linear entre as variáveis explanatórias torna imprecisa a interpretação da relação entre uma variável e a resposta porque, na presença destas relações, torna-se impossível alterar uma das variáveis , sem alterar de outra.

A partir das medidas fornecidas pelo STATISTICA vamos utilizar dois critérios para identificar a multicolinearidade: o Fator de Inflação da Variância (VIF) e a Tolerância.

O VIF é útil em parte por causa da facilidade de ser calculado e está disponível na maior parte dos pacotes estatísticos. Para construir a fórmula para o VIF primeiro definimos R_i^2 como sendo a estatística R^2 de um modelo de regressão considerando cada variável independente como sendo dependente e regredida sobre as restantes. Ou seja, esse novo R^2 é uma estatística de um modelo de regressão de X_i sobre as $p-2$ variáveis explanatórias. O VIF para a variável i é então dado por:

$$VIF_i = \frac{1}{1 - R_i^2}.$$

O VIF pode ser relacionado a variância do coeficiente estimado da seguinte forma:

$$Var[\hat{\beta}_i] = \frac{\sigma^2}{\sum x_i^2} VIF_i.$$

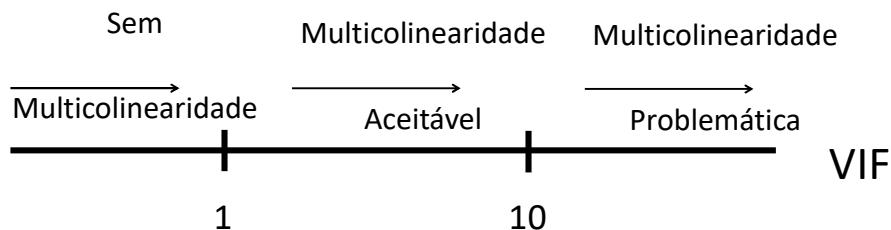
Uma outra medida também utilizada e que é construída a partir do VIF é denominada de

Tolerância é dada por:

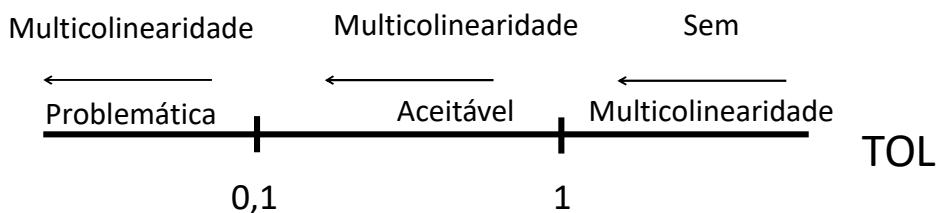
$$TOL_i = \frac{1}{VIF_i}.$$

A dificuldade de utilizar o VIF para detecção de colinearidade é que não valores de corte rigorosamente justificados. Muitos autores sugerem como pontos de corte para o VIF valores de 5 a 10 ($R^2 = 0,8$ a $0,9$; $TOL = 0,10$ a $0,20$) para determinar quando ações corretivas devem ser tomadas.

Uma regra prática para interpretar VIF pode ser obtida de Hair (1998) página 193, que considera que VIF até 1 não há multicolinearidade, de 1 até 10 a multicolinearidade é aceitável e VIF acima de 10 indica a multicolinearidade sendo um problema que ser resolvido. Também do Hair (1998) página 193 temos que uma Tolerância até 1 indica que não há multicolinearidade, Tolerância de 1 até 0,10 a multicolinearidade é aceitável e abaixo de 0,10 a multicolinearidade precisa ser corrigida. O esquema a seguir ilustra as orientações sobre como avaliar a multicolinearidade a partir das medidas VIF e Tolerância.



A direção da seta indica aumento na severidade da multicolinearidade considerando como critério a medida Fator de Inflação da Variância (VIF).



A direção da seta indica aumento na severidade da multicolinearidade considerando como critério a medida Tolerância (TOL).

No caso da regressão que estamos construindo podemos obter VIF e a Tolerância da seguinte forma:

- (i) primeiro construímos a regressão da forma como foi descrito anteriormente, de tal forma que teremos os resultados mostrados na Figura-R7;
- (ii) e selecionamos a opção *Partial correlations*.

A tabela gerada pela opção *Partial correlations* contém a medida Tolerance e R-square.

Por exemplo, na primeira linha a medida R-square da variável Educação considerou uma

regressão em que Educação é v.d. e as outras variáveis são consideradas como sendo variáveis independentes.

Na Figura-R8 verificamos que todos os valores de Tolerance estão no intervalo de 0,1 a 1, o que indica multicolinearidade aceitável nos dados. Como $TOL_i = 1/VIF_i$, temos todas as medidas de VIF concentradas no intervalo de 1 até 10, levando a mesma conclusão.

Autocorrelação

Uma das pressuposições da construção do modelo de regressão é que a correlação entre os resíduos gerados a partir das variáveis independentes seja igual a zero. O principal problema da autocorrelação é que os estimadores obtidos pelo método dos mínimos quadrados deixam de ser eficientes, ou seja, não apresentam variância mínima. Para verificar esta suposição utilizamos a estatística de Durbin-Watson em que as hipóteses testadas são:

$$H_0 : \rho = 0$$

$$H_a : \rho \neq 0.$$

A estatística de teste é dada por:

$$d = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=2}^n e_i^2},$$

Onde $e_i = y_i - \hat{y}_i$ e y_i e \hat{y}_i são, respectivamente, os valores observados e preditos para a variável resposta (v.d.) para um caso, ou indivíduo i . A medida d torna-se tanto menor quanto maior a correlação serial. Limites inferiores e superiores da estatística d , normalmente denotados por d_U e d_L , respectivamente, são tabulados para diferentes quantidades de variáveis explanatórias e n . Assim, dados d , d_U e d_L , temos as seguintes regras de decisão para o teste de autocorrelação:

Se $d < d_L$, rejeitar $H_0 : \rho = 0$.

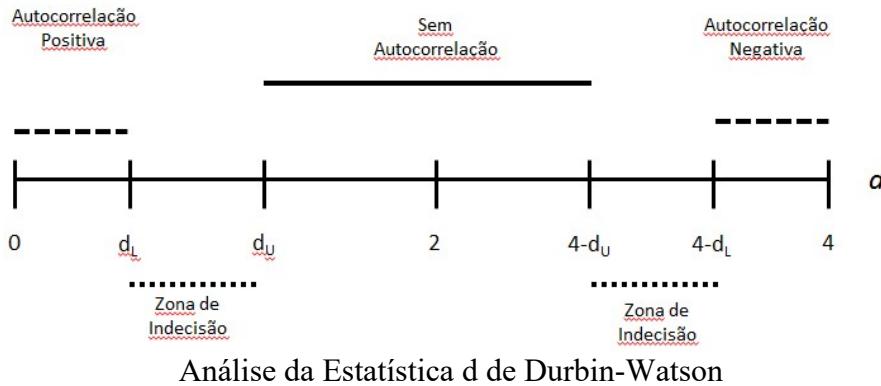
Se $d_L \leq d \leq d_U$, o teste é inconclusivo.

Se $4 - d_L < d < 4$, rejeitar $H_0 : \rho = 0$.

Se $4 - d_U \leq d \leq 4 - d_L$, o teste é inconclusivo.

Se $d_U < d < 4 - d_U$, não rejeitar $H_0 : \rho = 0$.

Normalmente, estas regras de decisão são apresentadas na forma de um esquema como o mostrado a seguir.



O STATISTICA fornece a estatística de Durbin-Watson, para obtê-la, assim que foi gerada a regressão deve-se realizar a seguinte sequencia de comandos: (i) construir a regressão, (ii) escolhe a opção Residuals/assumptions/prediction, (iii) clicar em Perform residual analysis e (iv) selecionar a opção Durbin-Watson statistic.

O resultado gerado pelo STATISTICA implica em uma estatística de Drbin-Watson igual a 1,57197. Para verificarmos se a hipótese nula é rejeitada ou não, precisamos de uma tabela que forneça os valores da estatística d_U e d_L . Nesse caso temos que:

- Tamanho da amostra, $n = 645$;
- Número de variáveis independentes, $p = 6$;
- Valor crítico inferior, $d_L = 1,85$;
- Valor crítico superior, $d_U = 1,88$;
- Estatística Durbin-Watson fornecida pelo STATISTICA, $d = 1,57197$.

Nesse caso a estatística de Durbin-Watson é menor do que 2 e está localizada na região esquerda entre 0 e d_L . Portanto, concluímos que os resíduos gerados apresentam autocorrelação positiva e isso reforça que sejam realizadas transformações nos dados e adequação do modelo.

Normalidade

Para verificar o pressuposto de que os resíduos gerados pela regressão têm distribuição normal, podemos utilizar o teste de Kolmogorov-Smirnov. Além desse teste, o STATISTICA também fornece a opção do Teste de Shapiro-Wilk.

A estatística W de Shapiro-Wilk é usada no teste de normalidade. De tal forma que, se a estatística W é significativa, a hipótese de que a respectiva distribuição é normal deve ser rejeitada. O teste W de Shapiro-Wilk é o teste de normalidade preferido por causa de suas propriedades em comparação com uma ampla gama de testes alternativos (Shapiro, Wilk, e Chen, 1968).

Dadas as opções do teste de Kolmogorov-Smirnov e Shapiro-Wilk, qual teste deve ser utilizado?

Kolmogorov-Smirnov

- Não é sensível a problemas nas caudas.

- Trabalha razoavelmente com conjuntos de dados com tamanho menor do que 50.

Shapiro-Wilk

- Não trabalha bem se muitos valores no conjunto de dados têm o mesmo valor.
- Trabalha bem para conjuntos de dados com tamanho maior do que 50, mas também pode ser utilizado em casos em que a amostra possui número de observações menor do que 30.

O STATISTICA implementa uma extensão para o teste de Shapiro-Wilk descrito por Royston (1982), o que permite que ele seja aplicado a amostras grandes (com até 2.000 observações).

Sendo assim, para realizar estes testes de normalidade no STATISTICA, seguimos a sequencia de comandos descrita a seguir.

Primeiro selecionamos a coluna de Resíduos Padronizados, depois escolhemos a opção *Basic Statistics* que permite selecionar as estatísticas descritivas em *Descriptive statistics*, tal como mostrado na Figura-R13.

Na aba *Normality*, selecionamos as opções *Kolmogorov-Smirnov & Lilliefors test for normality* e *Shapiro-Wilk's W test*. Mudamos a opção *Number of intervals* para 100. Feito isso, clicamos em *Histograms* que está na mesma janela e logo acima, tal como mostrado na Figura-R14. O resultado é mostrado na Figura-R15.

Box 6.1 Normality tests

An important (nonparametric) test for normality is the *one-sample Kolmogorov–Smirnov test*. We can use it to test whether or not a variable is normally distributed. Somewhat surprisingly, the test's null hypothesis is that the variable follows a specific distribution (e.g., the normal distribution). This means that only if the test result is insignificant, that is the null hypothesis is not rejected, can we assume that the data are drawn from the specific distribution against which it is tested. Technically, when assuming a normal distribution, the Kolmogorov–Smirnov test compares the sample scores with an artificial set of normally distributed scores that has the same mean and standard deviation as the sample data. However, this approach is known to yield biased results which are corrected for through the Lilliefors correction (1967). The Lilliefors correction considers the fact that we do not know the true mean and standard deviation of the population. An issue with the Kolmogorov–Smirnov test is that it is very sensitive when used on very large samples and often rejects the null hypothesis if very small deviations are present.

The *Shapiro–Wilk* test also tests the null hypothesis that the test variable under consideration is normally distributed. Thus, rejecting the Shapiro–Wilk test provides evidence that the variable is not normally distributed. It is best used for sample sizes of less than 50. A drawback of the Shapiro–Wilk test is that it works poorly if the variable you are testing has many identical values, in which case you should use the Kolmogorov–Smirnov test with Lilliefors correction.

To conduct the Kolmogorov–Smirnov test with Lilliefors correction and the Shapiro–Wilk test in SPSS, we have to go to ► Analyze ► Descriptive Statistics ► Explore ► Plots and choose the **Normality plots with tests** option (note that the menu option ► Analyze ► Nonparametric Tests ► Legacy Dialogs ► 1-Sample K-S will yield the standard Kolmogorov–Smirnov test whose results oftentimes diverge heavily from its counterpart with Lilliefors correction). We will discuss these tests using SPSS in this chapter's case study.

De acordo com os resultados apresentados na Figura-R15 rejeitamos a hipótese de que os resíduos gerados na regressão apresentam distribuição normal, tanto pelo teste de Kolmogorov-Smirnov, quanto pelo test de Shapiro-Wilk. Isso implica que deveremos alterar a especificação do nosso modelo e isso será possível através da transformação dos dados e análise dos *outliers*.

Homocedasticidade

Precisamos verificar outro pressuposto da regressão que afirma que a variância dos resíduos é constante para todos os valores X_i . Para isso, utilizaremos o teste desenvolvido por Mohammad Hashem Pesaran e Bahram Pesaran. As hipóteses a serem testadas são:

H_0 : Os resíduos são homocedásticos.

H_a : Os resíduos são heterocedásticos.

O teste de Pesaran-Pesaran é realizado construindo uma regressão tomando como variável dependente o quadrado dos resíduos padronizados e como variável independente o quadrado dos valores estimados padronizados. Ou seja,

$$(Erro Padronizado)^2 = \beta_0 + \beta_1 (Valores Estimados Padronizados)^2 + \varepsilon.$$

Se a estatística F da estimativa do coeficiente β_1 for significante indica a presença de heterocedasticidade. Lembre que dizer que um resultado é estatisticamente significante significa que as diferenças encontradas são grandes o suficiente para não serem atribuídas ao acaso. Podemos representar graficamente o comportamento dos resíduos conforme mostrado na Figura-R16. No primeiro caso os valores dos resíduos oscilam dentro de limites constante bem definidos, o que implica que a variância constante. Na Figura-R16b observamos uma forma típica de resíduos que não apresentam variância constante, caracterizada pela forma de cone do gráfico dos resíduos.

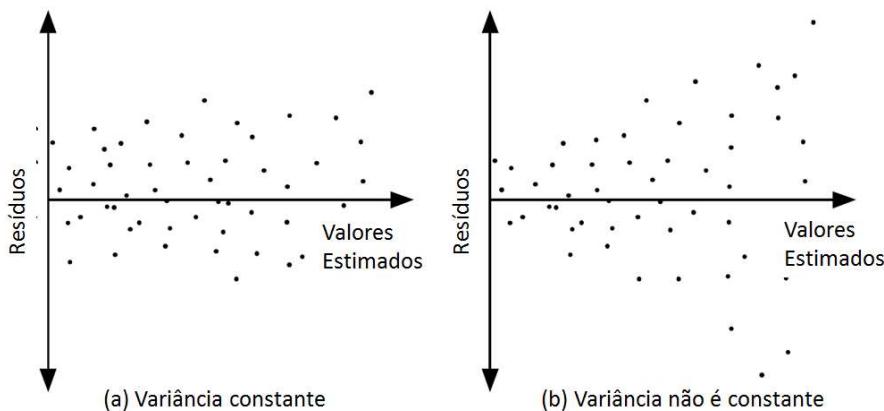


Figura-R16

De acordo com Hair (1998, página 37):

“Quando a variância dos termos de erro (ε) são constantes ao longo dos intervalos das variáveis preditoras, os dados são ditos homocedásticos. A suposição de variância igual do erro populacional E (onde E é estimado a partir de ε) é crítico para a aplicação adequada da regressão linear. Quando os termos de erro têm aumento ou variação modulante, os dados são ditos como sendo heterocedásticos.”

Uma referência para pesquisa sobre a etimologia das palavras homocedasticidade e heterocedasticidade pode ser encontrada no artigo de John Huston McCulloch publicado no volume 53, número 2 de março de 1985 na revista *Econometrica*. Este artigo também pode ser encontrado em www.ime.usp.br/~abe/lista/pdfZAptC9KazU.pdf.

Para obter do STATISTICA os resíduos padronizados e os valores estimados padronizados precisamos seguir a seguinte sequencia de comandos. Selecione a opção *Perform residual analysis* conforme mostrado na Figura-R17. Depois disso selecione a aba *Advanced* e clique na opção *Summary: Residuals & predicted*.

Essa sequencia de comandos irá gerar uma planilha com as colunas mostradas na Figura-R18. Se você for até o final na planilha vai verificar que as últimas quatro linhas contem os valores Mínimo, Máximo, Média e Mediana para cada uma das colunas. Por isso, deve tomar cuidado para não utilizar esses valores em futuros cálculos como se pertencessem aos conjuntos de dados da planilha.

Para realizar o teste de Pesaran-Pesaran precisamos construir a regressão usando como v.d. o quadrado do *Standard Residual* e como v.i. o quadrado do *Standard Pred. v.*, conforme descrito anteriormente. Para realizar essa regressão, construa uma nova *Speadsheet* e copie as colunas *Standard Residual* e *Standard Pred. v.*, conforme mostrado na Figura-R19. Vá até o final da coluna e apague as últimas 4 linhas que contêm os valores Mínimo, Máximo, Média e Mediana. Crie uma coluna com o nome coluna *Sqd Standard Residual* (para denotar resíduo padronizado ao quadrado) e uma outra coluna com o nome *Sqd Standard Pred. v.* (para denotar valor estimado padronizado), formate as colunas para *Number* e *Decimal places* igual a 5. Uma vez que as colunas foram criadas, clicamos na primeira célula da coluna *Sqd Standard Residual* e selecionamos a aba *Data* do menu principal do STATISTICA. Clique na primeira célula da coluna *Sqd Standard Residual*. Dentro da aba *Data* clicamos na opção *Transforms* que vai abrir a janela mostrada na Figura-R19 e onde devemos escrever a fórmula "*Sqd Standard Residual*"="*Standard Residual*"^2.

Depois de clicar em OK, o STATISTICA vai abrir uma janela de confirmação para o comando solicitado, tal como mostrado na Figura-R20.

Clicando em Sim, teremos todos os valores de *Sqd Standard Residual*. Não esqueça de verificar se foram apagadas as 4 últimas linhas que contêm os valores Mínimo, Máximo, Média e Mediana. A nova variável criada *Sqd Standard Residual* deve conter 645 valores. Repita o mesmo procedimento para obter *Sqd Standard Pred. v.* e depois construa a regressão tomando *Sqd Standard Residual* como variável dependente e *Sqd Standard Pred. v.* como variável independente. O resultado é mostrado na Figura-21.

Lembrando que a hipótese nula é de que os resíduos são homocedásticos e a estatística F é utilizada para tomar a decisão de rejeitar ou não H_0 , verificamos que o p-valor correspondente é dado por 0, 979525. Portanto, não rejeitamos a hipótese de que os resíduos são homocedásticos.

EXEMPLO DE REGRESSÃO MÚLTIPLA:

FAZER UM EXEMPLO COM

- DUAS V.I. PRA FAZER UM GRÁFICO 3D

- FAZER AS DERIVADAS PARA ACHAR OS ESTIMADORES DOS BETAS

PRECISO COLOCAR NO BACKGROUND AS PROPRIEDADES:

--> $\mathbf{M}^T \mathbf{N} = \mathbf{N}^T \mathbf{M}$ e $(\mathbf{M}\mathbf{N})^T \mathbf{M}\mathbf{N} = \mathbf{N}^T \mathbf{M}^T \mathbf{M}\mathbf{N}$

--> FALAR SOBRE POSTO, POSTO COMPLETO E COL

--> USAR PROPRIEDADES DE MATRIZES:

--> Como fazer Matriz ao Quadrado

--> Positiva Definida, Positiva Semidefinida

--> Variância do Produto de duas matrizes

COLOCAR AQUELA PARTE DA AULA SOBRE QUALIDADE DE ESTIMADOR:

- NÃO TENDENCIOSIDADE

- EFICIÊNCIA

- CONSISTÊNCIA

3.2 Modelos de Regressão Ridge, LASSO e Elastic Net

Para verificar se a combinação de variáveis independentes como um todo explica bem a variável dependente em um modelo linear, podemos construir um gráfico de dispersão entre os valores previstos (eixo X) e os resíduos (eixo Y) do modelo.

Então adicionamos uma função "Loess" ao gráfico de dispersão, o que nos dá uma melhor indicação da relação das variáveis que estamos procurando. Verificamos se a linha de Loess se move em torno da linha zero. Se isso acontecer, podemos assumir que temos um bom modelo linear. Se a linha de Loess mostrar um forte desvio da linha zero, temos que rever os livros novamente.

Livro ótimo 195

A suposição de linearidade assume que a relação entre a variável dependente (Y) e a(s) variável(is) independente(s) (X) é linear. Em outras palavras, mudanças em X devem resultar em mudanças constantes e proporcionais em Y. Para verificar essa suposição, você pode criar gráficos de dispersão e avaliar a linearidade dos dados.

- (i) linearidade e aditividade da relação entre variáveis dependentes e independentes:
 - (a) O valor esperado da variável dependente é uma função de linha reta de cada variável independente, mantendo as outras fixas.
 - (b) A inclinação dessa linha não depende dos valores das outras variáveis.
 - (c) Os efeitos de diferentes variáveis independentes no valor esperado da variável dependente são aditivos.

Idealmente, o gráfico residual não mostrará nenhum padrão ajustado. Ou seja, a linha vermelha deve ser aproximadamente horizontal em zero. A presença de um padrão pode indicar um problema com algum aspecto do modelo linear.

Regressão Auxiliar: Execute uma regressão OLS auxiliar dos resíduos calculados nas variáveis independentes originais e nos resíduos defasados.

Comandos do R

Causalidade e Regressão.

3.3 Modelo de Regressão usando Neurônio Único

O neurônio artificial constitui-se na unidade estrutural e funcional de um sistema de aprendizado de máquina (em inglês, Machine Learning - ML), apresenta uma complexidade inerente às tarefas que precisa executar, sendo sua formulação matemática sujeita às regras de implementação e organização de conjuntos interconectados de componentes. A viabilidade e as aplicações de uma rede neural com neurônio único mostraram-se uma tarefa desafiadora desde o início.

Sendo o início constituído por uma sequência de trabalhos seminais, o primeiro sendo a construção teórica do neurônio artificial, em 1943 com o trabalho de McCulloch e Pitts (*A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5 (4), 115-133*), a seguir, em 1949 é apresentada a aprendizagem Hebbiana (*The organization of behavior: A neuropsychological theory. New York, Wiley*) e o Perceptron de Rosenblatt em 1958 sendo a evolução seguinte (*The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65 (6), 386*).

Porém, em 1969, Papert e Minsky (*Perceptrons: An Introduction to Computational Geometry*) destacaram algumas limitações computacionais do perceptron de camada única. Isso levou muitos na comunidade de Inteligência Artificial (IA) a acreditar que as estruturas das redes neurais eram demasiado limitadas para lidar com problemas mais complexos do mundo real. Isso levou ao que é considerado o “Primeiro Inverno da Inteligência Artificial”, iniciando no final da década de 1960 e adentrando à década de 1970 em que as soluções apresentadas nessa área caracterizavam-se por soluções a problemas específicos e pouco uso comercial das pesquisas acadêmicas. Apesar do impacto gerado pelas críticas teóricas aos primeiros modelos matemáticos de representação do neurônio, outros projetos seguiram adiante evidenciando outras abordagens e a necessidade de sistemas inteligentes.

Dessa forma, vamos considerar o processo de aprendizagem para um modelo de neurônio único utilizando como algoritmo de treinamento o gradiente descendente.

Modelo de neurônio único: consiste em único neurônio ou nó, conforme mostrado na Figura-1. Esse tipo mais simples de rede neural, é caracterizado pela soma das multiplicações ponderadas dos respectivos sinais de entrada pelos pesos sinápticos, adicionados a um termo de bias (cujo efeito é aumentar ou diminuir a entrada líquida para o bloco elementar seguinte), passando por uma função de ativação (que determina a faixa de valores de ativação do neurônio artificial).

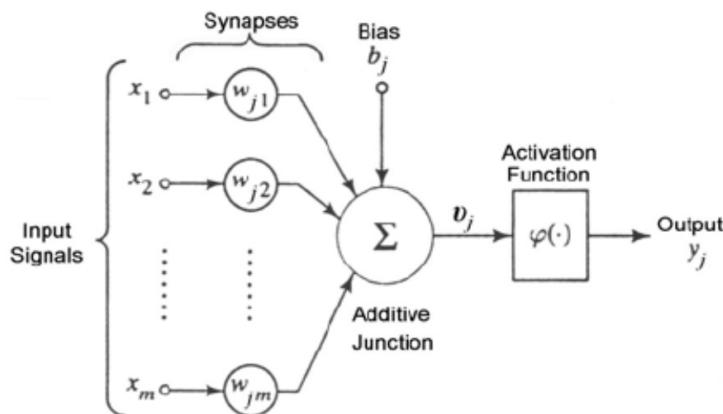


Figura-1 Modelo de Neurônio Único. Fonte: (<https://www.researchgate.net/figure/Artificial-Neuron-models->

and-its-parts-Source-Adapted-from-Haykin-1994_fig2_229036664, que é o Haykin na Página 36).

Um modelo de rede neural artificial (RNA) constituída por um único neurônio pode ser aplicado basicamente para duas tarefas: (i) **Regressão**, onde fazemos a predição de uma variável de interesse. Nesse caso, utilizamos uma função de ativação linear, ou seja, a função identidade; (ii) **Classificação**, em que é implementado um classificador binário, sendo as saídas exibindo rótulos que podem pertencer a uma de apenas duas classes. Para a tarefa de classificação, usaremos a função sigmóide.

De maneira simplificada, um neurônio único executa as operações de processamento mostradas nas equações a seguir.

$$u = \mathbf{w}^T \mathbf{x}, \quad (1)$$

$$v = \mathbf{w}^T \mathbf{x} + w_0, \quad (2)$$

$$y = f(v). \quad (3)$$

Como vimos anteriormente, um Modelo de Regressão Linear Múltipla corresponde a um modelo funcional em que procuramos explicar uma variável y a partir de algumas variáveis x , podendo ser escrito como:

$$y = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_m x_m + \varepsilon. \quad (4)$$

Onde,

- y é a variável resposta;
- w_0 , nos modelos de regressão é chamado de intercepto, é o valor de y quando a todas as outras variáveis x é atribuído o valor 0;
- w_i , corresponde ao i -ésimo coeficiente associado à variável explicativa x_i ;
- x_i , i -ésima variável explicativa;
- ε , é um termo de erro aleatório.

A equivalência do modelo de regressão com duas variáveis explicativas, em termos de um neurônio, pode ser conseguida utilizando a arquitetura mostrada na Figura-2.

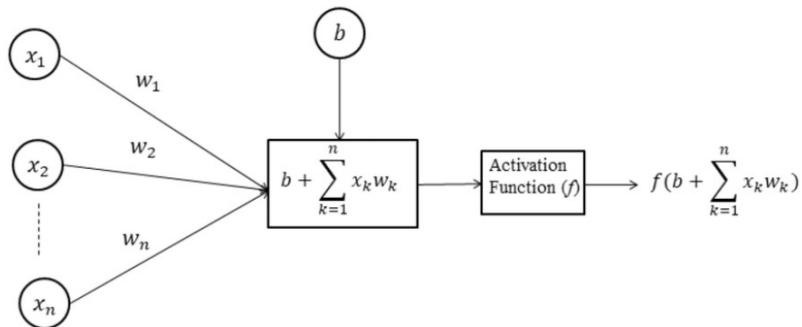


Figura-2 Neurônio único que gera um modelo de regressão. [ADAPTAR A FIGURA]

A seguir é apresentado um exemplo de código em Python que calcula a saída y a partir das entradas x_1 e x_2 na rede neural artificial apresentada na Figura-2.

EXEMPLO 12.1 Código Python para obter a saída do neurônio na Figura-2

Antes de escrever o código em Python, temos que:

$$u = w_1x_1 + w_2x_2. \quad (5)$$

Depois disso é adicionado o bias, produzindo:

$$v = w_0 + w_1x_1 + w_2x_2. \quad (6)$$

Aplicando a função de ativação identidade, resulta em:

$$y = f(v). \quad (7)$$

Considerando como sinais de entrada $x_1 = 3$, $x_2 = 5$, pesos das respectivas conexões sinápticas $w_1 = 1$ e $w_2 = 2$ e o bias $w_0 = -3$, o código Python para obter a saída y na Figura-2 é mostrado a seguir.

```
def neuronio_unico(w, w_0, x):
    u = 0
    v = 0
    for entrada_x, peso_w in zip(x, w):
        u += entrada_x * peso_w
    v = u + w_0
    y = linear(v)
    return y

# Considerando como entrada x1 = 3 e x2 = 5, com pesos iniciais para w iguais a
# w1 = 1 e w2 = 2
x = [3, 5]
w = [1, 2]
w_0 = -3

y = neuronio_unico(w, w_0, x)
print(y)
```

O que resulta na saída $y = 10$.

12.1.1 Modelo de Regressão e Gradiente Descendente

Utilizamos o método do gradiente descendente para determinar os pesos w_j e o bias w_0 que minimizam uma função de custo do modelo. Essa função de custo depende dos parâmetros que serão estimados, denotaremos por $C(\hat{y}^{(i)}, y^{(i)}) = C(w_0, w_1, \dots, w_m)$ e calculada como:

$$C(\hat{y}^{(i)}, y^{(i)}) = \frac{1}{2n} \sum_{i=1}^n (\hat{y}_{\text{predito}}^{(i)} - y_{\text{observado}}^{(i)})^2 = \frac{1}{2n} \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)})^2, \quad (8)$$

onde n corresponde ao número de observações consideradas, e está multiplicado por 2 por conveniência devido as derivadas que serão tomadas *a posteriori*. Em termos da linguagem de métodos de otimização podemos escrever:

função : $f_w(x) = w_0 + w_1x_1 + \dots + w_mx_m$

parâmetros : w_0, w_1, \dots, w_m

função custo : $C(w_0, w_1, \dots, w_m; \hat{y}^{(i)}, y^{(i)}) = \frac{1}{2n} \sum_{i=1}^n [f_w(x^{(i)}) - y^{(i)}]^2$ (9)

objetivo : $\underset{w_0, w_1, \dots, w_m}{\text{minimizar}} C(w_0, w_1, \dots, w_m)$

Em termos de nomenclatura vamos utilizar a terminologia do software Python, onde variáveis de uma função são chamadas de parâmetros ou argumentos, sendo que: (i) **funções:** são blocos de código que executam operações ou tarefas específicas. Ao definir uma função (ou regra), os programadores podem evitar a repetição de código, chamando funções previamente definidas; (ii) **parâmetros ou params:** são variáveis que ficam disponíveis entre parênteses de uma função definida e (iii) **argumentos ou args:** são os valores aribuídos a uma função quando ela é chamada. O termo parâmetro usado no Python não deve ser confundido com pesos ou coeficientes de modelos de aprendizado de máquina. Ou seja, por exemplo, definimos uma função $g(x)$ com um parâmetro x . O corpo da função verifica qual condição é atendida para executar o bloco de código correspondente. Avaliando a função $g(x)$ no valor 9, implica que esse valor 9 é o argumento da função. A função $g(x)$ é chamada e retorna o resultado.

Para implementar o método do gradiente descendente podemos utilizar o pseudocódigo mostrado a seguir aplicado ao treinamento do neurônio único da Figura-3.

Pseudocódigo do algoritmo da descida do gradiente

```

Iniciar os pesos  $w_0, w_1, \dots, w_m$ 
Setar o número de épocas  $num\_epocas$ 
Setar a taxa de aprendizagem  $\alpha$ 
  for  $k = 0$  to  $num\_epocas$  do
    Avalie  $grad^{(k)} = \nabla L(w_0^{(k)}, w_1^{(k)}, \dots, w_m^{(k)})$ 
     $w^{(k+1)} = w^{(k)} - \alpha \cdot grad^{(k)}$ 
  end for

```

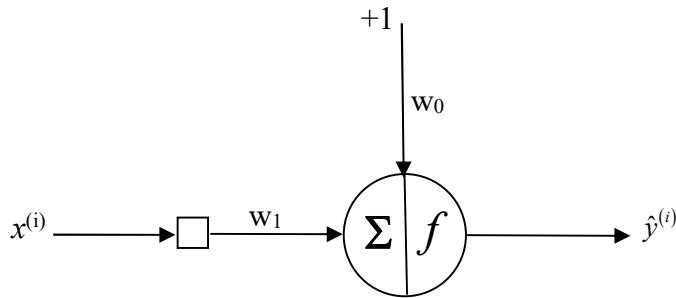


Figura-3 Neurônio único com uma entrada para modelo de regressão.

No caso do neurônio mostrado na Figura-3, executamos o treinamento utilizando o seguinte conjunto de equações:

$$v^{(i)} = w_0 + w_1 x^{(i)}. \quad (6)$$

$$\hat{y}^{(i)} = f(v^{(i)}), \quad (6)$$

sendo f , a função de ativação linear, temos que:

$$\hat{y}^{(i)} = w_0 + w_1 x^{(i)}. \quad (6)$$

O valor predito, que corresponde à saída do neurônio, tem sobreescrito (i) porque para cada valor do sinal de entrada x será gerado um valor de saída. A notação sobreescrita (i) indica qual das n observações de dados estamos considerando. Como a saída não será igual à entrada, essa diferença é um custo associado ao modelo que representa a variável y .

Quando consideramos somente um valor do sinal de entrada, digamos o valor (i) do sinal de entrada, o custo C associado a essa entrada será de

$$C(\hat{y}^{(i)}, y^{(i)}) = \frac{1}{2} [\hat{y}^{(i)} - y^{(i)}]^2. \quad (8)$$

Porém, é importante ressaltar que C é uma função que depende dos parâmetros que correspondem aos pesos sinápticos e ao bias da rede. Nesse caso, nossa função de custo é do tipo $C(w_0, w_1; \hat{y}^{(i)}, y^{(i)})$.

Como o sinal de entrada é constituído por n valores, a função custo que será calculada a cada época de treinamento é dada por

$$C(w_0, w_1; \hat{y}^{(i)}, y^{(i)}) = \frac{1}{2n} \sum_{i=1}^n [\hat{y}^{(i)} - y^{(i)}]^2. \quad (8)$$

Assim, temos uma função de custo calculada para toda a mostra de treinamento que é descrita em termos das diferenças (ou seja, os erros) de cada valor predito pelo modelo e o respectivo valor da variável de interesse y . Como queremos minimizar essa soma de erros, precisamos atualizar os parâmetros conforme alguma regra, que nesse caso será a descida do gradiente. Dessa forma, os pesos são atualizados através das seguintes regras:

$$\begin{aligned} w_0 &\leftarrow w_0 - \alpha \frac{\partial C(w_0, w_1; \hat{y}^{(i)}, y^{(i)})}{\partial w_0}, \\ w_1 &\leftarrow w_1 - \alpha \frac{\partial C(w_0, w_1; \hat{y}^{(i)}, y^{(i)})}{\partial w_1}. \end{aligned} \quad (13)$$

Sendo o parâmetro α usado para regular o tamanho do passo da descida do gradiente e denominado de taxa de aprendizagem.

A determinação das derivadas de C com relação a w_0 e w_1 é apresentada a seguir.

$$\frac{\partial C(w_0, w_1; \hat{y}^{(i)}, y^{(i)})}{\partial w_0} = \frac{\partial}{\partial w_0} \left\{ \frac{1}{2n} \sum_{i=1}^n [w_0 + w_1 x^{(i)} - y^{(i)}]^2 \right\} = \frac{1}{n} \sum_{i=1}^n [w_0 + w_1 x^{(i)} - y^{(i)}]. \quad (14)$$

$$\frac{\partial C(w_0, w_1; \hat{y}^{(i)}, y^{(i)})}{\partial w_1} = \frac{\partial}{\partial w_1} \left\{ \frac{1}{2n} \sum_{i=1}^n [w_0 + w_1 x^{(i)} - y^{(i)}]^2 \right\} = \frac{1}{n} \sum_{i=1}^n [w_0 + w_1 x^{(i)} - y^{(i)}] x^{(i)}. \quad (15)$$

Para continuar o processo de treinamento podemos basicamente utilizar duas estratégias, a primeira considerando que em cada época calculamos o gradiente total, ou seja, a soma de todos os valores de gradiente para cada valor de sinal de entrada, o que é denominado de aprendizado por lote (batch gradient descent).

EXEMPLO 12.X Modelo de Regressão Linear Simples

O conceito básico do aprendizado baseado na descida do gradiente, considera que estamos atualizando cada peso ao longo do gradiente da função de custo em relação a esse peso, para ajudar a rede neural a predizer de forma mais precisa.

Para comparar o Modelo de Regressão Linear com o Modelo do Neurônio Único sendo usado para a tarefa de regressão, vamos adaptar o exemplo de Mendenhall e Sincich [*William M. Mendenhall and Terry L. Sincich, Statistics for Engineering and The Sciences, Sixth Edition, CRC Press, 2016, página 486*]. Nesse caso, temos um conjunto de observações em que para a variável dependente y foram anotados os valores $\{1, 1, 2, 2, 4\}$ e para a variável idenpendente x temos os valores $\{1, 2, 3, 4, 5\}$. Na Tabela-1 temos a apresentação dos dados.

| n | x | y |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 2 | 1 |
| 3 | 3 | 2 |
| 4 | 4 | 2 |
| 5 | 5 | 4 |

Tabela-1 Variáveis x e y .

```

#####
# REGRESSÃO USANDO NEURÔNIO ÚNICO [RNA]
#
#####

import numpy as np
import matplotlib.pyplot as plt

# Definição do Modelo Linear
def modelo_linear(w, w_0, x):
    v = np.dot(w, x) + w_0
    return v

# Treinamento do Modelo de Regressão Linear
def treino_modelo_regressao(X, Y, alfa, epocas, imprimir):
    # Inicializar os pesos
    w_0 = 0
    w = np.zeros(len(X[0]))
    rmse = [] # Armazenar a medida de ajuste para cada época
    for epoca in range(epocas):
        custo_total = 0
        for x, y in zip(X, Y):
            v = modelo_linear(w, w_0, x)
            erro = v - y
            custo_total += (erro ** 2) / 2
        # Atualizar os Pesos
        w_0 -= alfa * erro * 1
        w -= alfa * erro * x.flatten()
        if imprimir:
            print("x:", x, "y:", y, "Diferença ou Desvio:", erro)
            print("Pesos:", w, w_0)
    acuracia = np.sqrt(custo_total / len(X)) # Calcula: Root Mean Squared Error [RMSE]
    rmse.append(acuracia)

    # Mostrar a Alteração da Função Custo a Medida que as Épocas Passam
    mostrar_a_cada = max(1, epocas // 10)
    if epoca % mostrar_a_cada == 0:
        print("Época", epoca, "Custo Total", custo_total)
    return w, w_0, rmse

# Exemplo [Mendenhall e Sincich, 2016, Pág. 486]
x_entrada = np.array([[1], [2], [3], [4], [5]])
y_saida = np.array([1, 1, 2, 2, 4])
alfa = 0.01
epocas = 1000
imprimir = False

# Treinando o Modelo
w, w_0, rmse = treino_modelo_regressao(x_entrada, y_saida, alfa, epocas, imprimir)
print("\nPesos Finais:")
print(w, w_0)
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(15, 5))

# Subplot: RMSE por Épocas
epocas = range(1, len(rmse) + 1)
ax1.plot(epocas, rmse, marker='o', linestyle='-', color='black', label='Root Mean Squared Error')
ax1.set_title('RMSE por Épocas')
ax1.set_xlabel('Épocas')
ax1.set_ylabel('RMSE')
ax1.legend()
ax1.grid(True)

# Plot do Ajuste Linear
def plot_ajuste_linear_rna(X, Y, w, w_0):
    ax2.scatter(X.flatten(), Y, color='black')
    ax2.set_xlabel('x')
    ax2.set_ylabel('y')
    ax2.set_title('Ajuste Linear - Neurônio Único')

```

```

    ax2.plot(X.flatten(), w * X.flatten() + w_0, color='black')
    ax2.grid(True)
plot_ajuste_linear_rna(x_entrada, y_saida, w, w_0)

#####
#          REGRESSÃO LINEAR SIMPLES [RLS]
#
#####

import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt

# Observações
x = np.array([1, 2, 3, 4, 5])
y = np.array([1, 1, 2, 2, 4])

# Adicionando o intercepto
x = sm.add_constant(x)

# Ajuste da Regressão usando o modelo OLS (Ordinary Least Squares - Mínimos Quadrados Ordinários)
modelo = sm.OLS(y, x).fit()

# Obtenção dos Valores Ajustados (Predições)
y_preditto = modelo.predict(x)

# Mostrando os Resultados da Regressão pelo OLS
print("Resultados da Regressão:")
print(modelo.summary())

# Inclinação e Intercepto
slope = modelo.params[1]
intercept = modelo.params[0]
fig, (ax3, ax4) = plt.subplots(1, 2, figsize=(15, 5))

# Plotando os Dados Originais e a Linha Ajustada
ax3.scatter(x[:, 1], y, label='Dados Originais - Observações')
ax3.plot(x[:, 1], y_preditto, color='red', label=f'Modelo Ajustado: y = {intercept:.2f} + {slope:.2f}x')
ax3.set_xlabel('x')
ax3.set_ylabel('y')
ax3.set_title('Ajuste Linear - Regressão Simples')
ax3.legend()
ax3.grid(True)

# Plot dos Ajustes da RNA e da RLS
ax4.scatter(x[:, 1], y, label='Dados Originais - Observações')
ax4.plot(x[:, 1], y_preditto, color='red', label=f'Modelo Ajustado RLS')
ax4.plot(x_entrada.flatten(), w * x_entrada.flatten() + w_0, color='black', label=f'Modelo Ajustado RNA')
ax4.set_xlabel('x')
ax4.set_ylabel('y')
ax4.set_title('Ajustes da RNA e da RLS')
ax4.legend()
ax4.grid(True)

```

```

    ➤ Época 0 Custo Total 9.158778596611885
    ➤ Época 100 Custo Total 0.6634318475621844
    ➤ Época 200 Custo Total 0.656325677470472
    ➤ Época 300 Custo Total 0.6555593631366254
    ➤ Época 400 Custo Total 0.6556219112491557
    ➤ Época 500 Custo Total 0.6557144536033169
    ➤ Época 600 Custo Total 0.6557646048141427
    ➤ Época 700 Custo Total 0.6557875296180113
    ➤ Época 800 Custo Total 0.6557974395945291
    ➤ Época 900 Custo Total 0.6558016331743399

    Pesos Finais:
    [0.73684377] -0.17576521601331777

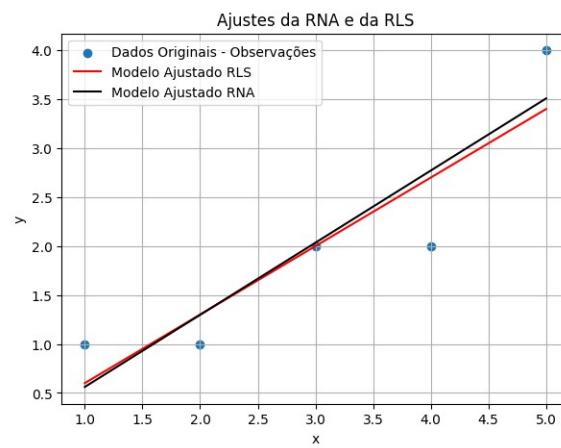
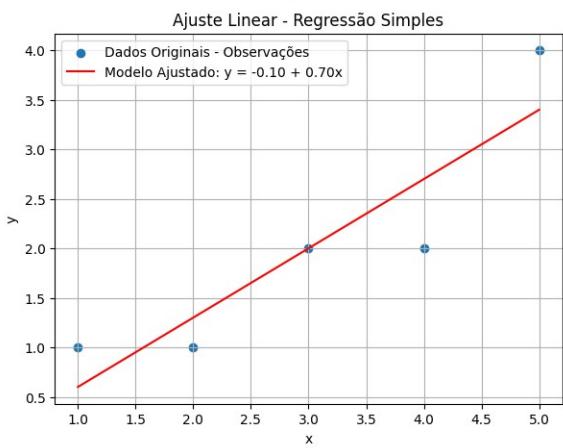
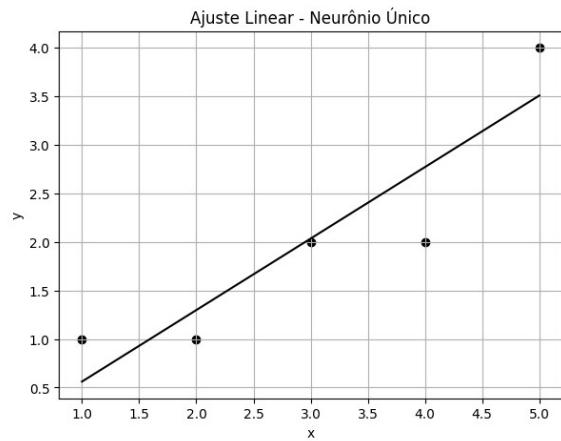
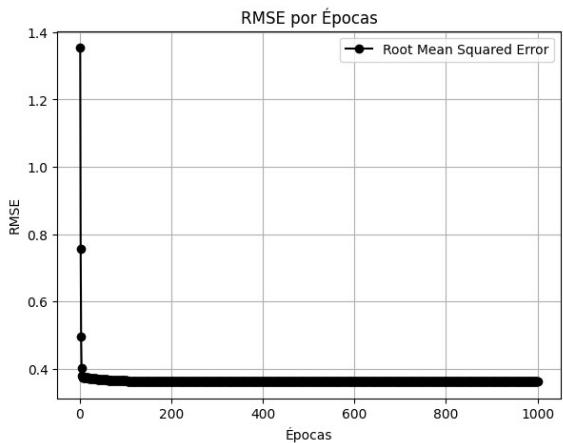
```

Figura-4 Custo total a medida que são geradas as épocas.

```

OLS Regression Results
=====
Dep. Variable:                      y      R-squared:                 0.817
Model:                            OLS      Adj. R-squared:            0.756
Method:                           Least Squares      F-statistic:             13.36
Date:                     Tue, 09 Jan 2024      Prob (F-statistic):       0.0354
Time:                         12:08:02      Log-Likelihood:          -3.3094
No. Observations:                  5      AIC:                   10.62
Df Residuals:                      3      BIC:                   9.838
Df Model:                           1
Covariance Type:                nonrobust
=====
      coef    std err        t      P>|t|      [0.025      0.975]
-----
const     -0.1000     0.635     -0.157     0.885     -2.121     1.921
x1        0.7000     0.191      3.656     0.035      0.091     1.309
=====
Omnibus:                          nan      Durbin-Watson:           2.509
Prob(Omnibus):                    nan      Jarque-Bera (JB):        0.396
Skew:                             -0.174      Prob(JB):                  0.821
Kurtosis:                          1.667      Cond. No.                   8.37
=====
```

Figura-5 Resultado do modelo de regressão linear simples.



3.4 Modelo de Classificação usando Neurônio Único

Como mencionado anteriormente, o aprendizado supervisionado é um tipo de aprendizado de máquina, implementado a partir de modelos matemáticos, em que você treina um algoritmo em um conjunto de dados rotulado. Dados rotulados significam que os dados de entrada que você usa para treinamento são emparelhados com a saída correta, essencialmente ensinando o algoritmo a aprender a relação entre a entrada (características, features) e a saída desejada (rótulos).

O algoritmo tenta aprender a função de mapeamento da entrada para a saída, generalizando a partir dos dados de treinamento rotulados. Então, quando recebe dados de entrada novos, dados ainda não vistos, ele pode predizer ou gerar a saída correta com base no que aprendeu durante a fase de treinamento. Os modelos preditivos podem ser classificados em dois grupos principais: (i) análise de regressão e (ii) modelos de classificação, usados predizer a classe (ou grupo) de dados.

FALAR QUE O PYTORCH USA O Gradient descent with negative log-likelihood
(NLL) loss

Para o problema de classificação, precisamos implementar apenas algumas alterações. Primeiro, alteramos a função de ativação para sigmóide; isso reduz a pré-ativação z a uma ativação (saída) que está entre 0 e 1. Também precisamos implementar o cálculo do gradiente, com esta função de ativação sigmóide.

Finalmente, para o problema de classificação, usaremos uma função de perda diferente: a perda de log-verossimilhança negativa (NLL). Isso permitirá que nosso código de treinamento anterior ainda funcione, mas para o problema de classificação e não de regressão.

REFERÊNCIAS

Angrist, J. D., & Pischke, J. S. (2008). **Mostly Harmless Econometrics: An Empiricist's Companion**. Princeton University Press.

Freedman, D. A. (2009). **Statistical Models: Theory and Practice**. Cambridge University Press.

Harrell, F. E. (2015). **Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis**. Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. Springer.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). **Introduction to Linear Regression Analysis**. Wiley.

Pearl, J. (2009). **Causality: Models, Reasoning and Inference**. Cambridge University Press.

Tibshirani, R. (1996). **Regression Shrinkage and Selection via the Lasso**. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Wooldridge, J. M. (2016). **Introductory Econometrics: A Modern Approach**. Cengage Learning.

Neter *et al.* (1983)

GUJARATI, D. N.. Econometria básica. 5 ed.. Porto Alegre. Bookman, 2011.

Multiple Regression in Hydrology /Cat No Mrh Capa comum – 1 dezembro 1985

Edição Inglês por [R. Holder](#) (Autor)

What Every Computer Scientist Should Know About Floating-Point Arithmetic,
David Goldberg (1991)

ACM Computing Surveys, Vol 23, No 1, March 1991

Econometric Methods with Applications in Business and Economics 1st Edition
by Christiaan Heij (Author), Paul de Boer (Author), Philip Hans Franses (Author), Teun Kloek (Author), & 1 more

(Johnson e Wichern, 1998).

Investimentos Capa comum – 21 agosto 2014
Edição Português por Zvi Bodie (Autor), Alex Kane (Autor), Alan Marcus (Autor),

MARKOWITZ, H. Portfolio Selection. *The Journal of Finance*, v.7, n. 1, p. 77-91, 1952.

SHARPE, W. F. Capital Asset Market Prices: a Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance*, v. 19, n. 3, p. 425–442, 1964.

Lintner, John (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets, *Review of Economics and Statistics*, 47 (1), 13-37.

Mossin, Jan. (1966). Equilibrium in a Capital Asset Market, *Econometrica*, Vol. 34, No. 4, pp. 768–783.

Cochran, W. G. (April 1934). "The distribution of quadratic forms in a normal system, with applications to the analysis of covariance".

Inferência Estatística, 2018 (3a. Reimpressão)
Edição Português por George Casella (Autor), Roger L. Berger (Autor)

An Introduction to Probability Theory and Its Applications, Volume 2: 81 Capa comum – 8 janeiro 1991
Edição Inglês por William Feller (Autor),

Probability, 1993 [MAURI: USAR 1993]
Edição Inglês por Jim Pitman (Autor)

Linear Regression Analysis: Theory and Computing Capa dura – Ilustrado, 1 julho 2009
Edição Inglês por Xin Yan (Autor), Xiaogang Su (Autor)

Gujarati, D. N., & Porter, D. C. (2009). *Basic Econometrics* (5th ed.). McGraw-Hill.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to Linear Regression Analysis* (6th ed.). Wiley.

Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach* (5th ed.). South-Western Cengage Learning.

- Draper, N., & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). Wiley.
- Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied Linear Regression Models* (4th ed.). McGraw-Hill.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.
- Weisberg, S. (2005). *Applied Linear Regression* (3rd ed.). Wiley.

The Signal and the Noise: Why So Many Predictions Fail-but Some Don't (English Edition) 1ª Edição

edição Inglês por [Nate Silver](#) (Autor)

Editora : Penguin Books; 1ª edição (27 setembro 2012)

Linear Models with R (Chapman & Hall/CRC Texts in Statistical Science) (English Edition) 2nd Edição, eBook Kindle

Edição Inglês por [Julian J. Faraway](#) (Autor)

Editora : Chapman and Hall/CRC; 2º edição (19 abril 2016)

- Gujarati, D. N., & Porter, D. C. (2009). *Basic Econometrics*. McGraw-Hill Education.
- Wooldridge, J. M. (2020). *Introductory Econometrics: A Modern Approach*. Cengage Learning.

Applied Linear Statistical Models 5ed (Pb 2013) Capa comum – 1 janeiro 2013

Edição Inglês por [KUTNER](#) (Autor)

Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47, pp. 987–1007.

Z. W. Birnbaum and Fred H. Tingey (1951). One-sided confidence contours for probability distribution functions. *The Annals of Mathematical Statistics*, 22/4, 592–596. doi:10.1214/aoms/1177729550.

William J. Conover (1971). Practical Nonparametric Statistics. New York: John Wiley & Sons. Pages 295–301 (one-sample Kolmogorov test), 309–314 (two-sample Smirnov test).

Durbin, J. (1973). Distribution theory for tests based on the sample distribution function. SIAM.

W. Feller (1948). On the Kolmogorov-Smirnov limit theorems for empirical distributions. *The Annals of Mathematical Statistics*, 19(2), 177–189. doi:10.1214/aoms/1177730243.

X.

George Marsaglia, Wai Wan Tsang and Jingbo Wang (2003). Evaluating Kolmogorov's distribution. *Journal of Statistical Software*, 8/18. doi:10.18637/jss.v008.i18.

Gunar Schröer (1991). Computergestützte statistische Inferenz am Beispiel der Kolmogorov-Smirnov Tests. Diplomarbeit Universität Osnabrück.

Gunar Schröer and Dietrich Trenkler (1995). Exact and Randomization Distributions of Kolmogorov-Smirnov Tests for Two or Three Samples. *Computational Statistics & Data Analysis*, 20(2), 185–202. doi:10.1016/0167-9473(94)00040-P.

Thomas Viehmann (2021). Numerically more stable computation of the p-values for the two-sample Kolmogorov-Smirnov test. <https://arxiv.org/abs/2102.08037>.

Patrick Royston (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, 31, 115–124. doi:10.2307/2347973.

Patrick Royston (1982). Algorithm AS 181: The W test for Normality. *Applied Statistics*, 31, 176–180. doi:10.2307/2347986.

Patrick Royston (1995). Remark AS R94: A remark on Algorithm AS 181: The W test for normality. *Applied Statistics*, 44, 547–551. doi:10.2307/2986146.

Royston, P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and computing*, 2(3), 117-119.

Blom, G. (1958), Statistical estimates and transformed beta variables, New York: John Wiley and Sons

Engle [Prêmio Nobel de Economia de 2003] com título: *Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation* [*Econometrica*, 50, pp. 987–1008].

O outro é de 1986 e foi publicado pelo Bollerslev: *Generalized Autoregressive Conditional Heteroskedasticity* [*Journal of Econometrics*, 31, pp. 307–327].

McCulloch publicou na Econometrica [Vol. 53, No. 2]

Breusch, T. S.; Pagan, A. R. (1979). "A Simple Test for Heteroskedasticity and Random Coefficient Variation". *Econometrica*. 47 (5): 1287–1294. doi:10.2307/1911963. JSTOR 1911963. MR 0545960.

Koenker, R. (1981). "A note on studentizing a test for heteroskedasticity". *Journal of Econometrics*. 17 (1): 107–112. doi:10.1016/0304-4076(81)90062-2.

[1] AITCHISON, J., AND S. D. SILVEY: "Maximum-Likelihood Estimation and Associated Tests of Significance," *Journal of the Royal Statistical Society, Series B*, 22 (1960), 154–171.

1973_a

[18] RAO, C. R.: *Linear Statistical Inference and Its Applications*, 2nd Ed. New York: John Wiley and Sons, 1973.

1973_b

[18] RAO, C. R.: *Linear Statistical Inference and Its Applications*, 2nd Ed. New York: John Wiley and Sons, 1973.

A History of Probability and Statistics and Their Applications before 1750

Por Anders Hald

William Petty, Political Arithmetick (London: Robert Clavel, 1691). Disponível em [[http://babel.hathitrust.org/cgi/pt?id=uc1.\\$b666115;view=1up;seq=7](http://babel.hathitrust.org/cgi/pt?id=uc1.$b666115;view=1up;seq=7)]

[Political Arithmetic: Simon Kuznets and the Empirical Tradition in Economics (National Bureau of Economic Research... by Robert William Fogel, Enid M. Fogel, Mark Guglielmo and Nathaniel Grotte (Apr 15, 2013)]

John A. Taylor, British Empiricism and Early Political Economy: Gregory King's 1696 Estimates of National Wealth and Population (Westport, Conn. 2005).

[William Petty: And the Ambitions of Political Arithmetic by Ted McCormick (Feb 8, 2010)]

Davenant, C. 1698. *Discourses on the Publick Revenues and on the Trade of England*, Vol. 1. London.

Bulhões de Carvalho, 2007

Gerald Caiden, *The Dynamics of Public Administration: Guidelines to Current Transformation in Theory and Practice*, Holt, Rinehart and Winston, New York, 1971, p. 25. 2. FM Marx (ed), *Elements of Public Administration*, Prentice-Hall of Indi&em visualiza&ao disponivel para esta página. [Comprar este livro.](#)

Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Thomson Wadsworth.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, 68, 155-165.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn and Bacon.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.

Statistics by Example

| | | |
|-------------------------------|------------|------------|
| Terry Sincich | B0034Q6O2A | 0024109819 |
|-------------------------------|------------|------------|

J. McCulloch has an article in *Econometrica* (1985)

Econometrica, Vol. 53, No. 2 (March, 1985)

MISCELLANEA
ON HETEROS*EDASTICITY
BY J. HUSTON MCCULLOCH¹

Applied Statistics for Public and Nonprofit Administration

Por Kenneth Meier, Jeffrey Bradney, John Bohte

Durbin, J. and Watson, G.S., "Testing for Serial Correlation in Least Squares Regression I", *Biometrika*, Vol. 37, 1950, pp. 409-428.

Durbin, J. and Watson, G.S., "Testing for Serial Correlation in Least Squares Regression II", *Biometrika*, Vol. 38, 1951, pp. 159-178.

Harvey, A.C., *The Econometric Analysis of Time Series*, Second Edition, MIT Press, 1990.

Savin, N.E. and White, K.J., "The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors", *Econometrica*, Vol. 45, 1977, pp. 1989-1996.

Theil, H., *Principles of Econometrics*, Wiley, 1971.

- Pesaran M.H.(1987) "The limits to Rational Expectations", Oxford, Basil Blackwell
- Pesaran M.H., Pesaran B.(1987) "Data-Fit: An interactive Econometric Software Packane,Oxford University Press, Oxford
- Pesaran M.H.(1990) "An econometric analysis of the exploration and extraction of oil in the U.K. Continental Shelf" The Economic Journal, 100, pp.367-391

Regression Diagnostics: Identifying Influential Data and Sources of Collinearity

| | | |
|----------------------------------|------------|------------|
| David A. Belsley | B001H6IQCK | 0471691178 |
| Edwin Kuh | B00288YEI6 | 0471691178 |
| Roy E. Welsch | B00288SRCU | 0471691178 |

Rousseeuw,P.J.- van Zomeren, B. C.: Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association* 85,1990. 633-639.

Rousseeuw, P.J.- Leroy,A.M.: *Robust Regression and Outlier Detection*. J.Wiley, New Jersey 2003.
ISBN 0-471-48855-0.

Gujarati, D. N., & Porter, D. C. (2009). *Econometria básica*. Editora Bookman.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control* (5th ed.). Wiley.

Wooldridge, J. M. (2020). *Introductory Econometrics: A Modern Approach* (7th ed.). Cengage Learning.

Statistical Modeling: Regression, Survival Analysis, and Time Series Analysis Capa comum – 17 março 2023

Edição Inglês por [Lawrence Mark Leemis](#) (Autor)

Lumley T, Emerson S. (2002) [The Importance of the Normality Assumption in Large Public Health Data Sets](#). *Annual Review of Public Health*. 23:151–69.

Annu Rev Public Health

• . 2002:23:151-69.

doi: 10.1146/annurev.publhealth.23.100901.140546. Epub 2001 Oct 25.

The importance of the normality assumption in large public health data sets

Thomas Lumley ¹, Paula Diehr, Scott Emerson, Lu Chen

Introductory Econometrics: A Modern Approach Capa dura – 4 janeiro 2019

Edição Inglês por [Jeffrey Wooldridge](#) (Autor)

Códigos dos Comandos em R para Construção de Figuras e Obtenção de Resultados Estatísticos

Capítulo-12

Figura-1

```
dev.off(dev.list()["RStudioGD"])
rm(list=ls())
cat("\f")

library(ggplot2)
library(pBrackets)
library(grid)
library(ggforce)

# Gerar os dados para a Curva Normal
mean <- 3.75 # Centro da curva
sd <- 0.4 # Desvio padrão
x_vals <- seq(2, 5.5, length.out = 100) # Intervalo da curva
y_vals <- dnorm(x_vals, mean = mean, sd = sd) # Densidade da curva

# Rotacionar os dados da Curva Normal em 90 graus
rotated_curve <- data.frame(
  x = y_vals + 1.25, # Ajustar a posição horizontal
  y = x_vals # A posição vertical permanece
)

# Base do gráfico
regressao <- qplot() +
  geom_point() +
  xlim(0, 10) +
  ylim(0, 10) +
  coord_fixed(ratio = 1) +
  geom_point(aes(x = 1.25, y = 3.75), fill = "black", size = 2.5) +
  geom_circle(aes(x0 = 1.25, y0 = 3.75, r = 0.3)) +
  geom_point(aes(x = 3.75, y = 5), fill = "black", size = 2.5) +
  geom_circle(aes(x0 = 3.75, y0 = 5, r = 0.3)) +
  geom_point(aes(x = 6.25, y = 6.25), fill = "black", size = 2.5) +
  geom_circle(aes(x0 = 6.25, y0 = 6.25, r = 0.3)) +
  geom_segment(aes(x = 0, y = 3.15, xend = 8.75, yend = 7.5), linewidth = 1) +
  geom_segment(aes(x = 1.25, y = 0, xend = 1.25, yend = 6.7), linewidth = 0.5, linetype = 2)
+
```

```

geom_segment(aes(x = 0, y = 3.75, xend = 2.5, yend = 3.75), linewidth = 0.5, linetype = 2)
+
geom_segment(aes(x = 3.75, y = 0, xend = 3.75, yend = 7.5), linewidth = 0.5, linetype = 2)
+
geom_segment(aes(x = 0, y = 5, xend = 5, yend = 5), linewidth = 0.5, linetype = 2) +
geom_segment(aes(x = 6.25, y = 0, xend = 6.25, yend = 9), linewidth = 0.5, linetype = 2)
+
geom_segment(aes(x = 0, y = 6.25, xend = 9, yend = 6.25), linewidth = 0.5, linetype = 2)
+

# Adicionar a Curva Normal rotacionada
geom_path(data = rotated_curve, aes(x = x, y = y), color = "black", linewidth = 1) +
geom_path(data = rotated_curve, aes(x = x + 2.5, y = y + 1.25), color = "black", linewidth
= 1) +
geom_path(data = rotated_curve, aes(x = x + 5, y = y + 2.5), color = "black", linewidth =
1) +

# Textos auxiliares
geom_point(aes(x = 0.5, y = 9), fill = "black", size = 2.5) +
geom_circle(aes(x0 = 0.5, y0 = 9, r = 0.3)) +
annotate("text", 2.8, 9, label = "Média Condicional", size = 3.5, parse = F) +
annotate("text", x = 9, y = 8, label = expression(E * group("[", Y ~ "|" ~ X[i], "]")),
size = 3.5, color = "black") +
theme(axis.text.x = element_blank(), axis.text.y = element_blank()) +
labs(x = "X", y = "Y")

# Exibir o gráfico
regressao

```

Figura-2

```
dev.off(dev.list()["RStudioGD"])
rm(list=ls())
cat("\f")

a <- 1
b <- 2
eps <- rnorm(50)

modelo_1 <- function(x) {a + b*x + 5*eps}
modelo_2 <- function(x) {a + b*(x*x) + 100*eps}
modelo_3 <- function(x) {- b*(x^3) + 5000*eps}
modelo_4 <- function(x) {exp(x/6) + 500*eps}

par(mfrow = c(2, 2))

library(ggplot2)
p1 <- ggplot(data.frame(x = c(1,50)), aes(x = x)) +
  stat_function(fun = modelo_1, geom = "point")

p2 <- ggplot(data.frame(x = c(1,50)), aes(x = x)) +
  stat_function(fun = modelo_2, geom = "point")

p3 <- ggplot(data.frame(x = c(1,50)), aes(x = x)) +
  stat_function(fun = modelo_3, geom = "point")

p4 <- ggplot(data.frame(x = c(1,50)), aes(x = x)) +
  stat_function(fun = modelo_4, geom = "point")

library(gridExtra)
grid.arrange(p1, p2, p3, p4, nrow = 2)
```

Capítulo-12

Figura-3

```
dev.off(dev.list()["RStudioGD"])
rm(list=ls())
cat("\f")

# Cria um espaço inicial e vai empilhando os comandos para as diversas
# formas da figura
library(ggplot2)
library(pBrackets)
library(grid)

desvios_regressao <- qplot() + geom_point() + xlim(0, 10) + ylim(0, 10) +
  geom_segment(aes(x = 1.25, y = 1.25, xend = 8.75, yend = 10), linewidth = 1 ) +
  geom_segment(aes(x = 0, y = 5, xend = 10, yend = 5), linewidth = 1 ) +
  geom_rect(aes(xmin = 6.1, xmax = 6.4, ymin = 6.8, ymax = 7.3), fill = "white", color =
"black") +
  geom_segment(aes(x = 6.25, y = 6.8, xend = 6.25, yend = 5), linewidth = 1 ) +
  geom_point(aes(x = 6.25, y = 8.75), fill = "black", size = 3.5) +
  geom_segment(aes(x = 6.25, y = 7.3, xend = 6.25, yend = 8.75), linewidth = 1 ) +
  annotate("text", 1, 3, label = "hat(Y) == hat(beta)[0] + hat(beta)[1]*X", size=5, parse= T) +
  annotate("text", 7.6, 6.1, label = "hat(Y)[i] - bar(Y)", size = 5, parse = T) +
  annotate("text", 7.6, 8, label = "Y[i] - hat(Y)[i]", size = 5, parse = T) +
  annotate("text", 5.1, 7, label = "Y[i] - bar(Y)", size = 5, parse = T) +
  annotate("text", 5.3, 9.5, label = "Y[i]", size = 5, parse = T) +
  annotate("text", 9.1, 7.1, label = "hat(Y)[i]", size = 5, parse = T) +
  annotate("text", 5.5, 3.75, label = "bar(Y)", size = 5, parse = T) +
  annotate("text", 0.5, 5.6, label = "bar(Y)", size = 5, parse = T) +
  annotate("text", 6, 2, label = "Y[i] - bar(Y) ==
Y[i] - hat(Y)[i] + hat(Y)[i] - bar(Y)", size = 5, parse = T) +
  geom_segment(aes(x = 0.6, y = 2.5, xend = 1.5, yend = 1.6),
    arrow = arrow(length = unit(0.3, "cm"), type = "closed")) +
  geom_segment(aes(x = 8.75, y = 7, xend = 6.4, yend = 7),
    arrow = arrow(length = unit(0.3, "cm"), type = "closed")) +
  geom_segment(aes(x = 5.5, y = 9.1, xend = 6.1, yend = 8.8),
    arrow = arrow(length = unit(0.3, "cm"), type = "closed")) +
  geom_segment(aes(x = 5.6, y = 3.6, xend = 6.25, yend = 4.9),
    arrow = arrow(length = unit(0.3, "cm"), type = "closed")) +
  theme(axis.text.x = element_blank(), axis.text.y = element_blank()) +
  labs(x = "X", y = "Y")

# Inserir as chaves na figura
desvios_regressao
```

```
grid.brackets(530, 230, 530, 85, lwd=2, col = "black")
grid.brackets(570, 170, 570, 230, lwd=2, col = "black")
grid.brackets(570, 80, 570, 140, lwd=2, col = "black")
```

Capítulo-12

EXEMPLO-12.6

```
dev.off(dev.list()["RStudioGD"])
rm(list=ls())
cat("\f")

library(tidyverse)
library(ggplot2)
library(quantmod)          # Para usar o "getSymbols"
library(nortest)           # Para realizar os testes de normalidade
library(stats)              # Para usar a função "anova"
library(lmtest)             # Para fazer o Teste de Durbin-watson
library(gridExtra)          # Para inserir tabela no "qqplot"
library(data.table)          # Para usar o "shift"
library(ggpubr)              # Para usar o "ggarrange" e "annotate_figure"
library(cowplot)             # Para fazer "qqplot" junto com histograma
library(MASS)                # Para ajustar "fitdistr"
library(car)                  # Para usar o residualPlots
library(olsrr)                # Para usar o ols_plot_resid_lev

# 1. BAIXAR OS DADOS DA EMBRAER ----
EMBR <- getSymbols("EMBR3.SA", src = "yahoo", from = "2021-01-01", to =
"2021-12-31", auto.assign = FALSE)
# O "auto.assign" é para armazenar com o símbolo no ambiente especificado

EMBR
EMBR3.SA <- na.omit(EMBR)
#View(EMBR3.SA)

# Cria o vetor de preço de fechamento
EMBR_fechamento <- EMBR3.SA$EMBR3.SA.Close
media_pf <- mean(EMBR_fechamento)
media_pf

# Retorno, com base no Preço de Fechamento
retorno_EMBR <- diff(EMBR_fechamento)/lag(EMBR_fechamento)
retorno_EMBR <- na.omit(retorno_EMBR)

tabela_retorno_EMBR <- cbind(EMBR_fechamento, retorno_EMBR)
head(tabela_retorno_EMBR)

# 2. BAIXAR OS DADOS DO IBOVESPA ----
IBOV <- getSymbols("^BVSP", src = "yahoo", from = "2021-01-01", to =
"2021-12-31", auto.assign = FALSE)

# O "auto.assign" é para armazenar com o símbolo no ambiente especificado
IBOV
IBOV.B3 <- na.omit(IBOV)
#View(IBOV.B3)

# Cria o vetor de variação do IBOVESPA
IBOV_fechamento <- IBOV.B3$BVSP.Close
media_ib <- mean(IBOV_fechamento)
media_ib

# Variação IBOVESPA
variaco_IBOV <- diff(IBOV_fechamento)/lag(IBOV_fechamento)
variaco_IBOV <- na.omit(variaco_IBOV)
```

```

tabela_variaco_IBOV <- cbind(IBOV_fechamento, variaco_IBOV)
head(tabela_variaco_IBOV)

# 3. SCATTER-PLOT E RETA DE REGRESSÃO ENTRE VARIAÇÃO DO IBOVESPA E
RETORNO DA EMBRAER ----
IBOV_EMBR <- data.frame(variaco_IBOV, retorno_EMBR)
grafico_dispersao <- ggplot(IBOV_EMBR, aes(x = BVSP.Close, y =
EMBR3.SA.Close)) +
  geom_point() +
  labs(x = "Variação IBOVESPA", y = "Retorno EMBR3", title = "")
grafico_dispersao

# Scatter Plot e Reta de Regressão
scatter_e_regressao <- grafico_dispersao +
  stat_smooth(method = "lm",
              formula = y ~ x,
              geom = "smooth") +
  stat_regline_equation(label.x = -0.03, label.y = 0.1) +
  stat_cor(label.x = -0.03, label.y = 0.07)

scatter_e_regressao

# Sumário da Regressão
regressao <- lm(formula = EMBR3.SA.Close ~ BVSP.Close,
                 data = IBOV_EMBR)
summary(regressao)

# 4. RESULTADOS DA REGRESSÃO A PARTIR DAS EQUAÇÕES DE AJUSTE DO MRLM ----
x <- variaco_IBOV
y <- retorno_EMBR
x_bar <- mean(x)
y_bar <- mean(y)
somatorio_x <- sum(x)
somatorio_y <- sum(y)
somatorio_x_quad <- sum(x*x)
somatorio_y_quad <- sum(y*y)
somatorio_x_y <- sum(x*y)

n <- length(x)
k <- 1           # Quantidade de v.i.'s no modelo

SS_xx <- sum((x - x_bar)^2)
SS_yy <- sum((y - y_bar)^2)
SS_xy <- sum((x - x_bar)*(y - y_bar))

beta_hat_1 <- SS_xy/SS_xx; beta_hat_1
beta_hat_0 <- somatorio_y/n - beta_hat_1*(somatorio_x/n); beta_hat_0

# 4.1 Cálculo do MSE ----
sse <- SS_yy - ((SS_xy)^2)/SS_xx; sse

mse <- (SS_yy - ((SS_xy)^2)/SS_xx)/(n - 2)
# mse <- ((n - 1)/(n - 2))*(var(y) - ((var(x,y))^2)/var(x)); mse
rmse <- mse^0.5; rmse
SSR <- SS_yy - sse; SSR
MSR <- SSR/k; MSR

# 4.2 Intervalo de Confiança para Beta_1 ----
alfa <- 0.05
confint(regressao)

s_beta_1 <- rmse/(SS_xx^0.5); s_beta_1
s_beta_0 <- rmse*((1/n) + x_bar^2/(SS_xx))^0.5; s_beta_0

```

```

LI_beta_1 <- beta_hat_1 - qt(alfa/2, n - 2, lower.tail = FALSE)*s_beta_1
LS_beta_1 <- beta_hat_1 + qt(alfa/2, n - 2, lower.tail = FALSE)*s_beta_1
LI_beta_1; LS_beta_1

# 4.3 Intervalo de Confiança para Beta_0 ----
LI_beta_0 <- beta_hat_0 - qt(alfa/2, n - 2, lower.tail = FALSE)*s_beta_0
LS_beta_0 <- beta_hat_0 + qt(alfa/2, n - 2, lower.tail = FALSE)*s_beta_0
LI_beta_0; LS_beta_0

# 4.4 Estatística de Teste e Teste de Hipóteses sobre o Beta_1 ----
# ANOVA
anova(regressao)
SSR; MSR; sse; mse

F_obs <- MSR/mse; F_obs

p_valor_F <- pf(F_obs,1, n-2, lower.tail = FALSE); p_valor_F

# 4.5 Teste de Hipótese sobre Beta_1 ----
# Cálculo da estatística t
# Vamos testar as seguintes hipóteses para Beta_1

# H0: Beta_1 = 0
# Ha: Beta_1 ≠ 0
t_obs_1 <- (beta_hat_1 - 0)/(rmse/(ss_xx^0.5)); t_obs_1

p_valor_1 <- 2*pt(t_obs_1, n-2, lower.tail = FALSE); p_valor_1

# No caso do CAPM também poderíamos fazer:
# H0: Beta_1 = 1
# Ha: Beta_1 > 1
t_obs_2 <- (beta_hat_1 - 1)/(rmse/(ss_xx^0.5)); t_obs_2

p_valor_2 <- pt(t_obs_2, n-2, lower.tail = FALSE); p_valor_2

# 4.6 Teste de Hipótese sobre Beta_0 ----
# Cálculo da estatística t
# Vamos testar as seguintes hipóteses para Beta_0
# H0: Beta_0 = 0
# Ha: Beta_0 ≠ 0

se_Beta_0 <- rmse*((1/n) + (x_bar^2)/ss_xx)^0.5; se_Beta_0

t_obs_3 <- beta_hat_0/se_Beta_0; t_obs_3

p_valor_3 <- 2*pt(t_obs_3, n-2, lower.tail = FALSE); p_valor_3

# 4.7 Coeficiente de Determinação [R2] ----
R_quadrado <- 1 - sse/SS_yy; R_quadrado

# R quadrado ajustado
R_quad_ajustado <- 1 - (sse/(n - k - 1))/(ss_yy/(n - 1)); R_quad_ajustado

# 4.8 Intervalo de Confiança para E[y|x] ----
# Primeiro consideramos a distribuição amostral de y dado xh
# Essa esperança pode ser escrita como E[yh] ou E[y|xh]
xh <- 0.01
media_y_hat_h <- beta_hat_0 + beta_hat_1*xh; media_y_hat_h #E[y|xh]
se_y_hat_h <- mse*((1/n) + ((xh - x_bar)^2)/ss_xx)^0.5; se_y_hat_h
LI_y_hat_h <- media_y_hat_h - (qt(alfa/2, n - 2, lower.tail = FALSE))*se_y_hat_h
LS_y_hat_h <- media_y_hat_h + (qt(alfa/2, n - 2, lower.tail = FALSE))*se_y_hat_h
LI_y_hat_h; LS_y_hat_h

```

```

# 4.9 Scatter Plot, Reta de Regressão e Intervalo de Confiança para E[y]
---
x <- seq(min(variaco_IBOV), max(variaco_IBOV), 0.0001)
LI_y <- (beta_hat_0 + beta_hat_1*x) -
  (qt(alfa/2, n - 2, lower.tail = FALSE))*(rmse*(
    ((1/n) + ((x - x_bar)^2)/SS_xx)^0.5))

#plot(x, LI_y, type = "l")
base_dados_1 <- data.frame(x, LI_y)

LS_y <- (beta_hat_0 + beta_hat_1*x) +
  (qt(alfa/2, n - 2, lower.tail = FALSE))*(rmse*(
    ((1/n) + ((x - x_bar)^2)/SS_xx)^0.5))

#plot(x, LS_y, type = "l")
base_dados_2 <- data.frame(x, LS_y)
adicionar_linha_1 <- function(dataset, varx, vary) {
  list(
    geom_line(data=dataset, aes_string(x=varx, y=vary)),
    geom_point(data=dataset, aes_string(x=varx, y=vary), col = "red",
               size = 0.25))}

IBOV_EMBR <- data.frame(variaco_IBOV, retorno_EMBR)
grafico <- ggplot(IBOV_EMBR, aes(x = BVSP.Close, y = EMBR3.SA.Close)) +
  geom_point()

scatter_e_regressao <- grafico +
  stat_smooth(method = "lm", formula = y ~ x, geom = "smooth") +
  stat_regrline_equation(label.x = -0.03, label.y = 0.1) +
  stat_cor(label.x = -0.03, label.y = 0.07)

scatter_e_regressao + adicionar_linha_1(base_dados_1, varx = "x", vary =
  "LI_y") +
  adicionar_linha_1(base_dados_2, varx = "x", vary = "LS_y") +
  labs(x = "Variação IBOVESPA", y = "Retorno EMBR3", title = "")

# 4.10 Predição de uma Nova Observação ----
adicionar_linha_2 <- function(dataset, varx, vary) {
  list(geom_line(data=dataset, aes_string(x=varx, y=vary), linetype =
'dashed'),
    geom_point(data=dataset, aes_string(x=varx, y=vary), col = "black",
               size = 0.10))}

LIP_y <- (beta_hat_0 + beta_hat_1*x) - (qt(alfa/2, n - 2, lower.tail =
FALSE))*(rmse*((1 + (1/n) + ((x - x_bar)^2)/SS_xx)^0.5))

#plot(x, LIP_y, type = "l")
base_dados_11 <- data.frame(x, LIP_y)

LSP_y <- (beta_hat_0 + beta_hat_1*x) +
  (qt(alfa/2, n - 2, lower.tail = FALSE))*(rmse*((1 + (1/n) + ((x -
x_bar)^2)/SS_xx)^0.5) )

#plot(x, LSP_y, type = "l")
base_dados_22 <- data.frame(x, LSP_y)
scatter_e_regressao + adicionar_linha_1(base_dados_11, varx = "x", vary =
  "LI_y") +
  adicionar_linha_1(base_dados_22, varx = "x", vary = "LS_y") +
  adicionar_linha_2(base_dados_11, varx = "x", vary = "LIP_y") +
  adicionar_linha_2(base_dados_22, varx = "x", vary = "LSP_y") +
  labs(x = "Variação IBOVESPA", y = "Retorno EMBR3", title = "")

# 4.11 Correlação ----
correlacao_1 <- R_quadrado^0.5

```

```

correlacao_2 <- ss_xy/((ss_xx*ss_yy)^0.5) # outra forma de obter

# 5. ANÁLISE DOS RESÍDUOS ----
# Verificação dos pressupostos do MRLS:
# i. Autocorrelação dos Resíduos
# ii. Homocedasticidade
# iii. Normalidade
# iv. Linearidade
# v. Identificação de outliers

# 5.1 Plot dos Resíduos ----
residuos <- resid(regressao)
plot(residuos)
par(mfrow = c(2, 2))
plot(regressao)

# 5.2 Autocorrelação dos Resíduos [Durbin-Watson] ----
# H0: Resíduos são não correlacionados
dwtest(regressao)
durbin_watson <- sum((diff(residuos))^2)/sum(residuos^2)
durbin_watson

# 5.3 Heterocedasticidade dos Resíduos [Breusch-Pagan] ----
# H0: Resíduos são Homocedásticos [Ou seja, os resíduos são distribuídos
# com igual varânci]
bptest(regressao) # Está na library "lmtest"

# 5.4 Normalidade dos Resíduos ----
# KOLMOGOROV-SMIRNOV
ks.test(residuos, "pnorm")

# SHAPIRO-WILK
shapiro.test(residuos)

# ANDERSON-DARLING
ad.test(residuos)

# LILLIEFORS
lillie.test(residuos) # Também precisa está na library "nortest"

# 5.5 Linearidade ----
y_hat <- regressao$fitted.values # Valores ajustados
ggplot(mapping = aes(x = residuos, y = y_hat)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color="red") +
  geom_smooth(method = "loess", se = FALSE)

# 5.6 Identificação de Outliers ----
residualPlots(regressao)

# Verificação de outliers
qqPlot(regressao,id.n=2)

# Plota a Leverage (Alavancagem) e Outliers
library(olsrr)
ols_plot_resid_lev(regressao)

# Plota a Distância de Cook
ols_plot_cooksd_chart(regressao)

```

Figura-DW1

```
library(ggplot2)
library(pBrackets)
library(grid)
```

```
dw_valores <- qplot() + geom_point() + xlim(-1, 5) + ylim(3, 4.8) +
  geom_segment(aes(x = 0, y = 4, xend = 4, yend = 4), linewidth = 1 ) +
  geom_segment(aes(x = 0, y = 3.95, xend = 0, yend = 4.05), linewidth = 1 ) +
  geom_segment(aes(x = 4, y = 3.95, xend = 4, yend = 4.05), linewidth = 1 ) +
  geom_segment(aes(x = 2, y = 3.95, xend = 2, yend = 4.05), linewidth = 1 ) +
  geom_segment(aes(x = 0.65, y = 3.95, xend = 0.65, yend = 4.05), linewidth = 1 ) +
  geom_segment(aes(x = 1.35, y = 3.95, xend = 1.35, yend = 4.05), linewidth = 1 ) +
  geom_segment(aes(x = 2.65, y = 3.95, xend = 2.65, yend = 4.05), linewidth = 1 ) +
  geom_segment(aes(x = 3.35, y = 3.95, xend = 3.35, yend = 4.05), linewidth = 1 ) +
  geom_segment(aes(x = 0, y = 4.1, xend = 0.65, yend = 4.1), linewidth = 1, linetype = 2) +
  geom_segment(aes(x = 0.65, y = 3.8, xend = 1.35, yend = 3.8), linewidth = 1, linetype = 2)
+
  geom_segment(aes(x = 1.35, y = 4.15, xend = 2.65, yend = 4.15), linewidth = 1, linetype =
= 1) +
  geom_segment(aes(x = 2.65, y = 3.8, xend = 3.35, yend = 3.8), linewidth = 1, linetype = 2)
+
  geom_segment(aes(x = 3.35, y = 4.1, xend = 4, yend = 4.1), linewidth = 1, linetype = 2) +
  annotate("text", 0, 3.9, label = "0", size = 5, parse = T) +
  annotate("text", 0.65, 3.9, label = "d[L]", size = 5, parse = T) +
  annotate("text", 1.35, 3.9, label = "d[U]", size = 5, parse = T) +
  annotate("text", 2, 3.9, label = "2", size = 5, parse = T) +
  annotate("text", 2.65, 3.9, label = "4 - d[U]", size = 5, parse = T) +
  annotate("text", 3.35, 3.9, label = "4 - d[L]", size = 5, parse = T) +
  annotate("text", 4, 3.9, label = "4", size = 5, parse = T) +
  annotate("text", x = 0.3, y = 4.25, label = expression(atop("Autocorrela o", ~Positiva)),
size = 4) +
  annotate("text", x = 0.95, y = 3.75, label = "Zona de Indecis o", size = 4) +
  annotate("text", x = 2.0, y = 4.25, label = "Sem Autocorrela o", size = 4) +
  annotate("text", x = 2.95, y = 3.75, label = "Zona de Indecis o", size = 4) +
  annotate("text", x = 3.65, y = 4.25, label = expression(atop("Autocorrela o", ~Negativa)),
size = 4) +
  theme(axis.text.x=element_blank(), axis.text.y=element_blank())
dw_valores
```