UNIVERSIDADE FEDERAL DA FRONTEIRA SUL
Campus CERRO LARGO

PROJETO DE EXTENSÃO
Software R:
Capacitação em análise estatística
de dados utilizando um software livre.

Fonte: https://www.r-project.org/

Módulo V
Modelos de regressão

Ministrante: Iara Endruweit Battisti

Blog do projeto: https://softwarelivrer.wordpress.com/equipe/

Equipe:
Coordenadora:
Profe. Iara Endruweit Battisti (iara.battisti@uffs.edu.br)
Colaboradores:
Profa. Denize Reis
Prof. Erikson Kaszubowski
Prof. Reneo Prediger
Profa. Tatiane Chassot
Mestrando Felipe Smolski
Bolsista:
Djaina Rieger - aluna de Engenharia Ambiental (djaina.rieger@outlook.com)

Sumário

1 Correlação e regressão linear simples

Muitas vezes há a necessidade de estudar duas ou mais variáveis ao mesmo tempo. Por exemplo, pode-se estudar investimento em comunicação e vendas para verificar se existe uma relação entre elas e o tipo da relação. Outro exemplo é verificar se sólidos removidos de um material estejam relacionados ao tempo de secagem.

Em outros casos, estudam-se conjuntamente duas variáveis para predizer uma variável em função da outra. Por exemplo, predizer as vendas para determinado investimento em comunicação. Outro exemplo, predizer a quantidade de sólidos removidos para cinco horas de secagem.

1.1 Análise de correlação

É a técnica mais simples para estudar a relação entre duas variáveis. Os dados compõem uma única amostra de pares de valores (x_i , y_i), correspondendo aos valores das variáveis X e Y, respectivamente, feitas em cada elemento da amostra.

Para analisar a existência de relação entre as duas variáveis, primeiramente pode-se fazer o Diagrama de Dispersão.

1.1.1 Diagrama de Dispersão

É um gráfico para verificar a existência de relação entre as variáveis X e Y. É composto por pontos, os quais correspondem aos pares de valores (x_i , y_i), sendo a variável X representada no eixo horizontal e a variável Y representada no eixo vertical.

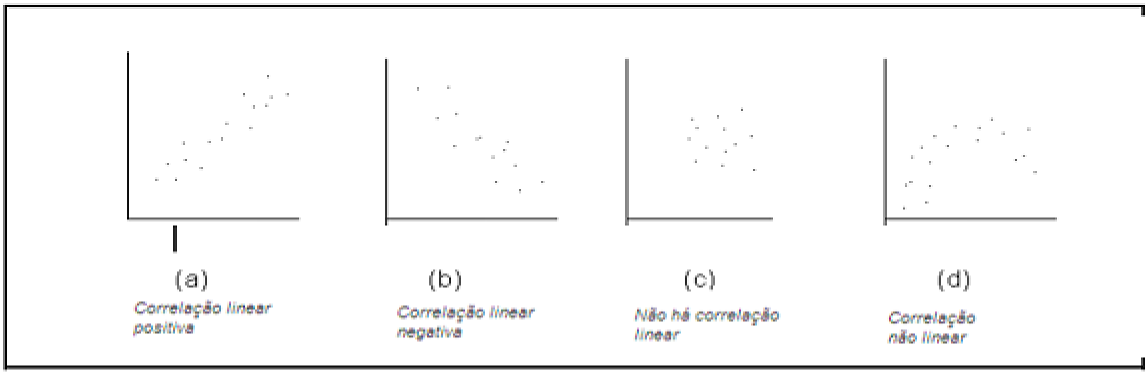


Figura 1 – Diagramas de Dispersão

O diagrama de dispersão fornece uma visualização gráfica do comportamento conjunto das duas variáveis em estudo. Na figura 1(a) percebe-se uma correlação (relação) linear positiva entre as variáveis X e Y, ou seja, os valores das duas variáveis crescem conjuntamente, já na figura 1(b) percebe-se uma correlação linear negativa entre as variáveis X e Y, neste caso, os valores de uma variável crescem enquanto os valores da outra variável decrescem. A figura 1(c) informa a ausência de relação entre as duas variáveis e, a figura 1(d) mostra uma relação não linear, a qual não será objeto de estudo nesta publicação.

Exemplo: Considere os dados referentes a tempo de estudo (X) e nota obtida na prova (y) de uma amostra aleatória de cinco estudantes.

Tempo (h)	4,0	7,0	3,5	1,5	9,0
Nota	4,5	7,5	4,7	4,0	9,5

Fonte: Dados simulados.

Sintaxe no software R:

```
tempo=c(4, 7, 3.5, 1.5, 9)
nota=c(4.5, 7.5, 4.7, 4, 9.5)

# Através do comando > plot(tempo, nota), o diagrama de dispersão
#   pode ser representado, como na Figura 2
```

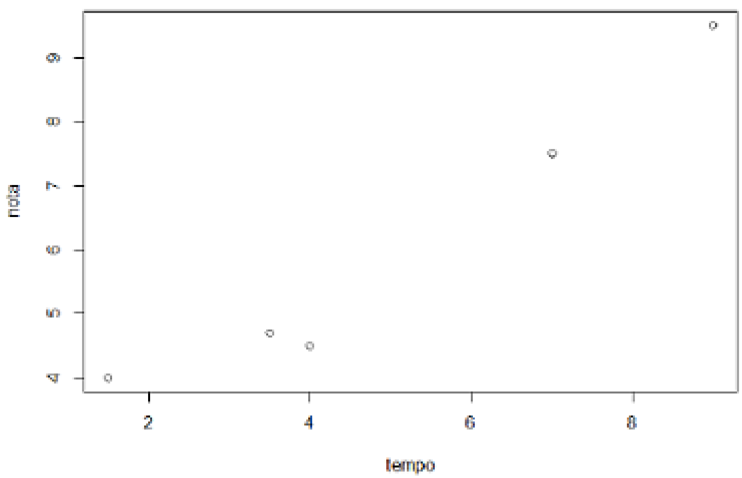


Figura 2 – Diagrama de dispersão da nota em relação ao tempo de estudo

1.2 Coeficiente de Correlação Linear

Mede o grau de relacionamento linear entre os valores emparelhados x e y em uma amostra. O coeficiente linear de Pearson (Karl Pearson 1857-1936) é obtido da seguinte forma:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

Em que:

n = número de pares na amostra

O coeficiente de correlação linear r é uma estatística amostral, representando a magnitude da relação entre duas variáveis. O parâmetro populacional é representado por "p". O valor de r está entre -1 e +1, inclusive. Se o valor de r está próximo de 0, conclui-se que não há correlação linear significativa entre X e Y. Se r está próximo de -1 ou +1, conclui-se pela existência de correlação linear significativa entre X e Y, sendo que o sinal indica uma relação linear positiva (direta) ou negativa (inversa).

Sintaxe no software R:

```
cor(tempo, nota)

## [1] 0.9727342
```

2 Análise de regressão

O estudo de regressão refere-se aos casos em que se pretende estabelecer uma relação entre uma variável Y considerada dependente (variável resposta) e uma ou mais variáveis x_1 , x_2 ,. .. , x_k (variáveis explicativas) consideradas independentes.

O objetivo da análise de regressão é ajustar uma equação que permita explicar o comportamento da variável resposta de maneira que o valor previsto possa estar próximo do que seria observado. A forma do modelo de regressão depende da relação entre as variáveis, expressa visualmente pelo diagrama de dispersão, conforme Figura 1.

A análise de regressão é uma técnica muito utilizada em variáveis quantitativas, como por exemplo:

- vendas em função do investimento em comunicação;

- altura de crianças em função da idade;

- nota obtida em função de horas de estudo;

- número de horas/homens em função do número de lotes.

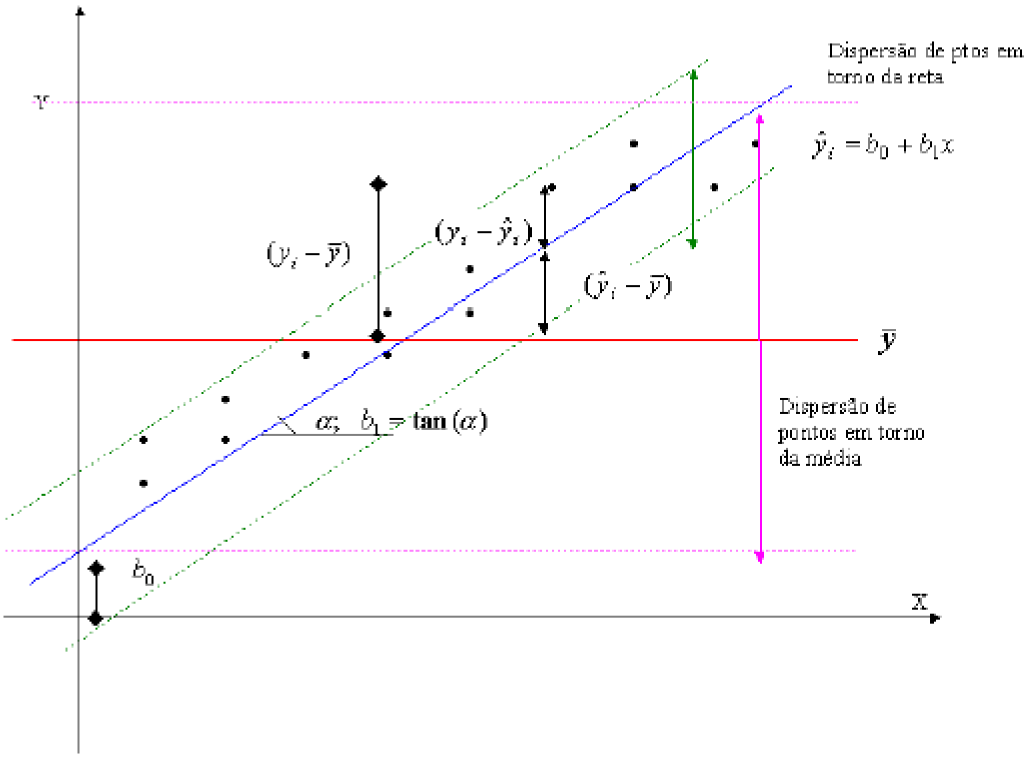


Figura 3 - Variação explicada e não explicada na análise de regressão

Conforme a Figura 3, fica estabelecida uma identidade na regressão, como segue, ou, na seguinte notação:

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

SQTotal = SQRegressão + SQResíduo

Percebe-se a partir da fórmula que o modelo de regressão será mais adequado na medida em que a proporção de SQRegressão seja mais alta em relação à SQTotal do que a SQResíduo.

2.1 Modelo de Regressão Linear Simples

O modelo de regressão linear simples é usado quando a resposta da variável dependente se expressa de forma linear (Figura 3) e neste caso com apenas uma variável explicativa, expresso da seguinte maneira (Hoffmann e Vieira, 1998):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Em que:

y_i : valores da variável resposta (dependente, desfecho), i = 1,2,...,n observações;

x_i : valores da variável explicativa (independente, preditora), i = 1,2,...,n observações;

β_0 : coeficiente linear (intercepto). Interpretado como o valor da variável dependente quando a variável independente é igual a 0;

β_1 : coeficiente angular (inclinação). Interpretado como acréscimo/decréscimo na variável dependente para a variação de uma unidade na variável independente;

ϵ_i : erros aleatórios supostamente de uma população normal, com média 0 e variância constante $\epsilon_i \ N(0, \sigma^2)$

2.2 Método dos mínimos quadrados

É utilizado para a obtenção dos coeficientes linear e angular. Consiste em minimizar a soma de quadrados de resíduos (SQResíduos), ou seja, minimizar

$$\sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

As expressões para os coeficientes, que minimizam SQResíduos são obtidas pela derivadas desta soma de quadrados em relação a b0 e em relação a b1 e podem ser descritas por (Hoffmann e Vieira,1998):

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}$$

Em que:
n : número de pares na amostra

$$b_0 = \bar{y} - b_1 \bar{x}$$

Em que:
$\bar{x}$: média aritmética dos valores de x
$\bar{y}$: média aritmética dos valores de y
b_1 : valor calculado do coeficiente angular

Obtendo-se a seguinte equação de regressão linear simples estimada:

$$\hat{y}=b_0+b_1 x$$

Em que:

b_0 : coeficiente linear estimado
b1 : coeficiente angular estimado
b_1 : valores da variável explicativa

Esta equação refere-se a reta de regressão, se b_1 é um valor positivo a reta é crescente, demonstrando uma relação positiva entre as variáveis e se b_1 é um o valor negativo, a reta é decrescente, demonstrando uma relação inversa entre as variáveis.

Sintaxe no software R:

```
regressao=lm(nota~tempo)
regressao

##
## Call:
## lm(formula = nota ~ tempo)
##
## Coefficients:
## (Intercept)        tempo
##      2.1738       0.7732
```

3 Análise de Variância

A análise de variância (técnica introduzida por Fisher, na década de 20) testa o ajuste da equação como um todo, ou seja, um teste para verificar se a equação de regressão obtida é significativa ou não. No caso de regressão linear simples, a análise de variância é definida como apresentada na Tabela 3.

As hipóteses testadas na Análise de Variância da Regressão são:

$$H_0: \beta_1 = 0 \quad (\text{a regressão não é significativa})$$
$$H_1: \beta_1 \neq 0 \quad (\text{a regressão é significativa})$$

Tabela 3 – Análise de variância para regressão linear simples

FV	GL	SQ	QM	F
Regressão	1	SQRegressão	QMRegressão	Fc
Desvios	n-2	SQResíduo	QMResíduo	-
Total	n-1	SQTotal	-	-

Em que:

SQResíduo = SQTotal - SQRegressão

QMRegressão = SQRegressão / gl regressão

QMResíduo = SQResíduo / gl resíduo

Fc = QMRegressão / QMResíduo

$$SQ\,\text{Re}\,gressão = \frac{\left(\sum xy - \frac{\sum x \sum y}{n}\right)^2}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}$$

$$SQTotal = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

Espera-se que o QMResíduo seja mínimo, assim o modelo de regressão estará bem ajustado. A distribuição de probabilidade para a razão de duas variâncias é conhecida como a distribuição F. Se a hipótese nula for rejeitada ao nível de significância α , rejeita-se a hipótese de regressão não significativa, portanto a regressão é significativa.

Sintaxe no software R:

```
anova(regressao)

## Analysis of Variance Table
##
## Response: nota
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## tempo      1 21.2254 21.2254  52.774 0.005382 **
## Residuals  3  1.2066  0.4022
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.1 Coeficiente de Determinação

Representa o percentual de variação total que é explicada pela equação de regressão, sendo obtido da seguinte forma:

$$R^2 = \frac{SQRegressão}{SQTotal}$$

Quanto mais próximo de 1 (ou 100%), melhor será o ajuste da equação de regressão. Os softwares apresentam também o R^2 ajustado, o qual considera o número de variáveis e o tamanho da amostra, sendo este o mais indicado para regressão múltipla.

Sintaxe no software R:

```
summary(regressao)

##
## Call:
## lm(formula = nota ~ tempo)
##
## Residuals:
##        1        2        3        4        5
## -0.76676 -0.08648 -0.18014  0.66634  0.36704
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1738     0.6031   3.605  0.03664 *
## tempo         0.7732     0.1064   7.265  0.00538 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.6342 on 3 degrees of freedom
## Multiple R-squared:  0.9462,Adjusted R-squared:  0.9283
## F-statistic: 52.77 on 1 and 3 DF,  p-value: 0.005382
```

Sintaxe no software R:

```
# Através do comando > abline(regressao), é possível
# fazer a apresentação da reta de regressão ajustada.
```

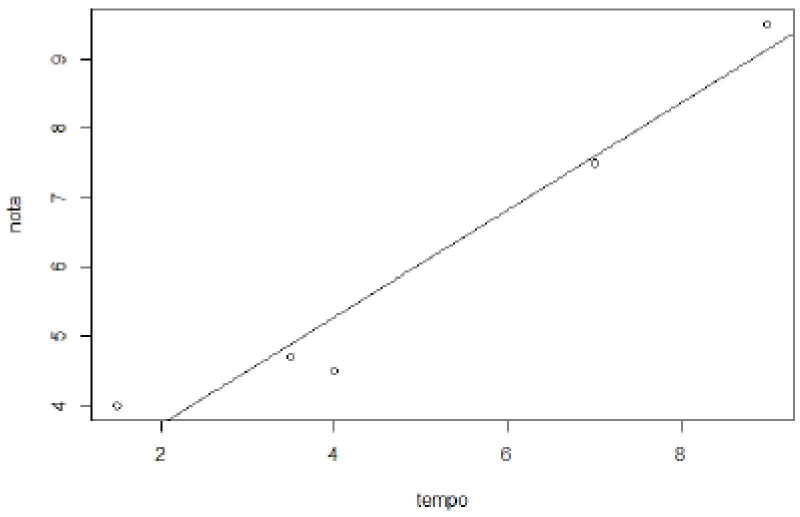


Figura 4 – Reta de regressão ajustada

A seguir é apresentado o comando para obter o intervalo de 95% de confiança para os coeficientes do modelo de regressão linear simples estimado.

Sintaxe no software R:

```
confint(regressao)

##                 2.5 %   97.5 %
## (Intercept) 0.2546198 4.092986
## tempo       0.4345015 1.111977
```

3.2 Intervalo de Predição

Se a equação de regressão se ajusta bem aos dados de acordo com o R^2 e a regressão for significativa (valor p do teste F), então a equação pode ser utilizada para predizer valores da variável Y (resposta) a partir de valores da variável X (explicativa). Caso a regressão não seja significativa a melhor predição para a variável Y é média dos valores de y, ou seja, $\bar{y}$.

A predição de valores só tem sentido nos seguintes casos:

- regressão significativa;

- os valores de X devem estar dentro dos limites inferior e superior dos dados amostrais;

- as inferências referem-se somente a população de onde a amostra aleatória foi extraída;

- as suposições sobre os resíduos devem ser satisfeitas de acordo com o item 2.5.

Quando tem-se um equação estimada do tipo $\hat{y} = b_0 + b_1 x$, $\hat{y}$ representa o valor predito da variável Y para um dado valor da variável X, ou seja, é uma predição pontual, porém esta não informa a sua precisão, a qual é contemplada no intervalo de predição, aqui a ideia é a mesma do intervalo de confiança, já visto em inferência estatística.

O intervalo de predição para um determinado Y é dado por:

$$\hat{y} \pm \varepsilon$$

Em que:

$$\varepsilon = t_{\left(n-2;\frac{\alpha}{2}\right)} . S_e . \sqrt{1 + \frac{1}{n} + \frac{n(x_p - \bar{x})^2}{n(\sum x^2) - (\sum x)^2}}$$

Sendo que:
x_p : o valor dado para x
s_e : o erro padrão da estimativa, definido por:

$$S_e = \sqrt{QM\,\mathrm{Re}\,síduo} = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$$

Assim, obtêm-se o intervalo de predição para um determinado Y, que também pode ser expresso da seguinte forma:

$$(\hat{y} - \varepsilon \;\; ; \;\; \hat{y} + \varepsilon)$$

Sintaxe no software R:

```
x0=data.frame(tempo=5.5)
predict(regressao, x0, interval="prediction")

##        fit      lwr      upr
## 1 6.42662 4.209244 8.643995
```

3.3 Análise dos Resíduos

Para a validade dos intervalos de confiança e teste de hipótese torna-se necessário supor que as observações de Y sejam independentes e o termo de erro tenha distribuição aproximadamente normal com média 0 e variância constante.

O método gráfico pode ser utilizado para testar estas suposições, descrevendo que após a estimação dos parâmetros do modelo, pode-se calcular os resíduos, através da diferença entre os valores observados y e os valores preditos $\hat{y}$, associados a cada x usado na análise. Faz-se então um gráfico com os pares (x , ϵ), sendo ϵ= y- $\hat{y}$. (Barbetta, 2001).

Se o modelo ajustado for apropriado para os dados, os pontos devem estar distribuídos de forma aleatória no gráfico dos resíduos, conforme figura 5(a). Caso a suposição não seja satisfeita, métodos alternativos podem ser utilizados como: método dos mínimos quadrados ponderados para o caso de não homocedasticidade; o método dos mínimos quadrados generalizados para o caso de erros correlacionados; e, métodos não-paramétricos para o caso de não normalidade.

Além da análise gráfica, existem testes para avaliar a homocedasticidade como o Teste de Bartlett e para avaliar a normalidade aplicam-se os testes de Shapiro Wilks ou Kolmogorov-Smirnov.

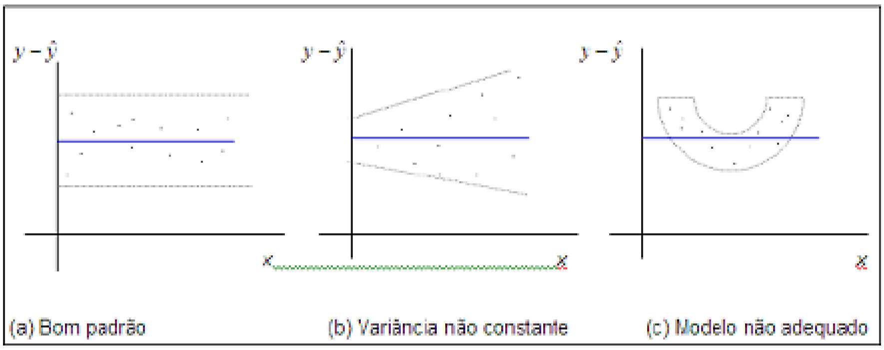
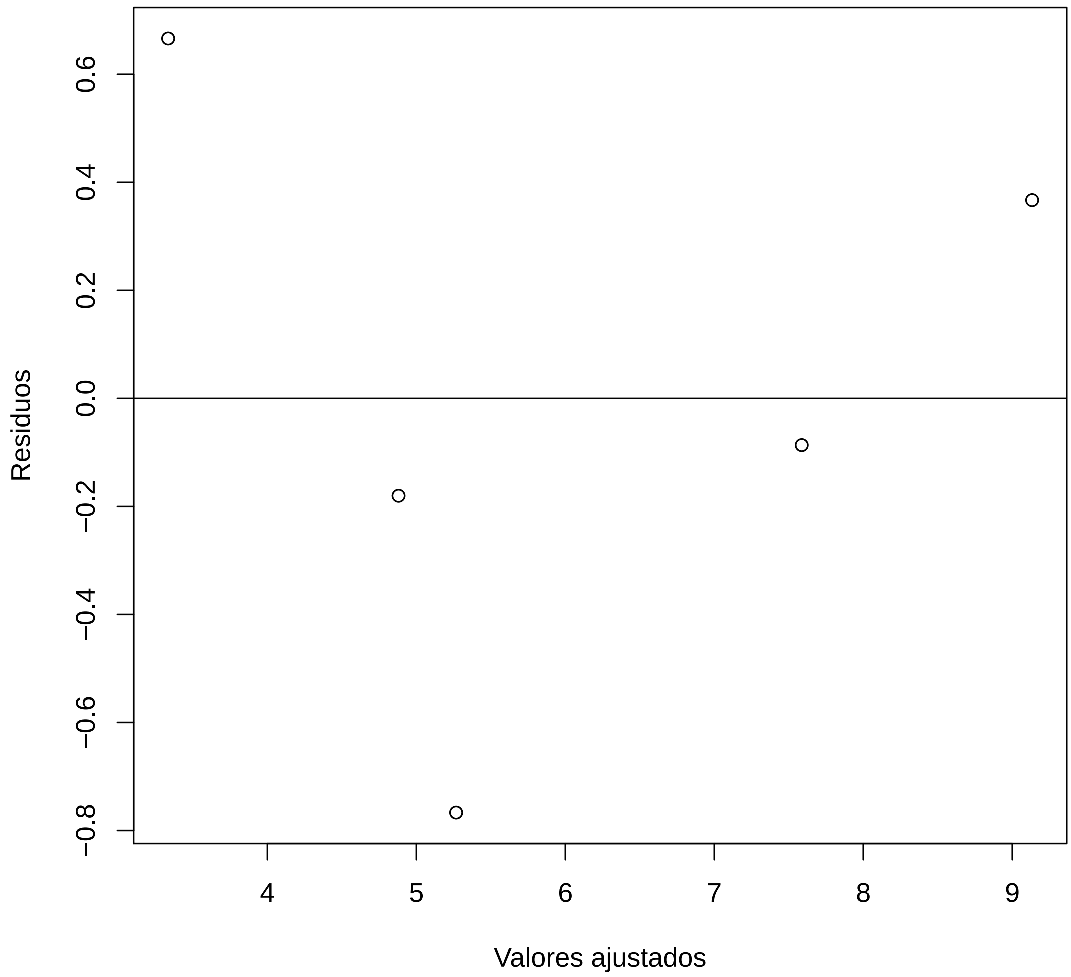


Figura 5 – Gráficos para análise de resíduos em regressão

Sintaxe no software R:

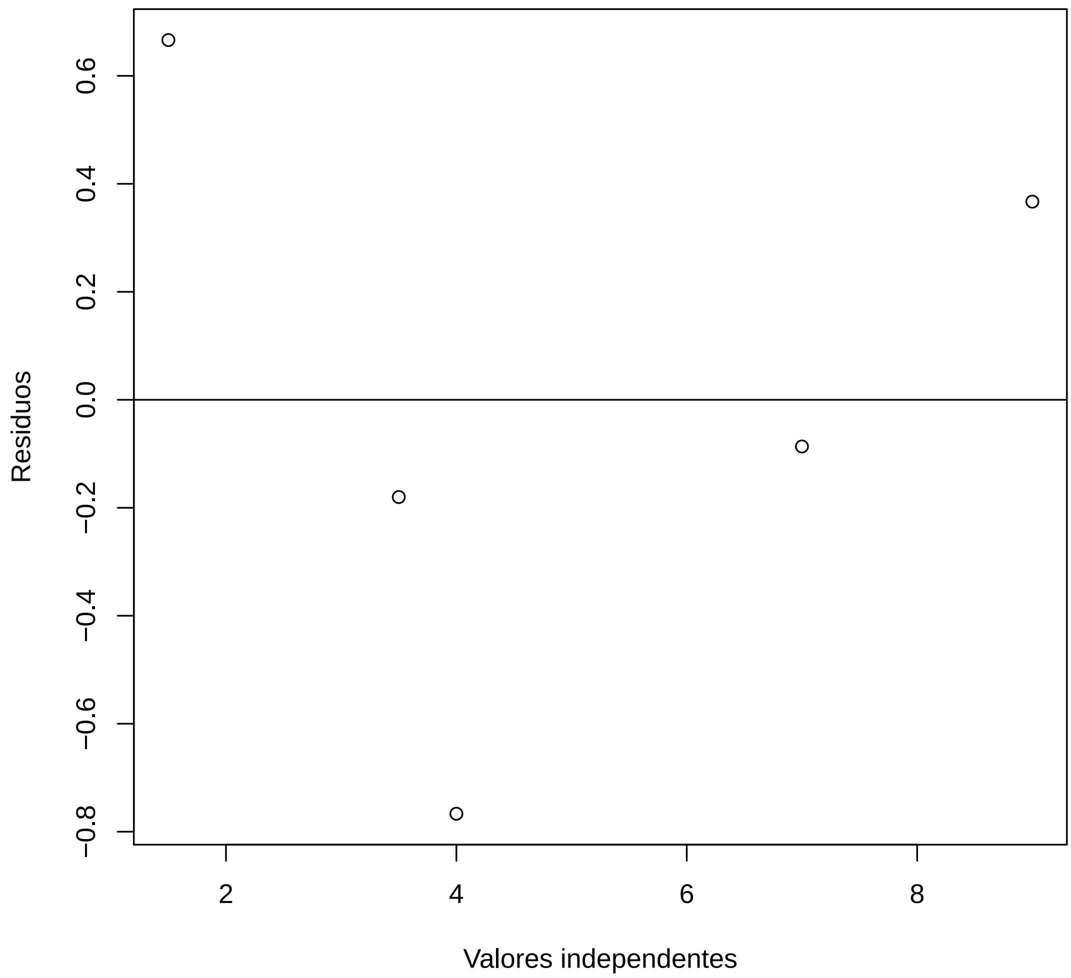
```
plot(fitted(regressao), residuals(regressao),
     xlab="Valores ajustados", ylab="Residuos")
abline(h=0)
```



Na figura acima é apresentado o gráfico de resíduo, em que no eixo y constam os resíduos e no eixo x constam os valores ajustados.

Sintaxe no software R:

```
plot(tempo, residuals(regressao), xlab = "Valores independentes",
     ylab="Residuos")
abline(h=0)
```



Na figura acima é apresentado o gráfico de resíduo, em que no eixo y constam os valores dos resíduos e no eixo x constam os valores da variável independente.

—Para exibir os Valores Ajustados e os Resíduos do ajuste:

Sintaxe no software R:

```
regressao$residuals

##           1           2           3           4           5
## -0.76676056 -0.08647887 -0.18014085  0.66633803  0.36704225

regressao$fitted.values

##        1        2        3        4        5
## 5.266761 7.586479 4.880141 3.333662 9.132958
```

Para testar a suposição que os erros aleatórios têm distribuição Normal, o teste de normalidade de Shapiro Wilk:

Sintaxe no software R:

```
shapiro.test(residuals(regressao))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(regressao)
## W = 0.97488, p-value = 0.9056
```

3.4 Modelo de Regressão Múltipla

Um modelo de regressão múltipla é expresso como:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

Em que:

y_i : valores da variável resposta, i = 1, 2,..., n observações;

y_{ki} : valores das variáveis explicativas, k = 1, 2,..., K variáveis;

β_k : parâmetros do modelo;

ϵ_i : erro aleatório.

A equação estimada para este modelo é definida como:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki}$$

Em que:

b_k : coeficientes estimados;

4 REFERÊNCIA BIBLIOGRÁFICA

BARBETTA, P. A. **Estatística Aplicada às Ciências Sociais**. UFSC. Florianópolis. SC. 1998;

HOFFMANN, R.; VIEIRA, S. **Análise de Regressão. Uma introdução à Econometria**. Hucitec. São Paulo. SP. 1998;