Modelos de Regressão Linear

Rodrigo R. Pescim

Universidade Estadual de Londrina

30 de junho de 2020

- Na inferência estatística é útil identificar se existe relação entre duas ou mais variáveis.

- Em muitos problemas existem duas ou mais variáveis (quantitativas) que são relacionadas, e tem-se o interesse em estudar e explorar essa relação.

- Exemplos
  - Idade e altura das crianças (ou adultos)
  - Tempo de prática de esportes e ritmo cardíaco
  - Taxa de desemprego e taxa de criminalidade
  - Temperatura e rendimento num processo industrial

Diagrama de Dispersão

- Quando deseja-se verificar se há relação entre duas variáveis
- Y - **variável dependente** ou **variável resposta**
- X é a **variável independente**, **variável explanatória** ou **covariável**
- O primeiro passo é fazer um **diagrama de dispersão**
- O padrão determinado pelos pontos no diagrama de dispersão sugere **se existe ou não relação** entre as variáveis

- Pode-se quantificar a relação existente entre duas variáveis utilizando o coeficiente de correlação linear.

- Uma relação entre duas variáveis pode ser identificada por meio de um gráfico de dispersão.

Gráfico de Dispersão

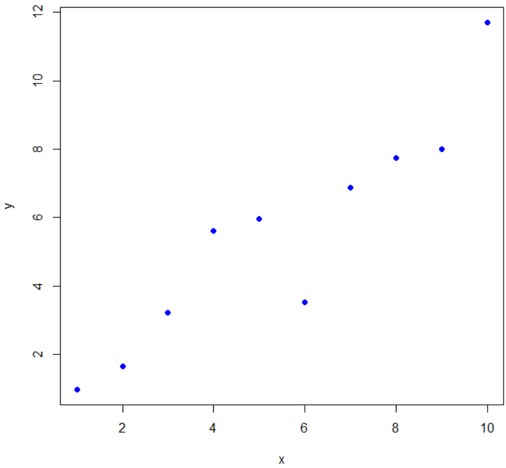


Figura: Correlação positiva

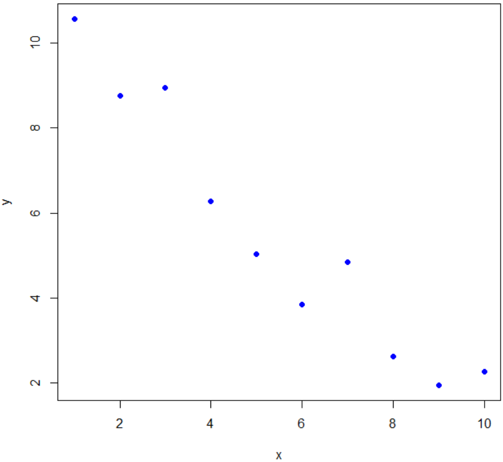


Figura: Correlação negativa

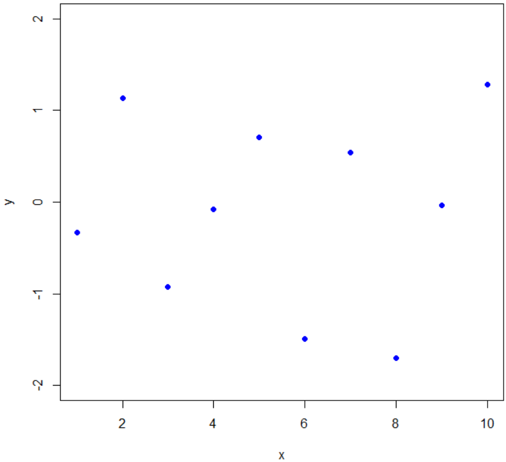


Figura: Não há correlação

Coeficiente de correlação

- O coeficiente de correlação linear tem por objetivo medir o grau de relação entre duas variáveis e é definida por:

$$r = \frac{Cov(X,Y)}{\sqrt{S_x^2\, S_y^2}}$$

em que

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$Cov(X,Y) = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i Y_i - n\,\bar{X}\,\bar{Y}\right]$$

- O coeficiente de correlação é denotado por r e somente pode assumir um valor entre -1 e 1 inclusive.
    - Se $r = +1$, existe uma correlação perfeita positiva entre as variáveis.
    - Se $r = -1$, existe uma correlação perfeita negativa entre as vaiáveis.
    - Se $r = 0$, não existe correlação entre as variáveis.

- **Exemplo 1:** Os dados a seguir correspondem à variável renda familiar e gasto com alimentação (em unidades monetárias) para uma amostra de 8 famílias.

Renda (Y)	3	5	10	20	30	40	50	60
Gasto (X)	2	3	6	10	15	10	20	20

(a) Construa o gráfico de dispersão.
(b) Calcular o coeficiente de correlação e interpretar o resultado.

- **Exemplo 2:** Os dados que se seguem referem-se a concentrações de $CO_2(X)$ aplicadas sobre folhas de trigo a uma temperatura de $35^o C$ e a quantias de $CO_2(Y, cm^3/dm^2/hora)$ absorvido pelas folhas.

X	Y
75	0.0
100	0.65
100	0.50
120	1.0
130	0.95
130	1.30
160	1.80
190	2.80
200	2.50
240	4.30
250	4.50

Existe relação entre as duas variáveis? Calcule o coeficiente de correlação e interprete.

É possível testar a hipótese que o coeficiente de correlação seja igual a zero, ou seja,

$$H_0: \quad \rho = 0$$

$$H_1: \quad \rho \neq 0$$

A estatística do teste apropriada para as hipóteses acima é dado

$$t_{cal} = r\sqrt{\frac{n-2}{1-r^2}}$$

que segue uma distribuição t com $n-2$ graus de liberdade.

- **Exemplo 3:** Estamos estudando se existe ou não correlação entre as notas de diversas disciplinas de um curso de mestrado. Analisando uma amostra de 12 alunos encontrou-se uma correlação de 0,6 entre as disciplinas de Estatística e Metodologia da Pesquisa. Teste a hipótese de não haver correlação entre as disciplinas. Utilize um nível de significância de 5%.

Análise de regressão

- A teoria de Regressão teve origem no século XIX com Galton
- Em um de seus trabalhos ele estudou a relação entre a altura dos pais e dos filhos, procurando saber como a altura do pai influenciava a altura do filho
- Ele concluiu que se o pai fosse muito alto ou muito baixo, o filho teria uma altura tendendo à média
- Por isso, ele chamou de regressão, ou seja, existe uma tendência dos dados "regredirem" à média
- A análise de regressão possibilita explorar a relação entre duas ou mais variáveis

- Em muitos problemas há duas ou mais variáveis quantitativas que são relacionadas, e é importante estudar e explorar essa relação

- Por exemplo, o efeito da temperatura de operação de um processo industrial pode estar relacionado (ou pode explicar) o rendimento do produto final

- Pode ser de interesse construir um modelo relacionando as temperaturas e os rendimentos para predição

- Em geral, suponha que haja uma única variável dependente, ou reposta, Y que depende de k variáveis independentes ou explicativas, $X_1, X_2, ..., X_k$.

- A relação entre essas variáveis é caracterizada por um modelo chamado equação de regressão, que é ajustado a um conjunto de dados amostrais.

- Em algumas situações, o pesquisador conhece a forma exata da relação funcional entre Y e $X_1, X_2, ..., X_k$ dada por $Y = f(X_1, X_2, ..., X_k)$.

- Entretanto, em muitos casos, essa relação é desconhecida, e o pesquisador escolhe uma função apropriada para aproximar f.

- Modelos de regressão são frequentemente usados para analisar dados de um experimento planejado, tal pode surgir de observações de um fenômeno não controlado ou registros históricos.

- Determinar a relação entre uma única variável explicativa (ou explanatória) X e uma variável resposta Y.

- É usual assumir que a variável explicativa (ou explanatória) X seja contínua e controlada pelo pesquisador, ou seja, se o experimento é planejado, escolhe-se os valores de X e observa-se as respostas Y.

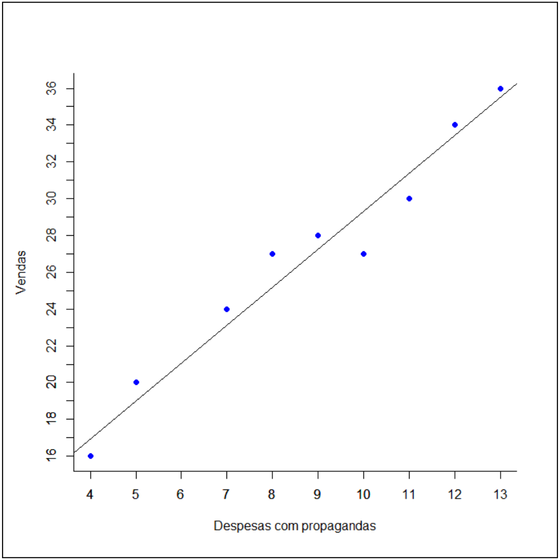
- Quando só existe uma variável explanatória, assume-se que cada observação Y pode ser descrita pelo modelo de regressão linear simples

$$Y_j = \beta_0 + \beta_1 X_j + \varepsilon_j, \qquad j = 1, \ldots, n$$

  em que ε é o erro aleatório em que $\varepsilon \sim N\left(0, \sigma^2\right)$

Ao estabelecer esse modelo pressupõe-se que

i) A relação entre Y e X é linear
ii) Os valores de X são fixos (ou controlados)
iii) A média do erro é nula
iv) Para um dado valor de X, a variância do erro ε_i é sempre σ^2
v) Os erros são independentes
vi) Os erros seguem distribuição normal



Vendas
36
34
32
30
28
26
24
22
20
18
16
4
5
6
7
8
9
10
11
12
13
Despesas com propagandas

Considere a estimação dos parâmetros do modelo, usando o método de mínimos quadrados. A função de mínimos quadrados é:

$$L = \sum_{j=1}^{n} \epsilon_j^2 = \sum_{j=1}^{n} (Y_j - \beta_0 - \beta_1 X_j)^2 \qquad (1)$$

Derivando-se L em relação aos parâmetros (β_0 e β_1) tem-se:

$$\begin{aligned} \frac{\partial L}{\partial \beta_0} &= 2\sum_{j=1}^{n}\left[y_j - \hat{\beta}_0 - \hat{\beta}_1 X_j\right] \times (-1) \\ \frac{\partial L}{\partial \beta_1} &= 2\sum_{j=1}^{n}\left[y_j - \hat{\beta}_0 - \hat{\beta}_1 X_j\right] \times (-X_j) \end{aligned}$$

igualando-se os resultados a zero e aplicando os somatórios, obtém-se o chamado **sistema de equações normais:**

$$\left\{ \begin{array}{lcl} \sum\limits_{j=1}^{n} Y_j & = & n\hat{\beta}_0 + \hat{\beta}_1 \sum\limits_{j=1}^{n} X_j \\ \sum\limits_{j=1}^{n} X_j Y_j & = & \hat{\beta}_0 \sum\limits_{j=1}^{n} X_j + \hat{\beta}_1 \sum\limits_{j=1}^{n} X_j^2 \end{array} \right. .$$

A solução para as equações normais é:

$$\hat{\beta}_0 = \frac{\sum\limits_{j=1}^{n} Y_j}{n} - \hat{\beta}_1 \frac{\sum\limits_{j=1}^{n} X_j}{n}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad (2)$$

Substituindo-se esse resultado na segunda equação do sistema de equações normais, tem-se:

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{n} X_j Y_j - n \bar{X} \bar{Y}}{\sum_{j=1}^{n} X_j^2 - \frac{\left(\sum_{j=1}^{n} X_j\right)^2}{n}}$$

O modelo de regressão linear simples ajustado é:

$$\hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_j$$

que fornece uma estimativa pontual da média de Y para cada valor de X

- Para $X = 0$, $\hat{\beta}_0$ representa o ponto em que a reta corta o eixo dos Y's e por isso é chamado **intercepto** (ou coeficiente linear)

- $\hat{\beta}_1$ é chamado **coeficiente de regressão ou coeficiente angular da reta**

- β_0 e β_1 são os parâmetros desconhecidos

- Após o ajuste do modelo de regressão, tem-se $\hat{\beta}_0$ e $\hat{\beta}_1$ que são as **estimativas dos parâmetros**

- Utilizando o método dos mínimos quadrados para estimar os parâmetros β_0 e β_1, temos:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_j Y_j - n\,\bar{X}\,\bar{Y}}{\sum_{j=1}^{n} X_j^2 - \frac{(\sum_{j=1}^{n} X_j)^2}{n}}$$

- A diferença entre o valor observado Y_i e o correspondente valor ajustado $\hat{Y}_j$ é chamado resíduo.

$$\text{res}_j = Y_j - \hat{Y}_j = Y_j - (\hat{\beta}_0 + \hat{\beta}_1 X_j), \qquad j = 1, 2, ..., n$$

Os resíduos têm papel importante **na verificação do ajuste do modelo** e nas suposições que são feitas.

Exemplo 4

Na Tabela 1 são apresentados a pureza do oxigênio produzido em um processo químico de destilação e a porcentagem de hidrocarboneto presentes no condensador principal da unidade de destilação.

Tabela 1: Níveis de Oxigênio e de Hidrocarbonetos

Observação	Nível de Hidrocarboneto (%)	Pureza (%)
1	0,99	90,01
2	1,02	89,05
3	1,15	91,43
4	1,29	93,74
5	1,46	96,73
6	1,36	94,45
7	0,87	87,59
8	1,23	91,77
9	1,55	99,42
10	1,40	93,65
11	1,19	93,54
12	1,15	92,52
13	0,98	90,56
14	1,01	89,54
15	1,11	89,85
16	1,20	90,39
17	1,26	93,25
18	1,32	93,41
19	1,43	94,98
20	0,95	87,33

a) Construa o diagrama de dispersão e calcule a correlação entre as variáveis.

b) Ajustar um modelo de regressão linear simples para os dados da Tabela 1 - sendo que a Pureza de Oxigênio é a variável resposta e o Nível de Hidrocarboneto é a variável explanatória.

(c) Determine a equação da reta de regressão linear.

(d) Realize uma análise de resíduos e conclua se o modelo estudado é adequado aos dados.

(e) Qual é o valor esperado da pureza de oxigênio quando o nível de hidrocarboneto for 1%.

(f) É possível fazer uma previsão para a pureza do oxigênio quando o nível de hidrocarboneto for 2% ? Justifique.

- O valor do coeficiente de correlação é $0,936$
- As variáveis pureza e nível de hidrocarboneto são positivamente altamente correlacionadas

O modelo de regressão linear simples ajustado é:

$$\hat{Y}_j = 74,28 + 14,95\, X_j$$

- Nesse ajuste, $\hat{\beta}_0 = 74,28$ e $\hat{\beta}_1 = 14,95$
- Intervalos de confiança e teste de hipóteses para os parâmetros do modelo

Coeficiente de Determinação

- A medida R^2 é chamada de coeficiente de determinação e seu campo de variação é $0 \leq R^2 \leq 1$ e indica a proporção da variação total que é "explicada" pela regressão.

- Se $R^2 = 1$, todos os pontos observados se situam "exatamente" sobre a reta de regressão, então, as variações de Y são 100% explicadas pelas variações de X através da função especificada.

- Por outro lado, um $R^2 = 0$ pode ou não indicar ausência de correlação entre X e Y .

- **Exemplo 5:** Um engenheiro químico está investigando o efeito da temperatura (X) de operação do processo no rendimento (Y) do produto. O estudo resultou nos dados da tabela seguinte:

Y	45	51	54	61	66	70	74	78	85	89
X	100	110	120	130	140	150	160	170	180	190

(a) Construa o gráfico de dispersão.

(b) Calcular o coeficiente de correlação e interpretar o resultado.

(c) Determine a equação da reta de regressão linear de Y em X.

(d) Estime o valor de Y para $X = 155$.