

Regressão Linear Simples

Elementos de Probabilidades e Estatística

1. Análise Bivariada
2. Regressão Linear Simples

1. Análise Bivariada

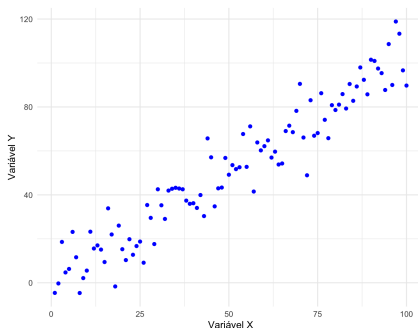
Análise bivariada é o estudo da relação entre duas variáveis quantitativas. Em Física, isso pode significar analisar como a velocidade de um corpo varia com o tempo ou como a energia cinética se relaciona com a massa.

Análise da associação entre duas variáveis

- 1 Começa pela recolha de uma amostra bivariada, ou seja, n pares de observações

$$(x_1, y_1); (x_2, y_2); \dots; (x_n, y_n)$$

- 2 Representação gráfica usando um **diagrama de dispersão**. Gráfico que mostra a relação entre duas variáveis. Cada ponto no gráfico representa uma observação.



Análise bivariada (cont.)

- 3 Verificar se a relação entre X e Y é linear. Isso é traduzido em verificar se parece existir ou não **correlação entre as variáveis** e em caso afirmativo, se essa correlação é:
- positiva ou negativa;
 - forte ou fraca;
 - tipo de correlação.
- 4 Se este for o caso, poderemos estudar a relação linear entre as duas variáveis usando um **modelo de regressão linear simples**, cuja equação é dada por

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Y – variável explicada ou dependente;
- X – variável explicativa ou independente;
- ϵ – variável de tipo residual que inclui outros fatores explicativos de Y não incluídos em X e possíveis erros de medição;
- β_0 e β_1 são os parâmetros da reta a ajustar (constantes);
- β_0 é a intersecção da reta com o eixo vertical;
- β_1 é o declive da reta.

Análise bivariada (cont.)

Covariância

A **covariância** entre duas variáveis x e y é uma **medida da associação linear** entre as duas variáveis:

$$\text{cov}[x, y] = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{n - 1}$$

O valor da covariância depende das unidades de medida utilizadas.

Análise da covariância

- Se $\text{cov}[x, y] > 0$, existe uma **associação linear positiva**, isto é, ambas as variáveis tendem a variar no mesmo sentido;
- Se $\text{cov}[x, y] < 0$, existe uma **associação linear negativa**, isto é, as variáveis tendem a variar em sentidos opostos;
- Se $\text{cov}[x, y] = 0$, **não existe associação linear** entre as variáveis.

A informação contida na covariância é principalmente sobre o sinal da associação entre x e y e não sobre a intensidade.

Coefficiente de correlação linear amostral

$$r = \frac{\text{cov}[x, y]}{s_x s_y}, \quad \text{com} \quad -1 \leq r \leq 1.$$

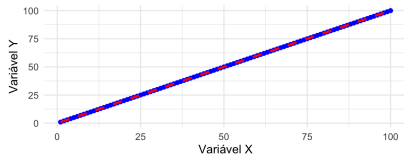
- Permite averiguar o grau de associação linear entre as duas variáveis.
- O seu valor não depende das unidades de medida que as variáveis estão expressas.

Análise do coeficiente de correlação

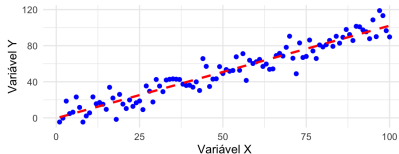
- $r = -1 \Rightarrow$ existe uma **correlação linear negativa perfeita** entre x e y ;
- $r = 1 \Rightarrow$ existe uma **correlação linear positiva perfeita** entre x e y ;
- $r = 0 \Rightarrow$ **não existe correlação linear** entre x e y ;
- $-1 < r < 0 \Rightarrow$ existe uma **correlação linear negativa** entre x e y (menos forte do que quando $r = -1$);
- $0 < r < 1 \Rightarrow$ existe uma **correlação linear positiva** entre x e y (menos forte do que quando $r = 1$).

Análise bivariada (cont.)

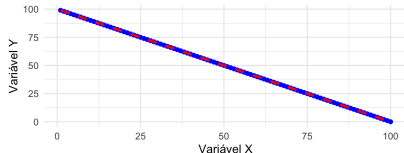
Correlação Positiva Perfeita



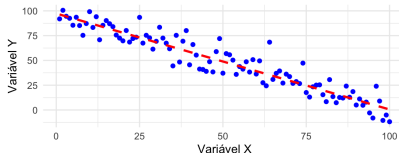
Correlação Positiva



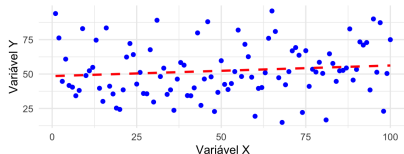
Correlação Negativa Perfeita



Correlação Negativa



Correlação Zero



Análise bivariada (cont.)

Exemplo 1: Investigadores de Geologia estão interessados em estudar a relação entre a porosidade de amostras de rocha e sua densidade. Entender essa relação pode ajudar a determinar a capacidade de armazenamento de fluidos das rochas, um fator crucial na exploração de petróleo e gás.

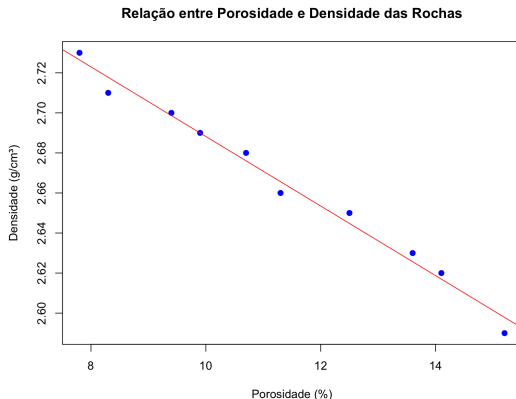
Os dados a seguir representam a porosidade (%) e a densidade (g/cm^3) de diferentes amostras de rochas coletadas em uma região específica. Com base nos dados coletados, calcule a covariância e o coeficiente de correlação entre a porosidade e a densidade das rochas.

Amostra	Porosidade (%)	Densidade (g/cm^3)
1	12.5	2.65
2	8.3	2.71
3	15.2	2.59
4	10.7	2.68
5	9.4	2.70
6	11.3	2.66
7	13.6	2.63
8	7.8	2.73
9	14.1	2.62
10	9.9	2.69

Análise bivariada (cont.)

Covariância entre Porosidade e Densidade = -0.1099778

Coefficiente de Correlação entre Porosidade e Densidade = -0.9931263



Análise bivariada no R

```
# Dados
porosidade <- c(12.5, 8.3, 15.2, 10.7, 9.4, 11.3, 13.6, 7.8, 14.1, 9.9)
densidade <- c(2.65, 2.71, 2.59, 2.68, 2.70, 2.66, 2.63, 2.73, 2.62, 2.69)

# Calcular a covariância
covariancia <- cov(porosidade, densidade)

# Calcular o coeficiente de correlação
correlacao <- cor(porosidade, densidade)

# Exibir os resultados
cat("Covariância entre Porosidade e Densidade:", covariancia, "\n")
cat("Coeficiente de Correlação entre Porosidade e Densidade:", correlacao)

# Gráfico de dispersão
plot(porosidade, densidade, main = "Relação entre Porosidade e Densidade
      das Rochas", xlab = "Porosidade (%)", ylab = "Densidade (g/cm³)",
      pch = 19, col = "blue")

# Adicionar linha de tendência
abline(lm(densidade ~ porosidade), col = "red")
```

2. Regressão Linear Simples

Regressão Linear Simples

Verificada a existência da relação/associação linear entre as variáveis, vamos definir agora nosso modelo de **regressão linear simples (RLS)** através da seguinte equação matemática

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \dots, n$$

onde:

- Y_i : variável de resposta (ou dependente) associada à i -ésima prova - **variável aleatória**;
- x_i : variável explicativa (regressora ou independente) - **constante conhecida, não-aleatória**;
- β_0 : ordenada na origem - **parâmetro a estimar**;
- β_1 : declive ou coeficiente angular - **parâmetro a estimar**;
- ϵ_i : **erro aleatório** associado à i -ésima prova, verificando:
 - $E(\epsilon_i) = 0, \forall_i$ (valor esperado nulo);
 - $V(\epsilon_i) = \sigma^2$ (variância constante independente de i) $\forall_i$, **parâmetro a estimar**;
 - $Cov(\epsilon_i, \epsilon_j) = 0, \forall_{i,j}$ com $(i \neq j)$ (não correlacionados);
 - ϵ_i é uma variável de tipo residual que inclui outros fatores explicativos de Y_i não incluídos em x_i e possíveis erros de medição.

Regressão Linear Simples

Observação

- 1 Como $E(\epsilon_i) = 0$ então

$$E(Y|x_i) = \beta_0 + \beta_1 x_i \quad (\text{reta de regressão})$$

Ou seja, para cada valor x_i o ponto sobre a reta tem ordenada $E(Y|x_i) = \beta_0 + \beta_1 x_i$.

- 2 Interpretação dos parâmetros β_0 e β_1 :

- $\beta_0 = E(Y|x = 0) \rightarrow$ ordenada na origem;
- $\beta_1 = E(Y|x = x^* + 1) - E(Y|x = x^*), \forall x^* \rightarrow$ declive;

- 3 O modelo RLS diz-se

- **simples** por só haver **uma variável explicativa** (em vez de por exemplo $Y_i = \beta_0 + \beta_1(x_1)_i + \beta_2(x_2)_i + \epsilon_i$).
- **linear** por ser **linear nos parâmetros** ($\beta_0, \beta_1, \dots$) (ou seja, modelos do tipo $Y = \beta_0 + \beta_1 x^2 + \epsilon$ ou $\log Y = \beta_0 + \beta_1 \log x + \epsilon$ também são lineares).

Regressão Linear Simples

A princípio, o modelo de regressão linear simples é dado por

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

onde Y é a variável resposta, x é a variável explicativa, β_0 e β_1 são parâmetros desconhecidos e ϵ é um termo de erro aleatório, que assumimos ter média zero e variância finita σ^2 .

O ponto de partida para a estimação pontual dos parâmetros β_0 e β_1 é o conjunto de dados (x_i, y_i) , $i = 1, \dots, n$. A obtenção das estimativas dos mínimos quadrados de β_0 e β_1 passa pela minimização da soma dos quadrados dos desvios (verticais) entre o que é observado, y_i , e o que é *esperado* de acordo com o modelo de RLS, $E(Y|x_i) = \beta_0 + \beta_1 x_i$. Com efeito, pretendemos encontrar estimativas de β_0 e β_1 (denotadas por $\hat{\beta}_0$ e $\hat{\beta}_1$) que minimizem

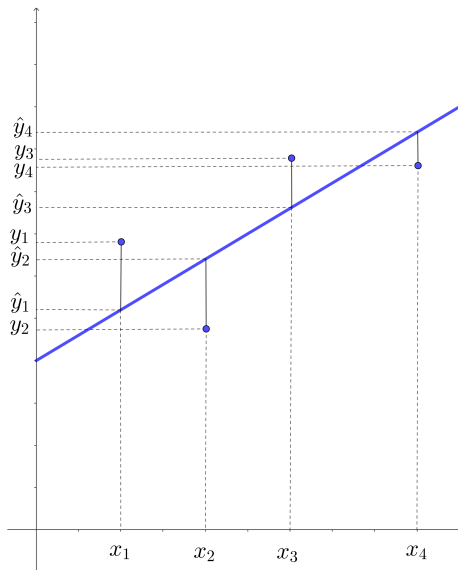
$$Q = Q(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - E(Y|x_i))^2 = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2.$$

Essa escolha dos coeficientes minimizadores define a **reta de regressão** ou **reta ajustada**, que é definida por

$$\hat{y} = \widehat{E(Y|x)} = \hat{\beta}_0 + \hat{\beta}_1 x, \quad \text{para } x \in [x_{(1)}, x_{(n)}],$$

onde $x_{(1)} = \min_{i=1, \dots, n} x_i$ e $x_{(n)} = \max_{i=1, \dots, n} x_i$.

Regressão Linear Simples (cont.)



Proposição (Estimativas de mínimos quadrados de β_0 e β_1)

Serão representados por $\hat{\beta}_0$ e $\hat{\beta}_1$ e são a solução do seguinte sistema de duas equações lineares:

$$(\hat{\beta}_0, \hat{\beta}_1) : \begin{cases} \frac{\partial Q}{\partial \beta_0} \Big|_{\beta_0=\hat{\beta}_0, \beta_1=\hat{\beta}_1} = 0 \\ \frac{\partial Q}{\partial \beta_1} \Big|_{\beta_0=\hat{\beta}_0, \beta_1=\hat{\beta}_1} = 0 \end{cases} \Leftrightarrow \begin{cases} 2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ 2 \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{cases}$$

Assim, as **estimativas de mínimos quadrados de β_0 e β_1** são iguais a

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}, \end{aligned}$$

com $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ e $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$.

Proposição (Estimadores de mínimos quadrados de β_0 e β_1)

Serão também representados por $\hat{\beta}_0$ e $\hat{\beta}_1$ e dados por:

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i Y_i - n \bar{x} \bar{Y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}.$$

Observações

- 1 Quando existe solução ela corresponde de facto a um mínimo (tem matriz Hessiana definida positiva).
- 2 Existe solução se e só se $\sum_{i=1}^n x_i^2 - n\bar{x}^2 = \sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$, ou seja, se na amostra existirem pelo menos dois valores distintos de x .

3

$$\hat{y} = \widehat{E(Y|x)} = \hat{\beta}_0 + \hat{\beta}_1 x$$

é a estimativa de mínimos quadrados da reta de regressão. A estimação pontual de $E(Y|x)$ deve restringir-se ao domínio dos valores observados na amostra da variável explicativa x , isto é, $x \in [x_{(1)}, x_{(n)}]$.

- 4 Às diferenças

$$e_i = y_i - \hat{y}_i, \quad \text{para } i = 1, \dots, n$$

chamam-se **resíduos**. A análise dos resíduos permite avaliar se o modelo assumido é adequado.

Regressão Linear Simples (cont.)

Exemplo 2: A perda percentual de massa (Y) de uma certa substância metálica (quando exposta a oxigénio seco a 500°C) depende do período de exposição (x ; em hora). Cinco medições conduziram a: $\sum_{i=1}^5 x_i = 12$, $\sum_{i=1}^5 x_i^2 = 32.5$, $\sum_{i=1}^5 y_i = 0.177$, $\sum_{i=1}^5 y_i^2 = 0.006789$, $\sum_{i=1}^5 x_i y_i = 0.4685$, onde

$$\left[\min_{i=1,\dots,5} x_i, \max_{i=1,\dots,5} x_i \right] = [1.0, 3.5].$$

Calcule as estimativas de mínimos quadrados dos parâmetros da reta de regressão linear simples de Y em x .

Regressão Linear Simples (cont.)

Temos

$$\hat{\beta}_1 = \frac{\sum_{i=1}^5 x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^5 x_i^2 - n \bar{x}^2} = \frac{0.4685 - 5(\frac{12}{5})(\frac{0.177}{5})}{32.5 - 5(\frac{12}{5})^2} = 0.01181081$$

$$\hat{\beta}_0 = \hat{y} - \hat{\beta}_1 \bar{x} = \frac{0.177}{5} - 0.01181081(\frac{12}{5}) = 0.007056$$

- A perda percentual de massa aumenta em média 0.01181 unidades para cada hora adicional de exposição ao oxigénio seco a 500°C.
- Quando $x = 0$ (ou seja, sem exposição), a perda estimada é muito próxima de zero: cerca de 0.00706.

Coeficiente de determinação

O coeficiente de determinação, R^2 , é uma medida descritiva indicadora da qualidade do ajustamento da reta estimada. Lembre que os **resíduos** são dados por $e_i = y_i - \hat{y}_i$ com $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Assim

$$\sum_{i=1}^n e_i = \sum_{i=1}^n y_i - \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i) = n\bar{y} - n\hat{\beta}_0 - \hat{\beta}_1 n\bar{x} = n\bar{y} - n(\bar{y} - \hat{\beta}_1 \bar{x}) - n\hat{\beta}_1 \bar{x} = 0$$

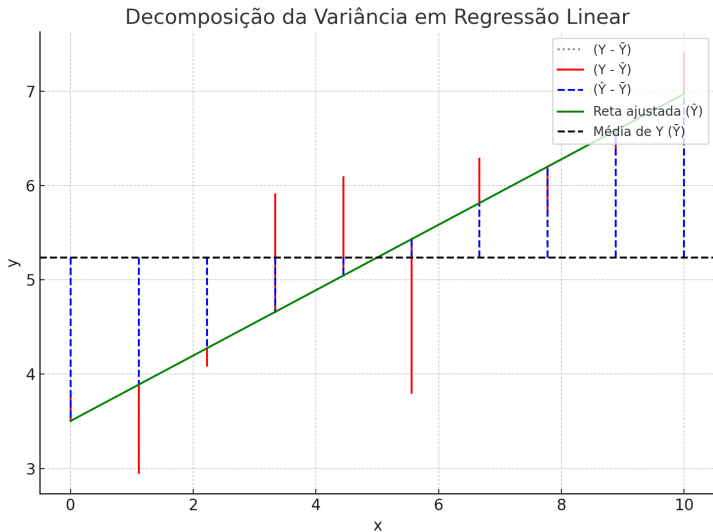
Mostra-se que:

$$\underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{SST} = \underbrace{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}_{SSE} + \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{SSR}$$

onde

- SST corresponde à soma de quadrados total
- SSE corresponde à soma dos quadrados dos resíduos
- SSR é a soma dos quadrados da regressão

Coeficiente de determinação



Coeficiente de determinação (cont.)

Definição

O coeficiente de determinação é definido por

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$
$$\Rightarrow R^2 = \frac{\left(\left(\sum_{i=1}^n x_i Y_i \right) - n\bar{x}\bar{Y} \right)^2}{\left(\left(\sum_{i=1}^n x_i^2 \right) - n\bar{x}^2 \right) \left(\left(\sum_{i=1}^n Y_i^2 \right) - n\bar{Y}^2 \right)}$$

Observação

- 1 $0 \leq R^2 \leq 1$ e $(R^2 \times 100)\%$ representa a percentagem da variabilidade total que é explicada pelo modelo de RLS,
- 2 $R^2 = 1 \Leftrightarrow \hat{y}_i = y_i, \quad \forall_i$. Todos os pontos da amostra estão sobre a reta, pelo que o modelo de RLS explica toda a variabilidade observada. O modelo é **ótimo!**.
- 3 $R^2 = 0 \Leftrightarrow \hat{y}_i = \bar{y}, \quad \forall_i \Leftrightarrow \hat{\beta}_1 = 0 \Leftrightarrow \hat{\beta}_0 = \bar{Y}$. A reta é horizontal, pelo que o modelo de RLS não explica nada da variabilidade observada. O modelo é **péssimo!**.
- 4 Observar que a expressão do coeficiente de determinação em RLS corresponde ao quadrado do coeficiente de correlação linear (r_{xy}) que vimos no Cap.1.

Exemplo 3: É geralmente aceite que a frequência cardíaca (Y , em batimentos por minuto) é influenciada pela temperatura corporal dos seres humanos (x , em $^{\circ}C$). Um conjunto de 130 medições independentes conduziu aos seguintes resultados:

$\sum_{i=1}^{130} x_i = 4784.7$, $\sum_{i=1}^{130} x_i^2 = 176121.67$, $\sum_{i=1}^{130} y_i = 9589$, $\sum_{i=1}^{130} y_i^2 = 713733$,
 $\sum_{i=1}^{130} x_i y_i = 353018.5$. Calcule o valor do coeficiente de determinação e comente a utilidade do modelo ajustado.

Coeficiente de determinação (cont.)

$$\sum_{i=1}^{130} x_i y_i - n \bar{x} \bar{y} = 353018.5 - 130 \times \frac{4784.7}{130} \times \frac{9589}{130} = 91.7$$

$$\sum_{i=1}^{130} x_i^2 - n \bar{x}^2 = 176121.67 - 130 \times \left(\frac{4784.7}{130} \right)^2 = 18.9$$

$$\sum_{i=1}^{130} y_i^2 - n \bar{y}^2 = 713733 - 130 \times \left(\frac{9589}{130} \right)^2 = 6433.6$$

$$\hat{\beta}_1 = \frac{91.7}{18.9} = 4.85 \quad \text{e} \quad \hat{\beta}_0 = \frac{9589}{130} - 4.85 \times \frac{4784.7}{130} = -104.74$$

$$R^2 = \frac{(91.7)^2}{18.9 \times 6433.6} = 0.069$$

O modelo só explica cerca de 7% da variabilidade, pelo que é muito pouco útil.

Estimação de $V(\epsilon_i) = \sigma^2$

Estimação de $\sigma^2 = V(\epsilon_i)$

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^n e_i^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ &= \frac{1}{n-2} \left[\left(\sum_{i=1}^n Y_i^2 - n\bar{Y}^2 \right) - (\hat{\beta}_1)^2 \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) \right]\end{aligned}$$

onde $\hat{Y}_i = \widehat{E[Y|x_i]} = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Nota:

- $\hat{\sigma}^2$ é também um **estimador da variância de Y_i** .
- $\hat{\sigma}$ = **erro padrão dos resíduos**.

Inferência sobre os parâmetros da reta de regressão

Para se fazerem inferências em RLS (i.e, realizar testes de hipóteses e calcular intervalos de confiança) é necessário admitir que os **erros tenham distribuição normal**.

As suposições do modelo de RLS são então:

- $E(\epsilon_i) = 0$,
- $V(\epsilon_i) = \sigma^2$ e
- $Cov(\epsilon_i, \epsilon_j) = 0, \quad \forall_{i,j} \quad (i \neq j)$.

Com a **nova hipótese de trabalho** tem-se agora que

$$\epsilon_i \underset{i.i.d.}{\sim} N(0, \sigma).$$

Inferência sobre os parâmetros da reta de regressão

É então possível mostrar que

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\right) \hat{\sigma}^2}} \sim t_{(n-2)}$$
$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}} \sim t_{(n-2)}$$

Isto permite calcular intervalos de confiança e realizar testes de hipóteses relativos a β_0 e β_1 .

Inferências relativas à ordenada na origem, β_0

Intervalo de confiança a $(1 - \alpha) \times 100\%$ para β_0

Variável fulcral:

$$T = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\right) \hat{\sigma}^2}} \sim t_{(n-2)}$$

Obtém-se, após dedução, o seguinte **intervalo para β_0** a $(1 - \alpha) \times 100\%$ de confiança

$$IC_{(1-\alpha) \times 100\%}(\beta_0) = \hat{\beta}_0 \pm F_{t_{(n-2)}}^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\right) \hat{\sigma}^2}$$

Inferências relativas à ordenada na origem, β_0

Exemplo 4: Um grupo de estudantes realizou uma experiência para estudar a relação entre a **temperatura ambiente** x (em $^{\circ}\text{C}$) e a **tensão gerada por um sensor térmico** y (em volts). Os dados coletados foram ajustados por um modelo de regressão linear simples:

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

Ajustando o modelo aos dados, obteve-se:

- $\hat{\beta}_0 = 0.42$
 - $\hat{\beta}_1 = 0.016$
 - Erro padrão dos resíduos: $\hat{\sigma} = 0.05$
 - Média das temperaturas: $\bar{x} = 22$
 - $\sum_{i=1}^{12} (x_i - \bar{x})^2 = 280$
 - Número de observações: $n = 12$
- (a) Construa um **intervalo de confiança a 95% para β_0** , assumindo que os erros do modelo são normais e independentes.
- (b) Interprete o intervalo de confiança encontrado. O que ele nos diz sobre a tensão gerada quando a temperatura é zero?

Inferências relativas à ordenada na origem, β_0

Sabemos que o IC para β_0 é da forma:

$$\text{IC}_{(1-\alpha) \times 100\%}(\beta_0) = \hat{\beta}_0 \pm F_{t_{(n-2)}}^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\right) \hat{\sigma}^2}$$

Assim,

$$\begin{aligned}\text{IC}_{95\%}(\beta_0) &= 0.42 \pm F_{t_{(10)}}^{-1}(0.975) \sqrt{\left(\frac{1}{12} + \frac{22^2}{280}\right) 0.05^2} \\ &= 0.42 \pm 2.228 \times 0.0673 \\ &= 0.42 \pm 0.1499 \\ &= [0.2701, 0.5699]\end{aligned}$$

(b) O intervalo de confiança nos diz que, com 95% de confiança, o valor verdadeiro da ordenada na origem β_0 — ou seja, a tensão prevista quando a temperatura é 0°C — está entre aproximadamente 0.27V e 0.57V.

Inferências relativas à ordenada na origem, β_0

Teste de hipóteses a β_0

- 1 Definir $H_0 : \beta_0 = \beta_{0,0}$ vs $H_1 : \beta_0 >, <, \neq \beta_{0,0}$
- 2 Estatística de teste:

$$T = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\right) \hat{\sigma}^2}} \underset{\text{Sob } H_0}{\sim} t_{(n-2)}$$

- 3 Região Crítica (ao nível de significância α):

- $H_1 : \beta_0 \neq \beta_{0,0}$, $RC : |T_{H_0}| > c$, com $c = F_{t_{(n-2)}}^{-1}\left(1 - \frac{\alpha}{2}\right)$
- $H_1 : \beta_0 > \beta_{0,0}$, $RC : T_{H_0} > c$, com $c = F_{t_{(n-2)}}^{-1}(1 - \alpha)$
- $H_1 : \beta_0 < \beta_{0,0}$, $RC : T_{H_0} < c$, com $c = F_{t_{(n-2)}}^{-1}(\alpha)$

- 4 Ou cálculo do valor-p.

Inferências relativas à ordenada na origem, β_0

Exemplo 5: Considere o mesmo contexto experimental do exercício anterior, em que se estuda a relação entre a **temperatura ambiente** x (em °C) e a **tensão gerada por um sensor térmico** y (em volts), por meio de um modelo de regressão linear simples:

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

Os valores estimados a partir dos dados experimentais foram:

- $\hat{\beta}_0 = 0.42$
- Erro padrão dos resíduos: $\hat{\sigma} = 0.05$
- Média das temperaturas: $\bar{x} = 22$
- $\sum_{i=1}^{12} (x_i - \bar{x})^2 = 280$
- Número de observações: $n = 12$

Pretende-se testar a hipótese de que a ordenada na origem β_0 é nula, ou seja, que a tensão gerada pelo sensor é zero quando a temperatura ambiente é zero utilizando um nível de significância de 5%.

Inferências relativas à ordenada na origem, β_0

Hipóteses

$$H_0 : \beta_0 = \beta_{0,0} = 0$$

$$H_1 : \beta_0 \neq 0 \text{ (teste bilateral)}$$

Nível de significância: $\alpha = 0.05$

Estatística de teste

$$T = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\right) \hat{\sigma}^2}} \underset{Sob H_0}{\sim} t_{(n-2)}$$

Região crítica

$$\begin{aligned} RC &=] -\infty, F_{t_{(n-2)}}^{-1}(1 - \alpha/2)[\cup] F_{t_{(n-2)}}^{-1}(1 - \alpha/2), +\infty[\\ &=] -\infty, F_{t_{(10)}}^{-1}(0.975)[\cup] F_{t_{(10)}}^{-1}(0.975), +\infty[\\ &=] -\infty, -2.228[\cup] 2.228, +\infty[\end{aligned}$$

Inferências relativas à ordenada na origem, β_0

Valor observado da estatística de teste

$$t_{obs} = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\right) \hat{\sigma}^2}} = \frac{0.42 - 0}{\sqrt{\left(\frac{1}{12} + \frac{22^2}{280}\right) \times 0.05^2}} = \frac{0.42}{0.0673} = 6.2407$$

Decisão

Como $t_{obs} \in RC$, então devemos rejeitar H_0 ao nível de significância de 5%.

Conclusão

Há evidência estatística suficiente suportando que a ordenada na origem é diferente de zero.

Decisão com base no valor-p

Como temos um teste bilateral, então

$$\begin{aligned} \text{valor-p} &= 2 \min\{P(T > t_{obs}), P(T < t_{obs})\} \\ &= 2 \min\{P(T > 6.2407), P(T < 6.2407)\} \\ &= 2 \min\{0.0001, 0.9999\} \\ &= 0.0002 \end{aligned}$$

Como $\alpha = 0.05 > 0.0002 = \text{valor-p}$ então deve-se rejeitar H_0 .

Inferências relativas ao declive, β_1

Intervalo de confiança a $(1 - \alpha) \times 100\%$ para β_1

Variável fulcral:

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\left(\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\right)}} \sim t_{(n-2)}$$

Obtém-se, após dedução, o seguinte intervalo para β_1 a $(1 - \alpha) \times 100\%$ de confiança

$$IC_{(1-\alpha) \times 100\%}(\beta_1) = \hat{\beta}_1 \pm F_{t_{(n-2)}}^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}$$

Inferências relativas ao declive, β_1

Exemplo 6: Um grupo de estudantes realizou uma experiência para estudar a relação entre a **temperatura ambiente** x (em $^{\circ}\text{C}$) e a **tensão gerada por um sensor térmico** y (em volts). Os dados coletados foram ajustados por um modelo de regressão linear simples:

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

Ajustando o modelo aos dados, obteve-se:

- $\hat{\beta}_0 = 0.42$
 - $\hat{\beta}_1 = 0.016$
 - Erro padrão dos resíduos: $\hat{\sigma} = 0.05$
 - Média das temperaturas: $\bar{x} = 22$
 - $\sum_{i=1}^{12} (x_i - \bar{x})^2 = 280$
 - Número de observações: $n = 12$
- (a) Construa um **intervalo de confiança a 95% para β_1** , assumindo que os erros do modelo são normais e independentes.
- (b) Interprete o intervalo de confiança encontrado. O que ele nos diz sobre a variação da tensão com o aumento da temperatura?

Inferências relativas ao declive, β_1

Sabemos que o IC para β_1 é da forma:

$$IC_{(1-\alpha) \times 100\%}(\beta_1) = \hat{\beta}_1 \pm F_{t_{(n-2)}}^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}$$

Assim,

$$\begin{aligned} IC_{95\%}(\beta_1) &= 0.016 \pm F_{t_{(10)}}^{-1}(0.975) \sqrt{\left(\frac{0.05^2}{280}\right)} \\ &= 0.016 \pm 2.228 \times 0.0029 \\ &= 0.016 \pm 0.0064 \\ &= [0.0096, 0.0224] \end{aligned}$$

(b) Com 95% de confiança, o verdadeiro valor de β_1 , ou seja, o aumento médio da tensão (em volts) para cada aumento de 1°C na temperatura, está entre aproximadamente 0.00934 V e 0.02266 V.

Inferências relativas ao declive, β_1

Teste de hipóteses a β_1

❶ Definir $H_0 : \beta_1 = \beta_{1,0}$ vs $H_1 : \beta_1 >, <, \neq \beta_{1,0}$

❷ Estatística de teste:

$$T = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}} \underset{Sob H_0}{\sim} t_{(n-2)}$$

❸ Região Crítica (ao nível de significância α):

- $H_1 : \beta_1 \neq \beta_{1,0}$, $RC : |T_{H_0}| > c$, com $c = F_{t_{(n-2)}}^{-1}(1 - \frac{\alpha}{2})$
- $H_1 : \beta_1 > \beta_{1,0}$, $RC : T_{H_0} > c$, com $c = F_{t_{(n-2)}}^{-1}(1 - \alpha)$
- $H_1 : \beta_1 < \beta_{1,0}$, $RC : T_{H_0} < c$, com $c = F_{t_{(n-2)}}^{-1}(\alpha)$

❹ Ou cálculo do valor-p.

Observações

- 1 Um teste importante em RLS é

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

pois não rejeitar H_0 significa que há evidência para a não existência de uma associação linear entre x e y , ou seja não há associação ou a associação não é linear. A este teste costuma designar-se por: **teste à significância da regressão**.

Exemplo 7: Considere o mesmo contexto experimental do exercício anterior, em que se estuda a relação entre a **temperatura ambiente** x (em °C) e a **tensão gerada por um sensor térmico** y (em volts), por meio de um modelo de regressão linear simples:

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

Os valores estimados a partir dos dados experimentais foram:

- $\hat{\beta}_1 = 0.016$
- Erro padrão dos resíduos: $\hat{\sigma} = 0.05$
- $\sum_{i=1}^{12} (x_i - \bar{x})^2 = 280$
- Número de observações: $n = 12$

Pretende-se testar a hipótese de que o coeficiente angular β_1 é nulo, ou seja, que **não existe relação linear entre temperatura e tensão gerada pelo sensor**, utilizando um nível de significância de 5%.

Inferências relativas ao declive, β_1

Hipóteses

$$H_0 : \beta_1 = \beta_{1,0} = 0$$

$$H_1 : \beta_1 \neq 0 \text{ (teste bilateral)}$$

Nível de significância: $\alpha = 0.05$

Estatística de teste

$$T = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}} \underset{\text{Sob } H_0}{\sim} t_{(n-2)}$$

Região crítica

$$\begin{aligned} RC &=] -\infty, F_{t_{(n-2)}}^{-1}(1 - \alpha/2)[\cup] F_{t_{(n-2)}}^{-1}(1 - \alpha/2), +\infty[\\ &=] -\infty, F_{t_{(10)}}^{-1}(0.975)[\cup] F_{t_{(10)}}^{-1}(0.975), +\infty[\\ &=] -\infty, -2.228[\cup] 2.228, +\infty[\end{aligned}$$

Inferências relativas ao declive, β_1

Valor observado da estatística de teste

$$t_{obs} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}} = \frac{0.016 - 0}{\sqrt{\frac{0.05^2}{280}}} = \frac{0.016}{0.0029} = 5.5172$$

Decisão

Como $t_{obs} \in RC$, então devemos rejeitar H_0 ao nível de significância de 5%.

Conclusão

Há evidência estatística suficiente suportando que declive é diferente de zero.

Decisão com base no valor-p

Como temos um teste bilateral, então

$$\begin{aligned}\text{valor-p} &= 2 \min\{P(T > t_{obs}), P(T < t_{obs})\} \\ &= 2 \min\{P(T > 5.5172), P(T < 5.5172)\} \\ &= 2 \min\{0.001, 0.999\} \\ &= 0.002\end{aligned}$$

Como $\alpha = 0.05 > 0.002 = \text{valor-p}$ então deve-se rejeitar H_0 .

Exemplo 8: Para descrever a relação existente entre o volume de uma massa de um gás ideal clássico e a respetiva pressão, registaram-se 10 valores do logaritmo de base 10 do volume, x (com o volume medido em polegadas ao quadrado), e os correspondentes valores experimentais do logaritmo de base 10 da pressão, Y (com a pressão medida em psi). Pretendendo avaliar-se a validade do modelo de regressão linear simples para descrever a relação existente entre o logaritmo da pressão do gás e o logaritmo do seu volume, efetuaram-se os seguintes cálculos:

$$\sum_{i=1}^{10} x_i = 19.4, \sum_{i=1}^{10} x_i^2 = 38.06, \sum_{i=1}^{10} y_i = 14.8, \sum_{i=1}^{10} y_i^2 = 22.76, \sum_{i=1}^{10} x_i y_i = 28.12.$$

- (a) Obtenha as estimativas de mínimos quadrados dos parâmetros da reta de regressão linear simples de Y em x e interprete o significado do sinal da estimativa do parâmetro β_1 do modelo.
- (b) Indicando as hipóteses de trabalho convenientes, obtenha um intervalo de confiança a 95% para o parâmetro β_1 do modelo de regressão linear simples de Y em x . O que pode concluir sobre a significância do modelo de regressão ao nível de significância de 5%?

Inferências relativas ao declive, β_1

1. Médias amostrais:

$$\bar{x} = \frac{19,4}{10} = 1.94, \quad \bar{y} = \frac{14.8}{10} = 1.48$$

2. Estimativa de $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} = \frac{28.12 - 10 \cdot 1.94 \cdot 1.48}{38.06 - 10 \cdot (1.94)^2} = \frac{-0.592}{0.4236} \approx -1.3973$$

3. Estimativa de $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 1.48 + 1.3973 \cdot 1.94 \approx 4.1918$$

Equação da reta de regressão:

$$\hat{y} = 4.1918 - 1.3973x$$

Inferências relativas ao declive, β_1

Assumindo como hipótese de trabalho que $\epsilon_i \underset{i.i.d.}{\sim} N(0, \sigma)$, $\forall i=1, \dots, 10$ vimos que

$$IC_{95\%}(\beta_1) = \left[\hat{\beta}_1 - a \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2}}, \hat{\beta}_1 + a \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2}} \right]$$

com $a = F_{t(8)}^{-1}(1 - \frac{0.05}{2}) = F_{t(8)}^{-1}(0.975) = 2.306$ (tabela)

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{10-2} \left[\left(\sum_{i=1}^{10} y_i^2 - 10\bar{y}^2 \right) - \hat{\beta}_1^2 \left(\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2 \right) \right] \\ &= \frac{1}{8} \left[(22.76 - 10(\frac{14.8}{10})^2) - (1.396)^2 \times 0.424 \right] \\ &= \frac{1}{8} [0.856 - 0.826] \\ &= 0.00375\end{aligned}$$

Portanto,

$$\begin{aligned} \text{IC}_{95\%}(\beta_1) &= \left[-1.396 - 2.306\sqrt{\frac{0.00375}{0.424}}, -1.396 + 2.306\sqrt{\frac{0.00375}{0.424}} \right] \\ &= [-1.396 - 0.217, -1.396 + 0.217] \\ &= [-1.613, -1.179] \end{aligned}$$

Nota: $R^2 = \frac{(-0.592)^2}{0.424 \times 0.856} = 0.966$.

Modelo RLS é bem ajustado à situação.

Tomar a decisão considerando a relação entre IC e teste de hipótese bilateral.

Exemplo

Os dados obtidos no experimento são:

T ($^{\circ}\text{C}$)	0	10	20	30	40	50	60	70	80	90	100
p (atm)	0.94	0.96	1.00	1.05	1.07	1.09	1.14	1.17	1.21	1.24	1.28

- 1 Ajuste uma reta de regressão linear entre T e p ;
- 2 Use a equação da reta para determinar o valor de T tal que $p = 0$;
- 3 Interprete fisicamente o significado desse valor de T .
- 4 Calcule o coeficiente de determinação.

Exemplo

Ajustamos um modelo de regressão linear da forma:

$$p = \hat{\beta}_0 + \hat{\beta}_1 T$$

Com os dados do experimento, obtemos:

$$\hat{\beta}_0 = \bar{p} - \hat{\beta}_1 \bar{T} = 1.104545 - 0.0034 \times 50 = 0.9345,$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{11} T_i p_i - n \bar{T} \bar{p}}{\sum_{i=1}^{11} T_i^2 - n \bar{T}^2} = \frac{645.1 - 11 \times 50 \times 1.104545}{38500 - 11 \times 50^2} = 0.0034$$

Queremos encontrar a temperatura T_0 para a qual $p = 0$:

$$0 = \hat{\beta}_0 + \hat{\beta}_1 T_0 \Rightarrow T_0 = -\frac{\hat{\beta}_0}{\hat{\beta}_1} = -273.14^\circ\text{C}$$

Interpretação: O valor $T_0 \approx -273.14^\circ\text{C}$ representa uma extrapolação do ponto onde a pressão seria zero. Este valor é uma estimativa experimental da temperatura do **zero absoluto**, ou seja, cerca de $-273,15^\circ\text{C}$.

Exemplo

$$R^2 = \frac{\left(\left(\sum_{i=1}^n x_i Y_i \right) - n\bar{x}\bar{Y} \right)^2}{\left(\left(\sum_{i=1}^n x_i^2 \right) - n\bar{x}^2 \right) \left(\left(\sum_{i=1}^n Y_i^2 \right) - n\bar{Y}^2 \right)} = 0.9957$$

Cerca de 99.57% da variabilidade da pressão é explicada à custa da temperatura, em °C.

- Manuel Cabral Morais (2020): Probabilidade e Estatística: Teoria, Exemplos & Exercícios. IST Press, 1a edição.