Estatística I

Prof. Fernando de Souza Bastos
fernando.bastos@ufv.br

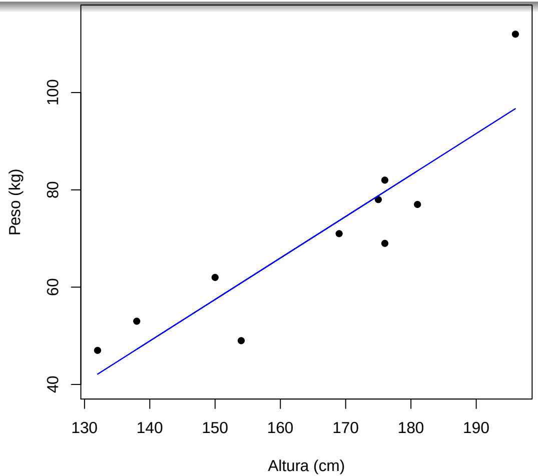
Departamento de Estatística
Universidade Federal de Viçosa
Campus UFV - Viçosa

Sumário

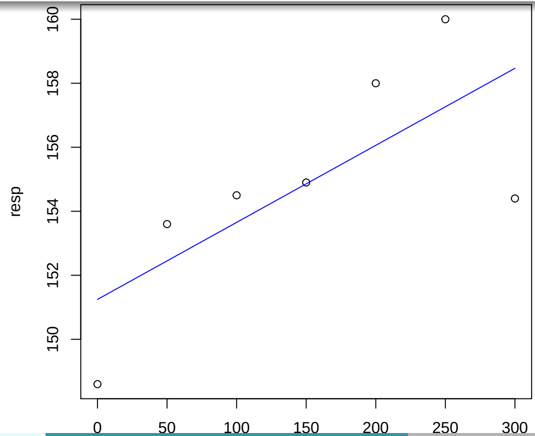
1. Regressão Linear Simples
2. Coeficiente de Correlação
3. Modelo de Regressão
4. Estimação por Mínimos Quadrados
5. Teste de Hipótese para Regressão Linear Simples
6. Análise de Variância (ANOVA) para a Regressão Linear Simples

Em certas situações podemos estar interessados em descrever a relação entre duas variáveis, e também predizer o valor de uma a partir de outra. Por exemplo, se sabemos a altura de um certo estudante, mas não o seu peso, qual seria um bom chute para o peso deste estudante?

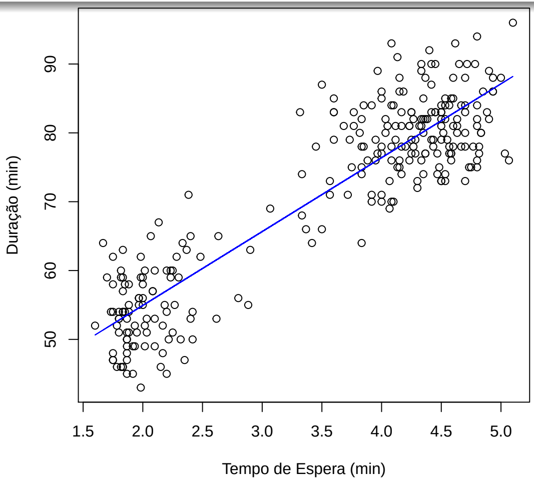
```r
x = c(176, 154, 138, 196, 132, 176, 181, 169, 150, 175)
y = c(82, 49, 53, 112, 47, 69, 77, 71, 62, 78)
plot(x,y,xlab="Altura (cm)",ylab="Peso (kg)",
      pch=16, ylim = c(40,115))
lines(x, fitted(lm(y ~ x)), col="blue")
```



```
doses <- c(0, 50, 100, 150, 200, 250, 300)
resp <- c(148.6, 153.6, 154.5, 154.9, 158, 160, 154.4)
reglin <- lm(resp ~ doses)
plot(doses, resp) #(variável indep. primeiro)
lines(doses, fitted(reglin), col="blue")#acrescenta
a reta de regressão ajustada
```



```r
#Tempo de espera entre erupções e a duração da erupção
fit <- lm(waiting~eruptions, data=faithful)
plot(faithful)
lines(faithful$eruptions, fitted(fit), col="blue")
```



Objetivos:

Estudar a relação linear entre duas variáveis quantitativas. Exemplos:

1. Altura dos pais e altura dos filhos;

Objetivos:

Estudar a relação linear entre duas variáveis quantitativas. Exemplos:

(i) Altura dos pais e altura dos filhos;

(ii) Renda semanal e despensas de consumo;

Objetivos:

Estudar a relação linear entre duas variáveis quantitativas. Exemplos:

(i) Altura dos pais e altura dos filhos;
(ii) Renda semanal e despensas de consumo;
(iii) Variação dos salários e taxa de desemprego;

Objetivos:

Estudar a relação linear entre duas variáveis quantitativas. Exemplos:

(i) Altura dos pais e altura dos filhos;
(ii) Renda semanal e despensas de consumo;
(iii) Variação dos salários e taxa de desemprego;
(iv) Demanda dos produtos de uma firma e publicidade;

Objetivos:

Estudar a relação linear entre duas variáveis quantitativas. Exemplos:

(i) Altura dos pais e altura dos filhos;
(ii) Renda semanal e despensas de consumo;
(iii) Variação dos salários e taxa de desemprego;
(iv) Demanda dos produtos de uma firma e publicidade;

Sob dois pontos de vista:

(i) Explicitando a forma dessa relação: regressão.

Objetivos:

Estudar a relação linear entre duas variáveis quantitativas. Exemplos:

1. (i) Altura dos pais e altura dos filhos;
2. (ii) Renda semanal e despensas de consumo;
3. (iii) Variação dos salários e taxa de desemprego;
4. (iv) Demanda dos produtos de uma firma e publicidade;

Sob dois pontos de vista:

1. (i) Explicitando a forma dessa relação: regressão.
2. (ii) Quantificando a força dessa relação: correlação.

Importante:

Uma relação estatística por sí propria não implica uma causa, para atribuir causa, devemos invocar alguma teoría!

Importante:

Uma relação estatística por sí propria não implica uma causa, para atribuir causa, devemos invocar alguma teoría!

Uma regressão espúria é uma relação estatística existente entre duas variáveis, mas onde não existe nenhuma relação causa-efeito entre elas. Essa relação estatística pode ocorrer por pura coincidência ou por causa de uma terceira variável. Ou seja, neste último caso, pode ocorrer que as variáveis X e Y sejam correlacionadas porque ambas são causadas por uma terceira variável Z.

Só porque (A) acontece juntamente com (B) não significa que (A) causa (B). Determinar se existe de fato uma relação de causalidade requer investigação adicional pois podem acontecer cinco situações:

(i) (A) causa realmente (B);

Só porque (A) acontece juntamente com (B) não significa que (A) causa (B). Determinar se existe de fato uma relação de causalidade requer investigação adicional pois podem acontecer cinco situações:

(i) (A) causa realmente (B);
(ii) (B) pode ser a causa de (A);

Só porque (A) acontece juntamente com (B) não significa que (A) causa (B). Determinar se existe de fato uma relação de causalidade requer investigação adicional pois podem acontecer cinco situações:

1. (A) causa realmente (B);
2. (B) pode ser a causa de (A);
3. Um terceiro factor (C) pode ser causa tanto de (A) como de (B);

Só porque (A) acontece juntamente com (B) não significa que (A) causa (B). Determinar se existe de fato uma relação de causalidade requer investigação adicional pois podem acontecer cinco situações:

(i) (A) causa realmente (B);
(ii) (B) pode ser a causa de (A);
(iii) Um terceiro factor (C) pode ser causa tanto de (A) como de (B);Pode ser uma combinação das três situações anteriores. Por exemplo, (A) causa (B) e ao mesmo tempo (B) causa também (A);

Só porque (A) acontece juntamente com (B) não significa que (A) causa (B). Determinar se existe de fato uma relação de causalidade requer investigação adicional pois podem acontecer cinco situações:

(i) (A) causa realmente (B);
(ii) (B) pode ser a causa de (A);
(iii) Um terceiro factor (C) pode ser causa tanto de (A) como de (B);Pode ser uma combinação das três situações anteriores. Por exemplo, (A) causa (B) e ao mesmo tempo (B) causa também (A);
(iv) A correlação pode ser apenas uma coincidência, ou seja, os dois eventos não têm qualquer relação para além do fato de ocorrerem ao mesmo tempo.

Exemplos:

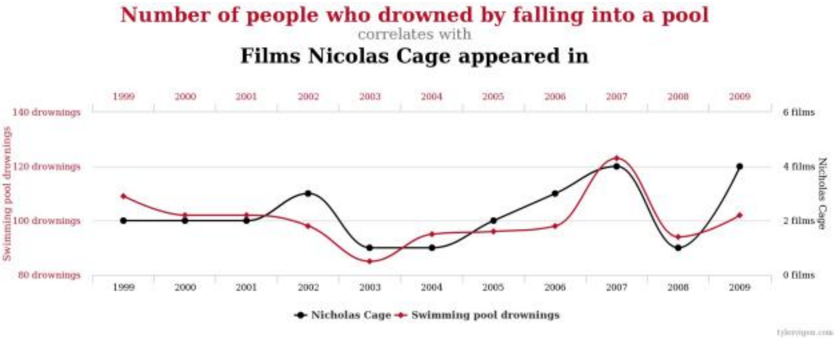
"Quanto maiores são os pés de uma criança, maior a capacidade para resolver problemas de matemática. Portanto, ter pés grandes faz ter melhores notas em matemática".

Exemplos:

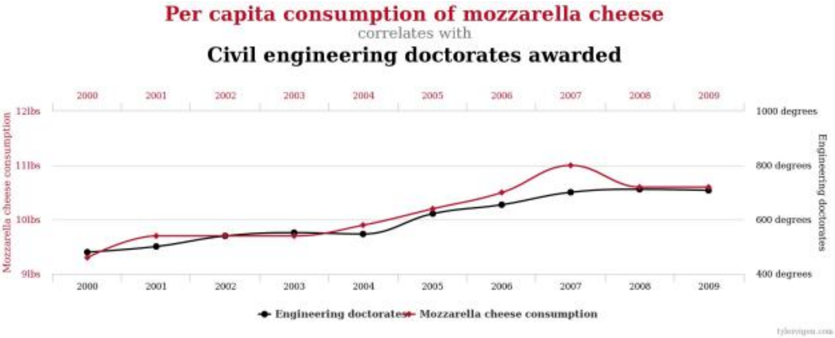
“Vários estudos apontavam inicialmente que as mulheres em menopausa que recebiam terapia de substituição hormonal (TSH) tinham também um menor risco de doença coronária, o que levou à ideia de que a TSH conferia protecção contra a doença coronária. No entanto, estudos controlados e randomizados (mais rigorosos), feitos posteriormente, mostraram que a TSH causava na verdade um pequeno mas significativo aumento do risco de doença coronária. Uma reanálise dos estudos revelou que as mulheres que recebiam a TSH tinham também uma maior probabilidade de pertencer a uma classe socioeconómica superior, com melhor dieta e hábitos de exercício. A utilização da TSH e a baixa incidência de doença coronária não eram causa e efeito, mas o fruto de uma causa comum".

“Quanto menos as pessoas se divorciam em Maine (EUA), menor fica o consumo de margarina naquele Estado".

Deveríamos banir Nicolas Cage do cinema para evitar o afogamento de pessoas? O primeiro gráfico nos dá o número de pessoas afogadas (linha vermelha) e as aparições do Nicolas Cage em filmes (linha preta).



Consumo de muçarela (linha vermelha) e doutorados obtidos em engenharia civil (linha preta)



Coeficiente de Correlação Amostral de Pearson

O coeficiente de correlação (r ou $\hat{\rho}$) mede o grau de associação linear entre duas variáveis aleatórias X e Y. Considere,

X_i	X_1	X_2	$\cdots$	X_n
Y_i	Y_1	Y_2	$\cdots$	Y_n

Assim, o coeficiente de correlação entre X e Y é dado por:

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_X^2 \cdot S_Y^2}} = \frac{\frac{SPD_{XY}}{n-1}}{\sqrt{\frac{SQD_X}{n-1} \cdot \frac{SQD_Y}{n-1}}} = \frac{SPD_{XY}}{\sqrt{SQD_X \cdot SQD_Y}}$$

Temos,

$$SPD_{XY} = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum\limits_{i=1}^{n} X_i\right)\left(\sum\limits_{i=1}^{n} Y_i\right)}{n}$$

Temos,

$$SPD_{XY} = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}$$

$$SQD_X = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n} \quad \text{e} \quad SQD_Y = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

Temos,

$$SPD_{XY} = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}$$

$$SQD_X = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n} \quad \text{e} \quad SQD_Y = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

$$-1 \leq r_{XY} \leq 1$$

Exemplo

Amostra A	4	8	3	9	7	5
Amostra B	1	5	2	14	3	11

Temos,

$$SPD_{AB} = \sum_{i=1}^{n} A_i B_i - \frac{\left(\sum_{i=1}^{n} A_i\right)\left(\sum_{i=1}^{n} B_i\right)}{n}$$

Exemplo

Amostra A	4	8	3	9	7	5
Amostra B	1	5	2	14	3	11

Temos,

$$SPD_{AB} = \sum_{i=1}^{n} A_i B_i - \frac{\left(\sum_{i=1}^{n} A_i\right)\left(\sum_{i=1}^{n} B_i\right)}{n}$$

$$SQD_A = \sum_{i=1}^{n} A_i^2 - \frac{\left(\sum_{i=1}^{n} A_i\right)^2}{n} \quad \text{e} \quad SQD_B = \sum_{i=1}^{n} B_i^2 - \frac{\left(\sum_{i=1}^{n} B_i\right)^2}{n}$$

Exemplo

Assim,

$$SPD_{AB} = 252 - \frac{(36)(36)}{6} = 36$$

Exemplo

Assim,

$$SPD_{AB} = 252 - \frac{(36)(36)}{6} = 36$$

$$SQD_A = 244 - \frac{(36)^2}{6} = 28 \quad \text{e} \quad SQD_B = 356 - \frac{(36)^2}{6} = 140$$

Exemplo

Assim,

$$SPD_{AB} = 252 - \frac{(36)(36)}{6} = 36$$

$$SQD_A = 244 - \frac{(36)^2}{6} = 28 \quad \mathrm{e} \quad SQD_B = 356 - \frac{(36)^2}{6} = 140$$

$$r_{AB} = \frac{SP_{AB}}{\sqrt{SQD_A SQD_B}} = \frac{36}{\sqrt{28 \cdot 140}} = 0,5750$$

Modelo Estatístico

De forma geral, um modelo estatístico pode ser escrito da seguinte forma:

$$Y = \text{componente determinística} + \text{componente aleatória}$$

existem diversas maneiras de específicar essas componentes. Começaremos com uma regressão linear simples.

Uma regresão linear simples tem como objetivo aproximar uma variável resposta Y através de uma função linear de uma variável de interesse, ou seja,

$$Y = f(X, \beta) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon$$

No qual assume-se que:

(i) $E(\varepsilon) = 0$

Uma regresão linear simples tem como objetivo aproximar uma variável resposta Y através de uma função linear de uma variável de interesse, ou seja,

$$Y = f(X, \beta) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon$$

No qual assume-se que:

1. $E(\varepsilon) = 0$
2. $V(\varepsilon) = \sigma^2$ (Homocedásticidade)

Uma regresão linear simples tem como objetivo aproximar uma variável resposta Y através de uma função linear de uma variável de interesse, ou seja,

$$Y = f(X, \beta) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon$$

No qual assume-se que:

(i) $E(\varepsilon) = 0$
(ii) $V(\varepsilon) = \sigma^2$ (Homocedásticidade)
(iii) $Cov(\varepsilon_i, \varepsilon_j) = 0$

em outras palavras, os erros tem média zero, variância constante e são não correlacionados.

A variável preditora X pode vir de diversas fontes:

- inputs quantitativos (valores reais, medidas)
- tranformação de variável quantitativas ($\log()$,$\sqrt{()}$,etc)
- inputs qualitativos (“dummy"e.x. genêro, classes)

Dessa forma, um modelos de regressão consiste em 4 passos:

1. Escolher o componente determinístico do modelo;

A variável preditora X pode vir de diversas fontes:

- inputs quantitativos (valores reais, medidas)
- tranformação de variável quantitativas ($\log()$,$\sqrt{()}$,etc)
- inputs qualitativos (“dummy"e.x. genêro, classes)

Dessa forma, um modelos de regressão consiste em 4 passos:

1. Escolher o componente determinístico do modelo;
2. Especificar a distribuição do erro;

A variável preditora X pode vir de diversas fontes:

- inputs quantitativos (valores reais, medidas)
- tranformação de variável quantitativas ($\log()$,$\sqrt{()}$,etc)
- inputs qualitativos (“dummy"e.x. genêro, classes)

Dessa forma, um modelos de regressão consiste em 4 passos:

1. Escolher o componente determinístico do modelo;
2. Especificar a distribuição do erro;
3. Utilizar os dados para estimar os parâmetros do modelo;

A variável preditora X pode vir de diversas fontes:

- inputs quantitativos (valores reais, medidas)
- tranformação de variável quantitativas ($\log()$, $\sqrt{()}$,etc)
- inputs qualitativos ("dummy"e.x. genêro, classes)

Dessa forma, um modelos de regressão consiste em 4 passos:

1. Escolher o componente determinístico do modelo;
2. Especificar a distribuição do erro;
3. Utilizar os dados para estimar os parâmetros do modelo;
4. Avaliar o modelo estatístico;

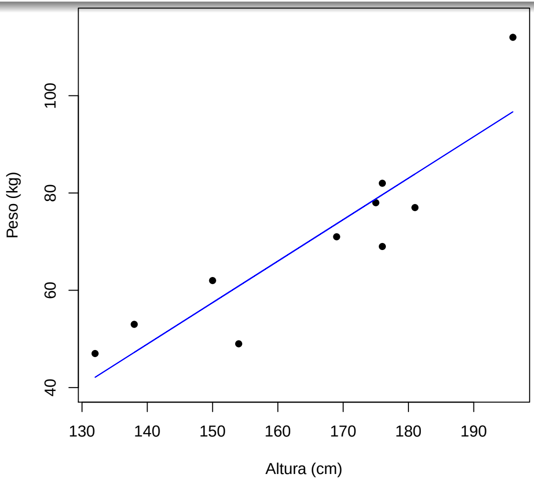
Os dados para a análise de regressão e correlação simples são da forma:

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

Com base nos dados constrói-se o diagrama de dispersão, que deve exibir uma tendência linear para que se possa usar a regressão linear. Este diagrama permite decidir empiricamente:

- Se um relacionamento linear entre as variáveis X e Y deve ser assumido;
- Se o grau de relacionamento linear entre as variáveis é forte ou fraco, conforme o modo como se situam os pontos em redor de uma reta imaginária que passa através do enxame de pontos.

```
#Encontre o modelo de Regressão Linear que melhor se
ajusta aos dados
x = c(176, 154, 138, 196, 132, 176, 181, 169, 150, 175)
y = c(82, 49, 53, 112, 47, 69, 77, 71, 62, 78)
```



```
#Encontre o modelo de Regressão Linear que melhor se
ajusta aos dados
x <-  c(176, 154, 138, 196, 132, 176, 181, 169, 150, 175)
y <-  c( 82,  49,  53, 112,  47,  69,  77,  71,  62,  78)
Reg <- lm(y~x)
Reg

Call:
lm(formula = y ~ x)


Coefficients:
(Intercept)            x
   -70.4627       0.8528
```

```
#library(texreg)
summary(Reg)

lm(formula = y ~ x)
Residuals:
     Min       1Q   Median       3Q      Max
-11.8746  -5.8428   0.7893   4.8001  15.3061
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -70.4627    24.0148  -2.934 0.018878 *
x             0.8528     0.1448   5.889 0.000366 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
Residual standard error: 8.854 on 8 degrees of freedom
Multiple R-squared:  0.8126, Adjusted R-squared:  0.7891
F-statistic: 34.68 on 1 and 8 DF,  p-value: 0.0003662
```

	Model 1
(Intercept)	-70.46^{*}
	(24.01)
x	0.85^{***}
	(0.14)
R^2	0.81
Adj. R^2	0.79
Num. obs.	10
RMSE	8.85

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Tabela: Statistical models

R^2

Representa a porcentagem de variação na resposta que é explicada pelo modelo. Ele é calculado como 1 menos a razão da soma dos quadrados dos erros (que é a variação que não é explicada pelo modelo) pela soma total dos quadrados (que é a variação total no modelo).

R^2

Representa a porcentagem de variação na resposta que é explicada pelo modelo. Ele é calculado como 1 menos a razão da soma dos quadrados dos erros (que é a variação que não é explicada pelo modelo) pela soma total dos quadrados (que é a variação total no modelo).

R^2

Use R^2 para determinar se o modelo se ajusta bem aos dados. Quanto mais alto o valor de R^2 melhor o modelo ajusta seus dados. O valor de R^2 está sempre entre 0 e 100%.

Considere as seguintes questões ao interpretar o valor de R^2 :

O R^2 sempre aumenta quando você adiciona mais preditores a um modelo. Por exemplo, o melhor modelo de cinco preditores terá sempre um R^2 que é pelo menos tão elevado quanto o melhor modelo de quatro preditores. Portanto, R^2 é mais útil quando for comparado a modelos do mesmo tamanho.

Considere as seguintes questões ao interpretar o valor de R^2 :

O R^2 sempre aumenta quando você adiciona mais preditores a um modelo. Por exemplo, o melhor modelo de cinco preditores terá sempre um R^2 que é pelo menos tão elevado quanto o melhor modelo de quatro preditores. Portanto, R^2 é mais útil quando for comparado a modelos do mesmo tamanho.

Amostras pequenas não fornecem uma estimativa precisa da força da relação entre a resposta e os preditores. Se você precisar que R^2 seja mais exato, deve usar uma amostra maior (geralmente, 40 ou mais).

R^2 é apenas uma medida de o quão bem o modelo ajusta os dados. Mesmo quando um modelo tem um R^2 elevado, você deve verificar os gráficos de resíduos para conferir se o modelo satisfaz os pressupostos do modelo.

Estimação por Mínimos Quadrados

$$e_i = Y_i - \beta_0 - \beta_1 X_i, i = 1, \cdots, n$$

$$\Rightarrow$$

$$e_i^2 = \left(Y_i - \beta_0 - \beta_1 X_i\right)^2, i = 1, \cdots, n$$

$$\Rightarrow$$

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)^2$$

Estimação por Mínimos Quadrados

Usando a primeira e a segunda derivada podemos obter os valores que minimizam a soma de quadrados, são eles:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} = \frac{SPD_{xy}}{SQD_x} \text{ e } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Estimação por Mínimos Quadrados

Usando a primeira e a segunda derivada podemos obter os valores que minimizam a soma de quadrados, são eles:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} = \frac{SPD_{xy}}{SQD_x} \text{ e } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Uma vez obtida estas estimativas, podemos escrever a equação estimada:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Estimação por Mínimos Quadrados

O R^2 é obtido por

$$R^2 = \frac{SQReg}{SQTotal} = \frac{\hat{\beta}_1 SPD_{xy}}{SQD_y}$$

Estimação por Mínimos Quadrados

O R^2 é obtido por

$$R^2 = \frac{SQReg}{SQTotal} = \frac{\hat{\beta}_1 SPD_{xy}}{SQD_y}$$

O R^2 indica a proporção (ou %) da variação de Y que é "explicada" pela regressão, ou quanto da variação na variável dependente Y está sendo "explicada" pela variável independente X.

Teste de Hipótese para Regressão Linear Simples

A equação de regressão estimada apenas estabelece uma relação funcional entre a variável dependente e a variável independente, para representar o fenômeno em estudo. Portanto, a simples obtenção da equação estimada não responde ao pesquisador se a variação da variável independente influencia significativamente na variação da variável dependente. Assim, faz-se necessário, realizar um teste estatístico para as estimativas dos coeficientes da regressão estimada. Um teste que pode ser realizado é o teste F da Análise de Variância e/ou teste t.

Teste de Hipótese para Regressão Linear Simples

Vamos proceder o teste t para avaliar a significância da regressão, ou seja, vamos testar se $\beta_1 = 0$ contra $\beta_1 \neq 0$.

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

$$t_{cal} = \frac{\hat{\beta}_1}{\sqrt{\hat{V}(\hat{\beta}_1)}}, \text{ em que } \hat{V}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{SQD_x} \text{ e } \hat{\sigma}^2 = \frac{SQResiduo}{n-2}$$

Regra de Decisão: Se $|t_{cal}| \geq |t_{\alpha/2}(n-2)| \Rightarrow$ rejeita-se H_0.

Teste de Hipótese para Regressão Linear Simples

Não rejeitar H_0 é equivalente a concluir que não existe relação linear entre X e Y. Por outro lado, rejeitar H_0 significa que X é importante para explicar a variabilidade de Y.

Análise de Variância (ANOVA) para a Regressão Linear Simples

O método da ANOVA consiste em fazer uma partição da variabilidade total da variável resposta Y em outros componentes de acordo com o modelo e o teste a ser feito. Assim, a seguinte identidade pode ser verificada:

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y})^2,$$

ou, em outras palavras

$$SQTotal = SQRegressao + SQResiduo,$$

Análise de Variância (ANOVA) para a Regressão Linear Simples

$$SQTotal = SQRegressao + SQResiduo,$$

em que,

- $SQTotal =$Variação Total em $Y = SQD_Y$

Análise de Variância (ANOVA) para a Regressão Linear Simples

$$SQTotal = SQRegressao + SQResiduo,$$

em que,

- $SQTotal =$Variação Total em $Y = SQD_Y$
- $SQRegressao$ é a variação em Y explicada pela regressão ajustada, ou seja:

Análise de Variância (ANOVA) para a Regressão Linear Simples

$$SQTotal = SQRegressao + SQResiduo,$$

em que,

- $SQTotal =$Variação Total em $Y = SQD_Y$
- $SQRegressao$ é a variação em Y explicada pela regressão ajustada, ou seja:
- $SQRegressao = \hat{\beta}_1 SPD_{XY}$ de tal forma que,

Análise de Variância (ANOVA) para a Regressão Linear Simples

$$SQTotal = SQRegressao + SQResiduo,$$

em que,

- $SQTotal =$Variação Total em $Y = SQD_Y$
- $SQRegressao$ é a variação em Y explicada pela regressão ajustada, ou seja:
- $SQRegressao = \hat{\beta}_1 SPD_{XY}$ de tal forma que,
- $SQResiduo$ é igual a variação não explicada pela regressão, isto é:

Análise de Variância (ANOVA) para a Regressão Linear Simples

$$SQTotal = SQRegressao + SQResiduo,$$

em que,

- $SQTotal =$Variação Total em $Y = SQD_Y$
- $SQRegressao$ é a variação em Y explicada pela regressão ajustada, ou seja:
- $SQRegressao = \hat{\beta}_1 SPD_{XY}$ de tal forma que,
- $SQResiduo$ é igual a variação não explicada pela regressão, isto é:
- $SQResiduo = SQD_Y - \hat{\beta}_1 SPD_{XY}$.

Análise de Variância (ANOVA) para a Regressão Linear Simples

Baseado nos cálculos acima podemos montar o seguinte quadro de ANOVA:

FV	GL	SQ	QM	F
Regressão	1	SQReg	QMReg=SQReg	$\dfrac{QMReg}{QMRes}$
Resíduo	$n-2$	SQRes	QMRes= $\dfrac{SQRes}{n-2}$	–
Total	$n-1$	SQTotal	–	–

Análise de Variância (ANOVA) para a Regressão Linear Simples

A estatística F acima do quadro da ANOVA serve para testar a significância da regressão, ou seja, testar:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Regra de Decisão: Se $F_{cal} \geq F_{\alpha}(1, n-2) \Rightarrow$ rejeita-se H_0.

- Bastos, Fernando de Souza (2025). *Apostila Interativa*. Disponível online: https://ufvest.shinyapps.io/ApostilaInterativa/.
- Ferreira, Eric Batista e Marcelo Silva de Oliveira (2020). *Introdução à Estatística com R*. Editora Universidade Federal de Alfenas. URL: https://www.unifal-mg.edu.br/bibliotecas/wp-content/uploads/sites/125/2021/12/32-EBR_Unifal.pdf.
- Montgomery, D. C. e G. C Runger (2016). *Estatística Aplicada E Probabilidade Para Engenheiros*. 6ª ed. São Paulo: Grupo Gen-LTC.
- Morettin, P.A. e W.O Bussab (2023). *Estatística básica*. 10ª ed. São Paulo: Editora Saraiva.
- Peternelli, Luiz Alexandre (s.d.). *Apostila (EST 106)*. Formato slide Disponível no PVANet - Moodle.