Correlação e Regressão

Anderson Rodrigo da Silva

Instituto Federal Goiano

Conteúdo

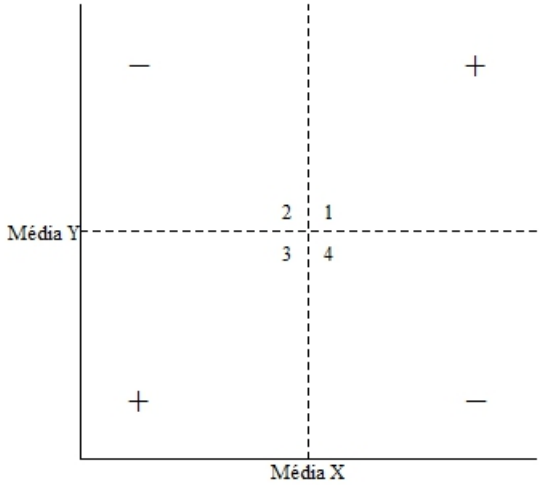
Covariância

- Associação linear entre duas variáveis resposta
- Exemplos: altura de planta e altura da espiga, teor de fósforo no solo e na folha, nº de lagartas por planta e área foliar, etc.

A covariância é uma medida de variabilidade conjunta, dada por:

$$Cov(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Diagrama de dispersão

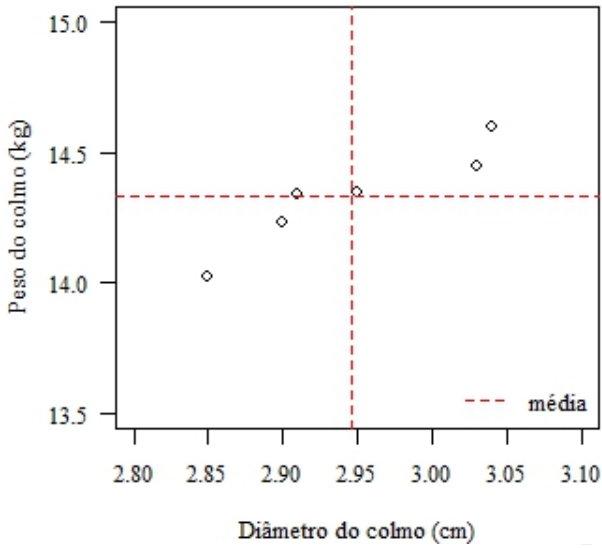


Exemplo

Tabela: Dados referentes ao diâmetro (cm) e ao peso do colmo (kg) de 6 plantas de cana-de-açúcar, medidas em uma pesquisa.

Índice (i)	Diâmetro (X)	Peso (Y)	XY
1	3,04	14,60	44,38
2	2,85	14,02	39,96
3	2,90	14,23	41,27
4	3,03	14,45	43,78
5	2,90	14,34	41,73
6	2,95	14,35	42,33
Soma	17,68	85,99	253,45

Exemplo


15.0
14.5
14.0
13.5
Peso do colmo (kg)
média
2.80
2.85
2.90
2.95
3.00
3.05
3.10
Diâmetro do colmo (cm)

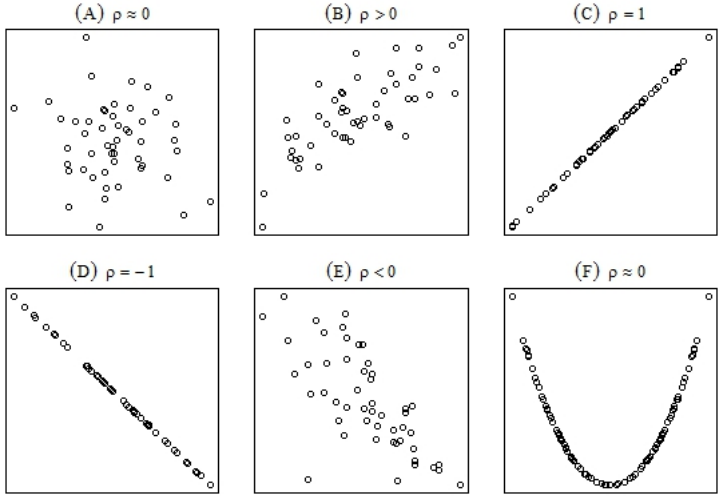
Correlação de Pearson (ρ)

- Também mede o grau de associação linear entre duas variáveis resposta
- É adimensional: não é afetado pela escala das variáveis

O coeficiente de correlação de Pearson é dado por:

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Ilustrando ρ



Teste da correlação

O teste da hipótese $H_0 : \rho_{XY} = 0$ versus $H_1 : \rho_{XY} \neq 0$ pode ser feito utilizando a estatística t-Student:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \sim t_{n-2}$$

desde que seja razoável adimitir que X e Y tenham distribuição normal.

Correlação de Spearman

- A correlação de Pearson é muito sensível à desvios da normalidade e assim como a média aritmética e a variância, também é sensível a valores discrepantes
- A correlação não paramétrica de Spearman é uma alternativa nesses casos
- É baseada nos ranks ou postos das observações
- É indicada quando a relação entre X e Y não é necessariamente linear
- O cálculo é feito usando a mesma equação do coeficiente de Pearson, substituíndo os valores observados por seus ranks

Regressão linear simples

- A correlação mede apenas o grau de associação entre duas variáveis, mas não nos informa nada sobre a relação de causa e efeito de uma variável sobre outra
- Na correlação, ambas as variáveis são supostas aleatórias (variáveis resposta)
- Exemplo: qual será o efeito na produção vegetal quando se aumentar em uma unidade a dose de um fertilizante?
- Exemplo 2: conhecendo-se a relação entre severidade de uma doença e tempo, a severidade pode ser predita num tempo específico
- A idéia consiste em ajustar um modelo para uma variável resposta (Y) em função de uma variável explicativa (X)
- Adimitindo que a relação entre ambas é linear, podemos ajustar o modelo:

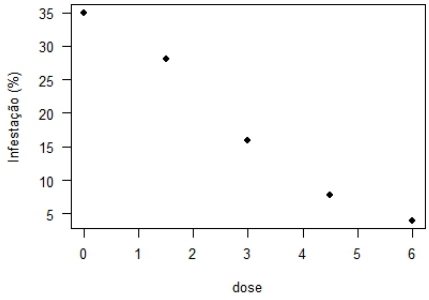
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

sendo β_0 e β_1 os parâmetros a serem estimados; ϵ é o erro aleatório associado a observação y

Exemplo

Tabela: Amostra de n = 5 parcelas experimentais nas quais foi avaliado o percentual de infestação por plantas daninhas monocotiledôneas após aplicação pós-emergencial de um herbicida seletivo.

Dose (L/ha)	0	1.5	3	4.5	6
Percentual	35	28	16	7.7	4



Objetivos

- Ajustar um modelo para predizer o grau de infestação (Y) em função da dose aplicada (X)
- Para tal precisamos: 1) estimar os parâmetros β_0 e β_1, 2) testar a significância dos parâmetros, 3) verificar o ajuste do modelo

Estimação de parâmetros

- Método dos mínimos quadrados
- Método da máxima verossimilhança

Mínimos Quadrados

O método consiste em obter estimativas para o vetor de parâmetros $\boldsymbol{\beta} = [\beta_0 \quad \beta_1]^T$ que tornem mínima a função (notação matricial)

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

Igualando as derivadas parciais de $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$ em relação à $\boldsymbol{\beta}$, obtemos:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Análise de variância da regressão

Admitindo que $\epsilon \sim Normal(0, \sigma^2)$, o teste da hipótese $H_0 : \beta_1 = 0$ pode ser feito através do teste F da ANOVA

Tabela: ANOVA da regressão

FV	GL	SQ	QM	F
Regressão	k	$\hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y} - C$		
Resíduo	$n - k - 1$	$\boldsymbol{y}^T \boldsymbol{y} - \hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y}$		
Total	$n - 1$	$\boldsymbol{y}^T \boldsymbol{y} - C$		

sendo k o nº de regressores ("x") no modelo. No caso da regressão linear simples, $k = 1$.

Coeficiente de determinação simples (r^2)

O coeficiente de determinação é utilizado para medir o grau de ajuste do modelo de regressão linear simples.

$$r^2 = \frac{SQreg}{SQtotal} \quad \in [0, 1]$$

quanto mais próximo da unidade, melhor o ajuste.

Teste t para cada parâmetro

O teste F da ANOVA apenas indica se o modelo de regressão é significativo, como um todo, ou não. O teste individual, para cada parâmetro do modelo, é construído da seguinte forma

$$t = \frac{\hat{\beta}_j}{\sqrt{Var(\hat{\beta}_j)}} \sim t_{GLres}$$

sendo

$$\hat{Var}(\hat{\beta}_j) = (\boldsymbol{X}^T \boldsymbol{X})^{-1}_{jj} QMres$$

Regressão linear múltipla

A idéia consiste em ajustar um modelo para uma variável resposta (Y) em função de dois ou mais regressores ($X_1, X_2, ...$)

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_1 x_{1i} + ... + \beta_k x_{ki} + \epsilon_i$$

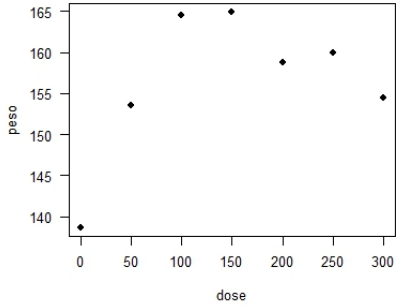
sendo β_0, β_1, ..., β_k os parâmetros a serem estimados; ϵ_i é o erro aleatório associado a observação y_i

Exemplo: modelar a produção vegetal em função das doses de N, P e K

Exemplo 1

Tabela: Peso de mil grãos de feijão sob efeito de doses de gesso (kg/ha)

Dose	0	50	100	150	200	250	300
Peso	138.6	153.6	164.5	164.9	158.7	159.9	154.4



Exercício 1

Ajuste um modelo de regressão aos dados de severidade de determinada doença em função da temperatura do ar. Depois, aplique o teste t para cada parâmetro do modelo.

Temperatura (°C)	2	1	5	5	20	20	23	10	30	25
Severidade (%)	1.9	3.1	3.3	4.8	5.3	6.1	6.4	7.6	9.8	12.4

Fonte: American Phytopathological Society (`http://www.apsnet.org/`)

Exercício 2

Ajuste um modelo de regressão linear múltipla para y em função de x_1 e x_2. Depois, aplique o teste t para cada parâmetro do modelo.

y	x_1	x_2
1.5	0	0
6.5	1	2
10	1	4
11	2	2
11.5	2	4
16.5	3	6