Regressão Linear

Anderson Rodrigo da Silva

Instituto Federal Goiano

Conteúdo

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)
- Admitindo que a relação entre ambas é linear, podemos ajustar o modelo:

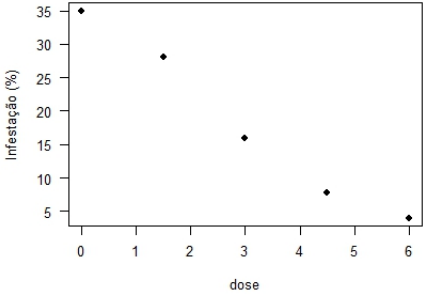
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

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

Exemplo 1 - sem repetição

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.

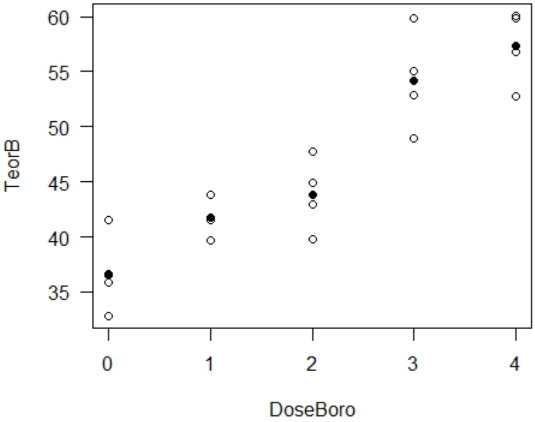
Exemplo 2 - com repetição

		Bloco				
Local	Dose	I	II	III	IV	Médias
Ipameri	0	35.9	41.5	32.8	36.5	36.7
	1	39.7	41.5	43.8	41.8	41.7
	2	43	39.8	44.9	47.8	43.9
	3	48.9	52.9	55.1	59.8	54.2
	4	52.8	56.8	59.8	60.1	57.4

```
Analysis of Variance Table

Response: TeorB
          Df  Sum Sq Mean Sq F value    Pr(>F)
Bloco      3   67.91  22.636  2.4734    0.1116
DoseBoro   4 1213.18 303.294 33.1403 2.127e-06 ***
Residuals 12  109.82   9.152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exemplo 2 - com repetição



Regressão linear múltipla

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

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

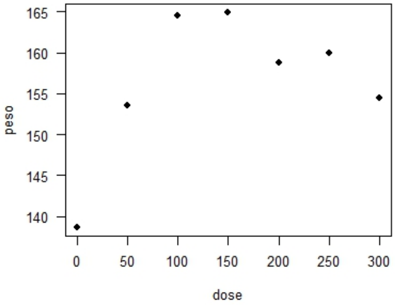
sendo β_0, β_1, ..., β_k os parâmetros a serem estimados; ϵ_i é o erro aleatório associado a observação y_i

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, realize a análise de variância da regressão.

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

Verifique se a dose de vinhaça (L/ha) afeta linearmente ($p < 0.05$) a produtividade (t/ha) de cada uma das três variedades de cana-de-açúcar. O modelo linear simples é adequado?

		Bloco		
Variedade	Dose de vinhaça	I	II	III
A	0	69	66	68
	500	72	70	71
	1000	70	73	71
	1500	66	64	67
B	0	65	67	64
	500	69	73	73
	1000	73	74	75
	1500	70	68	68
C	0	71	73	70
	500	76	79	77
	1000	77	79	76
	1500	74	75	76

Exercício 2

Dica: ao invés de montar a tabela convencional de ANOVA de um fatorial 3×4, monte a tabela da seguinte forma:

FV	GL	SQ	QM	F
Bloco	2			
Variedade	2			
Dose/var.A	3			
Dose/var.B	3			
Dose/var.C	3			
Resíduo	22			
Total	35			

Em que a fonte de variação **Dose/var.A** representa o efeito da dose de vinhaça sobre a produtividade da variedade A.