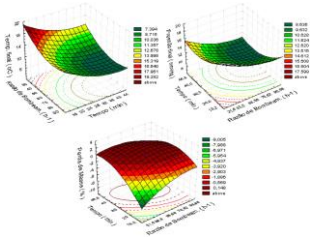
UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de ciências agrárias

1. Regressão múltipla: Introdução e obtenção de estimadores



Alcinei Mistico Azevedo
Professor Adjunto de Estatística e experimentação agrícola.

2.1 Introdução

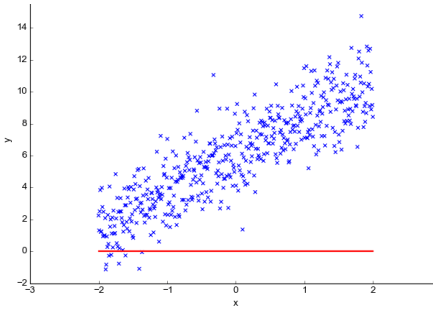
Métodos de análise: para a escolha do método de análise estatística à ser utilizado, deve-se considerar se as variáveis explicativas e variáveis respostas são qualitativas ou quantitativas.

Variáveis		Modelo de análise
Explicativas (X)	Resposta (Y)	
Qualitativa	Quantitativa	ANAVA + Teste de médias
Quantitativa	Qualitativa (binomial)	Regressão logit ou probit
Qualitativa	Qualitativa	Tabela de contigência
Quantitativa	Quantitativa	ANAVA + Regressão

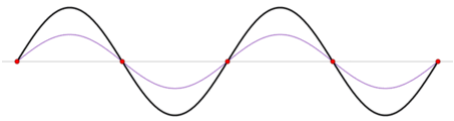
2.1 Introdução

Análise de regressão: Esta metodologia é utilizada quando tanto a(s) variável(eis) explicativas (independentes) quanto a(s) variável(eis) resposta (dependente) são quantitativas.

Objetivo Geral: Na análise de regressão, pretende-se modelar a variação da variável resposta em função da(s) variável(eis) explicativa (s) (independente) por meio da estimação de parâmetros.



2.1 Introdução



Objetivos específicos da análise de regressão:

- Estudar a existência de possíveis associações (relações) funcionais da(s) variável(eis) explicativa(s) sobre a variável resposta
- Predizer uma variável de interesse (variável dependente) por meio de variável(eis) explicativa(s)
- Descobrir quais variáveis explicativas são mais importantes em um determinado fenômeno.
- Descobrir valores da variável explicativa (ex.: dose, temperatura, umidade) que proporcionam melhores valores da variável resposta.

2.1 Introdução

Tipos de regressão linear

Número de		Modelo de análise
Explicativas (X)	Resposta (Y)	
1	1	Regressão simples
2 ou mais	1	Regressão múltipla
1	2 ou mais	Regressão multivariada
2 ou mais	2 ou mais	Regressão múltipla multivariada

2.1Introdução

- Em estudos de regressão múltipla são estudadas mais de uma variável explicativa simultaneamente.

- Lidar com mais de uma variável explicativa é mais difícil, pois:

  - ✓ É mais difícil escolher o melhor modelo, uma vez que diversas variáveis candidatas podem existir
  - ✓ É mais difícil visualizar a aparência do modelo ajustado, mais difícil a representação gráfica em mais de 3 dimensões
  - ✓ Às vezes, é difícil interpretar o modelo ajustado
  - ✓ Cálculos difíceis de serem executados sem auxílio de computador

2.2 Modelos de regressão múltipla

Exemplo: Supondo dados de comprimento e largura de folhas de acerola, é possível obter relações matemáticas (estimar coeficientes) que levem à obtenção da área foliar.

ID	Comp (x)	Larg (y)	AFO (z)
1	6.00	3.20	13.39
2	5.50	3.20	12.52
3	5.00	2.50	8.78
4	5.00	2.70	9.60
.	.	.	.
.	.	.	.
.	.	.	.
300	4.30	2.20	8.23

Modelos	Função (Notação algébrica)
1	$z_i = a + bx_i + e_i$
2	$z_i = a + bx_i + cx_i^2 + e_i$
3	$z_i = a + by_i + e_i$
4	$z_i = a + by_i + cy_i^2 + e_i$
5	$z_i = a + bx_i + cy_i + e_i$
6	$z_i = a + bx_i + cx_i^2 + dy_i + e_i$
7	$z_i = a + bx_i + cy_i + dy_i^2 + e_i$
8	$z_i = a + bx_i + cx_i^2 + dy_i + fy_i^2 + e_i$
9	$z_i = a + bx_i + cy_i + dx_iy_i + e_i$
10	$z_i = a + bx_i + cx_i^2 + dy_i + fx_iy_i + e_i$
11	$z_i = a + bx_i + cy_i + dy_i^2 + fx_iy_i + e_i$
12	$z_i = a + bx_i + cx_i^2 + dy_i + fy_i^2 + gx_iy_i + e_i$
13	$z_i = a + bx_i + cx_i^2 + dy_i + fy_i^2 + gx_iy_i + hx_i^2y_i + e_i$
14	$z_i = a + bx_i + cx_i^2 + dy_i + fy_i^2 + gx_iy_i + hx_iy_i^2 + e_i$
15	$z_i = a + bx_i + cx_i^2 + dy_i + fy_i^2 + gx_iy_i + hx_i^2y_i + jx_iy_i^2 + e_i$
16	$z_i = a + bx_i + cx_i^2 + dy_i + fy_i^2 + gx_iy_i + hx_i^2y_i + jx_iy_i^2 + k_i^2y_i^2 + e_i$

2.2 Modelos de regressão múltipla

Exemplo: Supondo dados de comprimento e largura de folhas de acerola, é possível obter relações matemáticas (estimar coeficientes) que levem à obtenção da área foliar.

ID	Comp (x)	Larg (y)	AFO (z)
1	6.00	3.20	13.39
2	5.50	3.20	12.52
3	5.00	2.50	8.78
4	5.00	2.70	9.60
.	.	.	.
.	.	.	.
.	.	.	.
300	4.30	2.20	8.23

Notação algébrica do modelo 5: $z_i = a + bx_i + cy_i + e_i$

Notação matricial do modelo 5: $\underset{\sim}{Y} = X\underset{\sim}{B} + \underset{\sim}{e}$

$$\underset{\sim}{Y} = \quad X \quad \underset{\sim}{B} + \underset{\sim}{e}$$

$$\begin{bmatrix} 13,39 \\ 12,52 \\ 8,78 \\ 9,60 \\ \vdots \\ 8,23 \end{bmatrix} = \begin{bmatrix} 1 & 6,00 & 3,20 \\ 1 & 5,50 & 3,20 \\ 1 & 5,00 & 2,50 \\ 1 & 5,00 & 2,70 \\ \vdots & \vdots & \vdots \\ 1 & 4,30 & 2,20 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ \vdots \\ e_{300} \end{bmatrix}$$

$z_i = a1 + bx_i + cy_i + e_i$

2.2 Modelos de regressão múltipla

Exemplo: Supondo dados de comprimento e largura de folhas de acerola, é possível obter relações matemáticas (estimar coeficientes) que levem à obtenção da área foliar.

ID	Comp (x)	Larg (y)	AFO (z)
1	6.00	3.20	13.39
2	5.50	3.20	12.52
3	5.00	2.50	8.78
4	5.00	2.70	9.60
.	.	.	.
.	.	.	.
.	.	.	.
300	4.30	2.20	8.23

Notação algébrica do modelo 6: $z_i = a + bx_i + cx_i^2 + dy_i + e_i$

Notação matricial do modelo 6: $\underset{\sim}{Y} = X\underset{\sim}{B} + \underset{\sim}{e}$

$$\underset{\sim}{Y} = \quad X \quad \underset{\sim}{B} + \underset{\sim}{e}$$

$$\begin{bmatrix} 13,39 \\ 12,52 \\ 8,78 \\ 9,60 \\ \vdots \\ 8,23 \end{bmatrix} = \begin{bmatrix} 1 & 6,00 & 6,00^2 & 3,20 \\ 1 & 5,50 & 5,50^2 & 3,20 \\ 1 & 5,00 & 5,00^2 & 2,50 \\ 1 & 5,00 & 5,00^2 & 2,70 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 4,30 & 4,30^2 & 2,20 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ \vdots \\ e_{300} \end{bmatrix}$$

$z_i = a1 + bx_i + cx_i^2 + dy_i + e_i$

2.2 Modelos de regressão múltipla

Exemplo: Supondo dados de comprimento e largura de folhas de acerola, é possível obter relações matemáticas (estimar coeficientes) que levem à obtenção da área foliar.

ID	Comp (x)	Larg (y)	AFO (z)
1	6.00	3.20	13.39
2	5.50	3.20	12.52
3	5.00	2.50	8.78
4	5.00	2.70	9.60
.	.	.	.
.	.	.	.
.	.	.	.
300	4.30	2.20	8.23

Notação algébrica do modelo 9: $z_i = a + bx_i + cy_i + dx_iy_i + e_i$

Notação matricial do modelo 9: $\underset{\sim}{Y} = X\underset{\sim}{B} + \underset{\sim}{e}$

$$\underset{\sim}{Y} = \quad X \quad \underset{\sim}{B} + \underset{\sim}{e}$$

$$\begin{bmatrix} 13,39 \\ 12,52 \\ 8,78 \\ 9,60 \\ \vdots \\ 8,23 \end{bmatrix} = \begin{bmatrix} 1 & 6,00 & 3,20 & 6,00*3,20 \\ 1 & 5,50 & 3,20 & 5,50*3,20 \\ 1 & 5,00 & 2,50 & 5,00*2,50 \\ 1 & 5,00 & 2,70 & 5,00*2,70 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 4,30 & 2,20 & 4,30*2,20 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ \vdots \\ e_{300} \end{bmatrix}$$

$z_i = a1 + bx_i + cy_i + dx_iy_i + e_i$

2.2 Modelos de regressão múltipla

Vantagens da notação matricial:

$$\underset{\sim}{Y} = X\underset{\sim}{B} + \underset{\sim}{e}$$

- Embora o modelo de regressão em sua forma algébrica muda de acordo com os parâmetros considerados, a forma matricial não se altera.
- Facilita os cálculos por meio de programação (recursos computacionais)
- Facilita a obtenção de estimadores.

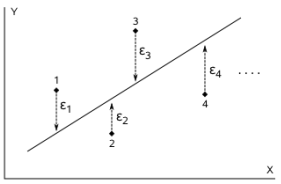
Como obter os coeficientes de regressão por álgebra de matriz???

2.3 Estimadores dos coeficientes de regressão

Desenvolvimento do raciocínio:

- Nosso objetivo é estimar coeficientes de um modelo de regressão que possibilite a predição da variável resposta com o menor erro possível.
- A partir do modelo de regressão em sua forma matricial é possível obter o erro por:

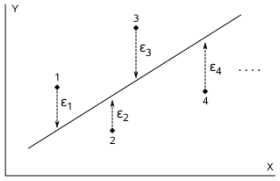
$$\underset{\sim}{Y} = X\underset{\sim}{B} + \underset{\sim}{e} \quad \Rightarrow \quad \underset{\sim}{e} = \underset{\sim}{Y} - X\underset{\sim}{B}$$



- Logo, queremos obter coeficientes do vetor B que tornem o erro o menor possível.
- Porém, o vetor de erros possui uma estimativa para cada unidade amostral. Consequentemente, precisamos de uma única medida de erro para representar todas as unidades amostrais.

2.3 Estimadores dos coeficientes de regressão

Desenvolvimento do raciocínio:



- Sendo o erro desvios com valores positivos e negativos, sua soma ou média nos levam ao valor zero. Logo, qual estratégia podemos usar para representar o erro associado à todas as unidades experimentais simultaneamente?

2.3 Estimadores dos coeficientes de regressão

Desenvolvimento do raciocínio:

- Uma estratégia é elevar cada desvio ao quadrado e depois somar. Com isso, obtemos o que chamamos de Soma de Quadrado dos Resíduos (SQR).

$$\underset{\sim}{e} = \underset{\sim}{Y} - X\underset{\sim}{B}$$

$$SQR = \underset{\sim}{e}'\underset{\sim}{e} = (\underset{\sim}{Y} - X\underset{\sim}{B})'(\underset{\sim}{Y} - X\underset{\sim}{B}) = \underset{\sim}{Y}'\underset{\sim}{Y} - 2\underset{\sim}{B}X'\underset{\sim}{Y} + \underset{\sim}{B}'X'X\underset{\sim}{B}$$

- Agora, precisamos apenas obter um estimador de B que minimize a SQR. Isso pode ser feito derivando a SQR em relação a B e depois igualando a zero:

$$\frac{\partial SQR}{\partial \underset{\sim}{B}} = -2X'Y + 2\underset{\sim}{B}'X'X \equiv 0$$

2.3 Estimadores dos coeficientes de regressão

Desenvolvimento do raciocínio:

$$-2X'Y + 2\underset{\sim}{B}'X'X = 0$$

$$2\underset{\sim}{B}'X'X = 2X'Y$$

$$\underset{\sim}{B}'X'X = X'Y$$

$$\hat{\underset{\sim}{B}} = (X'X)^{-1}X'Y$$

- Agora já temos o estimador para obtenção dos coeficientes de regressão.
- Este estimador foi obtido afim de minimizar os quadrados dos desvios (erros). Logo, podemos dizer que o estimador foi obtido pelo método dos quadrados mínimos ordinários.

2.3 Estimadores dos coeficientes de regressão

Voltando ao exemplo das folhas de acerola: Supondo dados de comprimento e largura de folhas de acerola, é possível obter relações matemáticas (estimar coeficientes) que levem à predição da área foliar.

ID	Comp (x)	Larg (y)	AFO (z)
1	6.00	3.20	13.39
2	5.50	3.20	12.52
3	5.00	2.50	8.78
4	5.00	2.70	9.60
.	.	.	.
.	.	.	.
.	.	.	.
300	4.30	2.20	8.23

Exemplo utilizando o modelo 6: $z_i = a + bx_i + cx_i^2 + dy_i + e_i$

Notação matricial : $\underset{\sim}{Y} = X\underset{\sim}{B} + \underset{\sim}{e}$

$$\underset{\sim}{Y} = \begin{bmatrix} 13,39 \\ 12,52 \\ 8,78 \\ 9,60 \\ \vdots \\ 8,23 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 6,00 & 6,00^2 & 3,20 \\ 1 & 5,50 & 5,50^2 & 3,20 \\ 1 & 5,00 & 5,00^2 & 2,50 \\ 1 & 5,00 & 5,00^2 & 2,70 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 4,30 & 4,30^2 & 2,20 \end{bmatrix} \quad B = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}$$

$$\hat{\underset{\sim}{B}} = (X'X)^{-1}X'Y = \begin{bmatrix} 0,1733 \\ 0,9523 \\ 0,3572 \\ 1,7323 \end{bmatrix}$$

$z_i = 0,1733^{ns} - 0,9523^{***}x_i + 0,3572^{***}x_i^2 + 1,7323^{***}y_i$

2.4 Representação por superfície resposta.

Após estimar o vetor *B* (coeficientes de regressão) conseguimos predizer a variável resposta (Área foliar) a partir das variáveis explicativas (comprimento e largura de folhas) por: $\hat{\underset{\sim}{Y}} = X\hat{\underset{\sim}{B}}$

Então podemos estabelecer diferentes valores da variável explicativa e, posteriormente, obter a variável resposta predita.

Largura (cm)	Comprimento da folha (cm)				
	1	2	3	...	10
1					
2					
3					
4					
5					

X= $\begin{pmatrix} 1 & 1 & 1^2 & 1 \\ 1 & 2 & 2^2 & 1 \\ 1 & 3 & 3^2 & 1 \\ 1 & 4 & 4^2 & 1 \\ 1 & 5 & 5^2 & 1 \\ 1 & 1 & 1^2 & 2 \\ 1 & 2 & 2^2 & 2 \\ 1 & 3 & 3^2 & 2 \\ 1 & 4 & 4^2 & 2 \\ 1 & 5 & 5^2 & 2 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & 1^2 & 10 \\ 1 & 2 & 2^2 & 10 \\ 1 & 3 & 3^2 & 10 \\ 1 & 4 & 4^2 & 10 \\ 1 & 5 & 5^2 & 10 \end{pmatrix}$ $\hat{\underset{\sim}{B}} = \begin{bmatrix} 0,1733 \\ 0,9523 \\ 0,3572 \\ 1,7323 \end{bmatrix}$

2.4 Representação por superfície resposta.

Após estimar o vetor *B* (coeficientes de regressão) conseguimos predizer a variável resposta (Área foliar) a partir das variáveis explicativas (comprimento e largura de folhas) por: $\hat{\underset{\sim}{Y}} = X\hat{\underset{\sim}{B}}$

Então podemos estabelecer diferentes valores da variável explicativa e, posteriormente, obter a variável resposta predita.

Largura (cm)	Comprimento da folha (cm)				
	1	2	3	...	10
1	1.3105	1.6031	2.6101	...	29.6623
2	3.0428	3.3354	4.3424	...	31.3946
3	4.7751	5.0677	6.0747	...	33.1269
4	6.5074	6.8000	7.807	...	34.8592
5	8.2397	8.5323	9.5393	...	36.5915

X= $\begin{pmatrix} 1 & 1 & 1^2 & 1 \\ 1 & 2 & 2^2 & 1 \\ 1 & 3 & 3^2 & 1 \\ 1 & 4 & 4^2 & 1 \\ 1 & 5 & 5^2 & 1 \\ 1 & 1 & 1^2 & 2 \\ 1 & 2 & 2^2 & 2 \\ 1 & 3 & 3^2 & 2 \\ 1 & 4 & 4^2 & 2 \\ 1 & 5 & 5^2 & 2 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & 1^2 & 10 \\ 1 & 2 & 2^2 & 10 \\ 1 & 3 & 3^2 & 10 \\ 1 & 4 & 4^2 & 10 \\ 1 & 5 & 5^2 & 10 \end{pmatrix}$ $\hat{\underset{\sim}{B}} = \begin{bmatrix} 0,1733 \\ 0,9523 \\ 0,3572 \\ 1,7323 \end{bmatrix}$

2.4 Representação por superfície resposta.

Após a predição de valores pelo modelo de regressão fica fácil obter gráficos de superfície resposta recorrendo-se à pacotes no software R ou no SigmaPlot, por exemplo.