Unidade 4. Regressão linear múltipla (duas variáveis)

4.1. Introdução e modelo estatístico

4.2. Estimação dos parâmetros do modelo

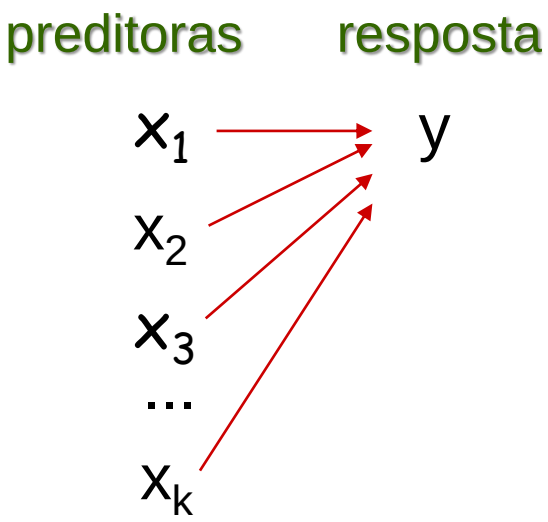
4.3. Inferências sobre os coeficientes de regressão parciais

4.4. Uso de software estatístico

Profa. Clause Piana

Análise de regressão linear múltipla

⇨ Os princípios básicos e os procedimentos da análise da regressão linear simples podem ser estendidos para situações que envolvem duas ou mais variáveis preditoras.



Essas são as circunstâncias mais comuns nas aplicações.

⇨ **Objetivo:** estudar o comportamento da variável resposta (Y) em função de duas ou mais variáveis preditoras (X_i).

Exemplo:

Um estudo foi realizado para identificar o modelo que melhor representa a relação entre a variável vazão mínima média (m^3/s) e as variáveis área de drenagem (km^2), declividade de drenagem (m/km) e densidade de drenagem (junções/km^2). Os dados observados em 10 estações fluviométricas da bacia do rio Paraopeba são apresentados na tabela abaixo.

Estação (j)	Vazão mínima média (y)	Área de drenagem (x_1)	Declividade de drenagem (x_2)	Densidade de drenagem (x_3)
1	2,60	461	2,69	0,098
2	1,49	291	3,94	0,079
3	1,43	244	7,20	0,119
4	3,44	579	3,18	0,102
5	1,37	293	2,44	0,123
6	28,53	5680	1,00	0,141
7	1,33	273	4,52	0,064
8	0,43	84	10,27	0,131
9	39,12	8734	0,66	0,143
10	45,00	10192	0,60	0,133

Análise de regressão linear múltipla

⇨ Inferências estatísticas são obtidas de uma amostra de n observações em cada uma das k+1 variáveis x_1, x_2, ..., x_k e y, ou seja, em um conjunto de observações:

$\{(x_{11}, x_{21},..., x_{k1}, y_1), (x_{12}, x_{22},..., x_{k2}, y_2),..., (x_{1n}, x_{2n},..., x_{kn}, y_n)\}$

⇨ Para uma observação j, a relação entre y e as variáveis preditoras x_1, x_2, ..., x_k é expressa pela seguinte equação, denominada equação do modelo amostral:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + ... + \beta_k x_{kj} + e_j,$$

onde:

$\beta_1, \beta_2,..., \beta_k$ são os coeficientes de regressão parciais

e_j é o erro aleatório

Análise de regressão linear múltipla

O conjunto das observações constitui um sistema de equações normais, com n equações e k+1 incógnitas.

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \ldots + \beta_k x_{k1} + e_1 \\ y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \ldots + \beta_k x_{k2} + e_2 \\ \qquad\qquad \ldots \\ y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \ldots + \beta_k x_{kn} + e_n \end{cases}$$

No exemplo: Predição de vazões mínimas (y) a partir das variáveis área de drenagem (x_1), declividade (x_2) e densidade de drenagem (x_3)

Vários modelos podem resultar desta análise:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + e_j \quad \leftarrow \text{modelo completo}$$

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + e_j \quad \leftarrow \text{exclusão da variável } x_3$$

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_3 x_{3j} + e_j \quad \leftarrow \text{exclusão da variável } x_2$$

$$\ldots$$

$$y_j = \beta_0 + \beta_2 x_{2j} + e_j \quad \leftarrow \text{exclusão das variáveis } x_1 \text{ e } x_3$$

$$y_j = \beta_0 + e_j \quad \leftarrow \text{nenhuma das variáveis tem efeito linear sobre y}$$

⇨ É possível que a relação entre as variáveis seja melhor representada por um modelo não linear.

Análise de regressão linear múltipla

⇨ O desenvolvimento algébrico dos fundamentos teóricos e os procedimentos computacionais da análise da regressão linear são relativamente simples para a situação de duas variáveis preditoras, constituindo, em geral, uma extensão simples da análise de regressão linear com uma única variável preditora.

$$y_i = \beta_0 + \beta_1 x_i + e_i$$
$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + e_j$$

⇨ Entretanto, para as situações de mais de duas variáveis preditoras, eles se tornam complexos e trabalhosos com os recursos da álgebra usual e de calculadoras comuns. Nestas situações, a fundamentação teórica torna-se consideravelmente facilitada com os recursos de sintetização simbólica propiciados pela álgebra linear matricial e a implementação dos procedimentos computacionais com os recursos da computação eletrônica facilita as aplicações.

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_k x_{kj} + e_j$$

⇨ Por conveniência didática, abordaremos nesta unidade a análise da regressão linear múltipla com duas variáveis preditoras.

Modelo de regressão linear múltipla com duas variáveis preditoras

Modelo estatístico

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + e_j$$

Exemplos:

✓ Relação entre quantidade de fósforo disponível na planta (Y) e quantidade de fósforo inorgânico (X_1) e de fósforo orgânico (X_2) disponível no solo;

✓ Relação entre peso (Y), em kg, altura (X_1), em cm, e idade (X_2), em meses, de um rebanho de cordeiros

✓ Relação entre produção de grãos de arroz (Y), em kg, altura da planta (X_1), em cm, e número de perfilhos (X_2)

Pressuposições

O modelo estatístico se completa com as seguintes pressuposições referentes a equação:

1. As variáveis X_i são fixas, isto é, observados sem erro.

2. Os erros (e_i) são aleatórios, têm média zero e variância constante, ou seja, $E(e_i) = 0$ e $V(e_i) = \sigma^2$.

3. Os erros (e_i) têm distribuição normal.

4. Os erros (e_i) são não correlacionados (o que implica em sua independência estatística, dado que têm distribuição normal).

Valores esperados de Y

Se $y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + e_j$, então

$E(e_j) = 0$

$$\mu_j = E(y_j) = E(\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + e_j)$$

$$\mu_j = E(y_j) = E(\beta_0) + E(\beta_1 x_{1j}) + E(\beta_2 x_{2j}) + E(e_j)$$

$$\mu_j = E(y_j) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j}$$

Erros

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + e_j$$

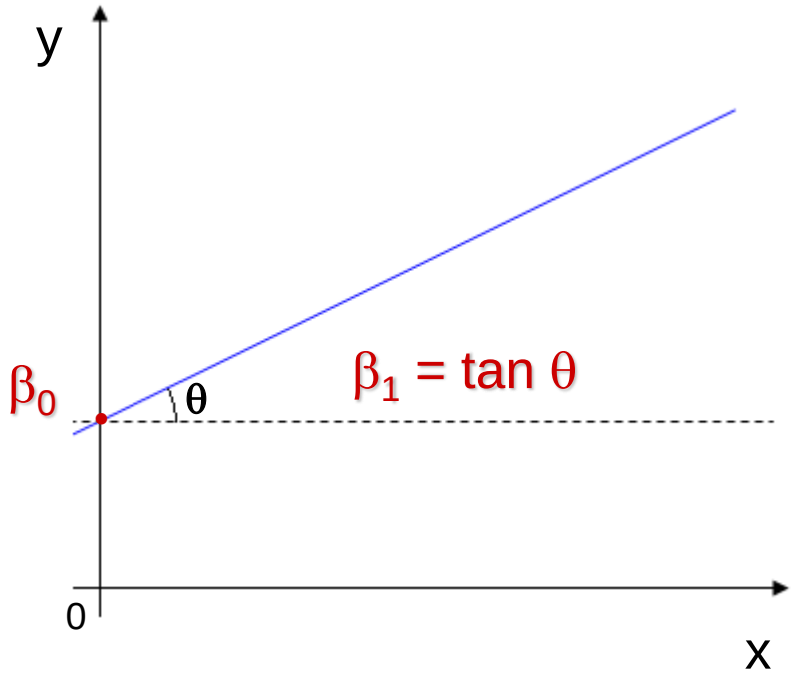
$$y_j = \mu_j + e_j$$

$$e_j = y_j - \mu_j$$

Modelo de regressão linear simples

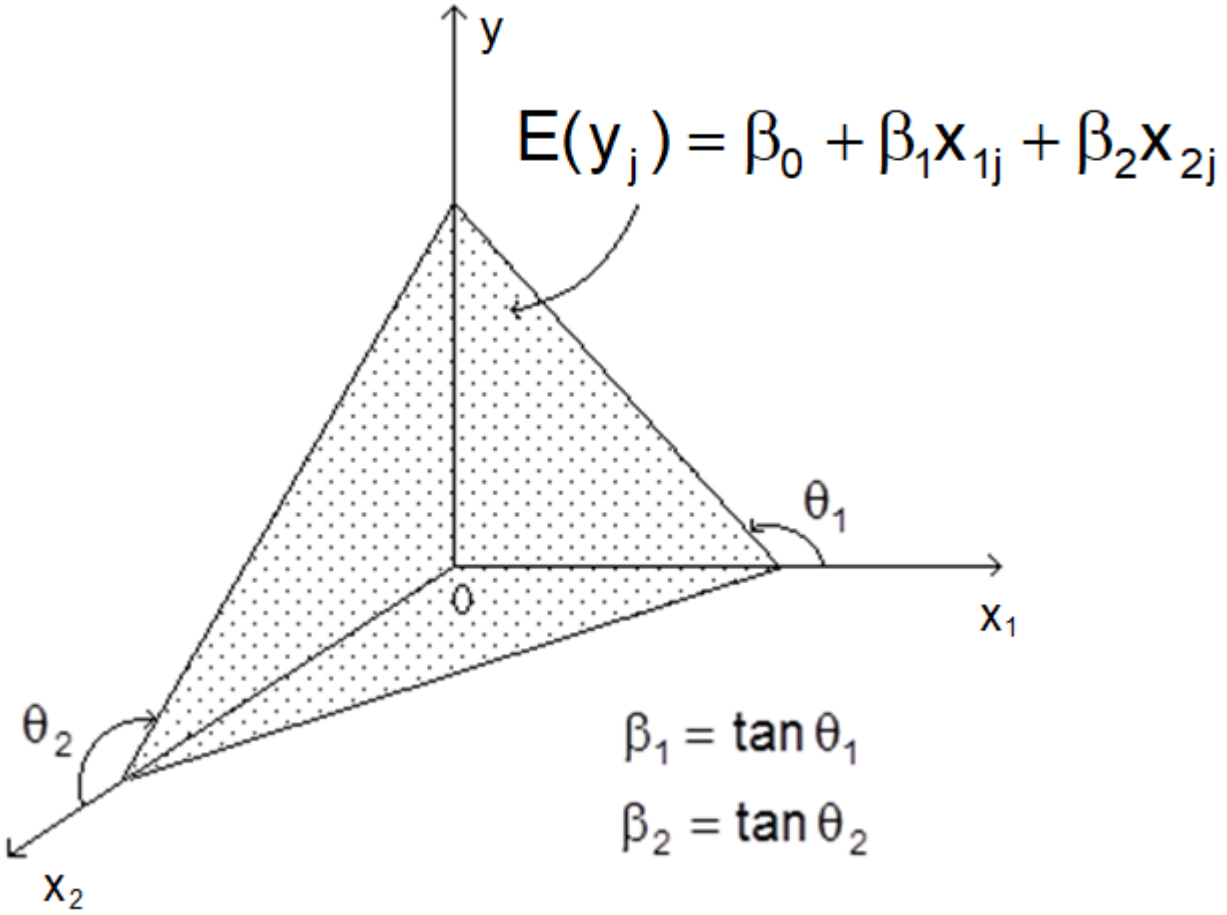
Representação geométrica

$$E(y_i) = \beta_0 + \beta_1 x_i$$



Modelo de regressão linear múltipla com duas variáveis preditoras

Representação geométrica



Modelo de regressão linear múltipla com mais de duas variáveis preditoras

$$E(y_j) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_k x_k$$

Esta equação não pode ser representada geometricamente nesses espaços em que o homem tem experiência.

Genericamente, uma equação de regressão linear múltipla com k (k>2) variáveis preditoras é a representação analítica de um hiperplano em um espaço de k+1 dimensões.

Análise de regressão linear múltipla com duas variáveis preditoras

Objetivo: **determinar a equação que melhor representa a relação existente entre as três variáveis** e, a partir desta equação, fazer predições para a variável resposta.

Para isso, uma sequência de passos deve ser seguida:

1. Obtenção das estimativas (pontuais) dos coeficientes β_0 , β_1 e β_2 para ajustar a equação da regressão.

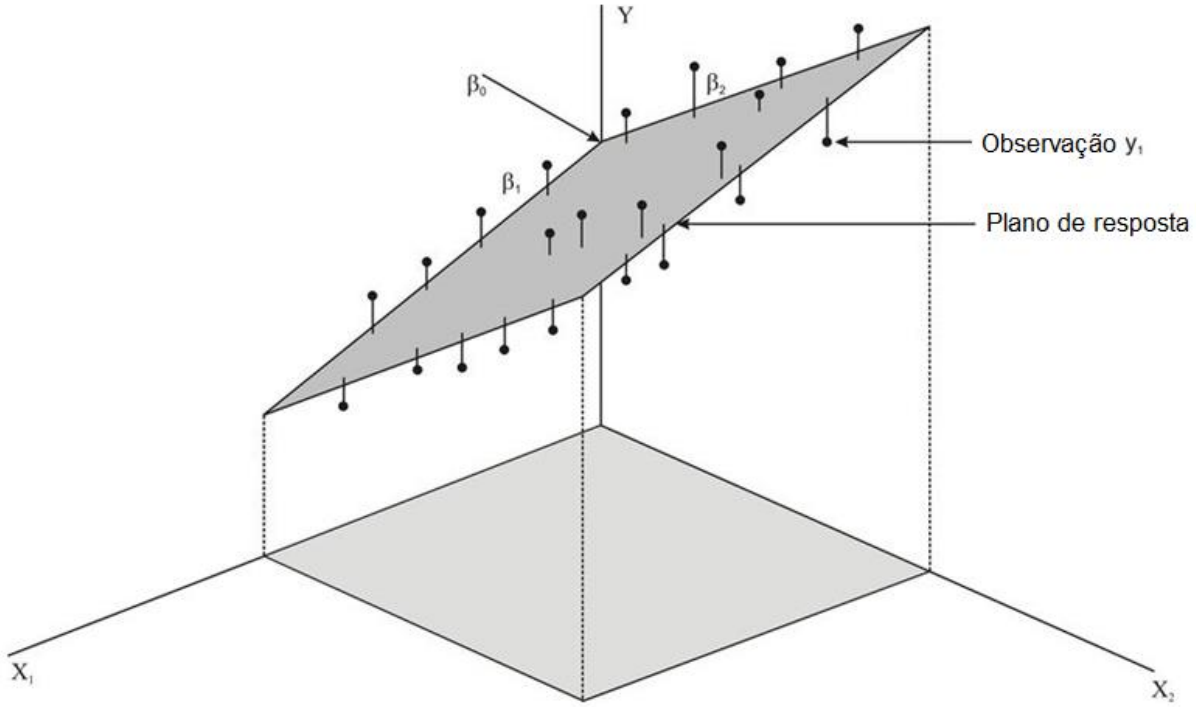
2. Aplicação de **testes de hipóteses** para as estimativas obtidas, a a fim de verificar se a equação de regressão é adequada.

3. Construção de **intervalos de confiança** para os valores estimados pela equação de regressão.

Estimação de parâmetros

A estimação pelo método dos quadrados mínimos consiste em determinar para estimadores dos parâmetros do modelo (β_0, β_1 e β_2) os valores $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$ que minimizam a soma dos quadrados dos erros, como função desses parâmetros:

$$f(\beta_0,\beta_1,\beta_2) = \sum e_j^2 = \sum (y_j - \mu_j)^2 = \sum \left(y_j - \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j}\right)^2$$



Melhor plano → $\sum e_i^2$ é mínima

Estimação de parâmetros

A estimação pelo método dos quadrados mínimos consiste em determinar para estimadores dos parâmetros do modelo (β_0, β_1 e β_2) os valores $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$ que minimizam a soma dos quadrados dos erros, como função desses parâmetros:

$$f(\beta_0,\beta_1,\beta_2) = \sum e_j^2 = \sum (y_j - \mu_j)^2 = \sum \left(y_j - \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j}\right)^2$$

O processo de minimização conduz às seguintes equações normais:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum x_{1j} + \hat{\beta}_2 \sum x_{2j} = \sum y_j \\ \hat{\beta}_0 \sum x_{1j} + \hat{\beta}_1 \sum x_{1j}^2 + \hat{\beta}_2 \sum x_{1j} x_{2j} = \sum x_{1j} y_j \\ \hat{\beta}_0 \sum x_{2j} + \hat{\beta}_1 \sum x_{1j} x_{2j} + \hat{\beta}_2 \sum x_{2j}^2 = \sum x_{2j} y_j \end{cases}$$

Estimadores dos coeficientes de regressão parciais

$$\begin{cases} \hat{\beta}_1 \sum (x_{1j} - \bar{x}_1)^2 + \hat{\beta}_2 \left[\sum (x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2) \right] = \sum (x_{1j} - \bar{x}_1)(y_j - \bar{y}) \\ \hat{\beta}_1 \left[\sum (x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2) \right] + \hat{\beta}_2 \sum (x_{2j} - \bar{x}_2)^2 = \sum (x_{2j} - \bar{x}_2)(y_j - \bar{y}) \end{cases}$$

$$\begin{cases} \hat{\beta}_1 SQX_1 + \hat{\beta}_2 SPX_1X_2 = SPX_1Y \\ \hat{\beta}_1 SPX_1X_2 + \hat{\beta}_2 SQX_2 = SPX_2Y \end{cases}$$

Estimador do intercepto

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2$$

Equação do plano ajustado

A equação do plano ajustado (também denominada equação predita, equação ajustada, ou equação de quadrados mínimos) é obtida da equação do modelo populacional substituindo os parâmetros β_0, β_1 e β_2 pelas respectivas estimativas de quadrados mínimos.

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_1 x_2$$

Estimativa de valores esperados de Y

O valor estimado da resposta (Y) para um par particular de valores das variáveis preditoras, x_1 e x_2, é obtido pela substituição destes valores nesta equação:

$$\hat{\mu}_{y(x_1,x_2)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_1 x_2$$

Esta substituição produz os valores preditos que correspondem aos respectivos valores observados da variável resposta.

Estimativas dos erros

$$e_j = y_j - \mu_j$$

$$\hat{e}_j = y_j - \hat{\mu}_j$$

Define-se como resíduo de uma observação y_j da variável resposta, denotado por $\hat{e}_j$, a diferença entre o valor observado y_j e o correspondente valor estimado $\hat{\mu}_j$.

Estimativa da variância do erro

$$S^2 = \frac{\sum_{j=1}^{n} \hat{e}_j^2}{n-3},$$

onde n-3 é o número de graus de liberdade (número de observações menos o número de parâmetros do modelo).

Exemplo:

Consideremos um experimento que teve como propósito estudar o efeito da suplementação de sal mineral e cálcio na dieta de ovinos sobre o peso ao abate. Os resultados obtidos são apresentados na tabela abaixo.

Animal (j)	Sal mineral (kg) (x_{1j})	Cálcio (kg) (x_{2j})	Peso (kg) (y_j)
1	0	0	1.5
2	1	2	6.5
3	1	4	10.0
4	2	2	11.0
5	2	4	11.5
6	3	6	16.5
Soma	9	18	57,0
Média	1,5	3	9,5

No exemplo:

Supondo-se a relação linear entre a variável resposta e as variáveis preditoras, cada valor observado da resposta pode ser expresso pela equação:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + e_j,$$

onde:

y_j é o peso do animal j, em kg;

x_{1j} é a quantidade suplementada de sal mineral, em kg;

x_{2j} é a quantidade suplementada de cálcio, em kg,

β_0 é o peso do animal, em kg, quando as quantidades suplementas de sal mineral e de cálcio são iguais a zero (X_1=0 kg e X_2=0 kg);

β_1 é a taxa de variação no peso do animal, em kg, para cada unidade (kg) suplementada de sal mineral, numa quantidade fixa qualquer de cálcio;

β_2 é a taxa de variação no peso do animal, em kg, para cada unidade (kg) suplementada de cálcio, numa quantidade fixa qualquer de sal mineral;

e_j é o erro (variação aleatória) associado ao animal j.

No exemplo: Estimação dos coeficientes de regressão parciais

$$\begin{cases} \hat{\beta}_1 SQX_1 + \hat{\beta}_2 SPX_1X_2 = SPX_1Y \\ \hat{\beta}_1 SPX_1X_2 + \hat{\beta}_2 SQX_2 = SPX_2Y \end{cases}$$

Quantidades necessárias

$$SQX_1 = \sum x_{1j}^2 - n\bar{x}_1^2$$

$$SPX_1Y = \sum x_{1j}y_j - n\bar{x}_1\bar{y}$$

$$SPX_1X_2 = \sum x_{1j}x_{2j} - n\bar{x}_1\bar{x}_2$$

$$SQX_2 = \sum x_{2j}^2 - n\bar{x}_2^2$$

$$SPX_2Y = \sum x_{2j}y_j - n\bar{x}_2\bar{y}$$

No exemplo: tabela auxiliar

j	x_{1j}	x_{2j}	y_j	y_j^2	x_{1j}^2	x_{2j}^2	$x_{1j}x_{2j}$	$x_{1j}y_j$	$x_{2j}y_j$
1	0	0	1,5	2,25	0	0	0	0	0
2	1	2	6,5	42,25	1	4	2	6,5	13
3	1	4	10	100	1	16	4	10	40
4	2	2	11	121	4	4	4	22	22
5	2	4	11,5	132,25	4	16	8	23	46
6	3	6	16,5	272,25	9	36	18	49,5	99
Soma	9	18	57	670	19	76	36	111	220
Média	1,5	3	9,5						

$$SQX_1 = \sum x_{1j}^2 - n\bar{x}_1^2 = 5{,}5$$

$$SPX_1Y = \sum x_{1j}y_j - n\bar{x}_1\bar{y} = 25{,}5$$

$$SPX_1X_2 = \sum x_{1j}x_{2j} - n\bar{x}_1\bar{x}_2 = 9$$

$$SQX_2 = \sum x_{2j}^2 - n\bar{x}_2^2 = 22$$

$$SPX_2Y = \sum x_{2j}y_j - n\bar{x}_2\bar{y} = 49$$

$$\begin{cases} \hat{\beta}_1 SQX_1 + \hat{\beta}_2 SPX_1X_2 = SPX_1Y \\ \hat{\beta}_1 SPX_1X_2 + \hat{\beta}_2 SQX_2 = SPX_2Y \end{cases}$$

$$\begin{cases} 5{,}5\hat{\beta}_1 + 9\hat{\beta}_2 = 25{,}5 \\ 9\hat{\beta}_1 + 22\hat{\beta}_2 = 49 \end{cases} \qquad \hat{\beta}_1 = 3 \qquad \hat{\beta}_2 = 1$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2$$

$$\hat{\beta}_0 = 9{,}5 - 3 \times 1{,}5 - 1 \times 3 = 2$$

No exemplo: tabela auxiliar

j	x_{1j}	x_{2j}	y_j	y_j^2	x_{1j}^2	x_{2j}^2	$x_{1j}x_{2j}$	$x_{1j}y_j$	$x_{2j}y_j$
1	0	0	1,5	2,25	0	0	0	0	0
2	1	2	6,5	42,25	1	4	2	6,5	13
3	1	4	10	100	1	16	4	10	40
4	2	2	11	121	4	4	4	22	22
5	2	4	11,5	132,25	4	16	8	23	46
6	3	6	16,5	272,25	9	36	18	49,5	99
Soma	9	18	57	670	19	76	36	111	220
Média	1,5	3	9,5						

Equação do plano ajustado

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\hat{\mu} = 2 + 3x_1 + 1x_2$$

Estimativas pontuais

No exemplo: Significado das estimativas dos parâmetros

$$\hat{\mu} = 2 + 3x_1 + 1x_2$$

As estimativas dos parâmetros β_0, β_1 e β_2 têm os seguintes significados referentes à relação de interesse entre a variável resposta Y e as variáveis preditoras X_1 e X_2:

$\hat{\beta}_0 = 2 \rightarrow$ estimativa do ganho de peso de ovinos (Y), em kg/animal, que não recebem suplementação de sal mineral e cálcio, ou seja, para X_1=0 e X_2=0;

$\hat{\beta}_1 = 3 \rightarrow$ estimativa do ganho de peso de ovinos(Y), em kg/animal, que corresponde a 1 kg de suplemento de sal mineral (X_1) na ração, para uma quantidade fixa, qualquer, de suplementação de cálcio (X_2);

$\hat{\beta}_2 = 1 \rightarrow$ estimativa do ganho de peso de ovinos, em kg/animal, que corresponde a 1 kg de suplemento de cálcio (X_2) na ração, para uma quantidade fixa, qualquer, de suplementação de sal mineral (X_1).

Equação do plano ajustado

$$\hat{\mu} = 2 + 3x_1 + 1x_2$$

Obtenção das estimativas de médias esperadas

$$\hat{\mu}_1 = \hat{\mu}_{y(x_1=0,x_2=0)} = 2 + 3 \times 0 + 1 \times 0 = 2$$

$$\hat{\mu}_2 = \hat{\mu}_{y(x_1=1,x_2=2)} = 2 + 3 \times 1 + 1 \times 2 = 7$$

j	x_{1j}	x_{2j}	y_j	y_j^2	x_{1j}^2	x_{2j}^2	$x_{1j}x_{2j}$	$x_{1j}y_j$	$x_{2j}y_j$
1	0	0	1,5	2,25	0	0	0	0	0
2	1	2	6,5	42,25	1	4	2	6,5	13
3	1	4	10	100	1	16	4	10	40
4	2	2	11	121	4	4	4	22	22
5	2	4	11,5	132,25	4	16	8	23	46
6	3	6	16,5	272,25	9	36	18	49,5	99
Soma	9	18	57	670	19	76	36	111	220
Média	1,5	3	9,5						

Equação do plano ajustado

$$\hat{\mu} = 2 + 3x_1 + 1x_2$$

Obtenção das estimativas de médias esperadas

$$\hat{\mu}_1 = \hat{\mu}_{y(x_1=0,x_2=0)} = 2 + 3 \times 0 + 1 \times 0 = 2$$

$$\hat{\mu}_2 = \hat{\mu}_{y(x_1=1,x_2=2)} = 2 + 3 \times 1 + 1 \times 2 = 7$$

$$\ldots$$

$$\hat{\mu}_6 = \hat{\mu}_{y(x_1=3,x_2=6)} = 2 + 3 \times 3 + 1 \times 6 = 17$$

Obtenção dos resíduos

$$\hat{e}_1 = 1{,}5 - 2 = -0{,}5$$

$$\hat{e}_2 = 7 - 6{,}5 = 0{,}5$$

$$\ldots$$

$$\hat{e}_6 = 17 - 16{,}5 = 0{,}5$$

Estimativas de médias esperadas e de resíduos

j	x_{1j}	x_{2j}	y_j	$\hat{\mu}_j$	$\hat{e}_j$	$\hat{e}_j^2$
1	0	0	1,5	2	-0,5	0,25
2	1	2	6,5	7	-0,5	0,25
3	1	4	10	9	1	1
4	2	2	11	10	1	1
5	2	4	11,5	12	-0,5	0,25
6	3	6	16,5	17	-0,5	0,25
Soma	9	18	57	57	0	3
Média	1,5	3	9,5	9,5		

Testes de hipóteses sobre os parâmetros

1. Testes da hipótese de linearidade da relação entre as variáveis

2. Testes das hipóteses parciais

Testes de hipóteses sobre os parâmetros

1. Testes da hipótese de linearidade da relação entre as variáveis

A primeira hipótese de interesse em análise de regressão linear múltipla é a hipótese geral referente à própria existência de relação linear entre a variável resposta e as variáveis preditoras. Essa hipótese pode ser expressa por:

$$\begin{cases} H_0 : \beta_i = 0, \text{ sendo } i = 1, 2 \\ H_1 : \beta_i \neq 0, \text{ para pelo menos um } i \ (i = 1 \text{ e/ou } 2) \end{cases}$$

H_0: nenhuma das variáveis preditoras tem efeito linear sobre a variável resposta (β_1=0 e β_2=0)

H_1: pelo menos uma das variáveis preditoras tem efeito linear sobre a variável resposta ($\beta_1 \neq 0$ e β_2=0 ou β_1=0 e $\beta_2 \neq 0$ ou $\beta_1 \neq 0$ e $\beta_2 \neq 0$)

Testes de hipóteses sobre os parâmetros

1. Testes da hipótese de linearidade da relação entre as variáveis

Primeira hipótese de interesse → **hipótese geral** referente à existência de relação linear entre a variável resposta e as variáveis preditoras.

$$\begin{cases} H_0 : \beta_i = 0, \text{ sendo } i = 1, 2 \\ H_1 : \beta_i \neq 0, \text{ para pelo menos um } i \ (i = 1 \text{ e/ou } 2) \end{cases}$$

H_0: nenhuma das variáveis preditoras tem efeito linear sobre a variável resposta (β_1=0 e β_2=0)

H_1: pelo menos uma das variáveis preditoras tem efeito linear sobre a variável resposta (β_1≠0 e β_2=0 ou β_1=0 e β_2≠0 ou β_1≠0 e β_2≠0)

Essa hipótese pode ser testada pela seguinte estatística F provida pela análise da variância:

$$F = \frac{s^2_{Reg}}{s^2} \sim F(\nu_{Reg}, \nu)$$

Análise da variância

A análise da variância decompõe a variação total das observações, representada pelos desvios $(y_j - \bar{y})$, em duas partes:

→ variação explicada pela equação de regressão → $(\hat{\mu}_j - \bar{y})$

→ variação aleatória, não explicada pela regressão → $(y_j - \hat{\mu}_j)$

Assim, a variação de cada observação pode ser representada pela seguinte expressão:

$$(y_j - \bar{y}) = (\hat{\mu}_j - \bar{y}) + (y_j - \hat{\mu}_j)$$

E a variação total das observações pode ser representada por:

$$\sum (y_j - \bar{y})^2 = \sum \left[(\hat{\mu}_j - \bar{y}) + (y_j - \hat{\mu}_j) \right]^2$$

$$\sum (y_j - \bar{y})^2 = \sum (\hat{\mu}_j - \bar{y})^2 + \sum (y_j - \hat{\mu}_j)^2$$

$$SQ_{Total} = SQ_{Regressão} + SQ_{Resíduo}$$

Tabela da análise da variância

Fonte de variação	GL (ν)	SQ	QM (S^2)	F
Regressão	ν_{Reg}=3-1	$\sum(\hat{\mu}_j - \bar{y})^2$	$\frac{SQ_{Reg}}{\nu_{Reg}}$	$\frac{S^2_{Reg}}{S^2}$
Resíduo	ν=n-3	$\sum(y_j - \hat{\mu}_j)^2$	$\frac{SQ}{\nu}$	-
Total	ν_{Total}=n-1	$\sum(y_j - \bar{y})^2$	-	-

Hipóteses estatísticas:

$$\begin{cases} H_0 : \beta_i = 0, \text{ sendo } i = 1, 2 \\ H_1 : \beta_i \neq 0, \text{ para pelo menos um } i \ (i = 1 \text{ e/ou } 2) \end{cases}$$

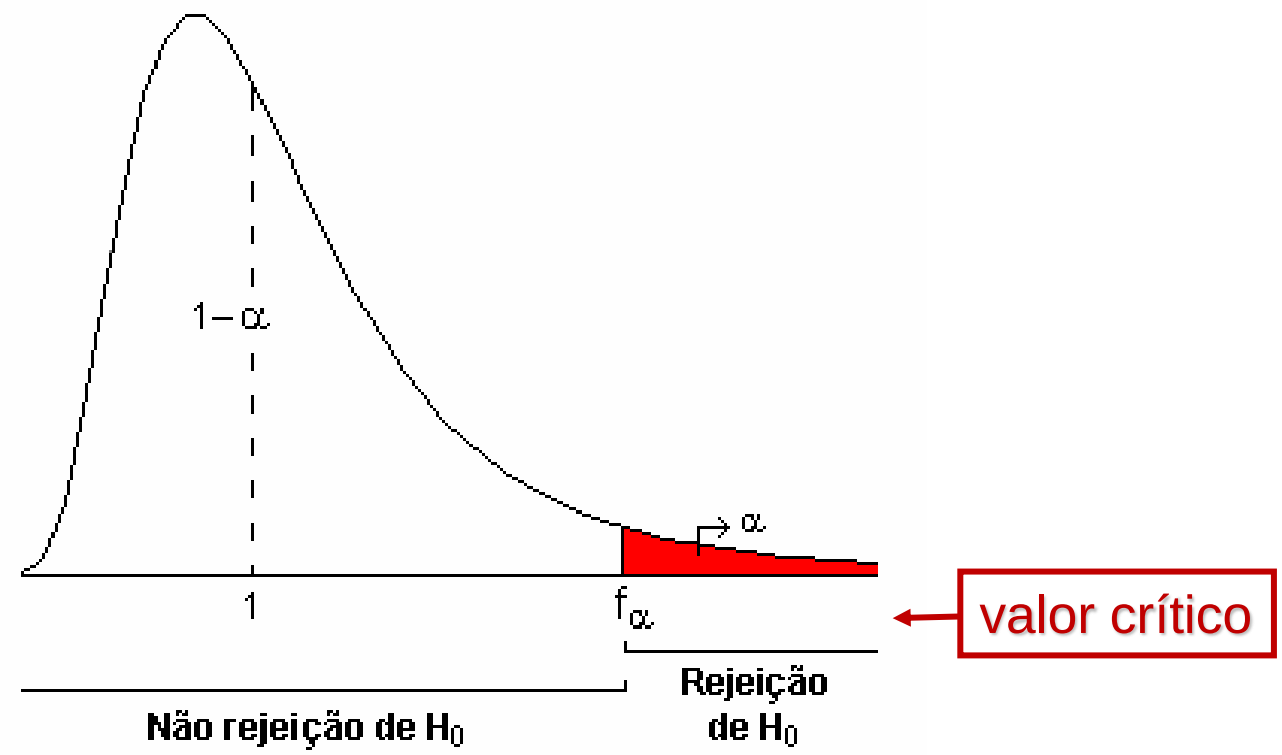
Estatística do teste: $F = \frac{s^2_{Reg}}{s^2} \sim F(\nu_{Reg}, \nu)$

Profa. Clause Piana

Como tomar a decisão a respeito de H_0?

Se H_0 é verdadeira, devemos esperar que o valor da estatística F seja próximo de 1.

Distribuição F



Critério de decisão

⇨ Se $f > f_\alpha$, rejeitamos H_0 ⇒ f é atípico

⇨ Se $f < f_\alpha$, não temos motivos para rejeitar H_0 ⇒ f é típico

Tabela da análise da variância

Fonte de variação	GL (ν)	SQ	QM (S^2)	F
Regressão	ν_{Reg}=3-1	$\sum(\hat{\mu}_j - \bar{y})^2$	$\frac{SQ_{Reg}}{\nu_{Reg}}$	$\frac{S^2_{Reg}}{S^2}$
Resíduo	ν=n-3	$\sum(y_j - \hat{\mu}_j)^2$	$\frac{SQ}{\nu}$	-
Total	ν_{Total}=n-1	$\sum(y_j - \bar{y})^2$	-	-

Obtenção das somas de quadrados:

$$SQ_{Total} = \sum(y_j - \bar{y})^2 = SQY$$

$$SQ_{Reg} = \sum(\hat{\mu}_i - \bar{y})^2 = \hat{\beta}_1 \sum(x_{1j} - \bar{x}_1)(y_j - \bar{y}) + \hat{\beta}_2 \sum(x_{2j} - \bar{x}_2)(y_j - \bar{y})$$

$$= \hat{\beta}_1 SPX_1Y + \hat{\beta}_2 SPX_2Y$$

$$SQ_{Res} = \sum(y_j - \hat{\mu}_j)^2 = \sum \hat{e}_j^2 \text{ (por diferença)}$$

Coeficiente de determinação (r^2)

- O coeficiente de determinação da regressão múltipla da variável resposta Y em relação às variáveis preditoras X_1 e X_2 é a proporção da variação total de Y que é "explicada" pela regressão de Y em relação à X_1 e X_2.

- O coeficiente de determinação é dado pela razão entre a soma dos quadrados da regressão (SQ_{Reg}) e a soma de quadrados total (SQ_{Total})

$$r^2 = \frac{SQ_{Reg}}{SQ_{Total}}$$

Coeficiente de determinação corrigido

- Incluir variáveis no modelo sempre aumenta o r^2. Por esta razão, recomenda-se ajustar o coeficiente para o número de parâmetros presentes no modelo, utilizando-se o coeficiente de determinação corrigido:

$$r_C^2 = r^2 - \frac{2}{n-3}\left(1 - r^2\right)$$

Exemplo: Consideremos um experimento que teve como propósito estudar o efeito da suplementação de sal mineral e cálcio na dieta de ovinos sobre o peso ao abate. Os resultados obtidos são apresentados na tabela abaixo.

Animal (j)	Sal mineral (kg) (x_{1j})	Cálcio (kg) (x_{2j})	Peso (kg) (y_j)
1	0	0	1,5
2	1	2	6,5
3	1	4	10,0
4	2	2	11,0
5	2	4	11,5
6	3	6	16,5
Soma	9	18	57,0
Média	1,5	3	9,5

Estimativas pontuais dos parâmetros do modelo

$$\hat{\beta}_0 = 2 \qquad \hat{\beta}_1 = 3 \qquad \hat{\beta}_2 = 1$$

Equação do plano ajustado: $\hat{\mu} = 2 + 3x_1 + 1x_2$

No exemplo: tabela auxiliar

j	x_{1j}	x_{2j}	y_j	y_j^2	x_{1j}^2	x_{2j}^2	$x_{1j}x_{2j}$	$x_{1j}y_j$	$x_{2j}y_j$
1	0	0	1,5	2,25	0	0	0	0	0
2	1	2	6,5	42,25	1	4	2	6,5	13
3	1	4	10	100	1	16	4	10	40
4	2	2	11	121	4	4	4	22	22
5	2	4	11,5	132,25	4	16	8	23	46
6	3	6	16,5	272,25	9	36	18	49,5	99
Soma	9	18	57	670	19	76	36	111	220
Média	1,5	3	9,5						

$$\hat{\mu} = 2 + 3x_1 + 1x_2$$

$$SPX_1Y = \sum x_{1j}y_j - n\bar{x}_1\bar{y} = 25{,}5$$

$$SPX_2Y = \sum x_{2j}y_j - n\bar{x}_2\bar{y} = 49$$

$$SQ_{Total} = SQY = \sum y_j^2 - n\bar{y}^2 = 128{,}5$$

$$SQ_{Reg} = \hat{\beta}_1 SPX_1Y + \hat{\beta}_2 SPX_2Y$$

$$= 3 \times 25{,}5 + 1 \times 49 = 125{,}5$$

No exemplo: Tabela da análise da variância

Fonte de variação	ν	SQ	S^2	F
Regressão	2	125,5	62,75	62,75
Resíduo	3	3,0	1,00	
Total	5	128,5		

$$\begin{cases} H_0 : \beta_i = 0, \text{ sendo } i = 1, 2 \\ H_1 : \beta_i \neq 0, \text{ para pelo menos um } i \ (i = 1 \text{ e/ou } 2) \end{cases}$$

Decisão: $f = 62,75 > f_{\alpha(2;\,3)} = 9,55 \rightarrow$ Rejeitamos H_0

Concluímos, ao nível de 5% de significância, que pelo menos uma das variáveis preditoras (quantidade de sal ou de cálcio) tem efeito linear sobre o peso de ovinos.

No exemplo: Tabela da análise da variância

Fonte de variação	ν	SQ	S^2	F	Prob.>F
Regressão	2	125,5	62,75	62,75	0,0036
Resíduo	3	3,0	1,00		
Total	5	128,5			

$$\begin{cases} H_0 : \beta_i = 0, \text{ sendo } i = 1, 2 \\ H_1 : \beta_i \neq 0, \text{ para pelo menos um } i \ (i = 1 \text{ e/ou } 2) \end{cases}$$

Decisão com base no valor p:

Valor p = 0,0036 < α = 0,05 $\rightarrow$ Rejeitamos H_0

No exemplo: Tabela da análise da variância

Fonte de variação	ν	SQ	S^2	F	Prob.>F
Regressão	2	125,5	62,75	62,75	0,0036
Resíduo	3	3,0	1,00		
Total	5	128,5			

Coeficiente de determinação corrigido

$$r_C^2 = r^2 - \frac{2}{n-3}\left(1 - r^2\right)$$

$$r_C^2 = 0{,}977 - \frac{2}{3}\left(1 - 0{,}977\right) = 0{,}961$$

Verifica-se que 96% da variação da resposta (ganho do peso dos ovinos) é explicada pelo efeito linear de pelo menos uma das variáveis preditoras uma das variáveis preditoras (quantidade de sal e quantidade de cálcio adicionadas a ração).

Testes de hipóteses sobre os parâmetros

2. Testes das hipóteses parciais

⇒ A hipótese parcial referente ao β_1 supõe o efeito linear da variável preditora X_1 sobre a variável resposta Y, em adição ao efeito da variável preditora X_2.

⇒ A hipótese parcial referente ao β_2 supõe o efeito linear da variável preditora X_2 sobre a variável Y, em adição ao efeito da variável preditora X_1.

As duas hipóteses parciais são especificadas por:

$$\begin{cases} H_0^1 : \beta_1 = 0 \\ H_1^1 : \beta_1 \neq 0 \end{cases} \quad \text{e} \quad \begin{cases} H_0^2 : \beta_2 = 0 \\ H_1^2 : \beta_2 \neq 0 \end{cases}$$

Essas hipóteses podem ser testadas pela estatística T:

$$T = \frac{\hat{\beta}_i}{S(\hat{\beta}_i)} \sim t(\nu = n - 3)$$

Variância da estimativa de β_1

$$V(\hat{\beta}_1) = \frac{\sum \left(x_{2j} - \bar{x}_2\right)^2}{\sum \left(x_{1j} - \bar{x}_1\right)^2 \sum \left(x_{2j} - \bar{x}_2\right)^2 - \left[\sum_{j=1}^{n} \left(x_{1j} - \bar{x}_1\right)\left(x_{2j} - \bar{x}_2\right)\right]^2} \sigma^2_{y(x_1,x_2)}$$

$$\hat{\sigma}^2_{y(x_1,x_2)} = S^2 = \frac{\sum_{j=1}^{n} \hat{e}_j^2}{n-3}$$

Estimativa desta variância

$$S^2(\hat{\beta}_1) = \frac{SQX_2}{SQX_1.SQX_2 - (SPX_1X_2)^2} \left(\frac{\sum_{j=1}^{n} \hat{e}_j^2}{n-3} \right)$$

Variância da estimativa de β_2

$$V(\hat{\beta}_2) = \frac{\sum (x_{1j} - \bar{x}_1)^2}{\sum (x_{1j} - \bar{x}_1)^2 \sum (x_{2j} - \bar{x}_2)^2 - \left[\sum_{j=1}^{n} (x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2)\right]^2} \sigma^2_{y(x_1,x_2)}$$

$$\hat{\sigma}^2_{y(x_1,x_2)} = S^2 = \frac{\sum_{j=1}^{n} \hat{e}_j^2}{n-3}$$

Estimativa desta variância

$$S^2(\hat{\beta}_2) = \frac{SQX_1}{SQX_1.SQX_2 - (SPX_1X_2)^2} \left(\frac{\sum_{j=1}^{n} \hat{e}_j^2}{n-3} \right)$$

Estimativas de variâncias dos coeficientes de regressão parciais

$$S^2(\hat{\beta}_1) = \frac{\sum x_{2j}^2 - n\bar{x}_2^2}{\left(\sum x_{1j}^2 - n\bar{x}_1^2\right)\left(\sum x_{2j}^2 - n\bar{x}_2^2\right) - \left(\sum x_{1j}x_{2j} - n\bar{x}_1\bar{x}_2\right)^2} \left(\frac{\sum_{j=1}^{n} \hat{e}_j^2}{n-3} \right)$$

$$S^2(\hat{\beta}_2) = \frac{\sum x_{1j}^2 - n\bar{x}_1^2}{\left(\sum x_{1j}^2 - n\bar{x}_1^2\right)\left(\sum x_{2j}^2 - n\bar{x}_2^2\right) - \left(\sum x_{1j}x_{2j} - n\bar{x}_1\bar{x}_2\right)^2} \left(\frac{\sum_{j=1}^{n} \hat{e}_j^2}{n-3} \right)$$

Estimativas de variâncias dos coeficientes de regressão parciais

$$S^2(\hat{\beta}_1) = \frac{SQX_2}{SQX_1.SQX_2 - (SPX_1X_2)^2} \left(\frac{\sum_{j=1}^{n} \hat{e}_j^2}{n-3} \right)$$

$$S^2(\hat{\beta}_2) = \frac{SQX_1}{SQX_1.SQX_2 - (SPX_1X_2)^2} \left(\frac{\sum_{j=1}^{n} \hat{e}_j^2}{n-3} \right)$$

No exemplo: tabela auxiliar

j	x_{1j}	x_{2j}	y_j	y_j^2	x_{1j}^2	x_{2j}^2	$x_{1j}x_{2j}$	$x_{1j}y_j$	$x_{2j}y_j$
1	0	0	1,5	2,25	0	0	0	0	0
2	1	2	6,5	42,25	1	4	2	6,5	13
3	1	4	10	100	1	16	4	10	40
4	2	2	11	121	4	4	4	22	22
5	2	4	11,5	132,25	4	16	8	23	46
6	3	6	16,5	272,25	9	36	18	49,5	99
Soma	9	18	57	670	19	76	36	111	220
Média	1,5	3	9,5						

Estimativas das variâncias dos coeficientes de regressão parciais

$$S^2(\hat{\beta}_1) = \frac{SQX_2}{SQX_1.SQX_2 - (SPX_1X_2)^2}\left(\frac{\sum_{j=1}^{n}\hat{e}_j^2}{n-3}\right) = \frac{22}{5{,}5\times 22 - 9^2}\times\frac{3}{6-3} = \frac{22}{40}\times 1 = 0{,}55$$

$$S^2(\hat{\beta}_2) = \frac{SQX_1}{SQX_1.SQX_2 - (SPX_1X_2)^2}\left(\frac{\sum_{j=1}^{n}\hat{e}_j^2}{n-3}\right) = \frac{5{,}5}{5{,}5\times 22 - 9^2}\times\frac{3}{6-3} = \frac{5{,}5}{40}\times 1 = 0{,}1375$$

No exemplo: tabela auxiliar

j	x_{1j}	x_{2j}	y_j	y_j^2	x_{1j}^2	x_{2j}^2	$x_{1j}x_{2j}$	$x_{1j}y_j$	$x_{2j}y_j$
1	0	0	1,5	2,25	0	0	0	0	0
2	1	2	6,5	42,25	1	4	2	6,5	13
3	1	4	10	100	1	16	4	10	40
4	2	2	11	121	4	4	4	22	22
5	2	4	11,5	132,25	4	16	8	23	46
6	3	6	16,5	272,25	9	36	18	49,5	99
Soma	9	18	57	670	19	76	36	111	220
Média	1,5	3	9,5						

Equação do plano ajustada

$$\hat{\mu} = 2 + 3x_1 + 1x_2$$

Estimativas dos coeficientes de regressão parciais

$$\hat{\beta}_1 = 3 \qquad \hat{\beta}_2 = 1$$

Estimativas das variâncias dos coeficientes de regressão parciais

$$S^2(\hat{\beta}_1) = \frac{22}{40} \times 1 = 0{,}55 \qquad S^2(\hat{\beta}_2) = \frac{5{,}5}{40} \times 1 = 0{,}1375$$

No exemplo: Testes das hipóteses parciais

Hipóteses estatísticas

$$\begin{cases} H_0^1 : \beta_1 = 0 \\ H_1^1 : \beta_1 \neq 0 \end{cases}$$

Efeito linear da suplementação de sal sobre o ganho de peso de ovinos, em adição ao efeito da suplementação de cálcio.

Estatística do teste

$$t = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} = \frac{3}{\sqrt{0{,}55}} = 4{,}045$$

Decisão e conclusão

$$t_{\alpha/2(n-3)} = 3{,}183$$

$$|t = 4{,}045| > t_{\alpha/2(n-3)} = 3{,}183 \leftarrow \text{Rejeitamos } H_0$$

Concluímos, ao nível de 5% de significância, que o coeficiente de regressão parcial populacional β_1 difere de zero. Portanto, existe efeito linear significativo da quantidade de sal mineral, adicional ao efeito da quantidade de cálcio, sobre o ganho de peso dos ovinos.

No exemplo: Testes das hipóteses parciais

Hipóteses estatísticas

$$\begin{cases} H_0^2 : \beta_2 = 0 \\ H_1^2 : \beta_2 \neq 0 \end{cases}$$

Efeito linear da suplementação de cálcio sobre o ganho de peso de ovinos, em adição ao efeito da suplementação de sal.

Estatística do teste

$$t = \frac{\hat{\beta}_2}{s(\hat{\beta}_2)} = \frac{1}{\sqrt{0{,}1375}} = 2{,}697$$

Decisão e conclusão

$$t_{\alpha/2(n-3)} = 3{,}183$$

$$|\, t = 2{,}697 \,| < t_{\alpha/2(n-3)} = 3{,}183 \quad \leftarrow \text{ Não rejeitamos } H_0$$

Concluímos, ao nível de 5% de significância, que o coeficiente de regressão parcial populacional β_2 não difere de zero. Portanto, não existe efeito linear significativo da quantidade de cálcio, adicional ao efeito da quantidade de sal mineral, sobre o ganho de peso dos ovinos.

Conclusão geral

A relação linear entre Y e (X_1, X_2) foi significativa, a contribuição adicional da variável X_1 para a explicação da variação de Y foi significativa e a contribuição adicional de X_2 não foi significativa.
Isso implica que a relação linear não pode prescindir da variável X_1, mas pode prescindir da variável X_2.
Assim, segundo os testes efetuados, o "melhor" modelo para exprimir a relação linear entre Y e (X_1, X_2) é:

$$\mu = \beta_0 + \beta_1 x_1$$

ou seja, o modelo de regressão linear simples de Y (ganho de peso) em relação a X_1 (suplementação de sal mineral).

As estimativas dos parâmetros β_0 e β_1 podem ser determinadas pelo procedimento da regressão linear simples. Obtém-se, assim, a equação da reta:

$$\hat{\mu} = 2{,}54 + 4{,}64 x_1$$

Seleção de variáveis

→ Em análise de regressão linear com mais de duas variáveis preditoras, a escolha do "melhor" modelo de relação linear não é tão simples.

→ Em algumas circunstâncias não existe um conhecimento mais objetivo sobre a importância relativa de variáveis sobre a resposta.

→ Nesse caso, é possível conduzir estudos com finalidades exploratórias considerando um conjunto de variáveis e utilizando a análise de regressão para auxiliar no processo de seleção das variáveis, eliminando aquelas que porventura não tenham efeito significativo sobre a resposta.

No exemplo: Predição de vazões mínimas (y) a partir das variáveis área de drenagem (x_1), declividade (x_2) e densidade de drenagem (x_3)

Vários modelos podem resultar desta análise:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + e_j \quad \leftarrow \text{modelo completo}$$

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + e_j \quad \leftarrow \text{exclusão da variável } x_3$$

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_3 x_{3j} + e_j \quad \leftarrow \text{exclusão da variável } x_2$$

$$\ldots$$

$$y_j = \beta_0 + \beta_2 x_{2j} + e_j \quad \leftarrow \text{exclusão das variáveis } x_1 \text{ e } x_3$$

$$y_j = \beta_0 + e_j \quad \leftarrow \text{nenhuma das variáveis tem efeito linear sobre y}$$

⇨ É possível que a relação entre as variáveis seja melhor representada por um modelo não linear.

Métodos de seleção de variáveis

⇒ Inclusão ascendente (*forward selection*): inicia-se com um modelo que possui somente o intercepto e, de acordo com o critério fixado, as variáveis preditoras são incluída no modelo, uma a uma. Uma vez incluída no modelo, a variável não sai mais.

$$E(Y) = \mu = \beta_0$$

⇒ Seleção descendente (*backward elimination*): começa com o modelo completo e, de acordo com o critério fixado, vai excluindo, uma a uma, as variáveis de menor contribuição não significativa, na presença das demais variáveis no modelo.

$$E(Y) = \mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

⇒ Seleção ascendente-descendente (*stepwise selection*) é uma aplicação conjunta dos critérios de inclusão e exclusão. O procedimento inicia do mesmo modo que a seleção ascendente, mas em cada passo verifica se, na presença das outras variáveis do modelo, alguma variável não agrega contribuição significativa à explicação da variação da resposta. Dentre as que não estão contribuindo significativamente, a de menor f parcial é eliminada. Por outro lado, uma variável que já foi excluída poderá retornar em um passo posterior.

$$E(Y) = \mu = \beta_0$$

Bibliografia consultada

SILVA, J.G.C. da **Estatística experimental: análise estatística de experimentos.** Pelotas, RS: Instituto de Física e Matemática, Universidade Federal de Pelotas, 2000. 318p.

NAGHETTINI, M.; PINTO, E. J. de A. **Hidrologia estatística**. Belo Horizonte: CPRM, 2007. 552 p.

Sistema Galileu de Educação Estatística. Disponível em: http://www.galileu.esalq.usp.br