Regressão Linear Simples e Correlação

tradução do Jay Davore e do livro Análise de variância (W.BUSSAB)

índice

Exemplo

Um psicólogo está investigando a relação entre o tempo que o indivíduo leva para reagir a certo estímulo e algumas de suas características tais como: sexo, idade e acuidade visual (em %). O resultado para 20 indivíduos é apresentado, onde

$i =$ indivíduo

$y_i =$ tempo de reação

$w_i =$ sexo

$x_i =$ idade

$z_i =$ acuidade visual

Exemplo

i	y_i	w_i	x_i	z_i	i	y_i	w_i	x_i	z_i
1	96	H	20	90	11	109	H	30	90
2	92	M	20	100	12	100	H	30	80
3	106	H	20	80	13	112	H	35	90
4	100	M	20	90	14	105	H	35	80
5	98	M	25	100	15	118	H	35	70
6	104	H	25	90	16	108	H	35	90
7	110	H	25	80	17	113	H	40	90
8	101	M	25	90	18	112	H	40	90
9	116	M	30	70	19	127	H	40	60
10	106	H	30	90	20	117	H	40	80

Introdução

Foi visto que a esperança condicional de Y dado que $X = x$, denotada por $E(Y|x)$ é uma função de x, ou seja:

$$E(Y|x) = \theta(x)$$

Consideramos que X e Y são definidas sobre a mesma população P. Por exemplo, X pode ser a idade e Y o tempo de reação a um estímulo.

Introdução

A análise do tempo de reação ao estímulo em função da idade sugere a existência de uma relação forte entre as duas variáveis. Percebemos que a média do tempo de reação aumenta conforme as pessoas envelhecem. Podemos sugerir um modelo do tipo:

$$\theta = E(Y|x) = \theta(x)$$

onde a relação pode ser uma função linear, que se escreve da seguinte forma:

$$E(Y|x) = \theta(x) = \beta_0 + \beta_1 x$$

(ou seja, o tempo de reação médio é uma função linear da idade).

Modelo de Regressão Linear

Figura 16.1: Gráfico de dispersão de idade e reação ao estímulo, com reta ajustada.

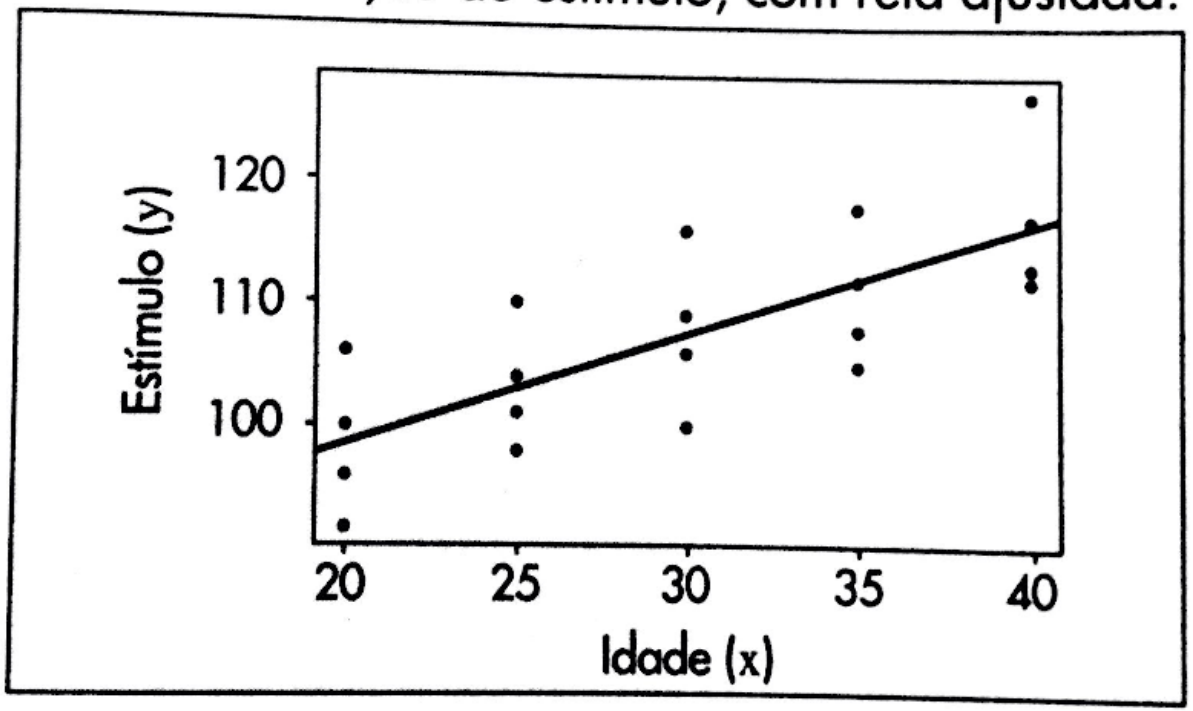


Figura: Dispersão de idade e reação ao estímulo

Introdução

Para o modelo, o tempo de reação da j-ésima pessoa do grupo de idade i, que era representado por:

$$y_{ij} = \theta_i + \epsilon_{ij}$$

passa a ter a representação:

$$y_{ij} = \theta(x_i) + \epsilon_{ij}$$

Entretanto, o modelo adotado será:

$$y_i = \theta_i + \epsilon_i$$

que indica o tempo de reação do i-ésimo indivíduo com x_i anos de idade.

Introdução

Estamos propondo um modelo de comportamento para as médias na população e não na amostra. Desta forma, o problema passa a ser o de estimar os parâmetros na função $\theta(x)$, baseandonos numa amostra de n observações.

No caso particular, o modelo procurado será :

$$y_i = E(\,Y|x_i) + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

devendo-se encontrar os valores mais prováveis para β_0 e β_1 a partir dos n pares (x, y).

Suposições

- 1. A variável auxiliar x é controlada e não sujeita a variações aleatórias.
- 2. Para um valor dado de x, os erros ϵ se distribuem em torno da média $\beta_0 + \beta_1 x$ ou seja:

$$E(\epsilon|x) = 0$$

- 3. Os erros mostram a mesma variabilidade em todos os níveis de x. Isto é, supomos que os dados são homocedásticos:

$$var(\epsilon|x) = \sigma_\epsilon^2$$

- 4. Os erros são não correlacionados.

Suposições

Colhida uma amostra de n indivíduos, teremos n pares (x_i, y_i), que devem satisfazer o modelo:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, 2, ..., n$$

Temos n equações e $n + 2$ incógnitas $(\beta_0, \beta_1, \epsilon_1, ..., \epsilon_n)$. Introduzimos então um novo critério envolvendo os valores de ϵ_i: O critério de mínimos quadrados.
Para cada observação, o desvio cometido pela adoção do modelo será:

$$\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$

A quantidade de informação perdida pelo modelo, é definida como a soma dos quadrados dos desvios:

$$SQ(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Método dos MQ

Derivando a expressão acima em relação a β_0 e β_1 e igualando a zero, teremos a solução (mostrada abaixo).
Substituindo os valores encontrados, teremos o estimador para as médias $\theta(x)$ dado por:

$$\hat{\theta}(x_i) = E(\hat{Y}|x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i = b_0 + b_1 x_i$$

Relação Linear

A forma mais simples de relação matemática deterministica entre duas variáveis x e y é uma relação Linear:

$$y = \beta_0 + \beta_1 x$$

o conjunto de pares (x, y) para os que $y = \beta_0 + \beta_1 x$ determina uma linha Reta.

Terminologia

A variável cujo valor é fixado, e denotado por x, é a variável independente (preditora, explicativa).

para um x fixado, a segunda variável será uma variável aleatória Y, cujo valor observado y, é referido como a variável dependente (resposta).

Terminologia

A variável cujo valor é fixado, e denotado por x, é a variável independente (preditora, explicativa).
para um x fixado, a segunda variável será uma variável aleatória Y, cujo valor observado y, é referido como a variável dependente (resposta).

Modelo de regressão Linear Simples

Existem parâmetros β_0, β_1 e σ^2 tais que para qualquer valor fixado de x, a variável dependente é relacionada com x através da equação modelo:

$$y = \beta_0 + \beta_1 x + \epsilon$$

onde ϵ é uma variável aleatória (chamada de desvio) com

$$E(\epsilon) = 0$$

$$Var(\epsilon) = \sigma^2$$

Modelo de Regressão Linear

Modelo de Regressão Linear

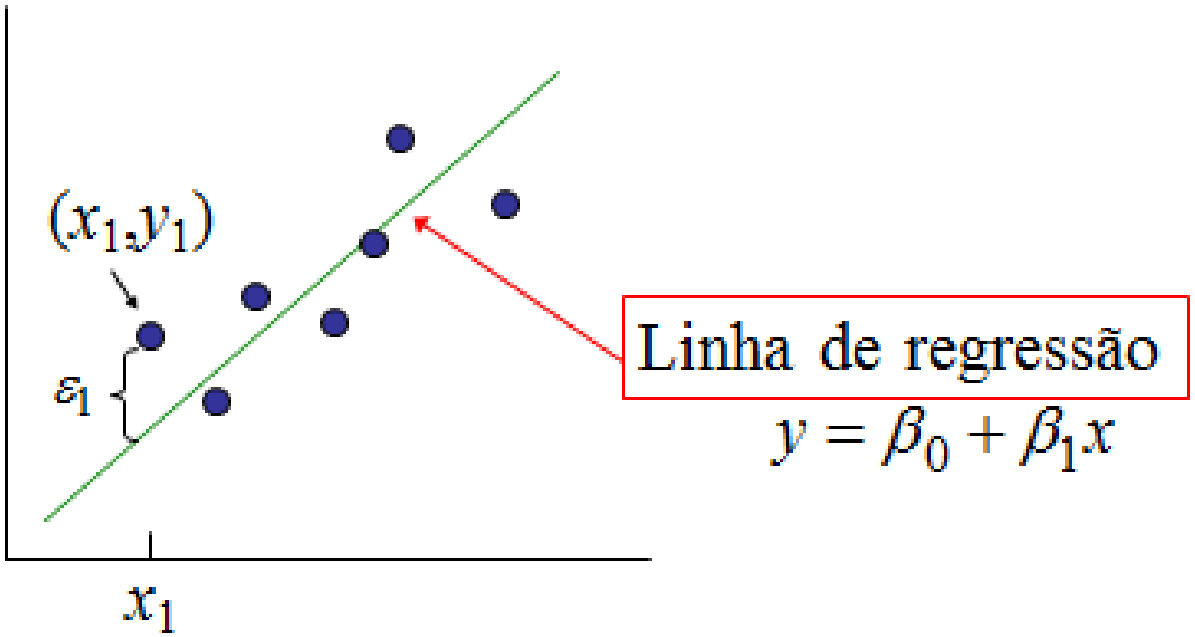


Figura: Linha de Regressão

Distribuição de ϵ

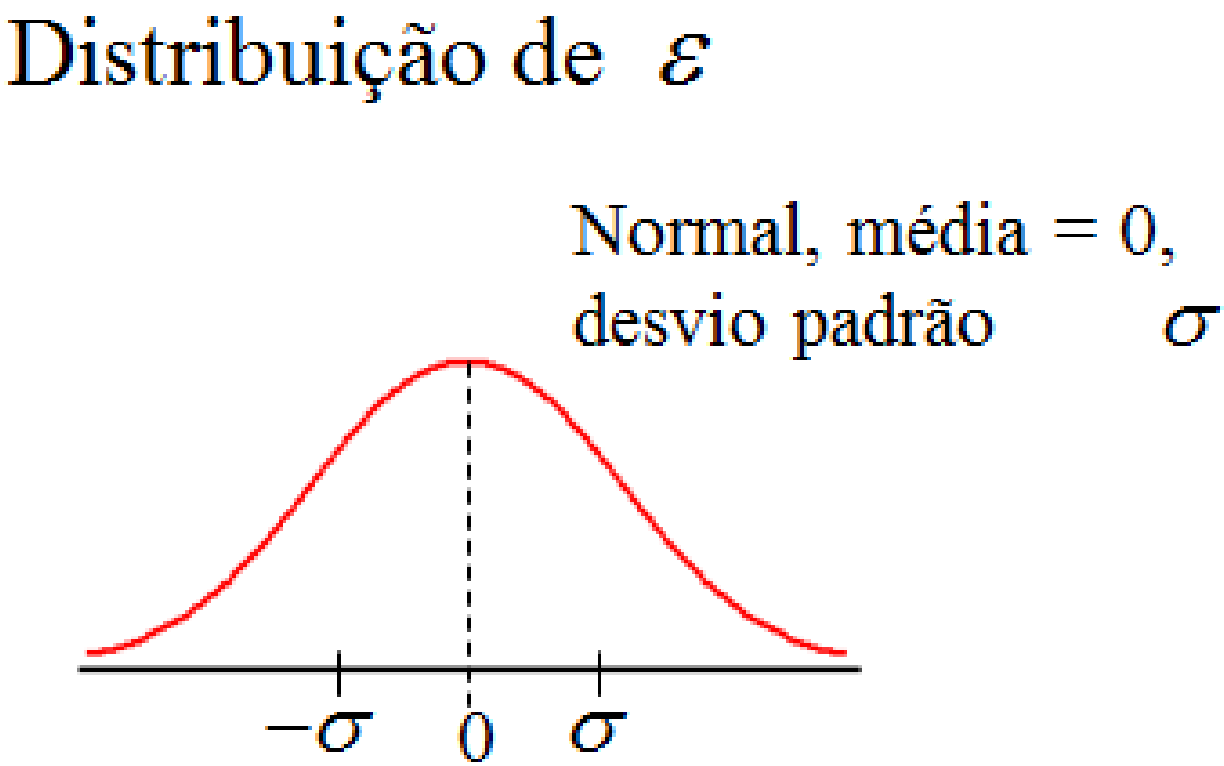


Figura: ϵ

Distribuição de Y

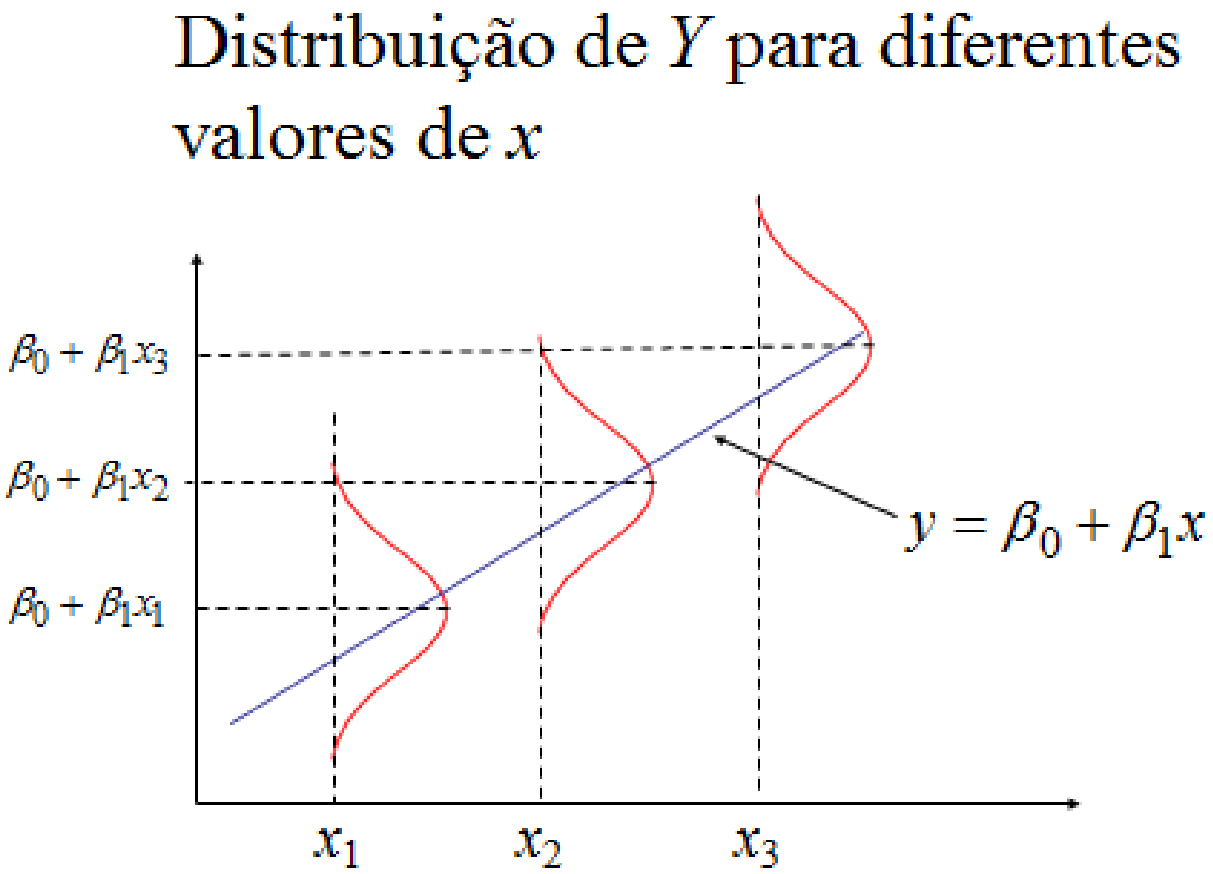


Figura: Y

O princípio dos Mínimos Quadrados

O desvio vertical do ponto (x_i, y_i) da reta $y = b_0 + b_1 x$ é

$$y_i - (b_0 + b_1 x_i)$$

a soma dos quadrados dos desvios desde os pontos $(x_1, y_1), ..., (x_n, y_n)$ à reta é :

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

O princípio dos Mínimos Quadrados

A reta de regressão de Mínimos quadrados para os dados é dada por

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

onde

$$b_1 = \hat{\beta}_1 = \frac{\sum x_i y_i - (\sum x_i)(\sum y_i)/n}{(\sum x_i^2) - (\sum x_i)^2/n}$$

e

$$b_0 = \hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1(\sum x_i)}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

Observações

Chamamos o modelo $= E(Y|x_i) = \beta_0 + \beta_1 x_i$ de linear, pois representa uma reta. Todavia, em casos mais gerais, o termo linear refere-se ao modo como os parâmetros entram no modelo, ou seja, de forma linear. Por exemplo, o modelo

$$E(Y|x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

embora gráficamente represente uma parábola, é linear em β_0, β_1 e β_2. Por outro lado,

$$E(Y|x) = \beta_0 e^{\beta_1 x}$$

não é linear em β_0 e β_1 .
Determinados modelos não lineares podem ser transformados em lineares por meio de transformações das variáveis. Por exemplo, ao modelo anterior aplicamos o logaritmo natural e obtemos:

$$ln(E(Y|x)) = ln(\beta_0) + \beta_1 x = \beta_0^{'} + \beta_1 x$$

que é linear em $\beta_0^{'}$ e β_1

exemplo

Encontre a equação de mínimos quadrados para os pares $(1, 2), (2, 3), (3, 7)$:

	x	y	xy	x^2
	1	2	2	1
	2	3	6	4
	3	7	21	9
soma	6	12	29	14

exemplo

$$\hat{\beta}_1 = \frac{3(29) - (6)(12)}{3(14) - (6)^2} = 2,5$$

$$\hat{\beta}_0 = \frac{12 - 2,5(6)}{3} = -1$$

portanto

$$y = -1 + 2,5x$$

Valores ajustados e residuais

Os valores ajustados (preditos) $\hat{y}_1, ..., \hat{y}_n$ são obtidos substituindo $x_1, ..., x_n$ na equação de regressão estimada :

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1$$
$$... = ...$$
$$\hat{y}_n = \hat{\beta}_0 + \hat{\beta}_1 x_n$$

Os residuos são os desvios verticais $y_1 - \hat{y}_1, ..., y_n - \hat{y}_n$ da reta estimada

Estimador de σ^2_ϵ

Para julgar a vantagem da adoção de um modelo mais complexo, usaremos a estratégia de compara-lo com o modelo mais simples, que é aquele visto na primeira seção:

$$y_i = \theta + \epsilon_i$$

A vantagem será medida pela redução dos erros de previsão, ou seja, da variância residual s^2_ϵ. Desta forma, para o modelo ajustado,

$$\hat{y}_i = b_0 + b_1 x_i$$

cada desvio é dado por:

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$$

Estimador de σ^2_ϵ

Quando estes desvios forem pequenos, o modelo está compensando. Para julgarmos se os desvios são pequenos ou não, o comparamos com os desvios do modelo alternativo, que serão: $y_i - \bar{y}$. Diante da dificuldade de compara-los um a um, trabalharemos com a soma dos desvios quadráticos, assim, teremos:

$$SQT = \sum (y_i - \bar{y})^2$$

e a Soma de quadrados dos erros:

$$SQRes = SQE = \sum (\hat{\epsilon}_i)^2 = \sum (y_i - \hat{y}_i)^2$$

A comparação dos dois resultados ajudará a verificar se houve ou não uma redução significativa nos resíduos

Exemplo

Se considerarmos o primeiro exemplo, e observamos o tempo de reação y_i em função da idade x_i tendo ϵ_i como desvio. Os resultados são:
$n = 20 \quad \sum y = 2150 \quad \sum x = 600 \quad \sum xy = 65400 \quad \bar{y} = 107,50$
$\bar{x} = 30 \quad \sum x^2 = 19000.$
Encontramos que

$$b_0 = \hat{\beta}_0 = 80,5 \quad e \quad b_1 = \hat{\beta}_1 = 0,9$$

que mostra o modelo ajustado:

$$\hat{y}_i = 80,5 + 0,9x_i$$

podemos prever, por exemplo, o tempo médio de reação para um indivíduo de 20 anos:

$$\hat{y}(20) = 80,5 + 0,9(20) = 98,5$$

Exemplo

Se quisermos estimar o tempo médio para um grupo de pessoas de 33 anos, teremos:

$$\hat{y}(33) = 80,5 + 0,9(33) = 110,2$$

A previsão pode ser melhorada se construirmos intervalos de Confiança. Os desvios poderiam ser vistos como:

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - (80,5 + 0,9x_i)$$

cuja soma elevada ao quadrado, será:

$$SQRes = SQE = 563$$

sabendo que

$$SQT = 1373$$

o que mostra uma redução de 810 unidades.

Exemplo

Entretanto, a comparação não parece justa, pois o segundo modelo contém mais parâmetros que o primeiro.
Para isto, comparamos as variâncias residuais.
Para o primeiro modelo, vimos que o estimador não viesado é:

$$s^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2 = \frac{SQT}{n-1}$$

vimos também que o estimador da variância residual era:

$$s_\epsilon^2 = \frac{SQDent}{n-k}$$

onde k é o número de grupos envolvidos (ou número de parâmetros do modelo). De um modo geral, perde-se um grau de liberdade para cada parâmetro envolvido no modelo.

Exemplo

Desse modo, é natural definir o estimador de σ_ϵ^2 por

$$s_\epsilon^2 = \frac{SQRes}{n-k}$$

no caso particular

$$s_\epsilon^2 = \frac{SQRes}{n-2} = \frac{SQE}{n-2}$$

sabemos que s_ϵ^2 é um estimador não viesado de σ_ϵ^2, isto é:

$$E(s_\epsilon^2) = \sigma_\epsilon^2$$

Exemplo

Do exemplo anterior, temos que

$$s^2 = \frac{1373}{19} = 72,26 \quad e \quad s = 8,5$$

$$s_\epsilon^2 = \frac{563}{18} = 31,28 \quad e \quad s_\epsilon = 5,59$$

Observe que perdendo um grau de liberdade, há redução de 813 unidades na soma de quadrados residual. O que implica uma vantagem em aplicar o segundo modelo.

Soma de quadrados do erro(SQE) ou (SSE)

A soma de quadrados do erro (SSE ou SQE) é :

$$SQE = \sum (y_i - \hat{y}_i)^2 = \sum \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right]^2$$

e a estimativa de σ^2 é

$$\hat{\sigma}_\epsilon^2 = s_\epsilon^2 = \frac{SQE}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

Fórmula para o cálculo de SQE e SQT

a fórmula para o cálculo de SSE é:

$$SQE = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$$

e para o cálculo da soma total de quadrados (SST ou SQT) é

$$SQT = S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

Decomposição da Soma de Quadrados

Ao passar de um modelo simples para o modelo de RL, a diminuição na soma de quadrados é dada por $SQT - SQRes(SQE)$, este lucro é devido à adoção do segundo modelo e será indicado por $SQReg(SQR)$:

$$SQR = SQT - SQE$$

ou

$$SQT = SQR + SQE$$

notamos que a relação:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) = \epsilon_i + (\hat{y}_i - \bar{y})$$

O desvio da observação em relação à média pode ser decomposto como o desvio da observação em relação ao valor ajustado pela regressão mais o desvio do valor ajustado em relação à média.

Decomposição da Soma de Quadrados

Elevando a quadrado ambos os membros, somando e observando que a soma do duplo produto se anula, obtemos:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum \hat{\epsilon}_i^2$$

$$SQT = \sum (\hat{y}_i - \bar{y})^2 + SQE$$

de onde

$$SQR = \sum (\hat{y}_i - \bar{y})^2$$

Decomposição da Soma de Quadrados

Da resolução pelo método MQ temos que

$$b_0 = \bar{y} - b_1 \bar{x}$$

e para o modelo

$$\hat{y}_i = b_0 + b_1 x_i$$

$$\hat{y}_i = \bar{y} - b_1 \bar{x} + b_1 x_i = \bar{y} + b_1 (x_i - \bar{x})$$

desse resultado temos:

$$\hat{y}_i - \bar{y} = \bar{y} + b_1 (x_i - \bar{x}) - \bar{y} = b_1 (x_i - \bar{x})$$

$$SQReg = SQR = b_1^2 \sum (x_i - \bar{x})^2$$

de onde observamos que quanto maior o valor de b_1, maior será a diminuição da soma dos quadrados dos resíduos.

Coeficiente de determinação

O coeficiente de determinação (ou o lucro relativo introduzido pelo modelo) é denotado por r^2 e é dado por

$$r^2 = 1 - \frac{SQE}{SQT} = \frac{SQR}{SQT}$$

é interpretado como a proporção da variação observada em y que pode ser explicado pelo modelo de regressão linear simples.

Soma de quadrados da regressão (SQReg ou SQR)

$$SQR = SQT - SQE$$

A soma de quadrados da regressão é interpretado como a quantidade de variação que é explicado pelo modelo. Temos que

$$r^2 = \frac{SQR}{SQT}$$

ANOVA

FV	gl	SQ	QM	F
Regressão	1	SQR	QMR	$\frac{QMR}{QME}$
Erro	$n-2$	SQE	$s^2_\epsilon = \frac{SQE}{n-2} = QME$	
Total	$n-1$	SQT	$s^2 = \frac{SQT}{n-1}$	

ANOVA para o exemplo

modelo: $\hat{y}_i = 80,5 + 0,9x_i$

FV	gl	SQ	QM	F
Regressão	1	810	810	25,9
Erro	18	563	31,28	
Total	19	1373	72,26	

$$R^2 = \frac{810}{1373} = 59\%$$

ANOVA para o exemplo

O modelo $\hat{y}_i = 80,5 + 0,9x_i$ diminui a variância residual em mais da metade e explica 59% da variabilidade total. (adicionar a idade melhora a explicação do fenômeno).
A estratégia para verificar se compensa ou não utilizar o modelo $y = \beta_0 + \beta_1 x + \epsilon$, consiste em verificar a diminuição do resíduo quando comparado com o do modelo $y = \mu + \epsilon$.
Se a redução for pequena, os modelos são equivalentes, isto ocorre quando a inclinação β_1 for zero ou muito pequena.
Estatísticamente equivale a testar $H_0 : \beta_1 = 0$, o que irá exigir uma estrutura de probabilidade sobre os erros.

Introdução

Queremos estudar as propriedades dos estimadores $b_0 = \hat{\beta}_0$ e $b_1 = \hat{\beta}_1$ para isso voltamos às suposições para a variável Y. x é uma variável auxiliar controlada e não sujeita a flutuações aleatórias, mas que estabelece uma relação com a média de Y.
Assim, para um dado nível de x, teriamos associada uma certa distribuição de probabilidade para Y como ilustrada abaixo.
A função que liga as médias de Y com x é linea, e mais ainda, as distribuições para nível de x têm a mesma dispersão (figura abaixo). Em nosso caso, introduziremos a condição adicional, de que estas distribuições sejam todas normais (figura).

Introdução

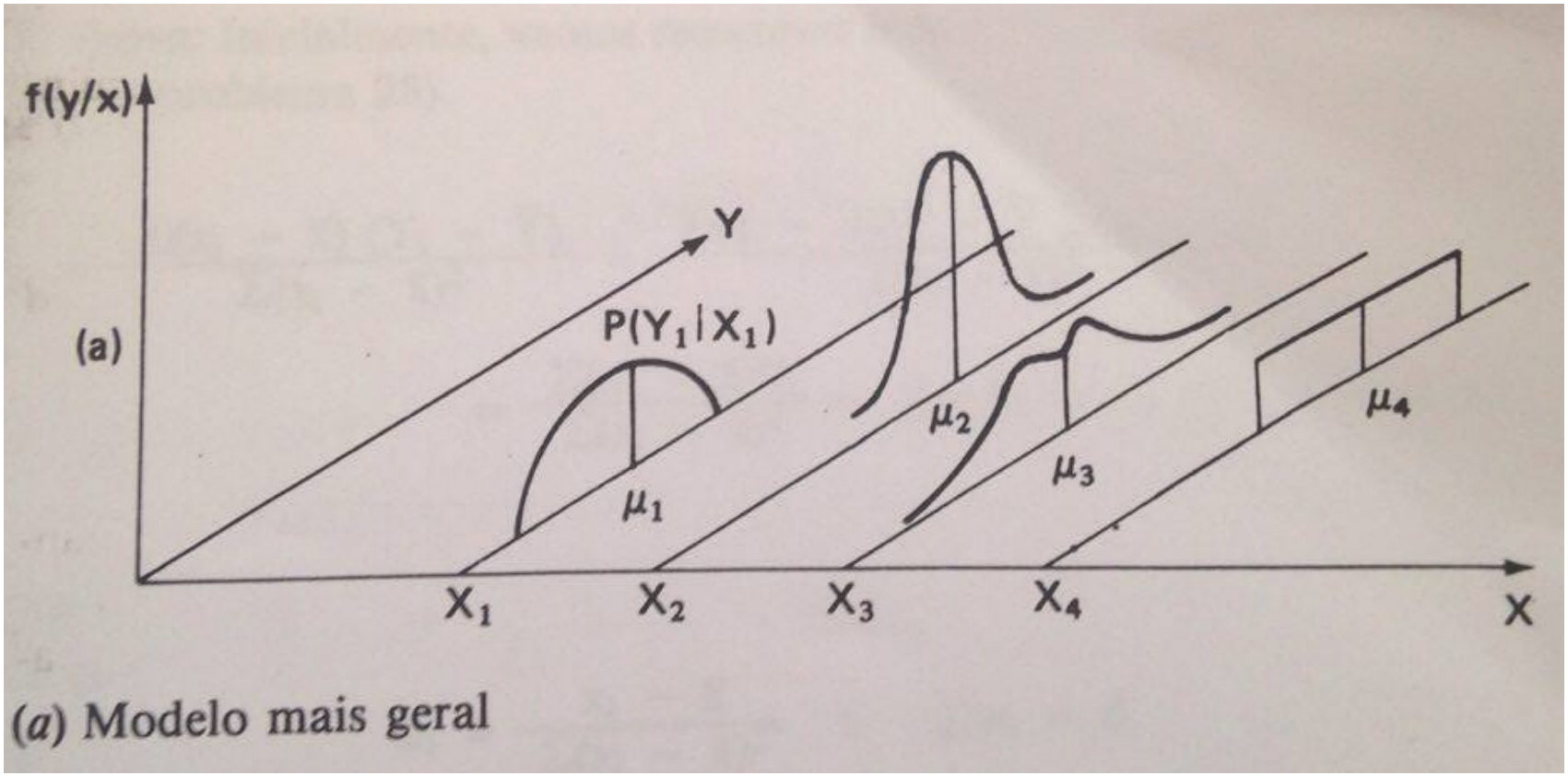
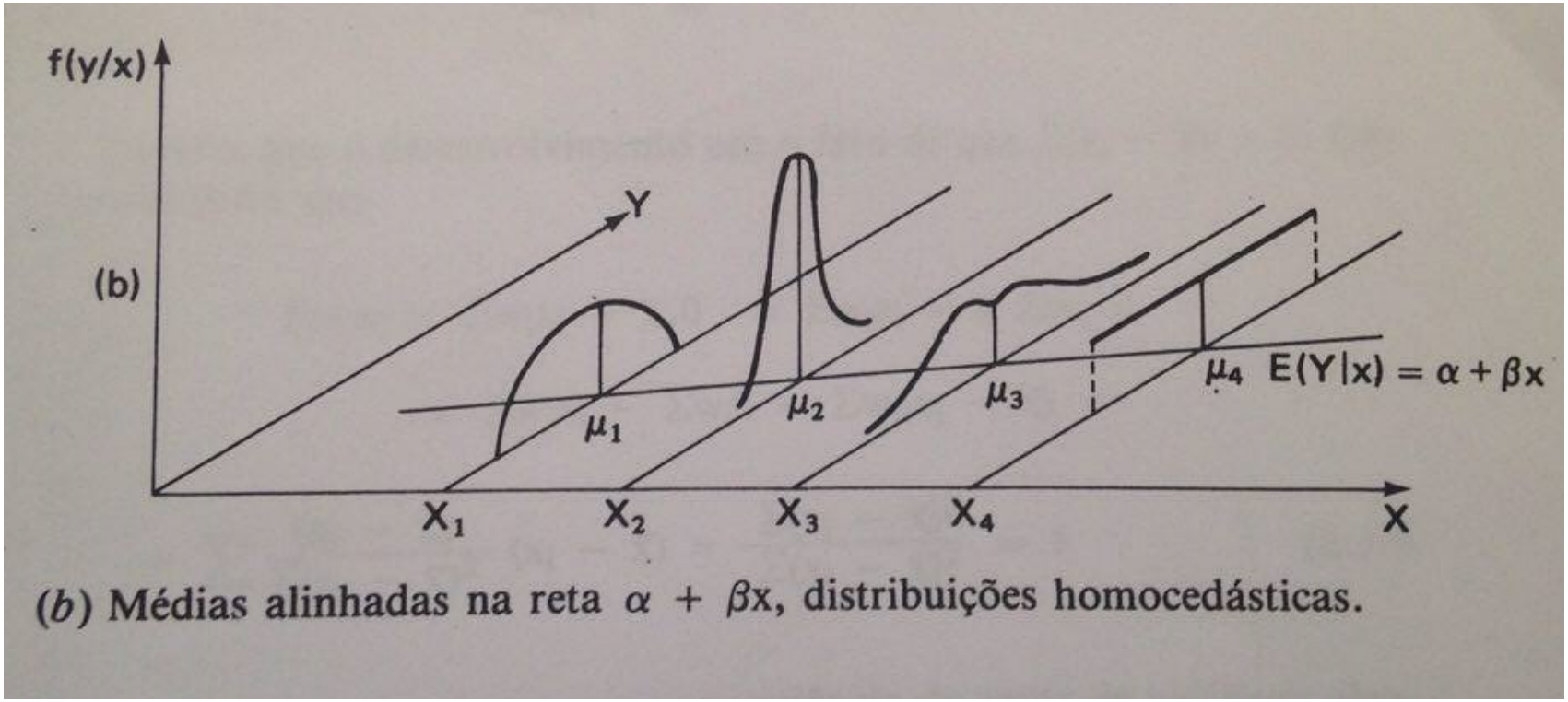


Figura: Modelo mais geral

Introdução



(*b*) Médias alinhadas na reta $\alpha + \beta x$, distribuições homocedásticas.

Figura: Médias alinhadas na reta $\beta_0 + \beta_1 x$, com distribuições homocedásticas

Introdução

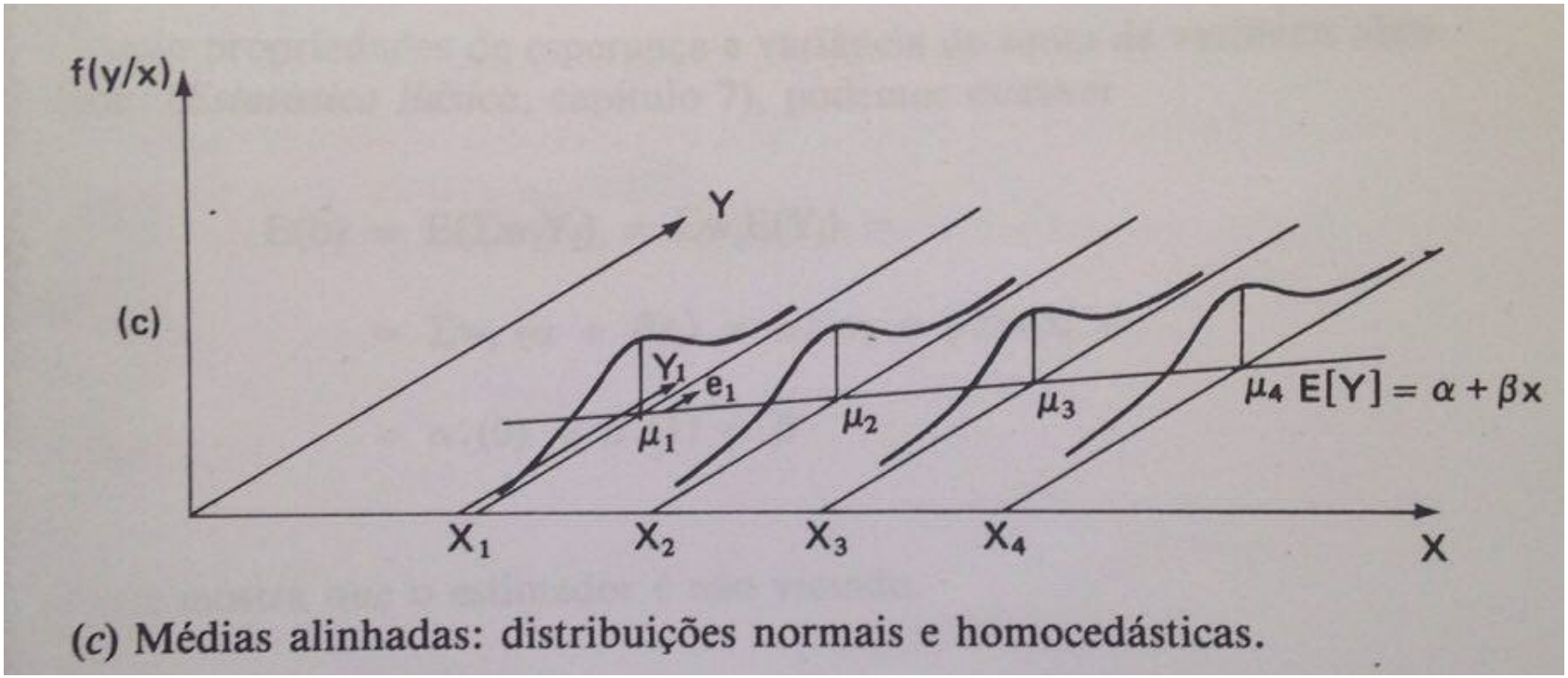


Figura: Médias alinhadas, distribuições normais e homocedásticas

$\hat{\beta}_1$

Pode-se mostrar que

$$E(\hat{\beta}_1) = E(b_1) = \beta_1 \quad \text{não viesado}$$

e

$$var(\hat{\beta}_1) = var(b_1) = \frac{\sigma_\epsilon^2}{\sum(x_i - \bar{x})^2}$$

Prova: Podemos escrever b_1 de um modo conveniente, considerando Y uma v.a e x não.

$$b_1 = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})(Y_i) - (\bar{Y})\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$$

$$= \frac{\sum(x_i - \bar{x})\, Y_i}{\sum(x_i - \bar{x})^2} = \sum w_i\, Y_i$$

onde

$$w_i = \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2} \quad e \quad \sum w_i = 0$$

$\hat{\beta}_1$

temos que:

$$\begin{aligned}\sum w_i x_i = & \sum w_i x_i - \bar{x}.0 = \sum w_i x_i - \bar{x}\sum w_i \\ = & \sum w_i x_i - \sum w_i \bar{x} = \sum w_i (x_i - \bar{x}) \\ = & \sum \frac{(x_i-\bar{x})}{\sum (x_i-\bar{x})^2}(x_i - \bar{x}) = \frac{\sum (x_i-\bar{x})^2}{\sum (x_i-\bar{x})^2} = 1\end{aligned}$$

$$\begin{aligned}E(b_1) = & E(\sum w_i Y_i) = \sum w_i E(Y_i) = \\ = & \sum w_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum w_i + \beta_1 \sum w_i x_i = \\ = & \beta_0(0) + \beta_1(1) = \beta_1\end{aligned}$$

$\hat{\beta}_1$

Por outro lado, como as observações são não correlacionadas

$$var(b_1) = var(\sum w_i Y_i) = \sum w_i^2 var(Y_i)$$

e

$$var(b_1) = \sum w_i^2 \sigma_\epsilon^2 = \sigma_\epsilon^2 \sum \left(\frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2} \right)^2 = \sigma_\epsilon^2 \frac{1}{\sum (x_i - \bar{x})^2}$$

$\hat{\beta}_1$

- 1. A média de $\hat{\beta}_1$ é

$$E(\hat{\beta}_1) = \mu_{\hat{\beta}_1} = \beta_1$$

- 2. A variância e o desvio padrão de $\hat{\beta}_1$ são:

$$Var(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_\epsilon}{S_{xx}}$$

$$\sigma_{\hat{\beta}_1} = \frac{\sigma_\epsilon}{\sqrt{S_{xx}}}$$

$$S_{\hat{\beta}_1} = \frac{s_\epsilon}{\sqrt{S_{xx}}}$$

- 3. $\hat{\beta}_1$ tem distribuição normal.

$\hat{\beta}_0$

$$E(b_0) = \hat{\beta}_0 = \beta_0$$

$$var(b_0) = \sigma^2 \frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}$$

Prova:

mostra-se que

$$cov(\bar{Y}, b_1) = 0 \quad \text{problema 31}$$

e

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$$

e também:

$$\bar{Y} = \frac{1}{n} \sum Y_i = \frac{1}{n} \sum (\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 \bar{x} + \frac{1}{n} \sum \epsilon_i$$

$\hat{\beta}_0$

$$E(\bar{Y}) = \beta_0 + \beta_1 \bar{x} + \frac{1}{n} \sum E(\epsilon_i) = \beta_0 + \beta_1 \bar{x}$$

$$var(\bar{Y}) = \frac{1}{n^2} \sum var(\epsilon_i) = \frac{n\sigma_\epsilon^2}{n^2} = \frac{\sigma_\epsilon^2}{n}$$

Então:

$$E(b_0) = E(\bar{Y} - b_1 \bar{x}) = E(\bar{Y}) - \bar{x}E(b_1) = \beta_0 + \beta_1 \bar{x} - \bar{x}\beta_1 = \beta_0$$

$\hat{\beta}_0$

$$
\begin{aligned}
var(b_0) = \quad & var(\bar{Y} - b_1\bar{x}) = var(\bar{Y}) + var(b_1\bar{x}) - 2Cov(\bar{Y}, b_1\bar{x}) \\
= \quad & var(\bar{Y}) + \bar{x}^2 var(b_1) - 2\bar{x}Cov(\bar{Y}, b_1) \\
= \quad & \frac{\sigma_\epsilon^2}{n} + \bar{x}^2 \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2} - 2\bar{x}(0) = \\
= \quad & \sigma_\epsilon^2 \{ \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \} = \sigma_\epsilon^2 \frac{\sum (x_i - \bar{x})^2 + n\bar{x}^2}{n \sum (x_i - \bar{x})^2} = \\
= \quad & \sigma_\epsilon^2 \frac{\sum x_i^2 - n\bar{x}^2 + n\bar{x}^2}{n \sum (x_i - \bar{x})^2} = \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}
\end{aligned}
$$

Distribuições amostrais de b_0 e b_1

Sabemos que

$$\epsilon_i \sim N(0, \sigma_\epsilon^2)$$

o que implica que

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma_\epsilon^2)$$

como b_0 e b_1 são combinações lineares normais independentes, teremos:

$$b_0 \sim N\left(\beta_0, \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}\right)$$

$$b_1 \sim N\left(\beta_1, \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2}\right)$$

Distribuições amostrais de b_0 e b_1

podemor concluir que:

$$\frac{b_1 - \beta_1}{\sigma_\epsilon} \cdot \sqrt{\sum (x_i - \bar{x})^2} \sim N(0,1)$$

$$\frac{b_0 - \beta_0}{\sigma_\epsilon} \cdot \sqrt{\frac{n \sum (x_i - \bar{x})^2}{\sum x_i^2}} \sim N(0,1)$$

substituindo σ_ϵ por seu estimador s_ϵ, a distribuição resultante será t de student com $n-2$ graus de liberdade

Distribuições amostrais de b_0 e b_1

as estatísticas:

$$t_{b_1} = \frac{b_1 - \beta_1}{s_\epsilon} \cdot \sqrt{\sum (x_i - \bar{x})^2} = \frac{\hat{\beta}_1 - \beta_1}{s_\epsilon} \cdot \sqrt{\sum (x_i - \bar{x})^2} \sim t_{n-2}$$

e

$$t_{b_0} = \frac{b_0 - \beta_0}{s_\epsilon} \cdot \sqrt{\frac{n \sum (x_i - \bar{x})^2}{\sum x_i^2}} = \frac{\hat{\beta}_0 - \beta_0}{s_\epsilon} \cdot \sqrt{\frac{n \sum (x_i - \bar{x})^2}{\sum x_i^2}} \sim t_{n-2}$$

Variável T

As suposições do modelo de Regressão linear simples implicam que a variável padronizada:

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_\epsilon / \sqrt{S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t(n-2)$$

Intervalo de Confiança

Intervalos de confiança de $100(1-\alpha)\%$ para $\hat{\beta}_0$ e $\hat{\beta}_1$ da linha de regressão são:

$$IC(\beta_0, 1-\alpha) = b_0 \pm t_{\alpha/2,(n-2)} * s_\epsilon * \sqrt{\frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}}$$

e para β_1:

$$\begin{aligned} IC(\beta_1, 1-\alpha) = & \quad b_1 \pm t_{\alpha/2,(n-2)} * s_\epsilon * \sqrt{\frac{1}{\sum(x_i-\bar{x})^2}} \\ = & \quad \hat{\beta}_1 \pm t_{\alpha/2,n-2}.S_{\hat{\beta}_1} \end{aligned}$$

Exemplo

Encontramos no exemplo inicial que:

$$\sum x_i^2 = 19000 \quad \sum (x_i - \bar{x})^2 = 1000 \quad \bar{x} = 30$$

Da ANOVA temos:

$$s_\epsilon^2 = 31,28 \quad s_\epsilon = 5,59 \quad t_{5\%,18} = 2,101$$

$$IC(\beta_0, 95\%) = 80,5 \pm 2,101 * 5,59 * \sqrt{\frac{19000}{1000 * 20}} = 80,5 \pm 11,45$$

$$IC(\beta_1, 95\%) = 0,9 \pm 2,101 * 5,59 * \sqrt{\frac{1}{1000}} = 0,9 \pm 0,3$$

Este último resultado reforça a idéia de que $\beta_1 \neq 0$

Procedimentos de testes de hipóteses

A hipótese:

$$H_0 : \beta_0 = \beta_{01}$$

ou a hipótese

$$H_0 : \beta_0 = 0$$

é testada por

$$t_{b_0} = \frac{b_0}{s_\epsilon}\sqrt{\frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2}}$$

Procedimentos de testes de hipóteses

Hipótese Nula:

$$H_0: \quad \beta_1 = \beta_{10}$$

Estatística de teste:

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$$

para testar $H_0 : \beta_1 = 0$

$$t_{b_1} = \frac{b_1}{s_\epsilon}\sqrt{\sum(x_i - \bar{x})^2}$$

cada uma com $t_{\alpha/2,n-2}$ graus de liberdade

Procedimentos de testes de hipóteses

Observe que

$$t_{b_1}^2 = \frac{b_1^2 \sum (x_i - \bar{x})^2}{s_\epsilon^2}$$

que resulta em :

$$t_{b_1} = \frac{SQR}{s_\epsilon^2} \sim F_{1,n-2}$$

Desta maneira, para testar $\beta_1 = 0$, pode-se usar $t_{b_1}^2 \sim F_{1,n.2}$

Exemplo

Para testar $\beta_0 = 0$ e $\beta_1 = 0$ as estatísticas serão:

$$t_{b_0} = \frac{80,5}{5,59}\sqrt{\frac{20 * 1000}{19000}} = 14,77$$

$$t_{b_1} = \frac{0,9}{5,59}\sqrt{1000} = 5,09$$

os quais comparados com o valor crítico de t ao nível 5% e com 18 g.l, que é $2,101$ nos levam à rejeição de que os parâmetros sejam iguais a zero. Também constata-se que $t^2_{b_1} = 25,9 = F$ é maior que o valor correspondente na tabela F.

Procedimentos de testes de hipóteses

Hipótese Alternativa	região de rejeição
$H_a: \beta_1 > \beta_{10}$	$t \geq t_{\alpha,n-2}$
$H_a: \beta_1 < \beta_{10}$	$t \leq -t_{\alpha,n-2}$
$H_a: \beta_1 \neq \beta_{10}$	$t \geq t_{\alpha/2,n-2}$ ou $t \leq -t_{\alpha/2,n-2}$

teste de hipóteses

O modelo teste de utilidade é o teste de

$$H_0: \quad \beta_1 = 0 \qquad\qquad H_a: \quad \beta_1 \neq 0$$

e o teste estatístico é a razão :

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

Introdução

O modelo linear pode ser utilizado tanto para fazer previsões da variável resposta para algúm nível desejado ou para estimar o tempo médio para um grupo.

Por exemplo, qual o tempo de reação aos 28 anos?
É importante saber se queremos estimar o tempo médio para o grupo de 28 anos ou para uma pessoa de 28 anos.
A estimativa pontual será a mesma, o que irá mudar será o intervalo de Confiança correspondente.

Introdução

O modelo definido é:

$$E(Y|x) = \beta_0 + \beta_1 x$$

que é estimado por

$$\hat{y} = b_0 + b_1 x = \bar{y} + b_1(x - \bar{x})$$

Um resultado obtido mostra que

$$E(\hat{y}|x) = \beta_0 + \beta_1 x = E(Y|x) = \mu(x)$$

$$var(\hat{y}|x) = \sigma_\epsilon^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

$\hat{Y}$

podemos escrever:

$$E(\hat{y}|x) = E(b_0) + xE(b_1) = \beta_0 + \beta_1 x$$

temos que:

$$\begin{aligned} var(\hat{y}|x) = & \quad var(\hat{Y}) + (x - \bar{x})^2 var(b_1) + 2(x - \bar{x}) Cov(\bar{Y}, b_1) \\ = & \quad \frac{\sigma_\epsilon^2}{n} + (x - \bar{x})^2 \frac{\sigma_\epsilon^2}{\sum (x - \bar{x})^2} = \sigma_\epsilon^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2} \right] \end{aligned}$$

$\hat{Y}$

Seja $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ onde x^* é um valor fixado de x.

1. A média de $\hat{Y}$ é

$$E(\hat{Y}) = \mu_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \beta_0 + \beta_1 x^*$$

2. Variância e Desvio padrão:

$$var(\hat{Y}) = \sigma^2_{\hat{Y}} = \sigma^2_{\epsilon} \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]$$

$$S_{\hat{Y}} = s_{\epsilon} * \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

3. $\hat{Y} \sim N()$

$\hat{Y}$

Um estimador não viesado para $\mu(x) = E(\,Y|x)$ é

$$\hat{y} = b_0 + b_1 x$$

e o IC é

$$\begin{aligned} IC(\mu(x), 1-\alpha) = & \quad \hat{y} \pm t_{\alpha/2,n-2}\sqrt{var(\hat{y}|x)} = \\ = & \quad \hat{y} \pm t_{\alpha/2,n-2} s_\epsilon \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x-\bar{x})^2}} \end{aligned}$$

Variável T

A variável

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*)}{S_{\hat{\beta}_0 + \hat{\beta}_1 x^*}} = \frac{\hat{Y} - (\beta_0 + \beta_1 x^*)}{S_{\hat{Y}}} \sim t_{n-2}$$

Intervalo de Confiança

Um IC $100(1-\alpha)\%$ para $\mu_{Y.x^*}$ ou valor esperado de Y quando $x = x^*$, é :

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} * S_{\hat{\beta}_0+\hat{\beta}_1 x^*} = \hat{y} \pm t_{\alpha/2,n-2} * S_{\hat{Y}}$$

Exemplo

Qual é a estimativa do tempo médio de reação para o grupo de 28 anos?

$$\hat{\mu}(28) = \hat{y} = 80,5 + 0,9(28) = 105,7$$

$$\begin{aligned} IC(\mu(28), 95\%) = & \quad 105,7 \pm 2,101 * 5,59 * \sqrt{\frac{1}{20} + \frac{(28-30)^2}{1000}} \\ = & \quad 105,7 \pm 2,73 = \\ = & \quad]102,98; 108,43[\end{aligned}$$

Intervalo de Predição

Vejamos agora como construir um IC para uma futura observação Y a um dado nível de x. Ela será estimada pela reta:

$$\hat{y} = b_0 + b_1 x = \bar{Y} + b(x - \bar{x})$$

e o desvio padrão será dado por $Y - \hat{y}$, cujas propriedades são:

$$E(Y - \hat{y}) = 0$$

e

$$var(Y - \hat{y}) = Var(Y) + var(\hat{y}) = \sigma_\epsilon^2 + \frac{\sigma_\epsilon^2}{n} + \sigma_\epsilon^2 \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}$$

podemos escrever o IC correspondente por :

$$IC(Y(x), 1 - \alpha) = \hat{y} \pm t_{\alpha/2, n-2} * s_\epsilon * \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

Intervalo de predição (IP)

Um valor futuro de Y não é um parâmetro senão uma variável aleatória, seu intervalo de valores plausíveis é referido como ***intervalo de predição***. Um $100(1-\alpha)\%$ IP para uma observação futura de Y quando $x = x^*$ é

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} * s_\epsilon * \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = \hat{y} \pm t_{\alpha/2,n-2} * \sqrt{s_\epsilon^2 + s_{\hat{Y}}^2}$$

Exemplo

Qual o tempo de reação esperado para uma pessoa de 28 anos que irá submeter-se ao teste?

$$\hat{y} = 80,5 + 0,9 * 28$$

o IC será:

$$\begin{aligned} IC(Y(28), 95\%) = & \quad 105,7 \pm 2,101 * 5,59 * \sqrt{1 + \frac{1}{20} + \frac{(28-30)^2}{1000}} \\ = & \quad 105,7 \pm 12,06 \\ = & \quad]93,64; 117,76[\end{aligned}$$

Este intervalo é bem maior do que aquele para a média

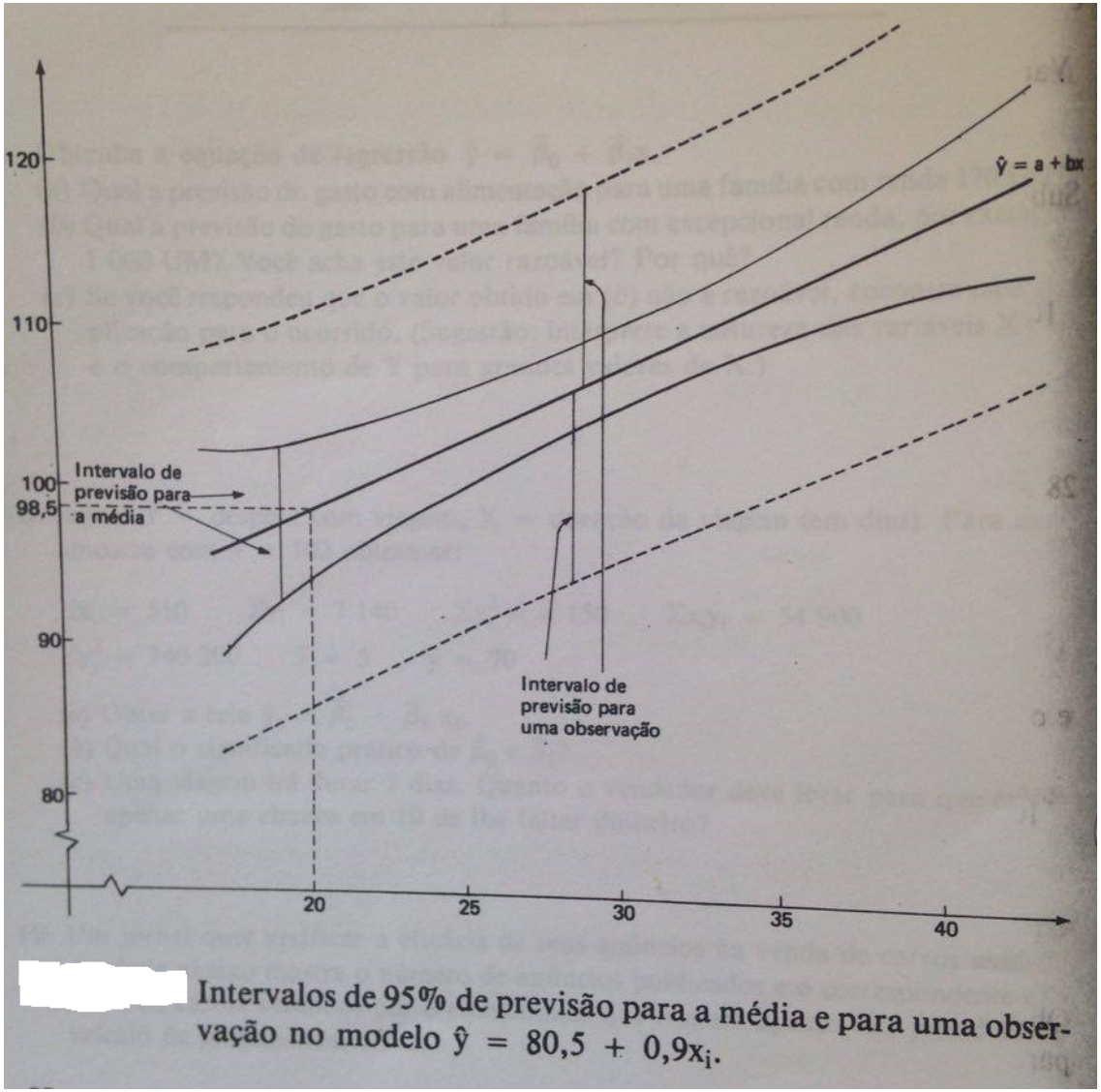
Exemplo

Podemos desenhar os limites superiores e inferiores, tanto do IC para a média como para uma futura observação, para cada x, obtendo uma região de confiança para a reta ajustada $\hat{y} = b_0 + b_1 x_i$.
Por exemplo, para $x = 30$ e $x = 40$, temos os seguintes ICt's para a média:

$$IC(\mu(30); 95\%) = 107,5 \pm 2,63$$

$$IC(\mu(40); 95\%) = 116,5 \pm 4,55$$

Exemplo



Intervalos de 95% de previsão para a média e para uma observação no modelo $\hat{y} = 80{,}5 + 0{,}9x_i$.

Correlação amostral

O coeficiente de correlação amostral r de n pares $(x_1, y_1), ..., (x_n, y_n)$ é:

$$r = \frac{S_{xy}}{\sqrt{\sum(x_i - \bar{x})^2} * \sqrt{\sum(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}} * \sqrt{S_{yy}}}$$

onde

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n$$

Exemplo

Encontre o coeficiente de correlação para a linha de mínimos quadrados dos pontos: $(1,2)(2,3)(3,7)$.

$$r = \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{n(\sum x^2)-(\sum)^2} * \sqrt{n(\sum y^2)-(\sum y)^2}}$$

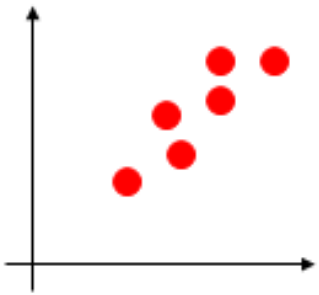
$$= \frac{3(29)-6(12)}{\sqrt{3(14)-(6)^2} * \sqrt{3(62)-(12)^2}} = 0,9449$$

Propriedades de r

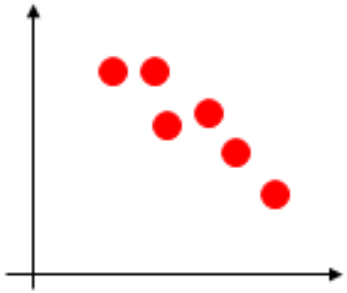
1. O valor de r não depende de qual variável foi rotulada com x ou y
2. o valor de r independe das unidades medidas para x ey
3. $-1 \leq r \leq 1$
4. $r = 1$ ssi todos os pares (x_i, y_i) estão em linha reta com inclinação positiva. $r = -1$ quando todos os pares estão em linha reta com inclinação negativa.
5. O quadrado do coeficiente de correlação amostral dá o valor do coeficiente de determinação que resultaria de ajustar o modelo de regressão linear

r

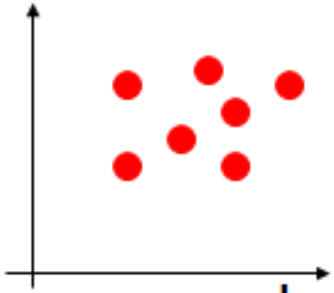
Valores diferentes para r



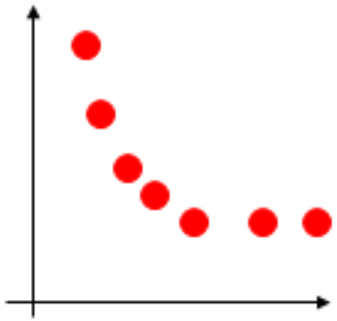
r prox. de 1



r prox. de -1



r prox. de 0,
sem relação



r prox. de 0,
relação não linear

Populacional

$$\rho = \rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X * \sigma_Y}$$

onde:

$$Cov(X, Y) = \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y) \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y) dxdy \end{cases}$$

dependendo se (X, Y) são discretos ou contínuos

Estimador

$$\hat{\rho} = R = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} * \sqrt{\sum(y_i - \bar{y})^2}}$$

Suposição

Para testar hipóteses sobre ρ, devemos fazer uma suposição análoga sobre a distribuição de pares de valores (x, y) na população. No momento, supomos que tanto X como Y são aleatórios, embora grande parte do trabalho concentre-se no x fixado pelo pesquisador.
A distribuição de probabilidade conjunta de (X, Y) é especificado por:

$$f(x,y) = \frac{e^{-[((x-\mu_1)/\sigma_1)^2 - 2\rho(x-\mu_1)(y-\mu_2)/\sigma_1\sigma_2 + ((y-\mu_2)^2/\sigma_2)^2]/[2(1-\rho^2)]}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

para $-\infty < x < \infty$, $-\infty < y < \infty$
$f(x, y)$ é chamada de distribuição de probabilidade normal bivariada.

Suposição

a superficie de $f(x, y)$ está acima do plano x, y e tem a aparência de um sino ou montanha. Se fatiarmos a superfície, cortando-a em qualquer plano perpendicular ao plano x, y e observarmos a forma da curva traçada no $plano$ de $corte$, o resultado será uma curva normal.
Se $X = x$, é possível mostrar que

$$f(\,Y|x) \sim N\left(\mu_{\,Y|x} = \mu_2 - \rho\mu_1\sigma_2/\sigma_1; (1-\rho^2)\sigma_2^2\right)$$

Esse é exatamente o modelo usado na RLS com $\beta_0 = \mu_2 - \rho\mu_1\sigma_2/\sigma_1$, $\beta_1 = \rho\sigma_2/\sigma_1$ e $\sigma^2 = (1-\rho^2)\sigma_2^2$ independentes de x.

Suposição

Se os pares (x_i, y_i) forem selecionados de uma distribuição normal bivariada, então o MRLS é uma forma apropriada de estudar o comportamento de Y para um x fixo.
Se $\rho = 0$, então $\mu_{Y|x} = \mu_2$ e independente de x. Isto é, quando $\rho = 0$,

$$f(x, y) = f(x)f(y)$$

Normal bivariada

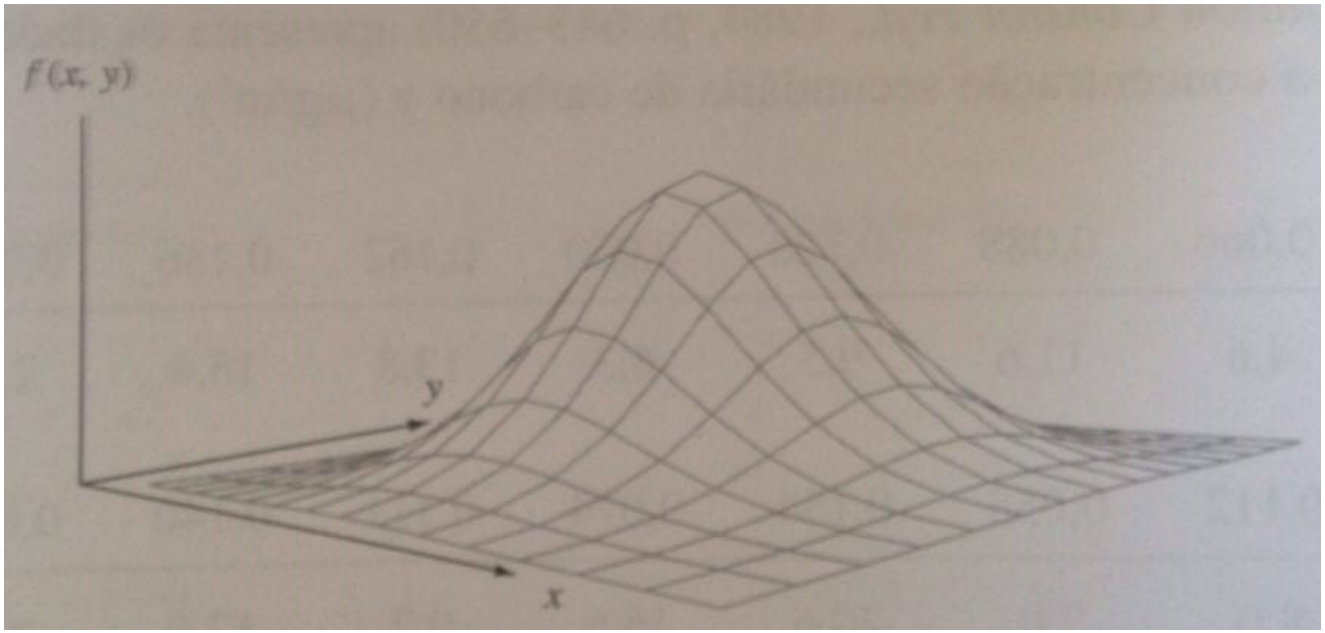


Figura: f.d.p. normal bivariada

Teste para ausência de correlação

Quando $H_0 : \rho = 0$ é verdadeiro, a estatística do teste

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t_{n-2}$$

Testes de hipóteses

Hipótese alternativa	Região de Rejeição para um nível α
$H_a : \rho > 0$	$t \geq t_{\alpha,n-2}$
$H_a : \rho < 0$	$t \leq -t_{\alpha,n-2}$
$H_a : \rho \neq 0$	$t \geq t_{\alpha,n-2}$ ou $t \leq -t_{\alpha,n-2}$

um p-valor baseado em $n-2$ graus de liberdade pode ser calculado como descrito previamente.

Outras inferências sobre ρ

O procedimento para testar $H_0 : \rho = \rho_0$ quando $\rho_0 \neq 0$ não é equivalente a qualquer procedimento da análise de regressão. A estatística baseia-se em uma transformação de R denominada transformação de FISHER. Quando $(X_1, Y_1), ..., (X_n, Y_n)$ é uma amostra de uma distribuição normal bivariada, a variável aleatória:

$$V = \frac{1}{2} ln \left(\frac{1+R}{1-R} \right)$$

tem distribuição normal aproximada com

$$\mu_V = \frac{1}{2} ln \left(\frac{1+\rho}{1-\rho} \right)$$

e

$$\sigma_V^2 = \frac{1}{n-3}$$

Teste estatístico

O teste para $H_0 : \rho = \rho_0$ é:

$$Z = \frac{V - \frac{1}{2} ln[\frac{1+\rho_0}{1-\rho_0}]}{\frac{1}{\sqrt{n-3}}}$$

Hipótese alternativa	Região de Rejeição para um nível α
$H_a : \rho > \rho_0$	$z \geq z_\alpha$
$H_a : \rho < \rho_0$	$z \leq -z_\alpha$
$H_a : \rho \neq \rho_0$	$z \geq z_\alpha$ ou $z \leq -z_\alpha$

IC para ρ

Para obter um IC para ρ derivamos primeiro um intervalo para

$$\mu_v = \frac{1}{2} ln \left[\frac{1+\rho}{1-\rho} \right]$$

Um IC $100(1-\alpha)\%$ para ρ é:

$$\left(\frac{e^{2c_1}-1}{e^{2c_1}+1}, \frac{e^{2c_2}-1}{e^{2c_2}+1} \right)$$

onde c_1 e c_2 são os limites esquerdo e direito.
O IC para μ_V é:

$$\left(v - \frac{z_{\alpha/2}}{\sqrt{n-3}}, v + \frac{z_{\alpha/2}}{\sqrt{n-3}} \right)$$

onde

$$v = \frac{1}{2} ln \left[\frac{1+r}{1-r} \right]$$