

Análise de Regressão Linear Simples e Múltipla

Departamento de Matemática
Escola Superior de Tecnologia de Viseu

Conceitos Iniciais

Introdução

A **análise de regressão** estuda o relacionamento entre uma variável chamada **variável dependente** e outras variáveis chamadas **variáveis independentes**.

Este relacionamento é representado por um modelo matemático, isto é, por uma equação que associa a variável dependente com as variáveis independentes.

Este modelo é designado por **modelo de regressão linear simples** se define uma relação linear entre a variável dependente e uma variável independente.

Se em vez de uma, forem incorporadas várias variáveis independentes, o modelo passa a denominar-se **modelo de regressão linear múltipla**.

Introdução

A **análise de correlação** dedica-se a inferências estatísticas das medidas de associação linear que se seguem:

- ▶ **coeficiente de correlação simples**: mede a “força” ou “grau” de relacionamento linear entre 2 variáveis;
- ▶ **coeficiente de correlação múltiplo**: mede a “força” ou “grau” de relacionamento linear entre uma variável e um conjunto de outras variáveis.

As técnicas de análise de correlação e regressão estão intimamente ligadas.

Diagrama de Dispersão

Os dados para a análise de regressão e correlação simples são da forma:

$$(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i), \dots, (x_n, y_n)$$

Com os dados constrói-se o **diagrama de dispersão**. Este deve exibir uma tendência linear para que se possa usar a regressão linear.

Portanto este diagrama permite decidir empiricamente se um relacionamento linear entre X e Y deve ser assumido.

Por análise do diagrama de dispersão pode-se também concluir (empiricamente) se o grau de relacionamento linear entre as variáveis é forte ou fraco, conforme o modo como se situam os pontos em redor de uma recta imaginária que passa através do enxame de pontos.

Diagrama de Dispersão

A correlação é tanto maior quanto mais os pontos se concentram, com pequenos desvios, em relação a essa recta.

Se o declive da recta é positivo, concluímos que a **correlação entre X e Y é positiva**, i.e., os fenómenos variam no mesmo sentido.

Ao contrário, se o declive é negativo, então a **correlação entre X e Y é negativa**, i.e., os fenómenos variam em sentido inverso.

Regressão Linear Simples

Diagrama de Dispersão

Sugerem uma regressão não linear
(i.e., a relação entre as duas variáveis poderá ser descrita por uma equação não linear)

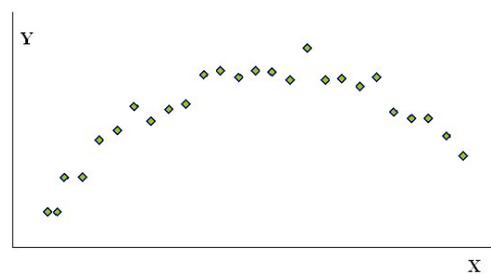
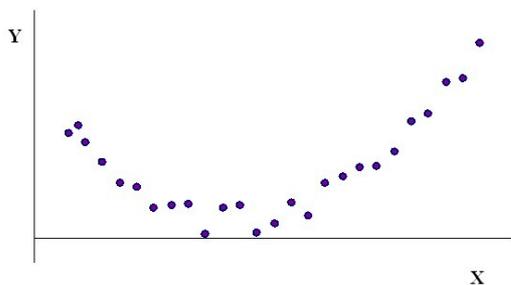
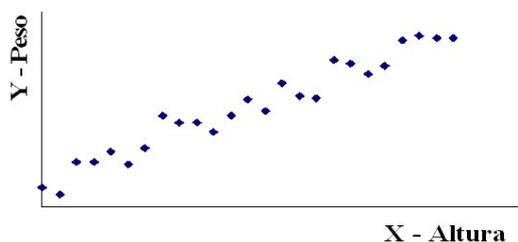
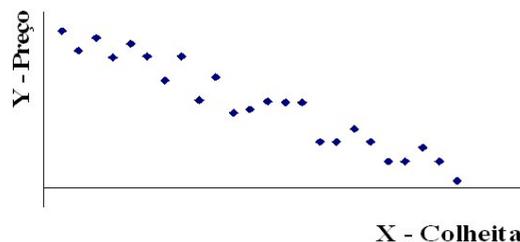


Diagrama de Dispersão

Sugerem uma regressão linear
(i.e., a relação entre as duas variáveis poderá ser descrita por uma equação linear)



Existência de correlação positiva (em média, quanto maior for a altura maior será o peso)



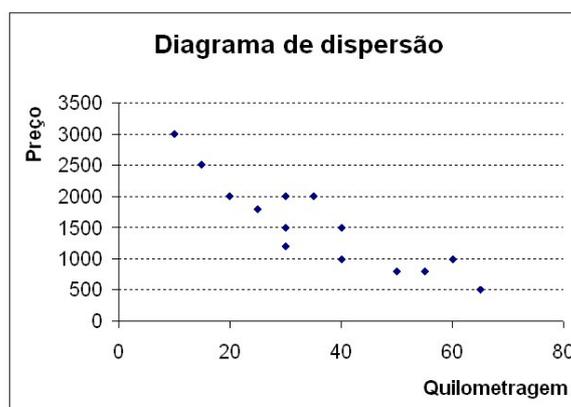
Existência de correlação negativa (em média, quanto maior for a colheita menor será o preço)

Regressão Linear Simples

Exemplo

Queremos estudar a relação entre a quilometragem de um carro usado e o seu preço de venda

Carros	Quilometragem X (1000 Km)	Preço de venda Y (dezena de Euros)
1	40	1000
2	30	1500
3	30	1200
4	25	1800
5	50	800
6	60	1000
7	65	500
8	10	3000
9	15	2500
10	20	2000
11	55	800
12	40	1500
13	35	2000
14	30	2000
Total	505	21600



Os dados sugerem uma relação linear entre a quilometragem e o preço de venda. Existe uma **correlação negativa**: em média, quanto maior for a quilometragem menor será o preço de venda.

O Modelo de Regressão Linear Simples

$$Y = \beta_0 + \beta_1 X + E$$

X – variável explicativa ou independente medida sem erro (não aleatória);
 E – variável aleatória residual na qual se procuram incluir todas as influências no comportamento da variável Y que não podem ser explicadas linearmente pelo comportamento da variável X ;
 β_0 e β_1 – parâmetros desconhecidos do modelo (a estimar);
 Y – variável explicada ou dependente (aleatória).

Exemplos

1. Relação entre o peso e a altura de um homem adulto (X : altura; Y : peso)
2. Relação entre o preço do vinho e o montante da colheita em cada ano (X : montante da colheita; Y : preço do vinho)

Regressão Linear Simples

Num estudo de regressão temos n observações da variável X : x_1, x_2, \dots, x_n (assume-se que estas observações são medidas sem erro).

Temos então n variáveis aleatórias Y_1, Y_2, \dots, Y_n tais que:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \quad i = 1, \dots, n$$

Admite-se que E_1, E_2, \dots, E_n são variáveis aleatórias independentes de média zero e variância σ^2 .

Para qualquer valor x_i de X , Y_i é uma variável aleatória de média $\mu_{Y_i} = \beta_0 + \beta_1 x_i$ e variância σ^2

Os dados para a análise de regressão e correlação simples são da forma: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ onde x_i é o valor da variável X e y_i a correspondente observação da variável aleatória Y_i ($i = 1, \dots, n$).

Cada observação satisfaz a seguinte relação:

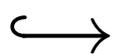
$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\mu_{Y_i}} + \varepsilon_i \quad i = 1, \dots, n$$



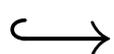
O valor observado de uma variável aleatória (y_i), usualmente difere da sua média (μ_{Y_i}) por uma quantidade aleatória ε_i .

A partir dos dados disponíveis estimamos β_0 e β_1 e substituímos estes parâmetros pelas suas estimativas para obter a **equação de regressão estimada**.

$$\hat{y} = \hat{\mu}_{Y|X} = b_0 + b_1 x$$

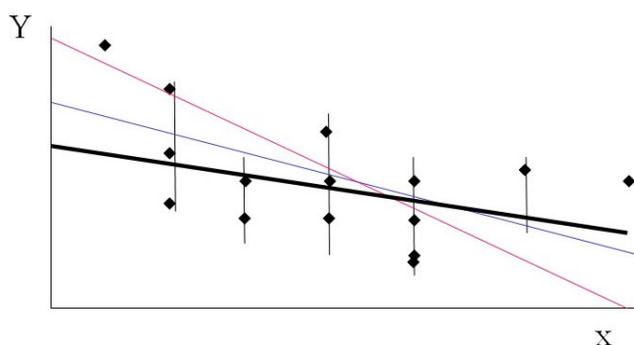


Esta equação estima o valor médio de Y para um dado valor x de X , mas é usada para estimar o próprio valor de Y .

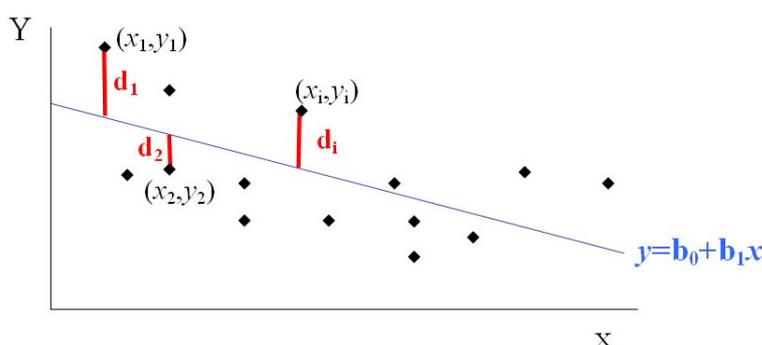


De facto, o senso comum diz-nos que uma escolha razoável para predizer o valor de Y para um dado x de X , é o valor médio estimado $\hat{\mu}_{Y|X}$.

Estimação pelo Método dos Mínimos Quadrados



Qual a recta que melhor se ajusta?



$\hat{y}_i = b_0 + b_1 x_i$
é o valor dado pela recta

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

resíduos

Estimação pelo Método dos Mínimos Quadrados

Iremos estimar os parâmetros usando o método dos mínimos quadrados.

Seja $d_i = y_i - \hat{y}_i \leftrightarrow$ i -ésimo resíduo.

O objectivo é escolher b_0 e b_1 de modo a minimizar a soma dos quadrados destes resíduos.

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

Estimação pelo Método dos Mínimos Quadrados

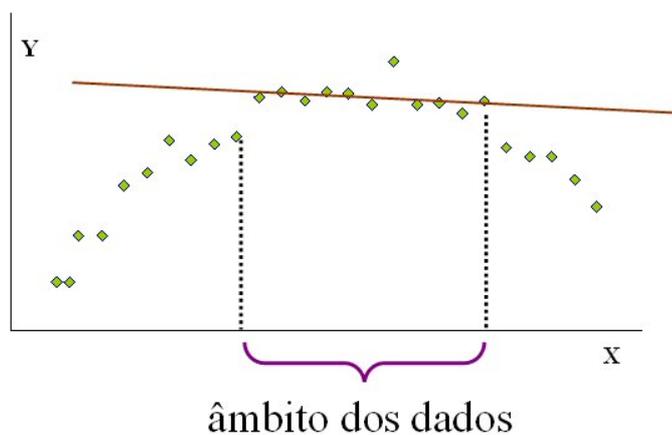
Para determinar b_0 e b_1 , de modo a minimizar SSE resolve-se o seguinte sistema de equações:

$$\begin{cases} \frac{\partial SSE}{\partial b_0} = 0 \\ \frac{\partial SSE}{\partial b_1} = 0 \end{cases} \Leftrightarrow \dots \Leftrightarrow \begin{cases} b_0 = \bar{y} - b_1 \bar{x} \\ b_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \end{cases}$$

Regressão Linear Simples

ATENÇÃO:

Um conjunto de pontos dá evidência de linearidade apenas para os valores de X cobertos pelo conjunto de dados. Para valores de X que saem fora dos que foram cobertos não há qualquer evidência de linearidade. Por isso é arriscado usar uma recta de regressão estimada para prever valores de Y correspondentes a valores de X que saem fora do âmbito dos dados.



O perigo de extrapolar para fora do âmbito dos dados amostrais é que a mesma relação possa não mais se verificar.

O Modelo de Regressão Linear Múltipla

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + E$$

X_1, \dots, X_k – variáveis explicativas ou independentes medidas sem erro (**não aleatórias**);

E – variável aleatória residual na qual se procuram incluir todas as influências no comportamento da variável Y que não podem ser explicadas linearmente pelo comportamento das variáveis X_1, \dots, X_k e os possíveis erros de medição;

β_0, \dots, β_k – parâmetros desconhecidos do modelo (a estimar);

Y – variável explicada ou dependente (**aleatória**).

O Modelo de Regressão Linear Múltipla

Exemplo (Ex. 3 da ficha nº8)

Relação entre o volume de vendas (Y) efectuadas durante um dado período de tempo por um vendedor, os seus anos de experiência (X_1) e o seu score num teste de inteligência (X_2).

Vendedores com 4 anos de experiência ($x_1 = 4$) e score 3 no teste de inteligência ($x_2 = 3$), podem apresentar volumes de vendas diferentes (Y 's diferentes).

Isto é, fixando a variável anos de experiência - X_1 - num valor, por exemplo 4 anos, e X_2 noutra valor, por exemplo 3, o volume de vendas vai variar devido a outras influências aleatórias.

Para x_1 e x_2 fixos, Y é uma variável aleatória.

O Modelo de Regressão Linear Múltipla

Num estudo de regressão temos n observações de cada variável independente:

$$\begin{array}{cccccc}
 & i = 1 & i = 2 & \dots & i = n \\
 X_1 & x_{11} & x_{12} & \dots & x_{1n} \\
 X_2 & x_{21} & x_{22} & \dots & x_{2n} \\
 \vdots & \vdots & \vdots & \ddots & \vdots \\
 X_k & x_{k1} & x_{k2} & \dots & x_{kn}
 \end{array}$$

Para cada i , i.e., para x_{1i}, \dots, x_{ki} fixos, Y_i é uma variável aleatória.

Temos então n variáveis aleatórias: Y_1, Y_2, \dots, Y_n :

$$Y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + E_i \quad i = 1, \dots, n$$

O Modelo de Regressão Linear Múltipla

$$Y_1 = \beta_0 + \beta_1 x_{11} + \dots + \beta_k x_{k1} + E_1$$

$$\vdots$$

$$Y_n = \beta_0 + \beta_1 x_{1n} + \dots + \beta_k x_{kn} + E_n$$

Admite-se que E_1, \dots, E_n são variáveis aleatórias independentes de média zero e variância σ^2

Então, para quaisquer valores x_{1i}, \dots, x_{ki} fixos, Y_i é uma variável aleatória de média

$$\mu_{Y_i} = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}$$

e variância σ^2 .

O Modelo de Regressão Linear Múltipla

Os dados para a análise de regressão e de correlação múltipla são da forma:

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

Cada observação obedece à seguinte relação:

$$y_i = \underbrace{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}}_{\mu_{Y_i}} + \varepsilon_i, \quad i = 1, \dots, n.$$



O valor observado de uma variável aleatória (y_i), usualmente difere da sua média (μ_{Y_i}) por uma quantidade aleatória ε_i .

O Modelo de Regressão Linear Múltipla

Temos então o seguinte sistema escrito em notação matricial:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \Leftrightarrow \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

y - Vector das observações da variável dependente;

X - Matriz significativa do modelo;

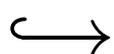
$\boldsymbol{\beta}$ - Vector dos parâmetros do modelo;

$\boldsymbol{\varepsilon}$ - Vector das realizações da variável aleatória residual.

Estimação pelo Método dos Mínimos Quadrados

A partir dos dados disponíveis estimamos $\beta_0, \beta_1, \dots, \beta_k$ e substituímos estes parâmetros pelas suas estimativas b_0, b_1, \dots, b_k para obter a equação de regressão estimada.

$$\hat{y} = \hat{\mu}_{Y|X_1, X_2, \dots, X_k} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$



Esta equação estima o valor médio de Y para um dado conjunto de valores x_1, x_2, \dots, x_k fixo, mas é usada para estimar o próprio valor de Y .

Estimação pelo Método dos Mínimos Quadrados

A cada observação $(y_i, x_{1i}, x_{2i}, \dots, x_{ki})$ está associado um **resíduo**

$$d_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i} + \dots + b_k x_{ki})$$

O objectivo é escolher b_0, b_1, \dots, b_k de modo a minimizar a **soma dos quadrados dos resíduos**.

$$SSE = \sum_{i=1}^n d_i^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \dots - b_k x_{ki})^2$$

Estimação pelo Método dos Mínimos Quadrados

Para determinar b_0, b_1, \dots, b_k , de modo a minimizar SSE resolve-se o seguinte sistema de equações:

$$\frac{\partial SSE}{\partial b_0} = 0 \quad \wedge \quad \frac{\partial SSE}{\partial b_1} = 0 \quad \wedge \quad \dots \quad \wedge \quad \frac{\partial SSE}{\partial b_k} = 0$$

Obtém-se $b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = (X^T X)^{-1} X^T y$ estimativa para $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$

O estimador é $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (X^T X)^{-1} X^T Y.$

Estimação pelo Método dos Mínimos Quadrados

Cada coeficiente de regressão estimado $b_i, i = 1, \dots, k$ (estimativa de β_i), **estima o efeito sobre o valor médio da variável dependente Y de uma alteração unitária da variável independente X_i** , mantendo-se constantes todas as restantes variáveis independentes.

No caso $k = 1$ (regressão simples) temos:

$$b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = (X^T X)^{-1} X^T y,$$

onde X tem apenas duas colunas.

Como já vimos, b_0 e b_1 podem também ser determinados pelas relações:

$$b_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \quad \text{e} \quad b_0 = \bar{y} - b_1 \bar{x}.$$

Exemplo (Ex. 3 da ficha nº8)

Os dados apresentados no quadro seguinte representam as vendas, Y , em milhares de Euros, efectuadas por 10 empregados de uma dada empresa, o nº de anos de experiência de cada vendedor, X_1 e o respectivo score no teste de inteligência, X_2 .

Vendedor	Vendas (Y)	Anos de experiência(X_1)	Score no teste de inteligência (X_2)
1	9	6	3
2	6	5	2
3	4	3	2
4	3	1	1
5	3	4	1
6	5	3	3
7	8	6	3
8	2	2	1
9	7	4	2
10	4	2	2

Regressão Linear Múltipla

Exemplo (Ex. 3 da ficha nº8 - cont.)

Pretende-se determinar se o sucesso das vendas pode ser medido em função das duas variáveis explicativas X_1 e X_2 através de um modelo linear .

Matriz significativa do modelo: $X =$

$$\begin{bmatrix} 1 & 6 & 3 \\ 1 & 5 & 2 \\ 1 & 3 & 2 \\ 1 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 3 & 3 \\ 1 & 6 & 3 \\ 1 & 2 & 1 \\ 1 & 4 & 2 \\ 1 & 2 & 2 \end{bmatrix}$$

Vector das observações da var. dependente: $y = [9 \ 6 \ 4 \ 3 \ 3 \ 5 \ 8 \ 2 \ 7 \ 4]^T$

Exemplo (Ex. 3 da ficha nº8 - cont.)

Vector das estimativas dos coeficientes de regressão:

$$b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = (X^T X)^{-1} X^T y = \begin{bmatrix} -0.262712 \\ 0.745763 \\ 1.338983 \end{bmatrix}$$

Equação de regressão estimada:

$$\hat{y} = \hat{\mu}_{Y|X_1, X_2} = -0.262712 + 0.745763x_1 + 1.338983x_2$$



Estima-se que o volume médio de vendas de um vendedor (em milhares de Euros) é igual a 0.745763 vezes os seus anos de experiência mais 1.338983 vezes o seu score no teste de inteligência menos 0.262712.

Exemplo (Ex. 3 da ficha nº8 - cont.)

Por exemplo, o volume médio de vendas para vendedores com 4 anos de experiência e com score 3 no teste de inteligência é estimado por:

$$\hat{y} = -0.262712 + 0.745763 \times 4 + 1.338983 \times 3 = 6.737289$$

$b_1 = 0.745763 \mapsto$ Em média, um ano extra de experiência entre vendedores com o mesmo score no teste de inteligência, conduz a um aumento no volume de vendas de uma quantidade que pode ser estimada em 745.763 Euros.

$b_2 = 1.338983 \mapsto$ Em média, um vendedor com score no teste de inteligência igual a 2 vende mais 1338.983 Euros (valor estimado) do que um vendedor com a mesma experiência e score 1, e menos 1338.983 Euros do que um vendedor com a mesma experiência e com score 3.

Exemplo (Ex. 3 da ficha nº8 - cont.)

Atenção:

- ▶ $b_0 = -0.262712$ não pode ser interpretado como sendo o volume médio de vendas de um vendedor hipotético sem experiência prévia e com score zero no teste de inteligência. Com efeito, vendas negativas são impossíveis. Note que valores nulos de X_1 e X_2 encontram-se fora do âmbito dos dados.
- ▶ Trata-se de uma relação média, assim um vendedor com determinados anos de experiência e determinado score no teste de inteligência não obterá necessariamente o volume de vendas exacto indicado pela equação.

Qualidade do Ajustamento

A equação de regressão estimada pode ser vista como uma tentativa para explicar as variações na variável dependente Y que resultam das alterações nas variáveis independentes X_1, X_2, \dots, X_k .

Seja \bar{y} a média dos valores observados para a variável dependente.

Uma medida útil associada ao modelo de regressão é o grau em que as predições baseadas na equação, \hat{y}_i , superam as predições baseadas em \bar{y} .

Se a dispersão (erro) associada à equação é muito menor que a dispersão (erro) associada a \bar{y} , as predições baseadas no modelo serão melhores que as baseadas em \bar{y} .

Qualidade do Ajustamento

Dispersão em torno de \bar{y} - **Variação total**:

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (\text{Soma dos quadrados totais})$$

Dispersão em torno da equação de regressão - **Variação não explicada**:

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (\text{Soma dos quadrados dos resíduos})$$

O ajustamento será tanto melhor quanto mais pequeno for **SSE** relativamente a **SST**.

Pode-se mostrar que:

$$\begin{array}{ccccc} \sum_{i=1}^n (y_i - \bar{y})^2 & = & \sum_{i=1}^n (y_i - \hat{y}_i)^2 & + & \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\ \downarrow & & \downarrow & & \downarrow \\ SST & = & SSE & + & SSR \end{array}$$

SST \longmapsto Soma dos quadrados totais - Variação total

SSE \longmapsto Soma dos quadrados dos resíduos - Variação não explicada

SSR \longmapsto Soma dos quadrados da regressão - Variação explicada

Isto é:

Variação Total de Y à volta da sua média	=	Variação que o ajustamento não consegue explicar	+	Variação explicada pelo ajustamento
--	---	--	---	-------------------------------------

Coeficiente de Determinação

O quociente entre SSR e SST dá-nos uma medida da proporção da variação total que é explicada pelo modelo de regressão. A esta medida dá-se o nome de **coeficiente de determinação** (r^2),

$$r^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = \frac{SST}{SST} - \frac{SSE}{SST} = 1 - \frac{SSE}{SST}$$

Note que:

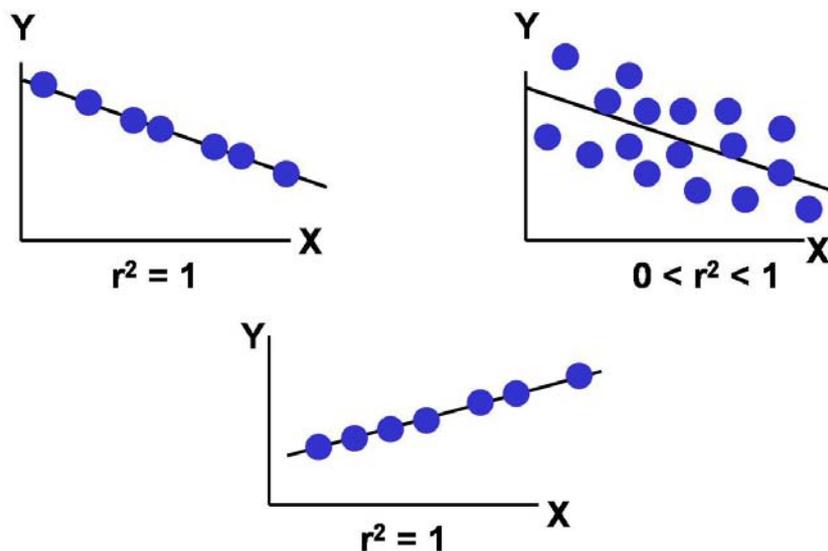
- ▶ $0 \leq r^2 \leq 1$;
- ▶ $r^2 \cong 1$ (próximo de 1) significa que grande parte da variação de Y é explicada linearmente pelas variáveis independentes;
- ▶ $r^2 \cong 0$ (próximo de 0) significa que grande parte da variação de Y não é explicada linearmente pelas variáveis independentes.

Coeficiente de Determinação

Este coeficiente pode ser utilizado como uma **medida da qualidade do ajustamento**, ou como medida da confiança depositada na equação de regressão como instrumento de previsão:

- ▶ $r^2 \cong 0 \implies$ modelo linear muito pouco adequado;
- ▶ $r^2 \cong 1 \implies$ modelo linear bastante adequado.

Exemplos de diagramas (Regressão simples)



Coeficiente de Correlação

À raiz quadrada de r^2 dá-se o nome de:

- ▶ **coeficiente de correlação simples** se está envolvida apenas uma variável independente;
- ▶ **coeficiente de correlação múltiplo** se estão envolvidas pelo menos duas variáveis independentes.

Coeficiente de Correlação Simples

$$r = \pm\sqrt{r^2} \text{ (com o sinal do declive } b_1)$$

Este coeficiente é uma medida do grau de relacionamento linear entre as variáveis X e Y .

- ▶ r varia entre -1 e 1 ;
- ▶ $r = -1$ e $r = 1$ indicam a existência de uma relação linear perfeita (negativa e positiva respectivamente) entre X e Y ;
- ▶ $r = 0$ indica a inexistência de qualquer relação ou tendência linear entre X e Y ;
- ▶ $r > 0$ indica uma relação linear positiva entre as variáveis X e Y , ou seja, as variáveis tendem a variar no mesmo sentido;
- ▶ $r < 0$ indica uma relação linear negativa entre as variáveis X e Y , ou seja, as variáveis tendem a variar em sentido inverso.

Coeficiente de Correlação Simples

O coeficiente de correlação simples r também pode ser calculado a partir da seguinte fórmula:

$$r = \pm \sqrt{\frac{b_0 \sum_{i=1}^n y_i + b_1 \sum_{i=1}^n y_i x_i - n\bar{y}^2}{\sum_{i=1}^n y_i^2 - n\bar{y}^2}} \text{ (com o sinal do declive } b_1)$$

Coeficiente de Correlação Múltiplo

É uma medida do grau de associação linear entre Y e o conjunto de variáveis X_1, X_2, \dots, X_k .

- ▶ r varia entre 0 e 1;
- ▶ $r = 1$ indica a existência de uma associação linear perfeita, ou seja, Y pode ser expresso como uma combinação linear de X_1, X_2, \dots, X_k ;
- ▶ $r = 0$ indica a inexistência de qualquer relação linear entre a variável dependente Y e o conjunto de variáveis independentes X_1, X_2, \dots, X_k .

Coeficiente de Correlação e de Determinação

Exemplo (Ex. 3 da ficha nº8 - cont.)

Para o exemplo em estudo temos a seguinte tabela

i	y_i	x_{1i}	x_{2i}	\hat{y}_i	d_i $= y_i - \hat{y}_i$	d_i^2	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$
1	9	6	3	8,22881	0,77119	0,59473
2	6	5	2	6,14407	-0,14407	0,02076
3	4	3	2	4,65254	-0,65254	0,42581
4	3	1	1	1,82203	1,17797	1,38760
5	3	4	1	4,05932	-1,05932	1,12216
6	5	3	3	5,99153	-0,99153	0,98312
7	8	6	3	8,22881	-0,22881	0,05236
8	2	2	1	2,56780	-0,56780	0,32239
9	7	4	2	5,39831	1,60169	2,56543
10	4	2	2	3,90678	0,09322	0,00869
Total	51					SSE =7.48305	SST =48.9	SSR =41.41695

Exemplo (Ex. 3 da ficha nº8 - cont.)

Coeficiente de determinação:

$r^2 = \frac{SSR}{SST} = \frac{41.41695}{48.9} = 0.84697 \implies$ 84.7% da variação nas vendas está relacionada linearmente com variações nos anos de experiência e no QI. Por outras palavras, as duas variáveis independentes utilizadas no modelo linear ajudam a explicar cerca de 84.7% da variação nas vendas. Ficam por explicar 15.3% das variações no volume de vendas, que se devem a outros factores não considerados, como por exemplo:

- ▶ a simpatia do vendedor;
- ▶ a reputação do vendedor;
- ▶ etc.

Exemplo (Ex. 3 da ficha nº8 - cont.)

Coeficiente de correlação múltiplo:

$r = \sqrt{0.84697} = 0.92031 \implies$ indica a existência de uma associação linear forte entre o volume de vendas e as variáveis independentes X_1 e X_2 , anos de experiência e score no teste de inteligência.

Podemos então concluir que o modelo linear se afigura bastante adequado para descrever o relacionamento entre a variável Y , volume de vendas, e as variáveis X_1 e X_2 .