

# ESTATÍSTICA II

**Michelle Hanne Soares de Andrade**

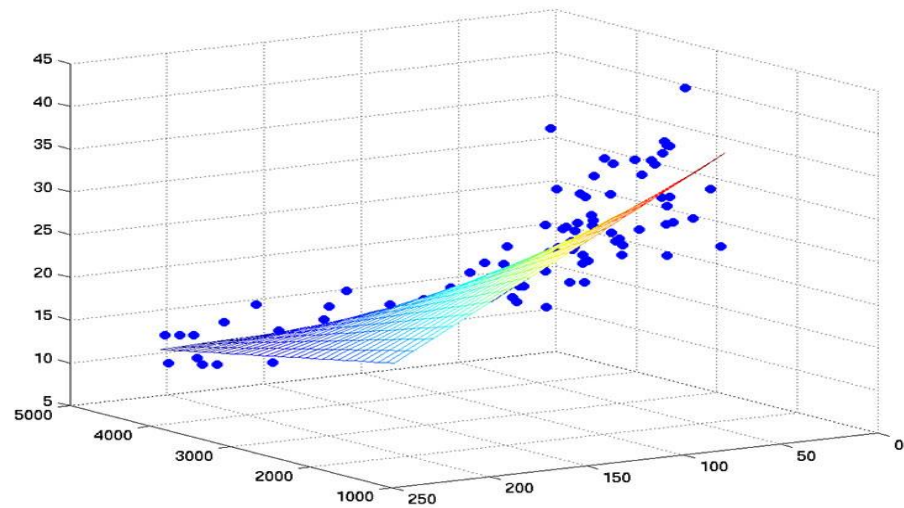
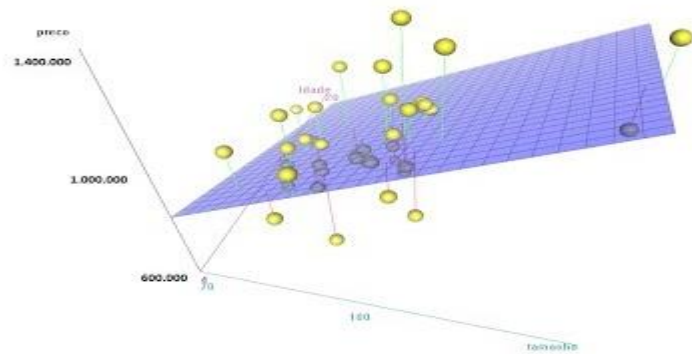
**[michellehanne.andrade@gmail.com](mailto:michellehanne.andrade@gmail.com)**

# REGRESSÃO LINEAR SIMPLES

# Introdução

- É a relação entre **duas ou mais variáveis quantitativas**: **uma variável dependente**, cujo valor deverá ser previsto e **uma variável (ou variáveis) independente(s)** ou explicativa(s), sobre a(s) qual(is) existe conhecimento teórico disponível.
- **Estimar uma equação é geometricamente equivalente a ajustar uma curva aos dados dispersos = REGRESSÃO.**

# Introdução



# Regressão Linear Múltipla

- A análise de uma regressão múltipla segue, basicamente, os mesmos critérios da análise de uma regressão simples.
- **A regressão múltipla envolve três ou mais variáveis, portanto, estimadores.** *Ou seja, ainda uma única variável dependente, porém duas ou mais variáveis independentes.*
- A finalidade das variáveis independentes adicionais é melhorar a capacidade de predição em confronto com a regressão linear simples.

## 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;

$\beta_0, \dots, \beta_k$  – parâmetros desconhecidos do modelo (a estimar);

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

# Modelo de Regressão Linear Múltipla

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

	$i = 1$	$i = 2$	$\dots$	$i = n$
$X_1$	$X_{11}$	$X_{12}$	$\dots$	$X_{1n}$
$X_2$	$X_{21}$	$X_{22}$	$\dots$	$X_{2n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$X_k$	$X_{k1}$	$X_{k2}$	$\dots$	$X_{kn}$

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

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

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

# 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  $\sigma^2$

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  $\sigma^2$ .



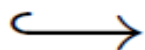
# Modelo de Regressão Linear Múltipla

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 ( $\mu_{Y_i}$ ) por uma quantidade aleatória  $\varepsilon_i$ .

## Modelo de Regressão Linear Múltipla

Temos então o seguinte sistema escrito em notação matricial:

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

$y$  - Vector das observações da variável dependente;

$X$  - Matriz significativa do modelo;

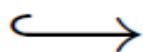
$\beta$  - Vector dos parâmetros do modelo;

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

# 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  $Y$ .

# Método dos Mínimos Quadrados

- O problema então é estimar o valor dos coeficientes  $\beta_i$  a partir de um conjunto de dados do tipo:

$Y$	$X_1$	$X_2$	...	$X_k$
$y_1$	$x_{11}$	$x_{12}$	...	$x_{1k}$
$y_2$	$x_{21}$	$x_{22}$	...	$x_{2k}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$y_n$	$x_{n1}$	$x_{n2}$	...	$x_{nk}$

## Método dos Mínimos Quadrados

- Novamente, o método dos Mínimos Quadrados é usado para minimizar a soma dos quadrados dos resíduos.

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

# 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:

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

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

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

# Método dos Mínimos Quadrados

Cada coeficiente de regressão estimado  $b_i$ ,  $i = 1, \dots, k$  (estimativa de  $\beta_i$ ), **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:

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

# Método dos Mínimos Quadrados

- Assim, temos:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 & X_{21} & X_{k1} \\ 1 & X_2 & X_{22} & X_{k2} \\ \dots & \dots & \dots & \dots \\ 1 & X_n & X_{2n} & X_{kn} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_k \end{bmatrix}$$

Que escrevendo ainda em outra em sua forma mais compacta temos:

$$Y = bX + \varepsilon$$

O estimador para  $b$  será dado por:

$$\hat{b} = (X'X)^{-1}(X'Y)$$

Há necessidade que o produto  $X'X$ , tenha uma matriz inversa, o que implica na condição obrigatória que nenhuma coluna da matriz  $X$  seja



## Exemplo 1

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

## Exemplo 1

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:  $X =$

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

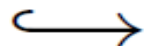
## Exemplo 1

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 1

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 1

### 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.

## Exemplo 2

- Pretende-se investigar a utilização de um modelo de regressão linear múltiplo para se tentar explicar a variação da viscosidade de um polímero (Y) em função da temperatura de reação,  $x_1$ , e da taxa de alimentação do catalisador,  $x_2$ . Realizando-se uma experiência, para os diferentes valores de  $x_1$  e  $x_2$ , obtiveram-se os valores de Y, respectivos, conforme tabela abaixo:

## Exemplo 2

N.º da observação	Viscosidade (y)	Temperatura ( $x_1$ , °C)	Catalisador ( $x_2$ , lb/h)
1	2256	80	8
2	2340	93	9
3	2426	100	10
4	2293	82	12
5	2330	90	11
6	2368	99	8
7	2250	81	8
8	2409	96	10
9	2364	94	12
10	2379	93	11
11	2440	97	13
12	2364	95	11
13	2404	100	8
14	2317	85	12
15	2309	86	9
16	2328	87	12

## Exemplo 2

O modelo a ser ajustado é do tipo  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ , onde se deve estimar os coeficientes de regressão. Em notação matricial,  $\hat{\beta} = (X^T X)^{-1} X^T Y$ , considerando a amostra obtém-se

$$X^T X = \begin{bmatrix} 16 & 1458 & 164 \\ 1458 & 133560 & 14946 \\ 164 & 14946 & 1726 \end{bmatrix} \text{ (matriz é simétrica),}$$

$$(X^T X)^{-1} = \begin{bmatrix} 14,176004 & -0,129746 & -0,223453 \\ -0,129746 & 1,429184 \times 10^{-3} & -4,763947 \times 10^{-5} \\ -0,223453 & -4,763947 \times 10^{-5} & 2,222381 \times 10^{-2} \end{bmatrix} \text{ e } X^T Y = \begin{bmatrix} 37577 \\ 3429550 \\ 385562 \end{bmatrix}, \text{ donde}$$

$$\hat{\beta}_0 = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 1566,07777 \\ 7,62129 \\ 8,58485 \end{bmatrix}.$$



## Exemplo 2

Assim, o modelo de regressão ajustado aos dados é, com quatro casas decimais,

$$y = 1566,0777 + 7,6213x_1 + 8,5848x_2.$$

A partir desta equação é possível obter os valores estimados (esperados através do modelo) de  $Y$  e prever observações futuras para a mesma variável. Por exemplo, para a primeira observação  $x_{11} = 80$  e  $x_{12} = 8$ , o valor ajustado será  $\hat{y}_1 = 1566,00777 + 7,6213x_{11} + 8,5848x_{12} = 2244,46$ , o valor observado correspondente é  $y_1 = 2256$ , o resíduo para esta observação é  $e_1 = y_1 - \hat{y}_1 = 11,54$ .

## Exemplo 2

Apresentam-se na tabela seguinte os valores ajustados (estimativas) da variável resposta a partir deste modelo de regressão e os respectivos erros de ajustamento para cada observação.

N.º da observação	$y_i$	$\hat{y}_i$	$e_i$
1	2256	2244,46	11,54
2	2340	2352,12	-12,12
3	2426	2414,06	11,94
4	2293	2294,04	-1,04
5	2330	2346,43	-16,43
6	2368	2389,26	-21,26
7	2250	2252,08	-2,08
8	2409	2383,57	25,43
9	2364	2385,50	-21,50
10	2379	2369,29	9,71
11	2440	2416,95	23,05
12	2364	2384,53	-20,53
13	2404	2396,89	7,11
14	2317	2316,91	0,09
15	2309	2298,77	10,23
16	2328	2332,15	-4,15

Tabela 2.3 – Observações e estimativas da variável resposta e respectivos resíduos

## Qualidade do Ajustamento do Modelo

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 previsões baseadas na equação,  $\hat{y}_i$ , superam as previsões baseadas em  $\bar{y}$ .

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

# Qualidade do Ajustamento do Modelo

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  $SST$ .

# Qualidade do Ajustamento do Modelo

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*  $\longrightarrow$  Soma dos quadrados totais - Variação total

*SSE*  $\longrightarrow$  Soma dos quadrados dos resíduos - Variação não explicada

*SSR*  $\longrightarrow$  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:

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

## Coefficiente 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  $X_1, X_2, \dots, X_k$ .



## Exemplo

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					<b>SSE</b> <b>=7.48305</b>	<b>SST</b> <b>=48.9</b>	<b>SSR</b> <b>=41.41695</b>

## Exemplo

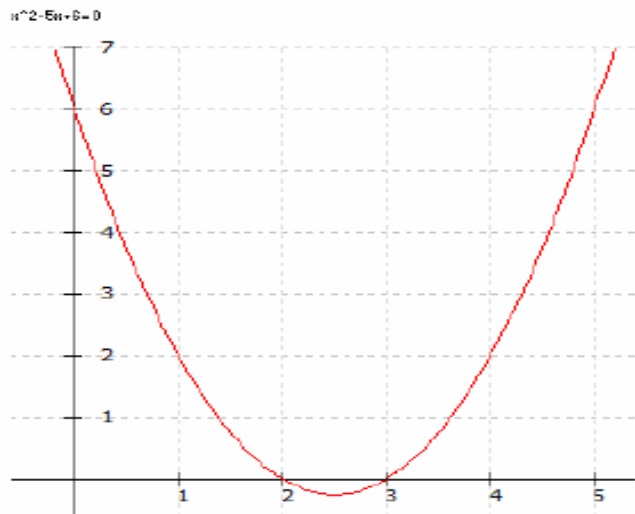
Coeficiente de determinação:

$r^2 = \frac{SSR}{SST} = \frac{41.41695}{48.9} = 0.84697 \rightarrow 84.7\%$  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:

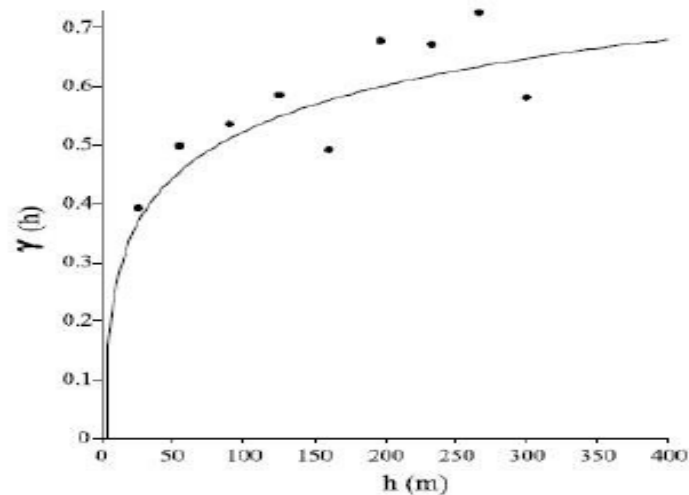
- ▶ a simpatia do vendedor;
- ▶ a reputação do vendedor;
- ▶ etc.

## Outros Tipos de Modelagem

### Modelo Quadrático

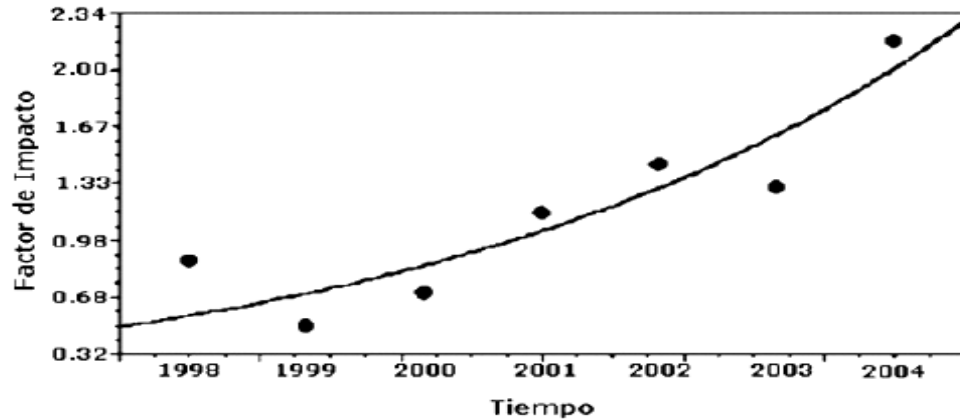


### Modelo Logarítmico

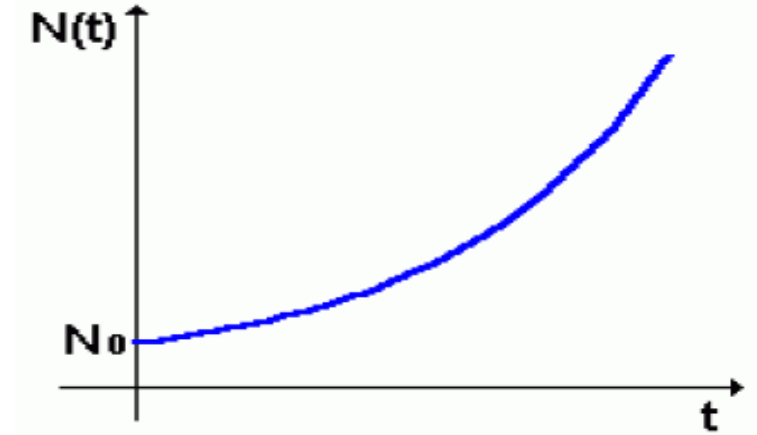


## Outros Tipos de Modelagem

### Modelo Exponencial

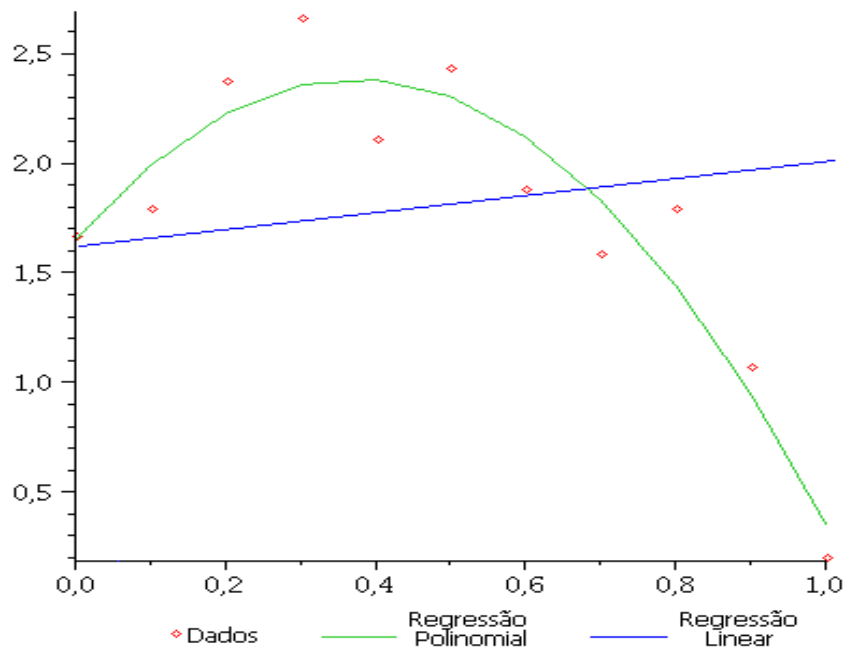


### Modelo Potencial



# Regressão Polinomial

- Existem muitos casos em que o modelo obedece a um comportamento polinomial. Para tais modelos é necessário adaptar o ajuste para uma função polinomial de grau superior.



# Regressão Polinomial

- A regressão polinomial pode ser tida como uma generalização da regressão linear. Para isso, podemos ver a regressão linear simples como a regressão polinomial de um polinômio de grau um. Assim, ao invés de ajustarmos a função

$$y = \alpha_0 + \alpha_1 x + \epsilon$$

- Utilizamos

$$y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \dots + \alpha_m x^m + \epsilon$$

# Regressão Polinomial

- Para ajustarmos os parâmetros dessa função, basta que resolvamos um sistema de  $m+1$  equações lineares simultâneas, tal o desenvolvimento da regressão linear múltipla. Assim, no caso da regressão polinomial, o erro padrão pode ser formulado como polinomial, o erro padrão pode ser formulado como:

$$s = \sqrt{\frac{S_r}{n-(m+1)}}$$

# Modelos de Regressão Polinomial

- As variáveis explanatórias devem ser quantitativas.
- Servem para representar modelos com resposta curvilínea.
- São fáceis de serem ajustados, pois são um caso especial do modelo de regressão linear múltipla.



## Quando Utilizar Regressão Polinomial

- Se a função de resposta curvilínea verdadeira é realmente uma função polinomial.
- Se a função de resposta curvilínea verdadeira é desconhecida (ou complexa), porém, uma função polinomial é uma boa aproximação para a verdadeira função.
- O principal problema com o uso de modelos polinomiais é com a extrapolação

## Variável preditora

- **Uma variável preditora - Modelo de segunda ordem**
- A variável preditora,  $x_i$ , é dada como desvio em relação a sua média. A razão para usar uma variável preditora centrada no modelo de regressão polinomial é que  $X$  e  $X^2$  são altamente correlacionadas.
- Isto pode causar sérias dificuldades para inverter a matriz  $\mathbf{X}'\mathbf{X}$  para estimar os coeficientes de regressão.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \quad \text{onde} \quad x_i = X_i - \bar{X}$$

## Variável preditora

- **Uma variável preditora - Modelo de segunda ordem**
- A variável preditora,  $x_i$ , é dada como desvio em relação a sua média. A razão para usar uma variável preditora centrada no modelo de regressão polinomial é que  $X$  e  $X^2$  são altamente correlacionadas.
- Isto pode causar sérias dificuldades para inverter a matriz  $\mathbf{X}'\mathbf{X}$  para estimar os coeficientes de regressão.

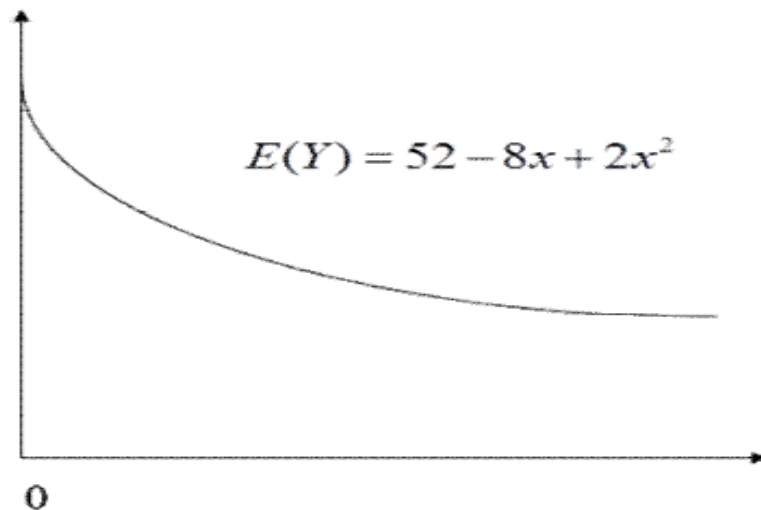
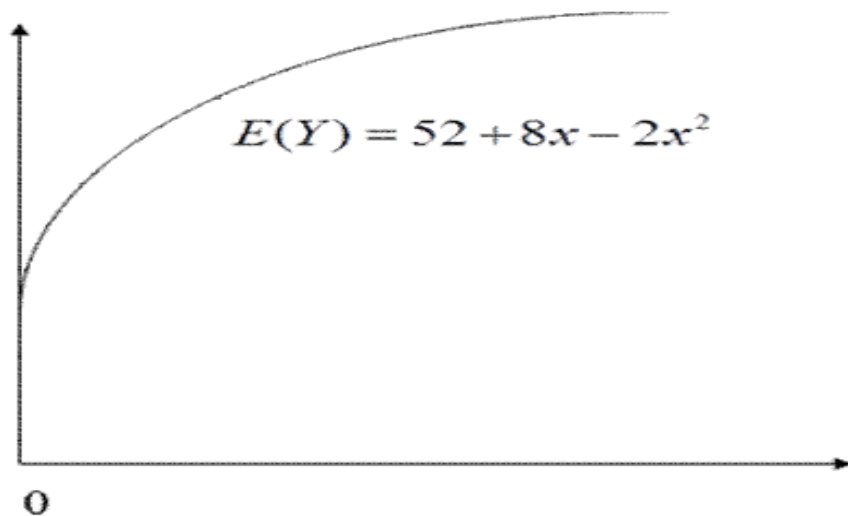
$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \quad \text{onde} \quad x_i = X_i - \bar{X}$$

# Variável preditora

- Exemplo:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \varepsilon_i$$


O gráfico desta função é uma parábola e denominada de função de resposta quadrática.



## Variável preditora

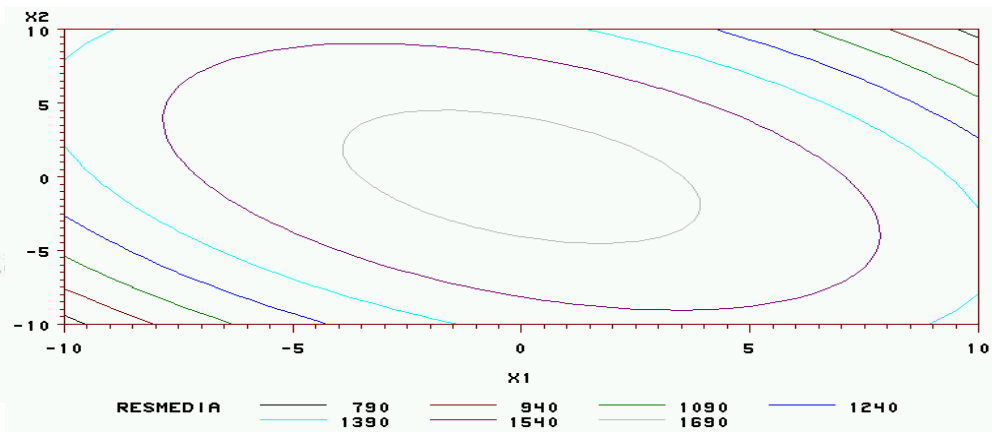
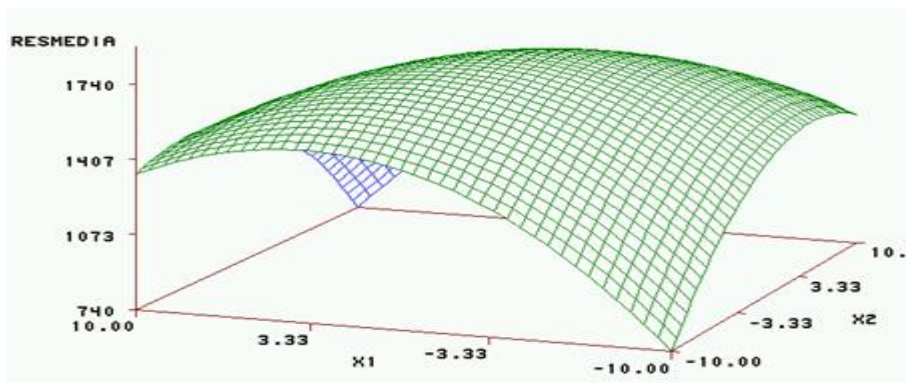
- **Duas variáveis preditoras - Modelo de segunda ordem**

Mostra as várias combinações dos níveis das 2 variáveis preditoras que resultam na mesma resposta

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \varepsilon_i$$


The diagram illustrates the components of the second-order model equation. A horizontal curly brace under the terms  $\beta_1 x_{i1} + \beta_2 x_{i2}$  is labeled "Linear". Another horizontal curly brace under the terms  $\beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2}$  is labeled "quadrático".

# Variável preditora



## Modelos de regressão polinomial

- Os modelos de regressão polinomial são casos especiais do modelo de regressão linear múltipla geral, assim, todos os resultados vistos para o ajuste de modelos e para inferência estatística são válidos.