Aula 05 – Regressão Linear, Múltipla e Não Linear

Luciana Rocha Pedro

GCC 1518 – Estatística e Probabilidade – CEFET Maracanã

24 de março de 2018

Correlação

Duas variáveis estão relacionadas quando a alteração no valor de uma varíavel (independente) provoca alterações no valor da outra variável (dependente).

Exemplos:

Fabricação: número de peças produzidas e número de peças defeituosas.

Construção: dias de atraso de entrega e número de dias chuvosos.

Correlação

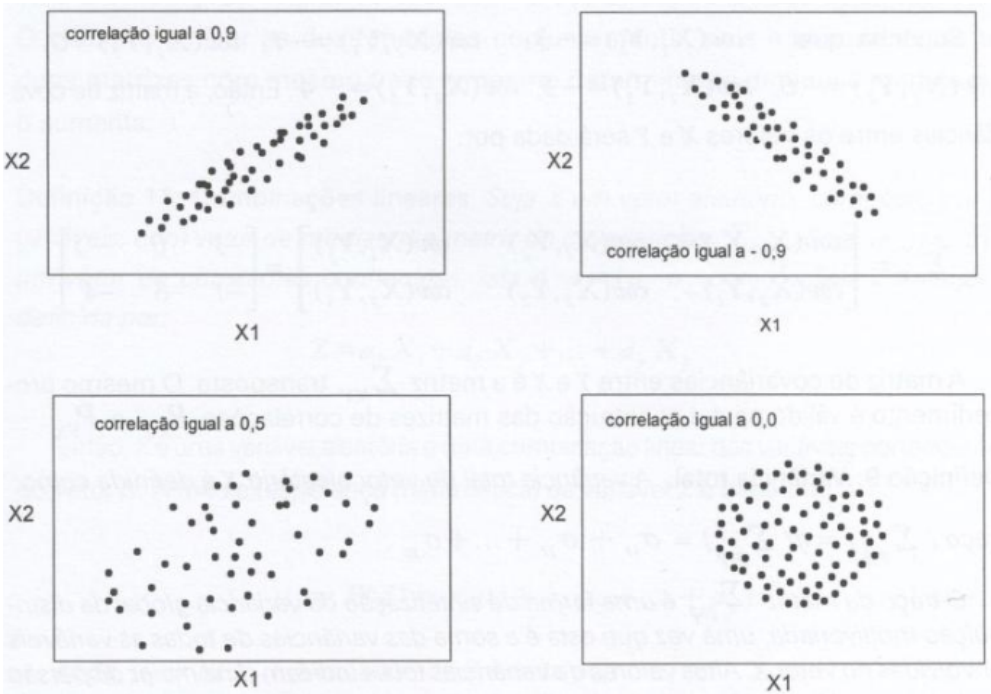
As variáveis podem estar correlacionadas porque uma delas depende da outra (há uma relação de causalidade).

As variáveis podem estar correlacionadas porque são interdependentes. Exemplo: idade do marido, idade da esposa.

As duas variáveis podem estar correlacionadas porque ambas são influenciadas por uma terceira variável e é o fato de ambas responderem a variações nessa variável que explica a correlação. Exemplo: número de insolações e produção de trigo.

Análise de Correlação

A **análise de correlação** é uma medida da relação entre dois atributos que indica a força e a direção do relacionamento linear entre estes atributos.



Análise de Correlação

Em muitas aplicações, duas ou mais variáveis estão relacionadas, sendo necessário explorar a natureza desta relação.

A correlação muito próxima de 1, ou de -1, implica que existe uma relação linear entre os dois atributos e permite verificar se é possível ajustar um modelo que expresse esta relação.

Esse é o objetivo da **análise de regressão**.

Correlação e Causalidade

Pesquisadores freqüentemente são tentados a inferir uma relação de **causa** e **efeito** entre dois atributos quando eles ajustam um modelo de regressão ou realizam uma análise de correlação.

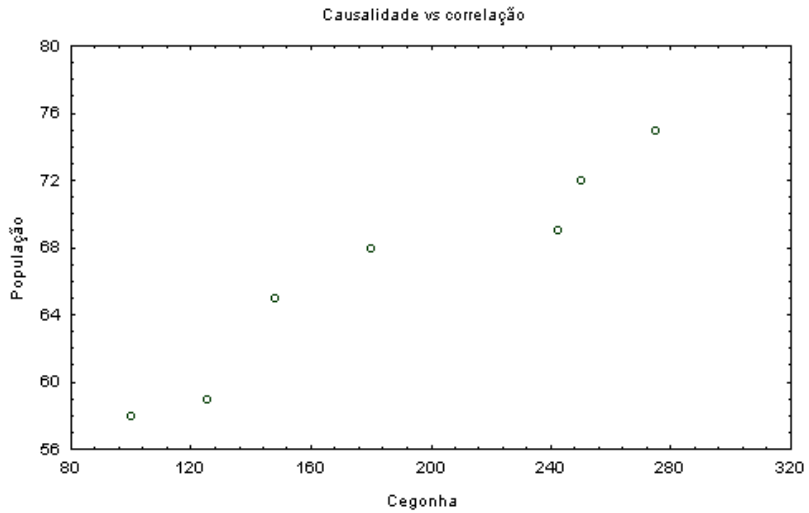
Uma associação significativa entre dois atributos em ambas as situações não necessariamente implica numa relação de causa e efeito.

Duas variáveis podem estar altamente correlacionadas e não existir relação de causa e efeito entre elas.

Correlação não necessariamente implica em causalidade.

Correlação e Causalidade

O gráfico a seguir mostra a população de Oldemberg, Alemanha, no final de cada um de sete anos contra o número de cegonhas (pássaros) naquele ano.



Correlação e Causalidade

Interpretação: existe associação entre os dois atributos.

Freqüentemente, quando os dois atributos parecem estar fortemente associados, pode ser porque estes atributos estão, de fato, associados a um terceiro atributo.

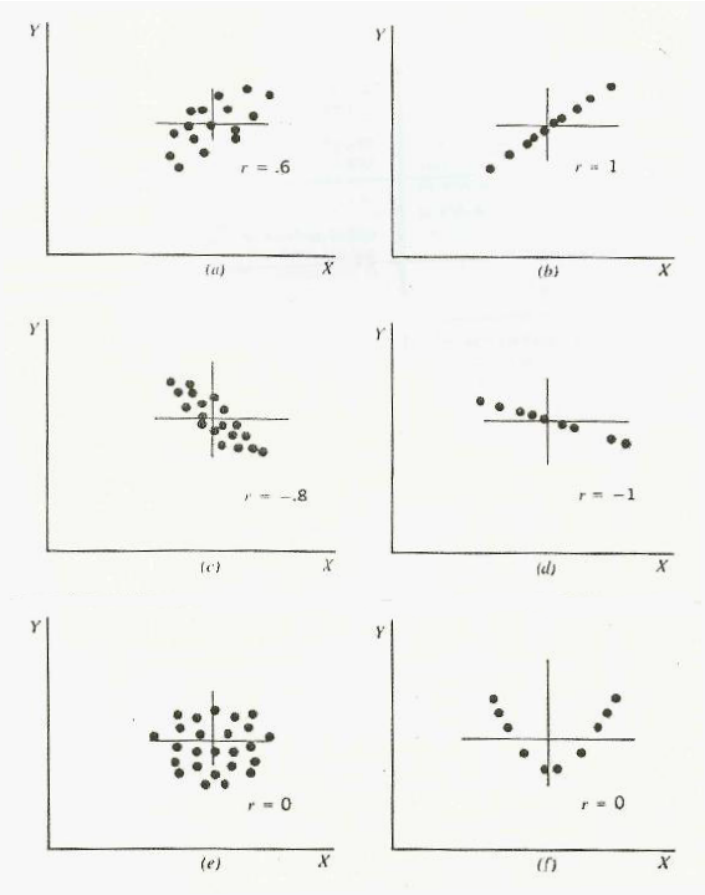
No exemplo, os dois atributos aumentam com o tempo.

Coeficiente de Correlação Linear de Pearson

O coeficiente de correlação linear de Pearson r só mede a intensidade ou grau de relacionamentos **lineares**.

Este coeficiente não serve para medir intensidade de relacionamentos não-lineares.

Exemplos



Coeficiente de Correlação Linear de Pearson

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \Rightarrow S_{xx} = n(\sum x_i^2) - (\sum x_i)^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \Rightarrow S_{yy} = n(\sum y_i^2) - (\sum y_i)^2$$

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

$$r = \frac{Sxy}{\sqrt{S_{xx}.S_{yy}}}$$

$-1 \leq r \leq 1$

Coeficiente de Correlação Linear de Pearson

$$r = \frac{n\sum(x_i \cdot y_i) - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2} \cdot \sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

Coeficiente de Correlação Linear de Pearson

O valor r sempre será um valor entre $-1 \leq r \leq 1$.

Quanto mais próximo de -1, maior a correlação negativa.

Quanto mais próximo de 1, maior a correlação positiva.

Quanto mais próximo de 0, menor a correlação linear.

Análise de Regressão

A **análise de regressão** engloba uma série de técnicas voltadas para a modelagem e a investigação de relações entre dois ou mais atributos.

Na **análise de correlação**, o objetivo é determinar o **grau** de relacionamento entre duas variáveis.

Já na **análise de regressão**, o objetivo é determinar o **modelo** que expressa esta relação (equação de regressão), a qual é ajustada aos dados.

Análise de Regressão

A análise de regressão permite construir um modelo matemático que represente dois atributos x e y:

$$y = f(x),$$

em que $f(\cdot)$ é a função que relaciona x e y.

- x é a variável **independente** da equação.
- $y = f(x)$ é a variável **dependente** das variações de x.

Análise de Regressão

Podemos usar este modelo para predizer o valor de y para um dado valor de x e realizar previsões sobre o comportamento futuro de algum fenômeno da realidade.

Neste caso, extrapolamos para o futuro as relações de causa-efeito, já observadas no passado, entre as variáveis.

Análise de Regressão

Na maioria dos casos, a função $f(\cdot)$ é desconhecida.

Cabe ao usuário escolher uma função apropriada para aproximar $f(\cdot)$.

Normalmente, usamos um modelo polinomial.

Análise de Regressão

A análise de regressão compreende quatro tipos básicos de modelos:

- Linear simples
- Linear multivariado
- Não linear simples
- Não linear multivariado

Regressão Simples

Na **regressão simples** existe apenas uma variável de saída (y) e uma variável de entrada (x).

Exemplo: $y = f(x)$

Regressão Múltipla

Na **regressão múltipla** existe apenas uma variável de saída (y) e várias variáveis de entrada $(x_i, i = 1, \ldots, n)$.

Exemplo: $y = f(x_1, x_2, \ldots, x_n)$

Regressão Linear

A **regressão linear** considera que a relação da entre as variáveis é descrita por uma função linear (equação da reta ou do plano).

Exemplo: $y = \alpha + \beta x$

Regressão Não Linear

Na **regressão não linear**, a relação entre as variáveis não pode ser descrita por uma função linear.

Pode ser uma função exponencial ou logarítmica, por exemplo.

Exemplo: $y = \alpha e^{\beta x}$

Gráfico de Dispersão

Um **gráfico de dispersão** é uma representação puramente visual dos dados, dado pelo gráfico cartesiano dos pares de informação x e y referente a cada observação.

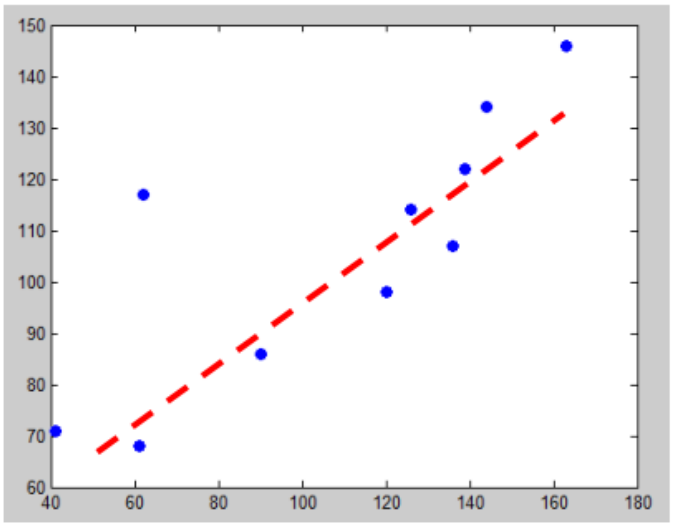
Consiste de uma *nuvem* de pontos que, por sua vez, define um eixo ou direção que caracterizará o padrão de relacionamento entre as variáveis x e y.

Gráfico de Dispersão

A regressão será linear se observada uma tendência linear na nuvem de pontos.

Em **modelos de regressão** é sempre importante verificar o gráfico de dispersão para saber que modelo usar.

y	x
122	139
114	126
86	90
134	144
146	163
107	136
68	61
117	62
71	41
98	120



Regressão Linear

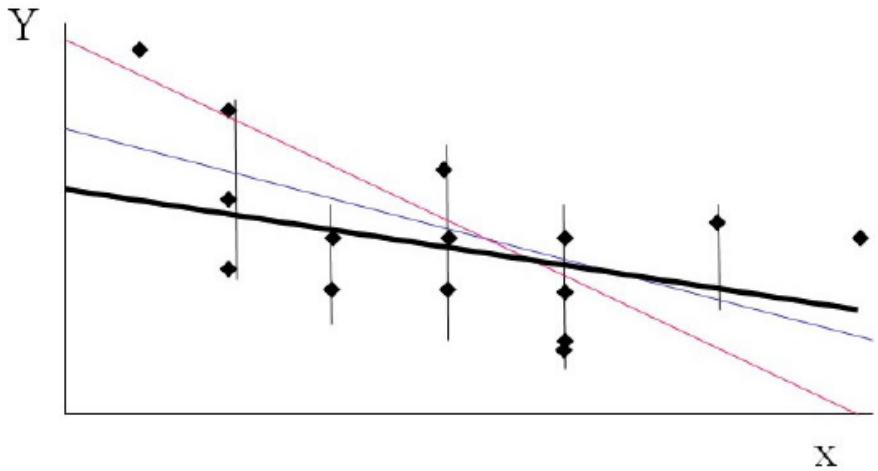
A **regressão linear** implica no ajuste de uma reta que represente de forma **adequada** a estrutura dos dados.

Diferentes retas podem ser traçadas, a olho nu, em um diagrama de dispersão. Cada pessoa terá uma tendência diferente.

Nenhuma reta passará exatamente por todos os pontos (se a correlação não for máxima).

Precisamos encontrar uma reta que esteja tão próxima dos pontos quanto possível.

Reta de Regressão Linear



Reta de Regressão Linear

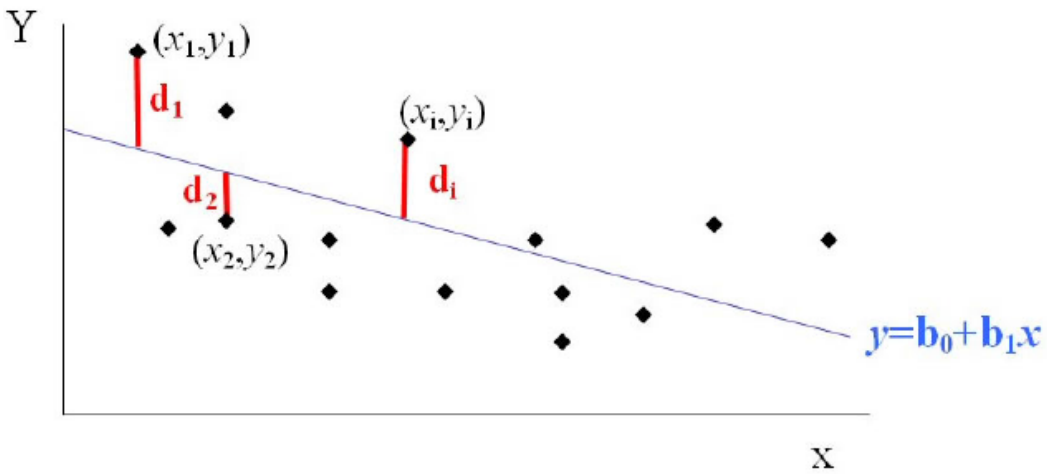
Se um diagrama de dispersão sugere uma relação linear, é de interesse representar este padrão através de uma reta.

Usamos o **método dos mínimos quadrados** para ajustar uma reta de regressão ao conjunto de pontos do diagrama.

A reta de regressão descreve como uma variável resposta y (dependente) varia em relação a uma variável explanatória x (independente).

Os erros de predição para a reta são erros em y (direção vertical).

Reta de Regressão Linear



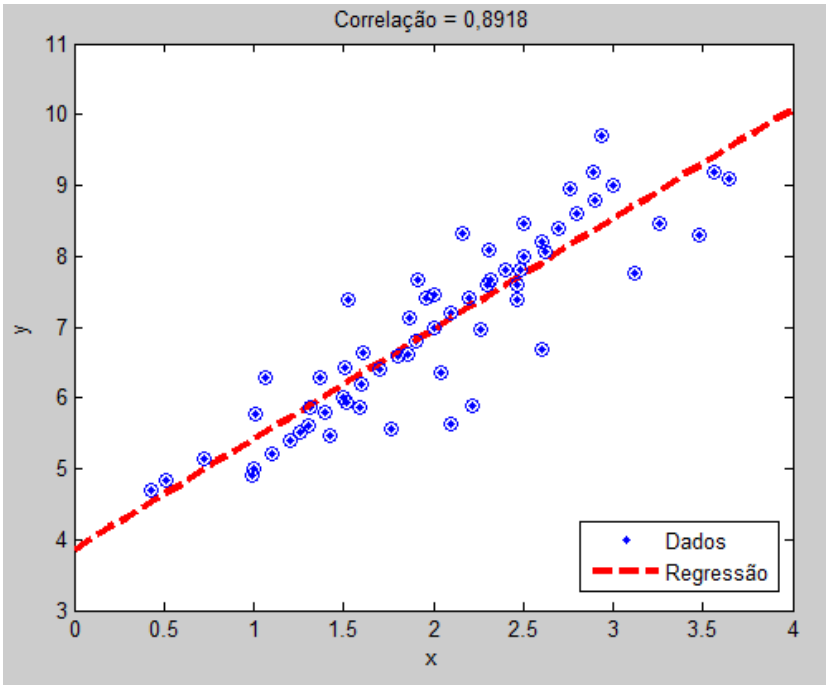
Regressão Linear Simples

Definições Básicas

- Existe uma única variável de saída, y (variável dependente).
- Existe uma única variável de entrada, x (variável independente).
- Assumimos que as variáveis de entrada são medidas com erro (i.e. ruído) desprezível.
- Exemplo: $y = \alpha + \beta x + \epsilon$

Exemplo de Regressão Linear Simples

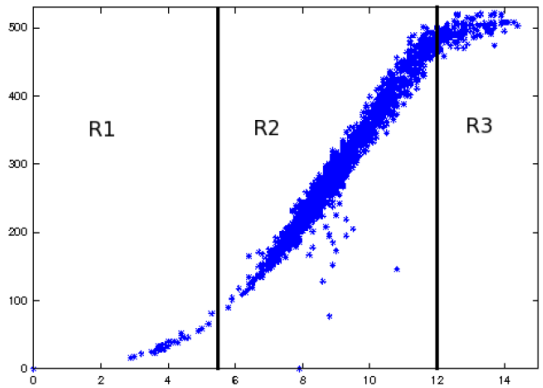
$$y = 1.55x + 3.86$$



Dados Não Lineares

O que fazer quando o modelo de regressão linear não é apropriado?

Solução 1: Podemos dividir o domínio original dos dados em sub-domínios e aplicar o modelo linear dentro de cada sub-domínio.



Dados Não Lineares

Solução 2: Aplicar uma linearização dos dados e continuar usando a regressão linear.

Solução 3: Podemos utilizar um modelo de regressão polinomial de ordem maior do que um ou um modelo não linear.

Método dos Mínimos Quadrados

O **método dos mínimos quadrados** foi proposto por Carl Friedrich Gauss em 1795.

O método foi utilizado no cálculo de órbitas de planetas e cometas a partir de medidas obtidas por telescópios.

Adrien Marie Legendre publicou o método dos mínimos quadrados primeiro, em 1806. O mesmo método foi desenvolvido de forma independente.

Método dos Mínimos Quadrados

O método dos mínimos quadrados é uma técnica de otimização matemática que procura o melhor ajuste para um conjunto de dados

$$(x(1), y(1)), (x(2), y(2)), \ldots, (x(n), y(n)).$$

ao mesmo tempo em que tenta minimizar a soma dos quadrados das diferenças entre o valor estimado e os dados observados,

$$\sum_{i=1}^{n} \epsilon_i^2.$$

Método dos Mínimos Quadrados

No método dos mínimos quadrados, procuramos por parâmetros α e β que minimizem a soma dos quadrados dos erros:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y(i) - \alpha - \beta x(i))^2 = \mathcal{J}(\alpha, \beta).$$

Isso equivale a fazer com que a soma dos quadrados dos **resíduos** entre os valores medidos (observações) e a reta de regressão seja mínima.

Método dos Mínimos Quadrados

A equação de regressão é calculada a partir das derivadas parciais da soma dos quadrados dos erros.

Derivadas parciais com relação aos parâmetros α e β:

$$\frac{\partial \mathcal{J}(\alpha, \beta)}{\partial \alpha} = -2 \sum_{i=1}^{n} (y(i) - \alpha - \beta x(i)),$$

$$\frac{\partial \mathcal{J}(\alpha, \beta)}{\partial \beta} = -2 \sum_{i=1}^{n} (y(i) - \alpha - \beta x(i)) x(i).$$

Método dos Mínimos Quadrados

Algumas deduções matemáticas e substituições e temos que

$$\alpha = \bar{y} - \beta\bar{x},$$

$$\beta = \frac{\sum_{i=1}^{n}(x(i) - \bar{x})(y(i) - \bar{y})}{\sum_{i=1}^{n}(x(i) - \bar{x})^2},$$

em que $\bar{x}$ e $\bar{y}$ são as médias amostrais de x e y, respectivamente.

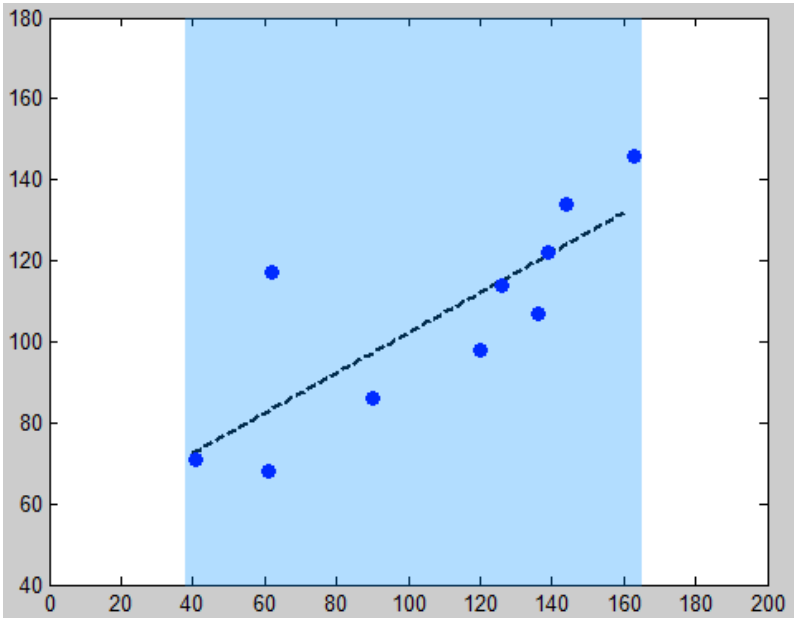
Método dos Mínimos Quadrados

Normalmente, a relação linear $y = \alpha + \beta x$ é considerada válida apenas para $x \in [x_{\text{min}}, x_{\text{max}}]$.

Modelos de regressão linear não costumam ser válidos para fins de **extrapolação**, apenas de **interpolação**.

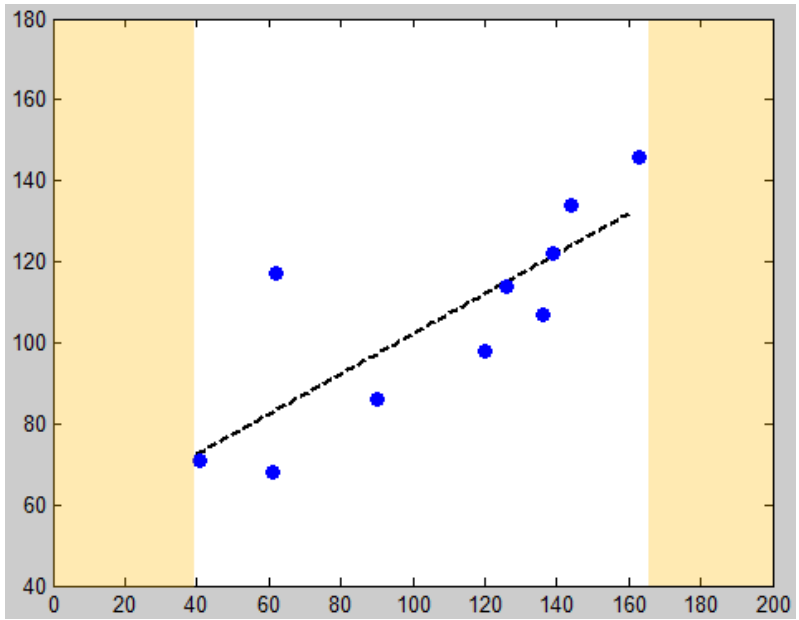
Interpolação

A **interpolação** consiste em calcular um valor de uma equação ou função, em um lugar **dentro** da zona conhecida.



Extrapolação

A **extrapolação** consiste em calcular um valor de uma equação ou função, em um lugar **fora** da zona conhecida.



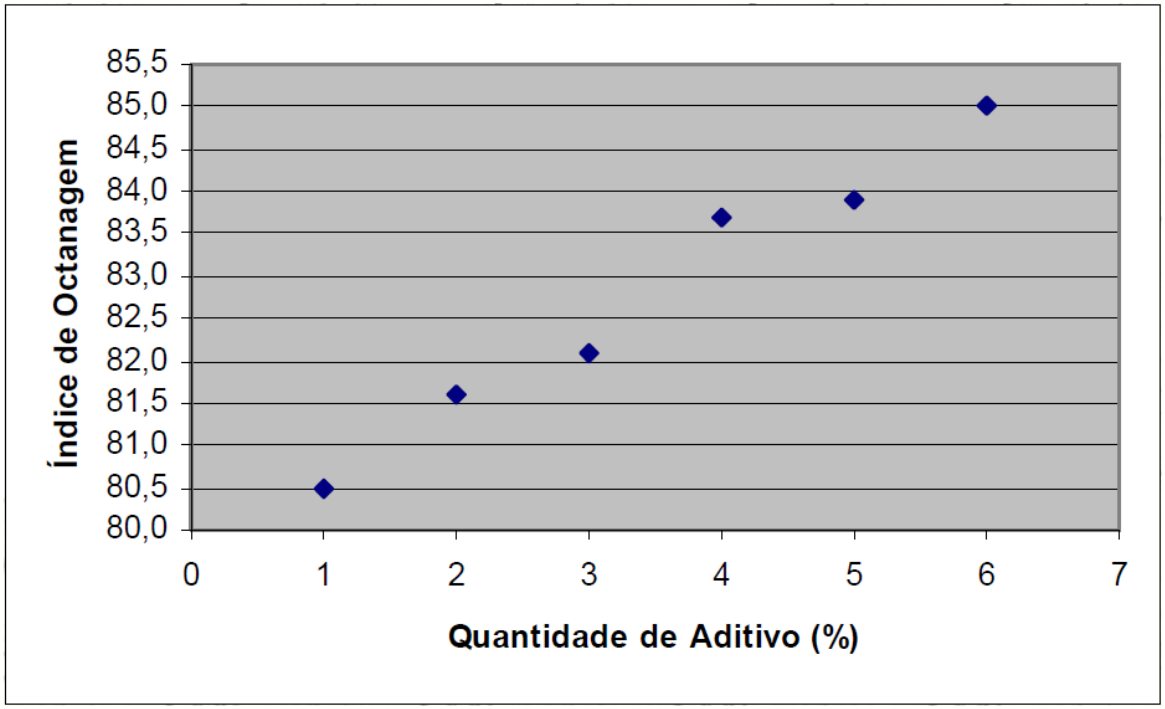
Gasolina × Aditivo

Considere um experimento em que se analisa a octanagem da gasolina (Y) em função da adição de um aditivo (X).

Para isto, foram realizados ensaios com os percentuais de 1, 2, 3, 4, 5 e 6% de aditivo.

Gasolina × Aditivo

X	Y
1	80,5
2	81,6
3	82,1
4	83,7
5	83,9
6	85,0



Gasolina × Aditivo

	x_i	y_i	x_i^2	$x_i y_i$
	1	80,5	1	80,5
	2	81,6	4	163,2
	3	82,1	9	246,3
	4	83,7	16	334,8
	5	83,9	25	419,5
	6	85,0	36	510,0
Soma	21	496,8	91	1.754,3

Gasolina $\times$ Aditivo

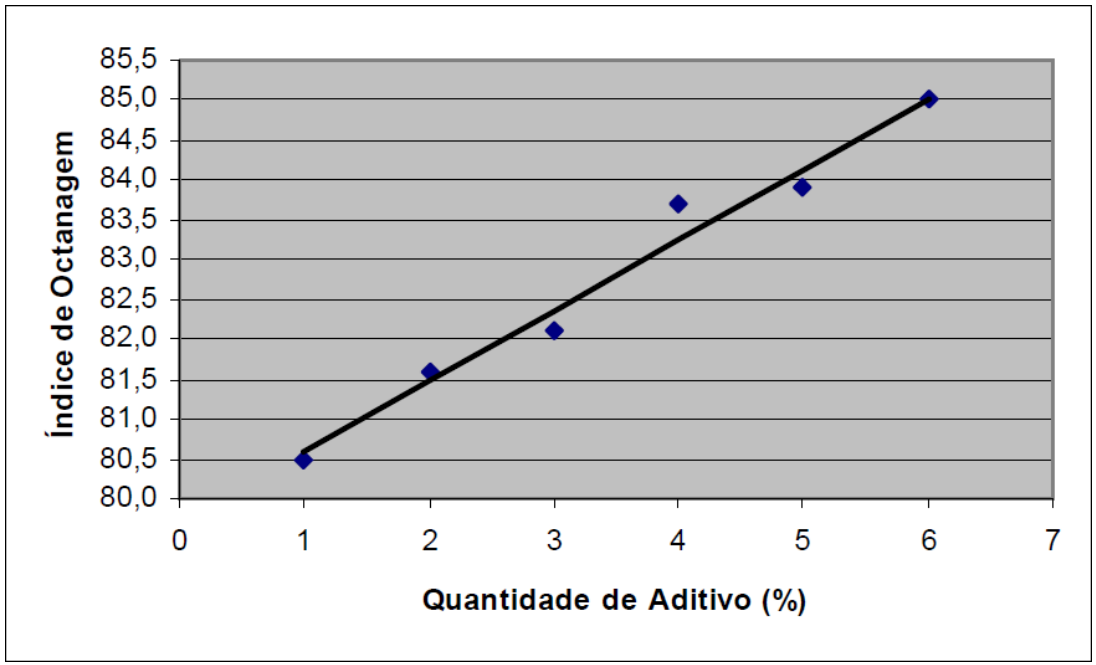
$$\beta = \frac{6 \cdot (1754.3) - (21) \cdot (496.8)}{6 \cdot (91) - (21)^2} = \frac{93}{105} = 0.886$$

$$\alpha = \frac{496.8 - (0.886) \cdot (21)}{6} = 79.7$$

$$y = 79.7 + 0.886x$$

Gasolina × Aditivo

$$y = 79.7 + 0.886x$$



Regressão Linear Múltipla

A intuição nos diz que, geralmente, podemos melhorar uma predição se incluirmos novas variáveis independentes ao modelo (equação) de regressão.

Uma reta é um polinômio de ordem um.

Podemos, então, considerar o uso de modelos polinomiais de ordem maior que um.

Regressão Linear Múltipla

Antes de tudo, devemos buscar o equilíbrio entre o número de parâmetros e a capacidade preditiva do modelo.

Um número excessivo de parâmetros pode causar um **sobreajuste**, tornando o modelo muito específico.

Um número reduzido de parâmetros pode causar um **subajuste**, tornando o modelo pouco preditivo.

Regressão Linear Múltipla

A **regressão linear múltipla** funciona de forma parecida com a regressão linear simples.

Basicamente, ela leva em consideração diversas variáveis de entrada x_i, $i = 1, \ldots, p$, influenciando ao mesmo tempo uma única variável de saída, y.

Exemplo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon.$$

Regressão Linear Múltipla

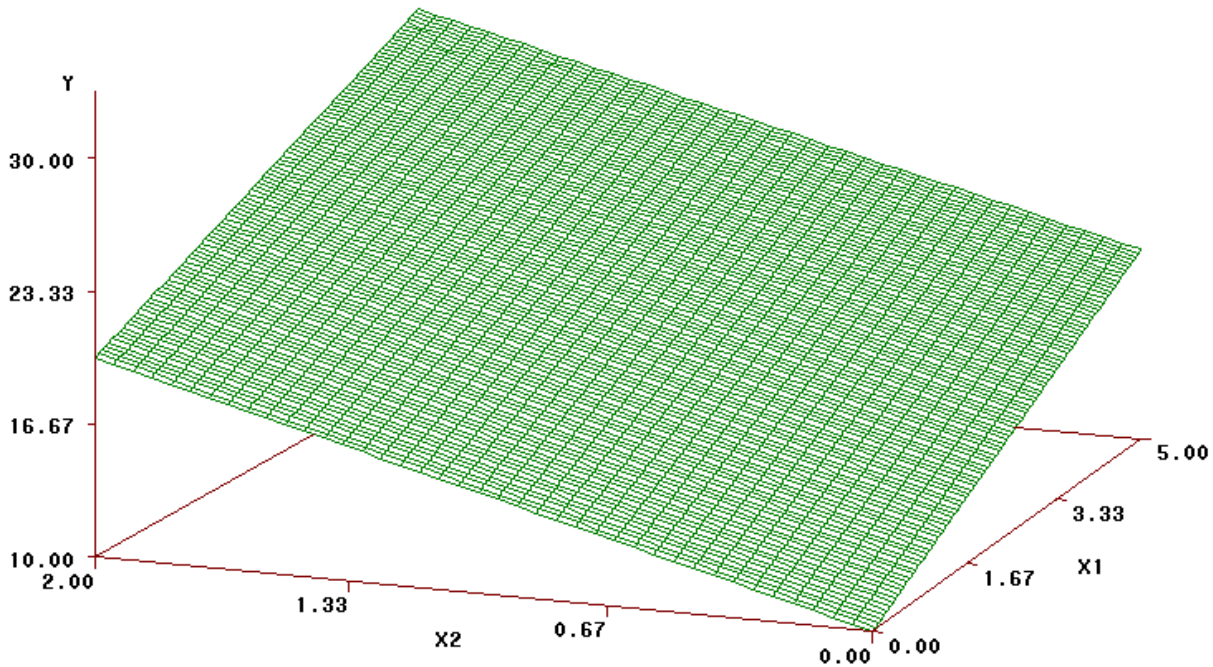
A função de regressão na regressão múltipla é chamada de **superfície de resposta**.

Ela descreve um hiperplano no espaço p-dimensional das variáveis de entrada x_i.

Os parâmetros β_i, $i = 0, \ldots, p$ são os **coeficientes de regressão**.

Regressão Linear Múltipla

$$y = 10 + 2x_1 + 5x_2$$



Regressão Linear Múltipla

Para calcularmos a superfície de regressão, usamos o **método dos mínimos quadrados**, como feito com a regressão linear simples, para estimar os coeficientes de regressão β_i, $i = 0, \ldots, p$.

Temos, então, n equações na forma $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$, uma para cada observação dos dados.

Regressão Linear Múltipla

Podemos expressar as operações matemáticas utilizando notação matricial

$$y = X\beta + \epsilon,$$

ou seja

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{12} & \dots & x_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \dots & x_{pn} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Regressão Linear Múltipla

Nosso objetivo é fazer com que a soma dos quadrados dos resíduos entre os valores medidos (observações) e a superfície de regressão seja mínima.

A solução continua a mesma: procurar pelos parâmetros β_i, $i = 0, \ldots, p$ que minimizem a soma dos quadrados dos resíduos

$$\sum_{i=1}^{n} \epsilon_i^2.$$

Regressão Linear Múltipla

Algumas deduções matemáticas e substituições e temos que

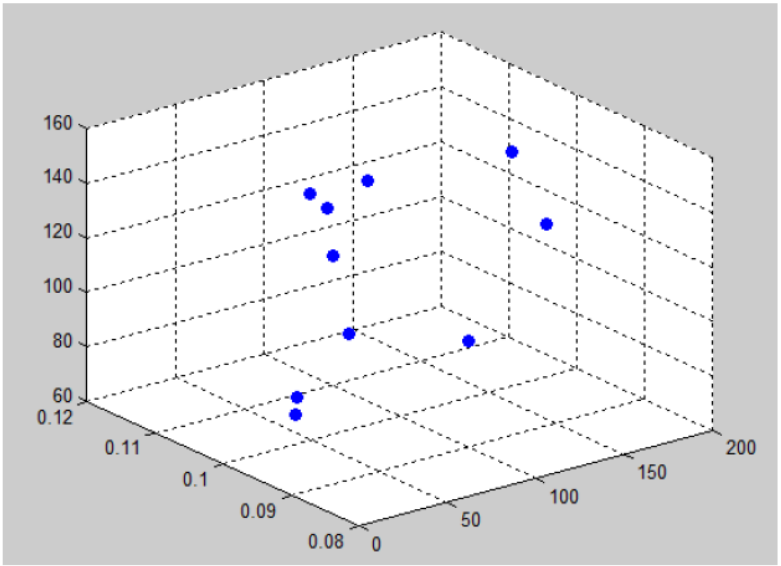
$$\beta = (X^T X)^{-1} X^T y,$$

em que A^{-1} representa a matriz inversa da matriz A.

Regressão Linear Múltipla

Calcular a regressão para o seguinte conjunto de dados.

y	x_1	x_2
122	139	0,115
114	126	0,12
86	90	0,105
134	144	0,09
146	163	0,1
107	136	0,12
68	61	0,105
117	62	0,08
71	41	0,1
98	120	0,115



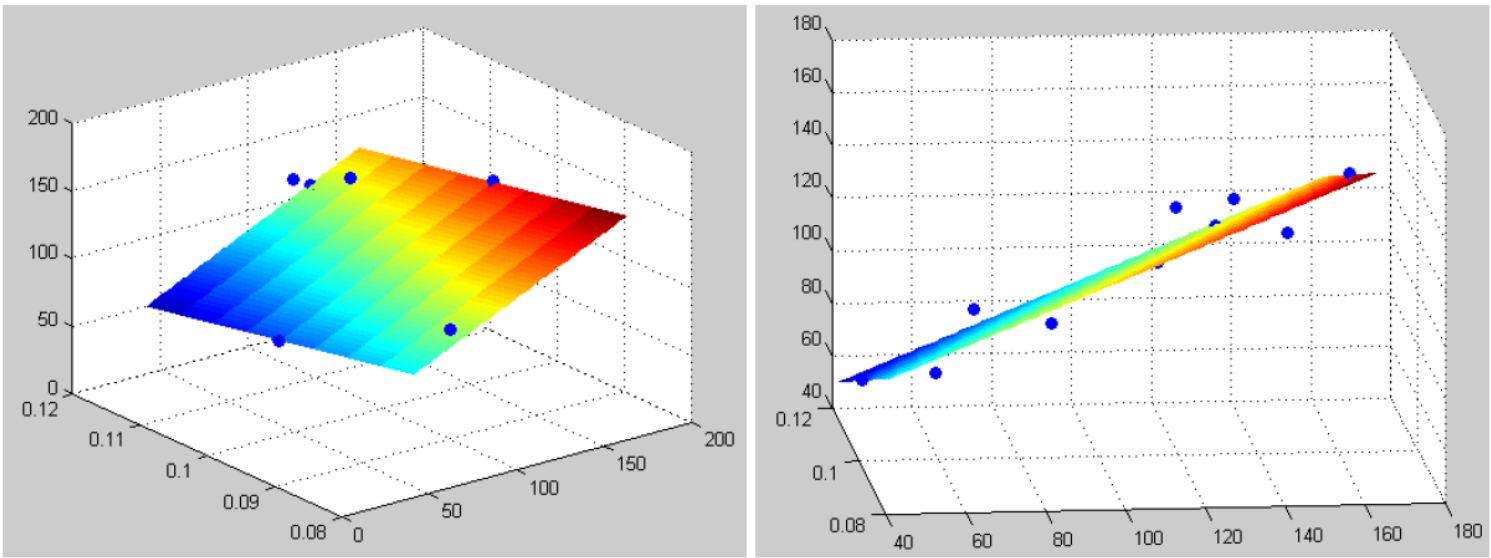
Regressão Linear Múltipla

Solução do sistema:

$$\beta = (148.52, 0.6136, -1034.41),$$

$$y = 148.52 + 0.6136x - 1034.41x^2.$$

Regressão Linear Múltipla



Observações

Nem sempre é possível calcular a inversa da matriz $X^T X$.

Seu determinante muitas vezes é zero ou quase igual a zero.

Isto geralmente ocorre quando as variáveis de entrada são intercorrelacionadas.

Se a intercorrelação for grande, existe multicolinearidade: as linhas da matriz $(X^T X)$ não são linearmente independentes.

Regressão Não Linear

A **regressão não linear** é uma forma de regressão em que os dados são modelados por uma função que é uma combinação não linear de parâmetros.

Pelo menos um dos seus parâmetros deve estar na forma não linear.

Exemplos:

- Função exponencial: $y = \alpha e^{\beta x}$
- Função logarítmica: $y = \alpha + \beta \log x$
- Função potência: $y = \alpha x^{\beta}$

Regressão Não Linear

- Aplicações biológicas são modeladas por meio de relações não lineares.
- Modelos de crescimento.
- Modelos de rendimento.

Regressão Não Linear

Podemos tentar transformar uma relação não linear em linear (transformação linearizante). Em seguida, resolvemos o problema linear.

Exemplo:

- Relação exponencial: $y = \alpha e^{\beta x}$
- Modelo linear:

$$y' = \alpha' \beta x,$$

em que $y' = \log y$ e $\alpha' = \log \alpha$.

Regressão Não Linear

Nem sempre é possível fazer essa transformação. Algumas relações não lineares são não linearizáveis.

Além disso, estimar os parâmetros na relação linearizada não produz os mesmos resultados que estimar os parâmetros na relação não linear original.

Regressão Não Linear

Como na regressão linear, os dados são ajustados geralmente pelo **método dos mínimos quadrados**. Isso vale para relações linearizadas ou não.

Podemos, também, usar um método de aproximações sucessivas, como o **método de Gauss-Newton**.

Regressão Não Linear

O modelo de regressão não precisa ser uma linha reta.

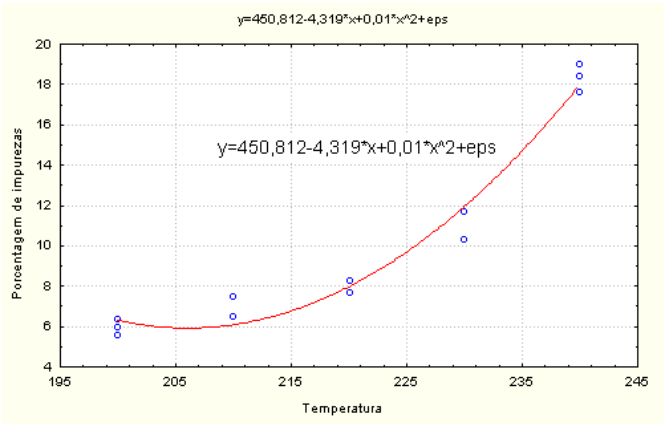
O modelo a seguir chama-se modelo quadrático, ou de segundo grau, e sua figura é uma parábola.

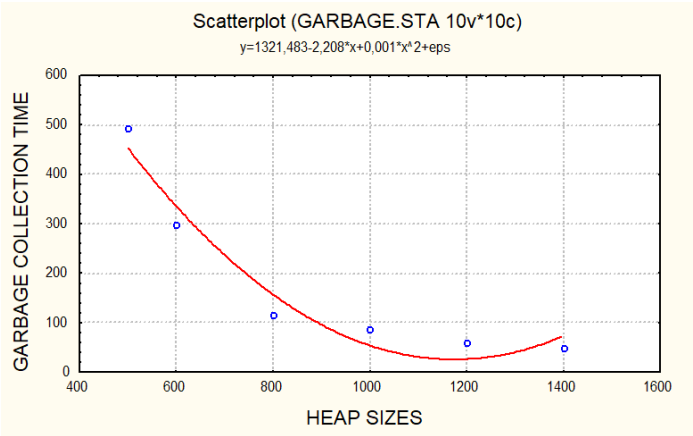
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

Esse modelo, embora não seja uma linha reta, continua sendo um modelo linear nos parâmetros.

O método discutido para o modelo de regressão linear simples aplica-se diretamente aos demais modelos lineares nos parâmetros.

Exemplos


y=450,812-4,319*x+0,01*x^2+eps
y=450,812-4,319*x+0,01*x^2+eps
Porcentagem de impurezas
Temperatura
20
18
16
14
12
10
8
6
4
195
205
215
225
235
245


Scatterplot (GARBAGE.STA 10v*10c)
y=1321,483-2,208*x+0,001*x^2+eps
GARBAGE COLLECTION TIME
HEAP SIZES
600
500
400
300
200
100
0
400
600
800
1000
1200
1400
1600

Regressão Não Linear

Existem, porém, modelos não lineares nos parâmetros.

Exemplo: modelo de crescimento logístico, em que x é o tempo.

$$y = \frac{\beta_1}{1 + \beta_2 e^{\beta_3 x}} + \epsilon$$