Análise de Regressão

Tópicos em Avaliação de Desempenho de Sistemas

Aline Oliveira aso2@cin.ufpe.br
Camila Araujo cga2@cin.ufpe.br
Iure Fé isf2@cin.ufpe.br
Janailda jbs4@cin.ufpe.br

UNIVERSIDADE FEDERAL DE PERNAMBUCO

Agenda

Parte I:

- Contextualização
- Modelo de Regressão
- Regressão Linear
- Linear Simples
  - Métodos mínimos quadrados
- Linear Múltipla
- Inferência

Parte II:

- Exercícios práticos em sala

Agenda

Objetivos

- Use regressão linear simples para a construção de modelos empíricos para engenharia e dados científicos
- Entenda como o método dos mínimos quadrados é usado para estimar os parâmetros de uma forma linear modelo de regressão
- Use o modelo de regressão para fazer uma previsão de uma observação futuro e
- Representar graficamente a relação entre as variáveis de um estudo e a reta de regressão a partir da equação de regressão obtida.
- Testar a significância do coeficiente de correlação obtido em um estudo de regressão linear.

Metodologia:

- Minitab
- Excel
- Mathematica
- Lista de exercício n. 04, com entrega para dia 12/10/2015

Referências

Regression Analysis, F. Graybill, H. K. Iyer, Duxbury Press, 1994.

Applied Statistics and Probability for Engineers, Third Edition, Douglas C. Montgomery,George C. Runger, John Wiley & Sons, Inc.

Contextualização

História:

Este modelo teve origem nos trabalhos de astronomia elaborados por Gauss no período de 1809 a 1821. O termo regressão foi utilizado pela primeira vez por Galton, por volta de 1885, quando investigava relações entre características antropométricas de sucessivas gerações. Ele observou, dentre outros fatos, que os filhos apresentavam as mesmas características dos seus pais, porém em uma intensidade menor. Por exemplo: pais com estatura baixa têm filhos de estatura baixa, mas, em média, a estatura destes é maior. O mesmo ocorre, mas em direção contrária, para pais com estatura alta. Este fenômeno, da altura dos filhos moverem-se em direção a altura média de todos os homens, ele denominou de regressão.

Atualmente, a análise de regressão é uma das mais importantes técnicas estatísticas, sendo utilizada em aplicações de diversas áreas como: Engenharia, Medicina, Economia, etc.

Contextualização

Definição:

(RAJ JAIN, 1991):
O modelo de regressão é um dos métodos estatísticos mais usados para investigar a relação entre variáveis.

(GRAYBILL & IYER, 1994, p. 1):
Área da Estatística que lida com métodos para investigação da existência de associações entre várias quantidades observáveis e, se presente, a natureza das associações.

Contextualização

Relação entre as variáveis

Modelos de regressão são modelos matemáticos que relacionam o comportamento de uma variável Y com outra X. Quando a função f que relaciona duas variáveis é do tipo f (X) = a + b X temos o modelo de regressão simples. A variável X é a variável independente da equação enquanto Y = f (X) é a variável dependente das variações de X. O modelo de regressão é chamado de simples quando envolve uma relação causal entre duas variáveis. O modelo de regressão é multiplo quando envolve uma relação causal com mais de duas variáveis. Isto é, quando o comportamento de Y é explicado por mais de uma variável independe X1, X2, ...Xn.

Contextualização

Relação entre as variáveis

Para que serve determinar a relação entre duas variáveis?

1 - Para realizar previsões sobre o comportamento futuro de algum fenômeno da realidade. Neste caso extrapola-se para o futuro as relações de causa-efeito – já observadas no passado – entre as variáveis. Pode-se, por exemplo, prever a população futura de uma cidade simulando a tendência de crescimento da população no passado.
2 - Pesquisadores interessados em simular os efeitos sobre uma variável Y em decorrência de alterações introduzidas nos valores de uma variável X também usam este modelo. Por exemplo: de que modo a produtividade (Y) de uma área agrícola é alterada quando se aplica certa quantidade (X) de fertilizante sobre a terra.

Contextualização

Diagrama de dispersão

A maneira mais simples de se estudar a relação entre duas variáveis é fazendo um gráfico denominado Diagrama de Dispersão.

- Coletar pares de dados das variáveis x e y que se pretende estudar;
- Traçar um sistema de eixos cartesianos e represente uma variável em cada eixo;
- Estabeleça as escalas de maneira a dar ao diagrama o aspecto de um quadrado;
- Escreva os nomes das variáveis nos respectivos eixos e depois faça as graduações;
- Fazer um ponto para representar cada par de valores x e y;
- Escreva o título e complemente com uma legenda.

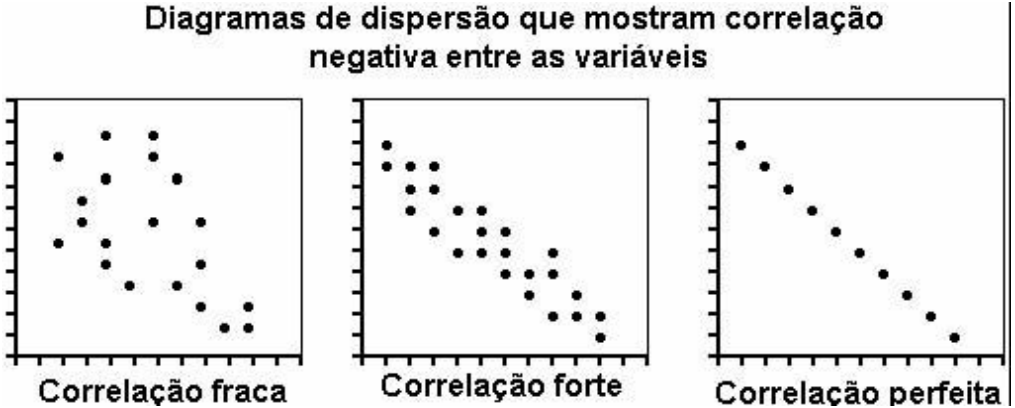
Contextualização

Diagrama de Dispersão

Se x e y crescem no mesmo sentido, existe uma correlação positiva entre as variáveis, que será tanto maior quanto menor fora a dispersão dos pontos.

Contextualização

Diagrama de Dispersão

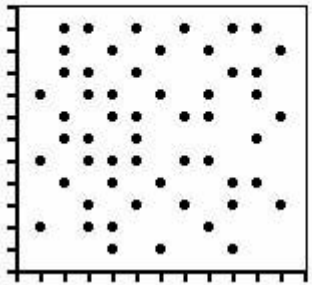


Se x e y variam em sentidos contrários, existe correlação negativa entre as variáveis. Essa correlação é tanto maior quanto menor é a dispersão dos pontos.

Contextualização

Diagrama de Dispersão

Diagram ade dispersão que mostra correlação nula entre variáveis



Se x cresce e y varia ao acaso, não existe correlação entre as variáveis ou o que é o mesmo a correlação entre elas é nula.

Contextualização

Modelos de Regressão são construídos com os objetivos:

i) Predição - Uma vez que esperamos que grande parte da variação da variável de saída seja explicada pelas variáveis de entrada, podemos utilizar o modelo para obter valores de Y correspondentes a valores de X que não estavam entre os dados. Esse procedimento é chamado de predição e, em geral, usamos valores de X que estão dentro do intervalo de variação estudado. A utilização de valores fora desse intervalo recebe o nome de extrapolação e deve ser usada com muito cuidado, pois, o modelo adotado pode não ser correto fora do intervalo estudado. Acredita-se que a predição seja a aplicação comum dos modelos de regressão;

Contextualização

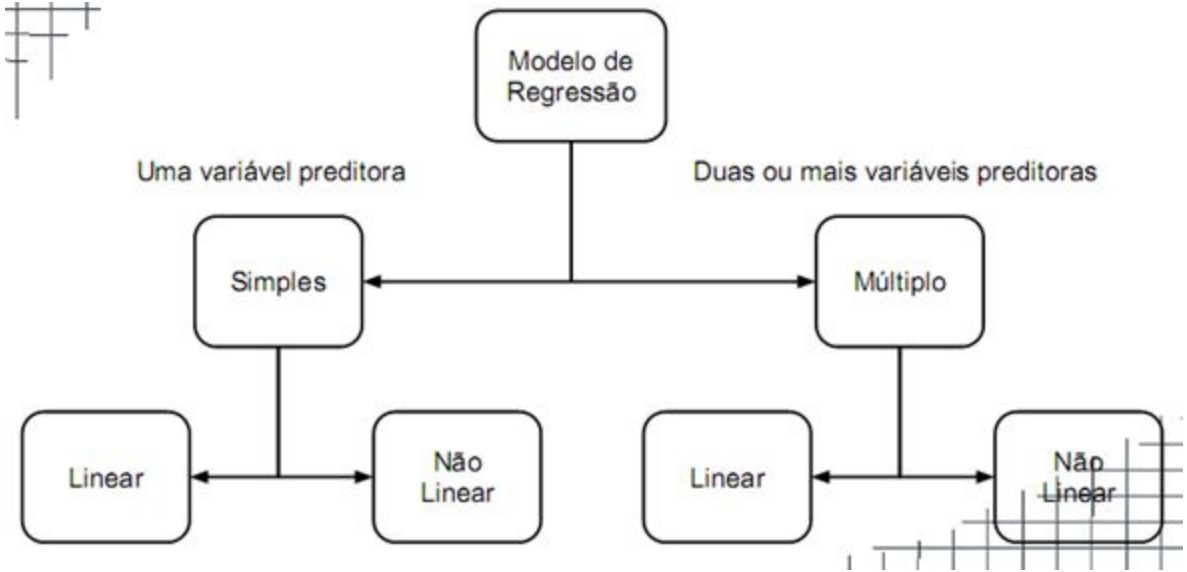
Modelos de Regressão são construídos com os objetivos:

ii) Seleção de variáveis - Frequentemente, não se tem idéia de quais são as variáveis que afetam significativamente a variação de Y. Para responder a esse tipo de questão, estudos são realizados com um grande número de variáveis. A análise de regressão pode auxiliar no processo de seleção de variáveis eliminando aquelas cuja contribuição não seja importante;

iii) Estimação de parâmetros - Dado um modelo e um conjunto de dados referente às variáveis resposta e preditoras, estimar parâmetros ou ajustar um modelo aos dados significa obter valores ou estimativas para os parâmetros, por algum processo, tendo por base o modelo e os dados observados;

iv) Inferência - O ajuste de um modelo de regressão em geral tem por objetivos básicos, além de estimar os parâmetros, realizar inferências sobre eles, tais como, testes de hipóteses e intervalos de confiança.

Modelos de Regressão



Regressão Linear Simples

A análise de **regressão linear *simples*** consiste em achar uma reta que **relacione *duas* variáveis quantitativas**;

Relação entre a variável de **resposta Y** e uma variável **preditora X;**

Exemplos:

- Relação entre nível de escolaridade e renda? Renda (Y) e Escolaridade (X)
- Relação entre anos de estudos e salário? Salário (Y) e Anos de Estudos (X)
- Associação entre tempo de estudo e nota na prova? Nota (Y) e Tempo de Estudo (X)
- Prever a satisfação de um aluno dado o seu desempenho acadêmico? Satisfação (Y) e Desempenho (X)

Duas variáveis estão relacionadas, se a mudança de uma provoca a mudança na outra.

Investigaremos a presença ou ausência de relação linear sob dois pontos de vista:

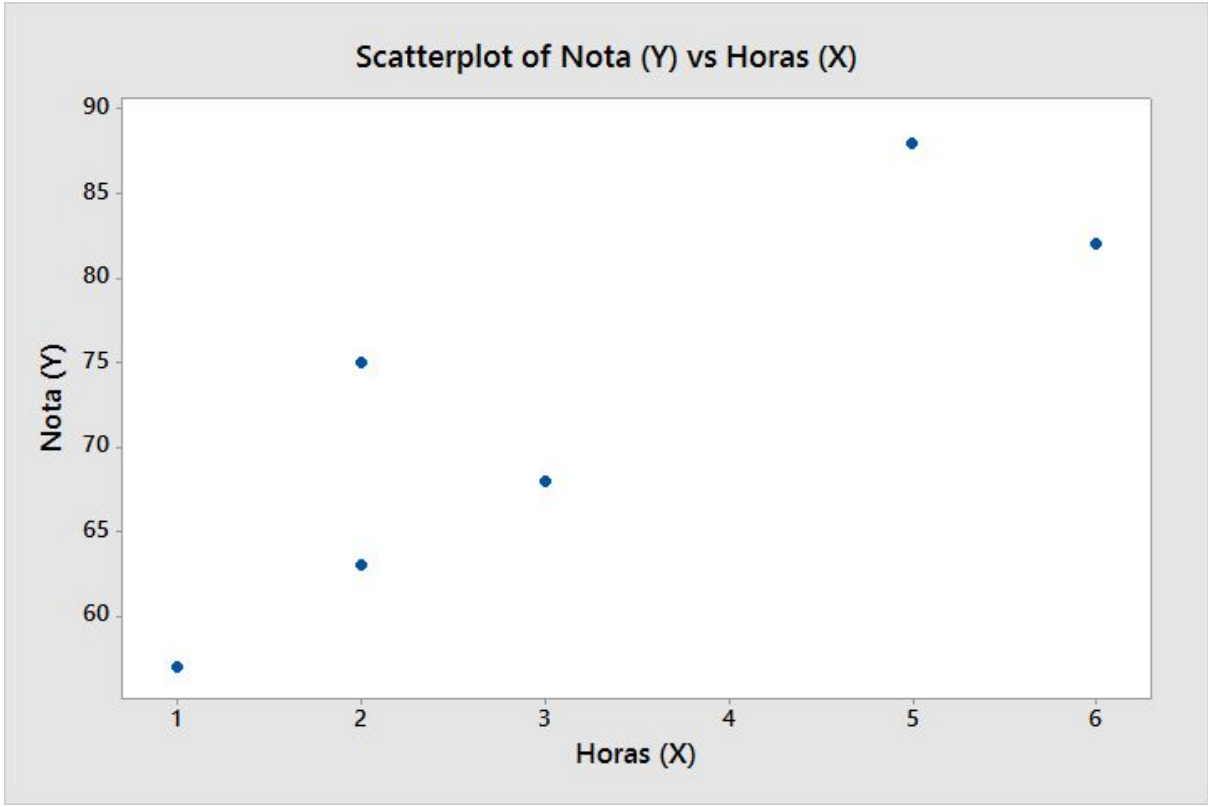
A CORRELAÇÃO mede a força, ou grau, de relacionamento entre duas variáveis; a REGRESSÃO dá uma equação que descreve o relacionamento em termos matemáticos.

Regressão Linear Simples

Exemplo 1: Relação entre tempo de estudo e nota na prova?

- Y: nota na Prova (Variável Resposta)
- X: horas de Estudo (Variável Preditora)

Aluno	Horas (X)	Nota (Y)
A	6	82
B	2	63
C	1	57
D	5	88
E	3	68
F	2	75



Regressão Linear Simples

Exemplo 2: O rendimento do produto está relacionado com a temperatura do processo?

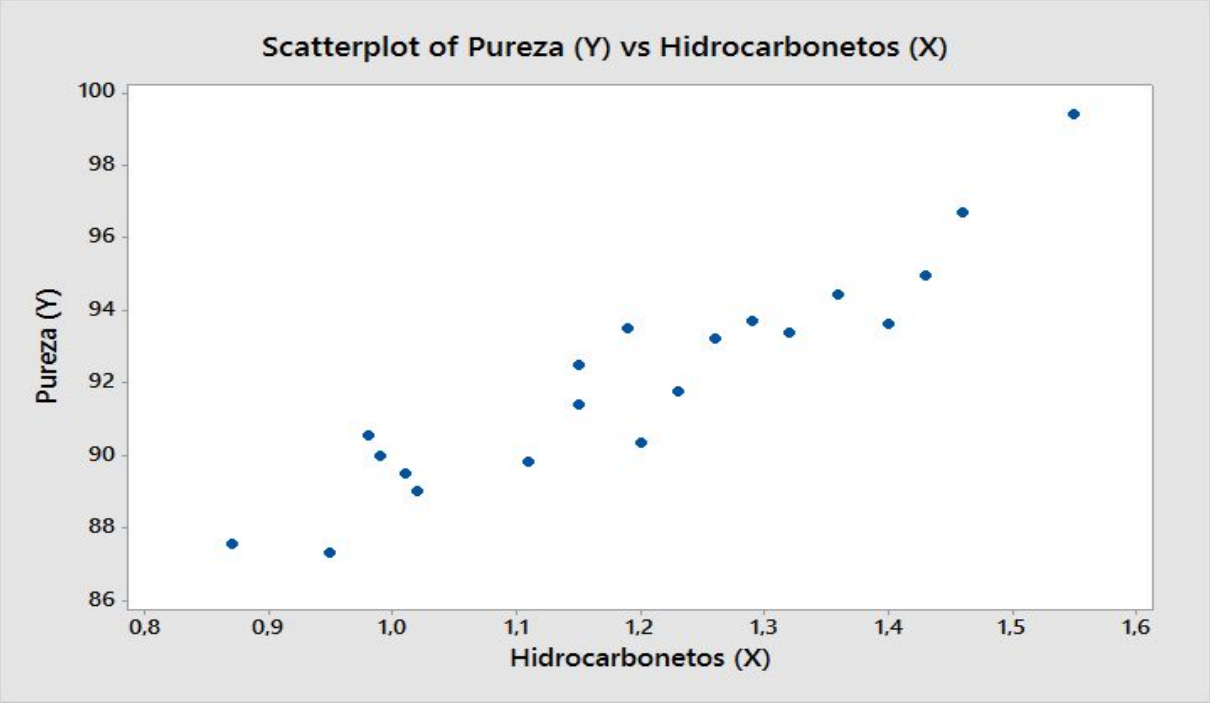
Por exemplo, em um processo químico, suponha que o rendimento do produto está relacionada com a temperatura do processo operacional. A análise de regressão pode ser utilizado para construir um modelo para predizer o rendimento num dado nível de temperatura. Este modelo também pode ser utilizado para otimização de processos, encontrando o nível de temperatura que maximiza o rendimento, ou para fins de controlo do processo.

É possível prever rendimento para uma dada temperatura ?
Esse modelo pode ser usado na otimização do processo?

Regressão Linear Simples

Exemplo 2: O rendimento do produto está relacionado com a temperatura do processo?

- Y: pureza do oxigênio produzido em processo químico de destilação
- X: porcentagem de hidrocarbonetos presentes no condensador



N.	Hidrocarbonetos	Pureza
1	0,99	90,01
2	1,02	89,05
3	1,15	91,43
4	1,29	93,74
5	1,46	96,73
6	1,36	94,45
7	0,87	87,59
8	1,23	91,77
9	1,55	99,42
10	1,40	93,65
11	1,19	93,54
12	1,15	92,52
13	0,98	90,56
14	1,01	89,54
15	1,11	89,85
16	1,20	90,39
17	1,26	93,25
18	1,32	93,41
19	1,43	94,98
20	0,95	87,33

Regressão Linear Simples

COEFICIENTE DE CORRELAÇÃO

Mede a intensidade e a direção da relação linear entre duas variáveis quantitativas. Chamado também de Coeficiente de Correlação de Pearson (Karl Pearson, 1857-1936).

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

r - mede o grau de relacionamento linear entre valores emparelhados x e y em uma amostra.

Quanto mais próximo de –1: correlação negativa (X↑ Y ↓)
Quanto mais próximo de 1: maior correlação positiva (X↑ Y↑)
Quanto mais próximo de 0: menor a correlação linear

Regressão Linear Simples

COEFICIENTE DE CORRELAÇÃO

Exemplo 3: nota da prova e tempo de estudo

X : tempo de estudo (em horas)
Y : nota da prova

Pares de observações (X_i , Y_i) para cada estudante

Tempo(X)	Nota(Y)
3	4,5
7	6,5
2	3,7
1,5	4,0
12	9,3

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Regressão Linear Simples

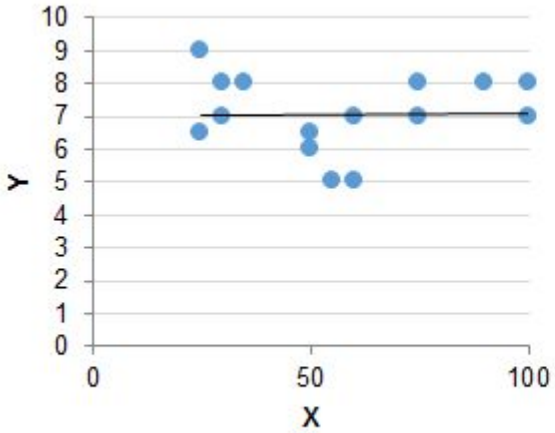
Tempo X	Nota Y	(X-médiaX)	(Y-médiaY)	(X-médiaX)*(Y-médiaY)	(X-média X)^2	(Y-média Y)^2
3	4,5	-2,1	-1,1	2,31	4,41	1,21
7	6,5	1,9	0,9	1,71	3,61	0,81
2	3,7	-3,1	-1,9	5,89	9,61	3,61
1,5	4	-3,6	-1,6	5,76	12,96	2,56
12	9,3	6,9	3,7	25,53	47,61	13,69
25,5	28			41,2	78,2	21,8

Média X = 5,1
Média Y = 5,6

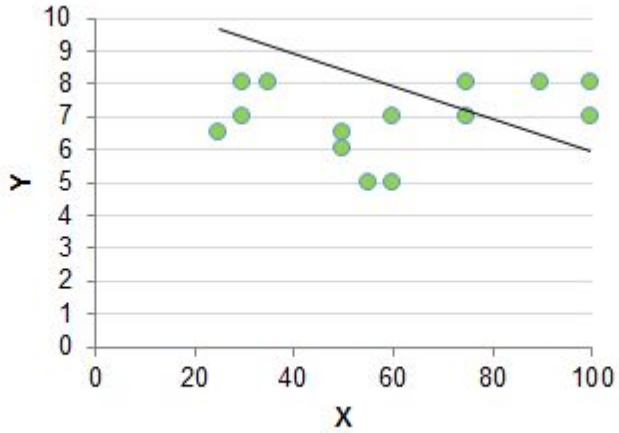
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$r = \frac{41{,}2}{\sqrt{78{,}2 * 21{,}8}} = 0{,}996$$

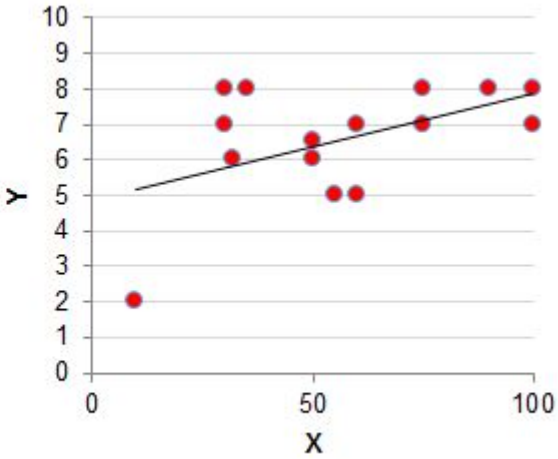
Regressão Linear Simples



Menor Correlação Linear:
R=0,0123

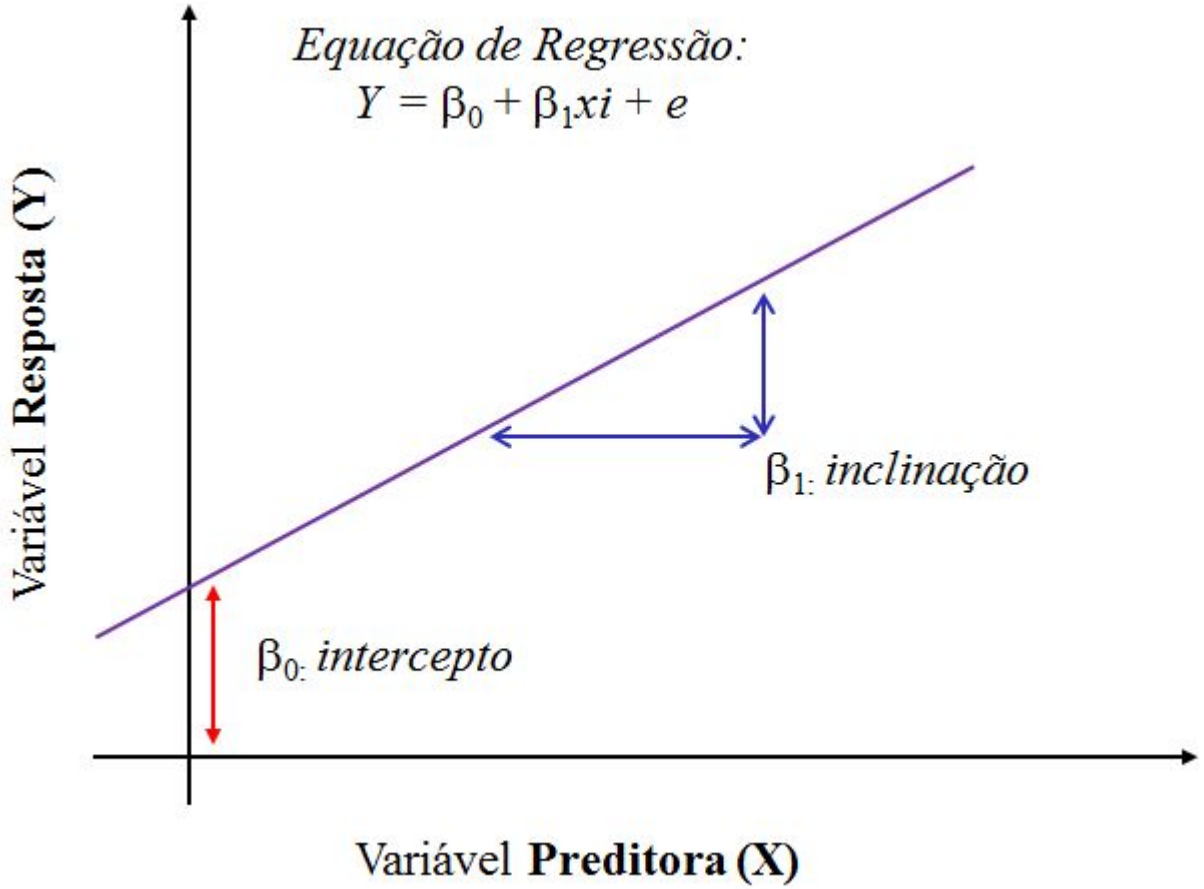


Correlação negativa (X↑ Y ↓):
R= -0,2902

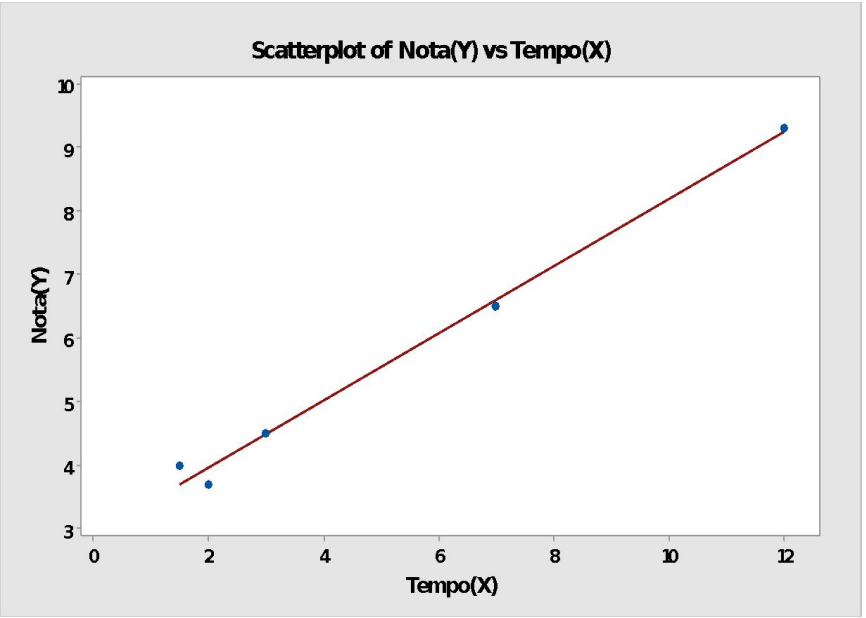


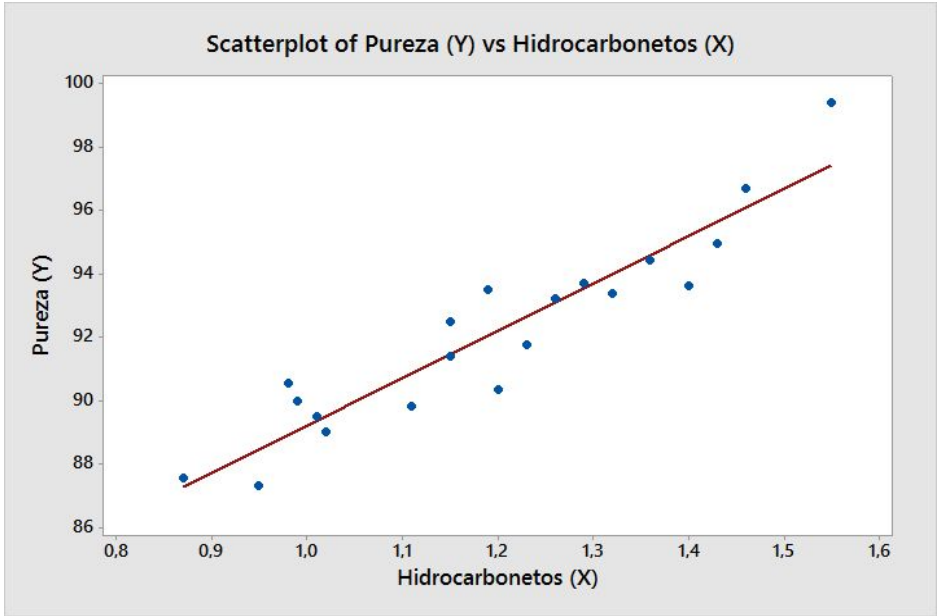
Correlação positiva (X↑ Y↑):
R=0,5458

Regressão Linear Simples



Regressão Linear Simples





Métodos dos Mínimos Quadrados

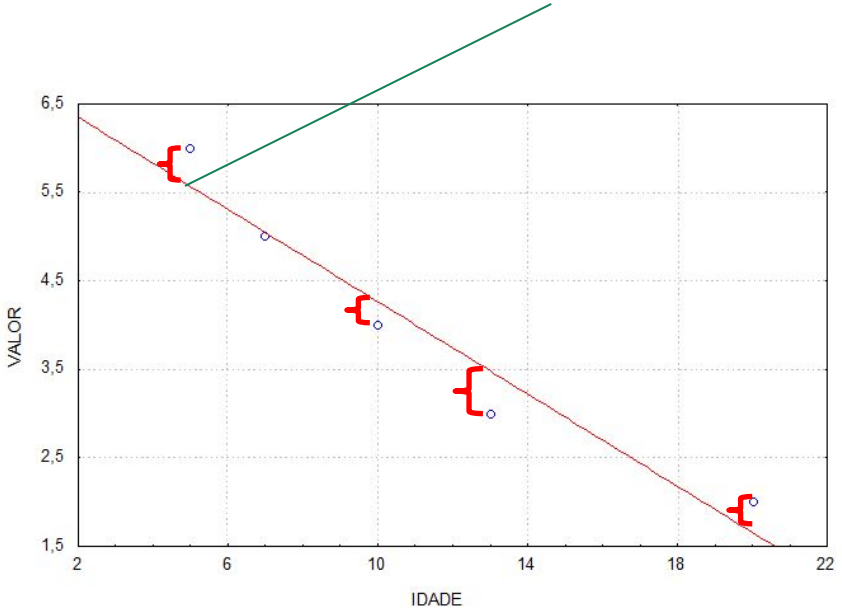
Para observações (X_i, Y_i) $i=1,..,n$, temos o modelo

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \mathrm{i} = 1,.., \mathrm{n}$$

Minimizar

$$Q = \sum_{i=1}^{n} \left(Y_i - \hat{y}\right)^2 = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)^2$$

$$Y_i - (\beta_0 + \beta_1 X_i)$$



Métodos dos Mínimos Quadrados

Derivando-se em relação a B0 e B1, igualando-se a 0 para encontrar os valores que minimizam Q.

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i(Y_i - \beta_0 - \beta_1 X_i)$$

Métodos dos Mínimos Quadrados

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

$$b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right) = \bar{Y} - b_1 \bar{X}$$

Métodos dos Mínimos Quadrados

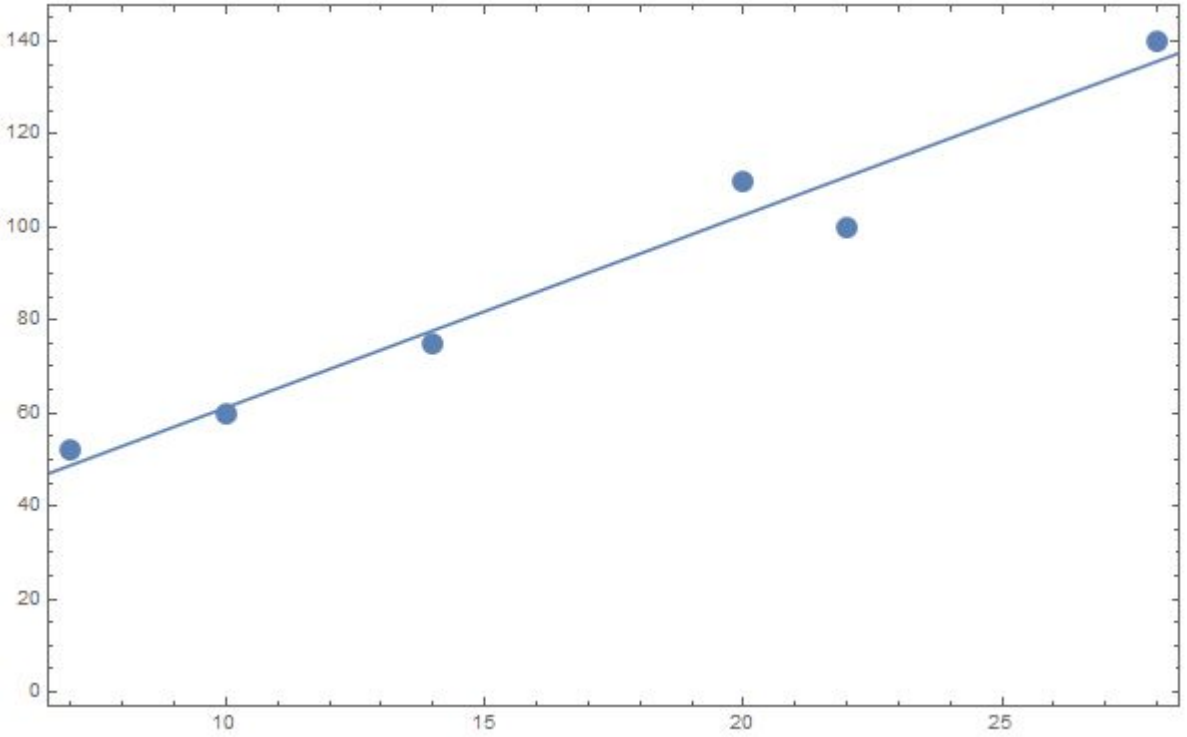
Considerando os valores da variável da variável preditora X a memória RAM e Y a quantidade de programas suportados.

x	y
28	140
20	110
22	100
14	75
10	60
7	52

Métodos dos Mínimos Quadrados

	x	y	Xi - X	yi-y	(Xi - X)(yi-	(Xi - X)^2
	28	140	11,16667	50,5	563,9167	124,6944
	20	110	3,166667	20,5	64,91667	10,02778
	22	100	5,166667	10,5	54,25	26,69444
	14	75	-2,83333	-14,5	41,08333	8,027778
	10	60	-6,83333	-29,5	201,5833	46,69444
	7	52	-9,83333	-37,5	368,75	96,69444
Total	101	537				
Média	16,83333	89,5				
			b1	4,137986		
			b1	19,8439		

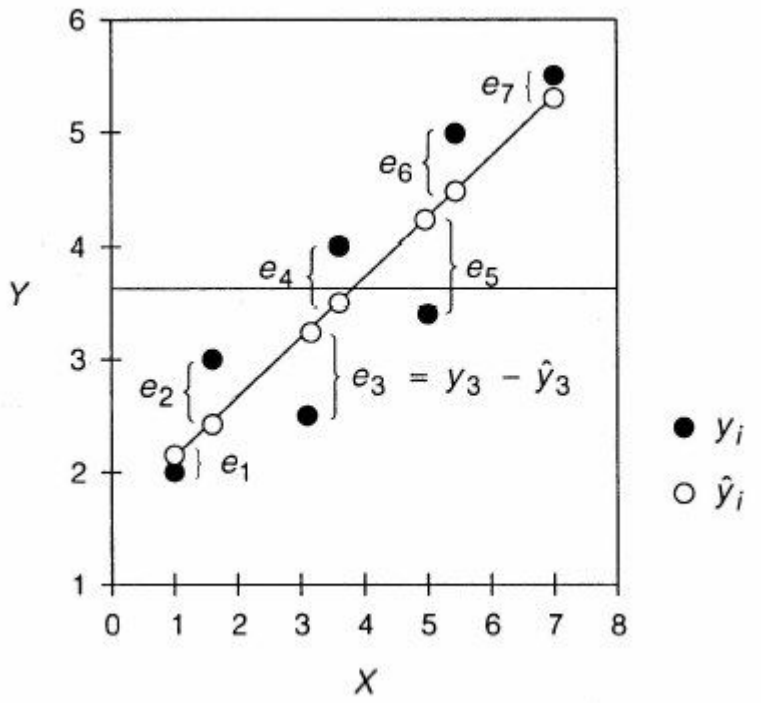
Métodos dos Mínimos Quadrados



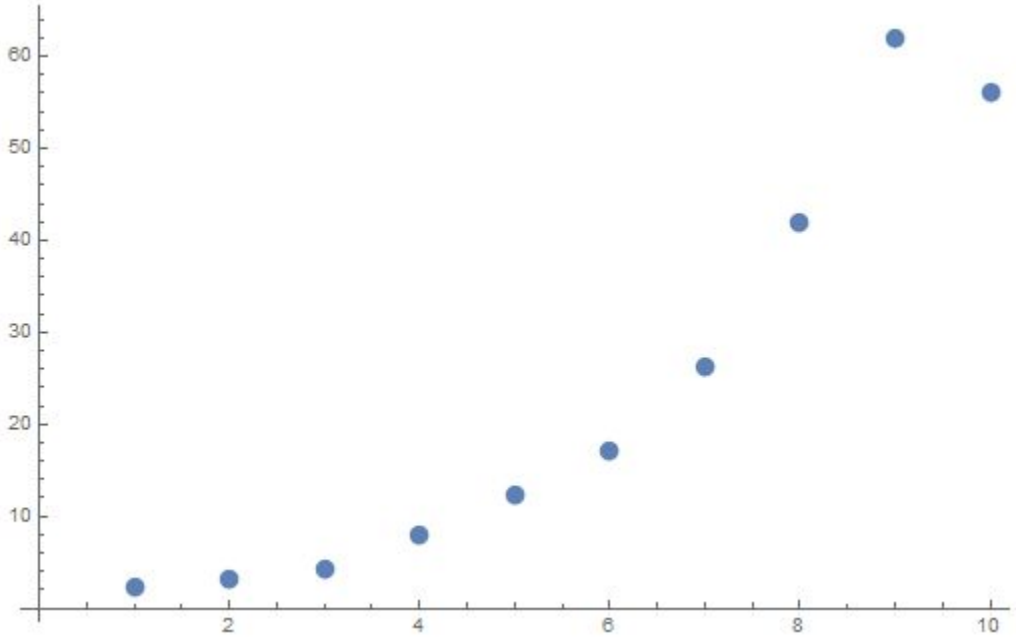
Regressão Linear Simples

Residuos: a diferença entre o valor observado e o estimado pela função

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

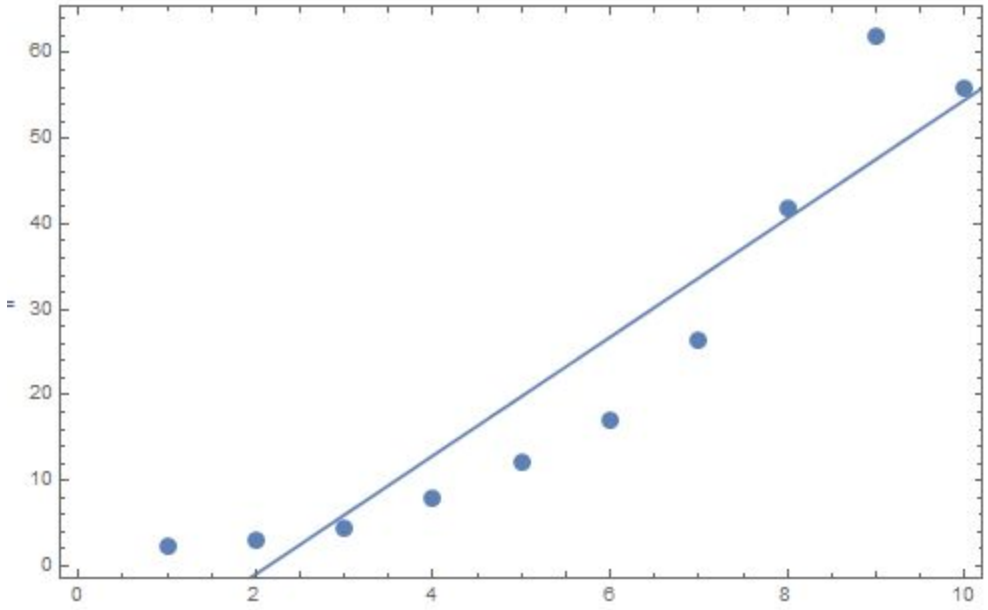


Regressão NÃO Linear Simples

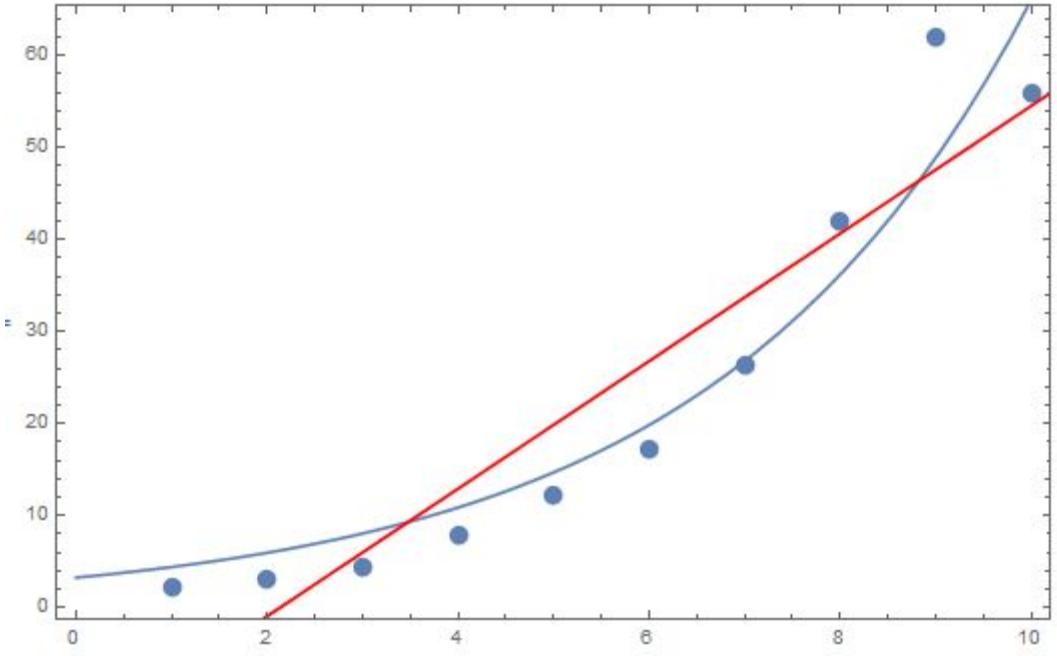


$$Y_i = \gamma_0 \exp(\gamma_1 X_i) + \varepsilon_i$$

Regressão NÃO Linear Simples



Regressão NÃO Linear Simples



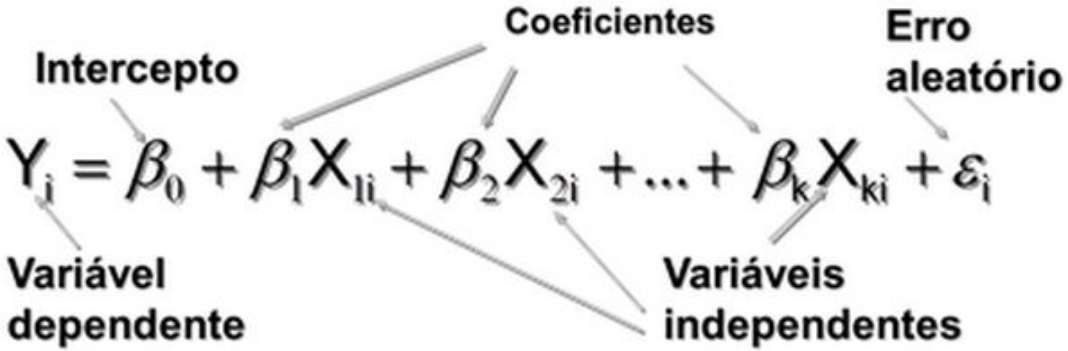
Regressão Linear múltipla

Porque usar a Linear Multipla:

- Para reduzir os resíduos. Reduzindo-se a variância residual (erro padrão da estimativa) aumenta a força dos testes de significância;

- Para eliminar a tendenciosidade que poderia resultar simplesmente ignorássemos uma variável que afeta Y substancialmente.

Regressão Linear Múltipla



i = indexador do indivíduo

Regressão Linear Múltipla

Em uma representação tabular para o modelo expresso na equação:

Observation Number	Response Y	Predictors X_1	X_2	...	X_p
1	y_1	x_{11}	x_{12}	...	x_{1p}
2	y_2	x_{21}	x_{22}	...	x_{2p}
3	y_3	x_{31}	x_{32}	...	x_{3p}
⋮	⋮	⋮	⋮	⋮	⋮
n	y_n	x_{n1}	x_{n2}	...	x_{np}

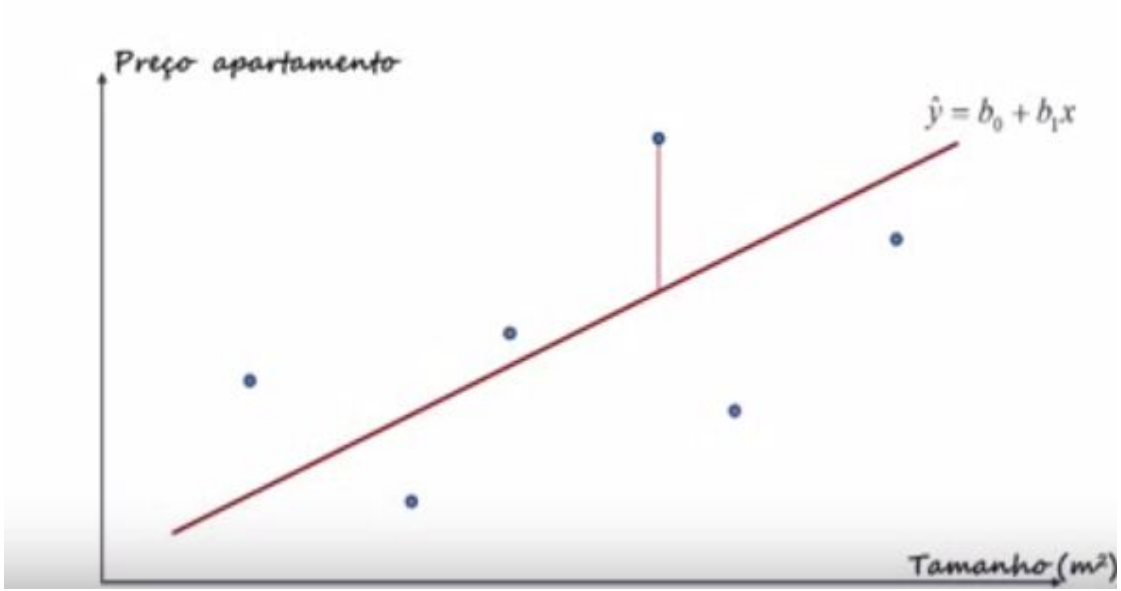
Regressão Linear Múltipla

- Para efetuar a descoberta do valor para os parâmetros (coeficientes de regressão), é necessário aplicar o método dos quadrados mínimos (assim como na regressão linear simples).

$$\mathrm{SSE} = \sum_{i=1}^{n} \varepsilon_i^2$$

Regressão Linear múltipla

Relembrando:



Regressão Linear múltipla

A diferença entre Linear e Multipla é:

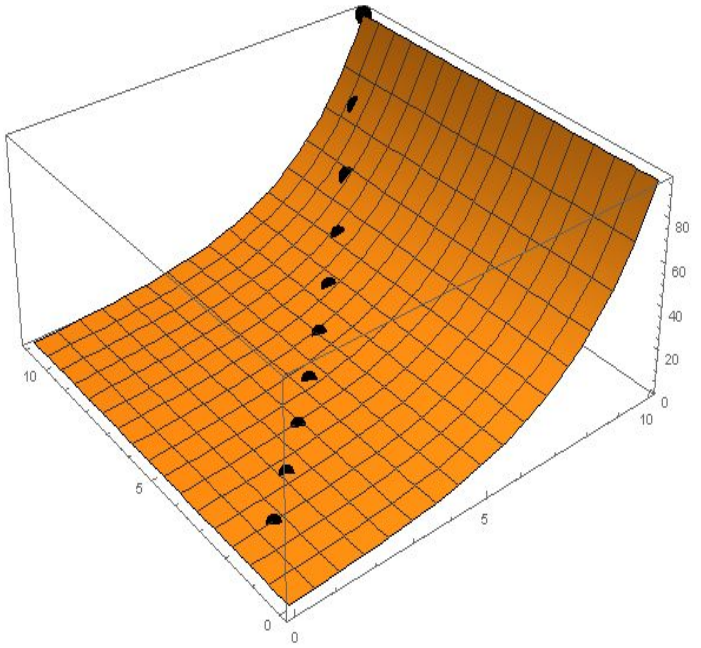
A regressão múltipla envolve **três** ou **mais variáveis**

- 1 **dependente (Apartamento)**
- 3 **ou mais dependentes (Idade,tamanho,Localização)**

Regressão Não Linear Múltipla

Os parâmetros entram na equação de forma não linear:

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\gamma}) + \varepsilon_i$$



Coeficiente de Determinação

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

onde,

$$SQT = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

recebe o nome de coeficiente de determinação que é usado para julgar a adequação do modelo de regressão.

Intervalo de confiança

$$QME = \frac{SQE}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n-2} \quad (\text{é o } Quadrado\ Médio\ dos\ Erros\ (dos\ Resíduos)).$$

$$\widehat{Y}(x_0) + t_{\left(1-\frac{\alpha}{2};n-2\right)} \sqrt{QME\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)},$$