

Métodos Quantitativos

Aula 06

Regressão e Predição (Parte 1)

Roberto Massi de Oliveira
Alex Borges Vieira

Revisão: Coeficiente de Correlação

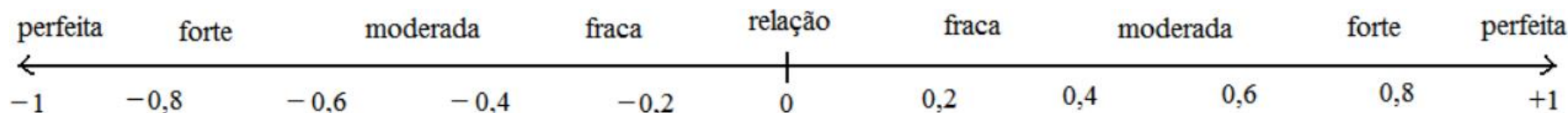
- Quando dispomos das amostras \mathbf{x} e \mathbf{y} de dados bivariados (e.g., peso e altura de um grupo de indivíduos), o coeficiente de correlação é dado por:

$$r_{xy} = \frac{Cov(X, Y)}{S_x S_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

```
1 import numpy as np
2
3 x = [1.6, 1.7, 1.8, 1.9]
4 y = [60, 70, 80, 90]
5 xy = [x, y]
6
7 r = np.corrcoef(xy)
```

Revisão: Coeficiente de Correlação

- Varia de -1,0 a 1,0.



- Quando $r > 0$, à medida que **x** cresce também cresce **y** (em média)
- Quando $r < 0$, à medida que **x** cresce, **y** decresce (em média)

Revisão: Coeficiente de Correlação

- Ex. 07: Um aluno, com bastante dificuldade numa dada disciplina, foi estudando cada vez mais para melhorar suas notas a cada prova e evitar a reprovação. A tabela abaixo resume o número horas estudadas por dia antes de cada uma das 3 provas realizadas e a nota tirada nas mesmas. Qual o coeficiente de correlação entre as horas estudadas por dia e as notas?

Estudo/Dia	1 h	2 h	3 h
Nota	3,0	7,0	9,0

```
1 import numpy as np
2
3 t = [1, 2, 3]
4 n = [3, 7, 10]
5 tn = [t, n]
6
7 r = np.corrcoef(tn)
```

r =	r(t,t)	r(t,n)
	r(n,t)	r(n,n)
r =	1	0,997
	0,997	1

Diagrama de Dispersão

- Ex. 01: Vamos imaginar uma amostra com variáveis **dependentes** resumidas na tabela a seguir:

x (investimento)	y (lucro)
95	1200
125	1000
150	1350
160	1700
200	1800
240	2390

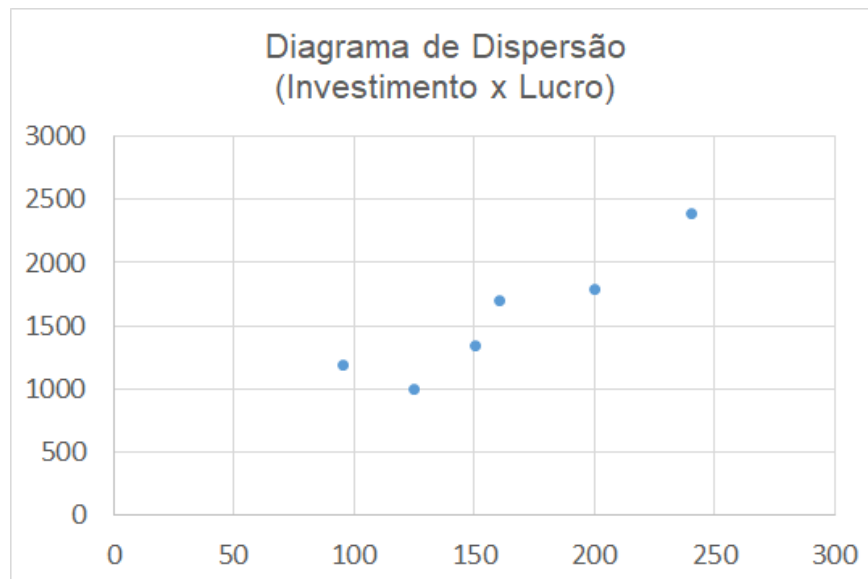


Diagrama de Dispersão

- Ex. 01: Vamos imaginar uma amostra com variáveis **dependentes** resumidas na tabela a seguir (correlação positiva e forte):

x (investimento)	y (lucro)
95	1200
125	1000
150	1350
160	1700
200	1800
240	2390

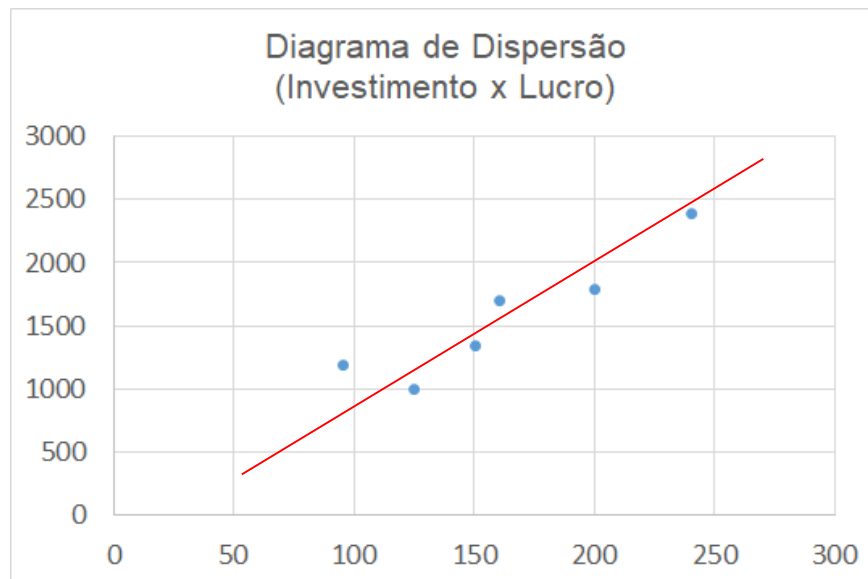


Diagrama de Dispersão

- Ex. 02: Vamos ver como seria o diagrama de dispersão para a amostra resumida na tabela a seguir:

x (preço)	y (vendas)
110	18
120	15
130	13
140	15
150	12
160	10

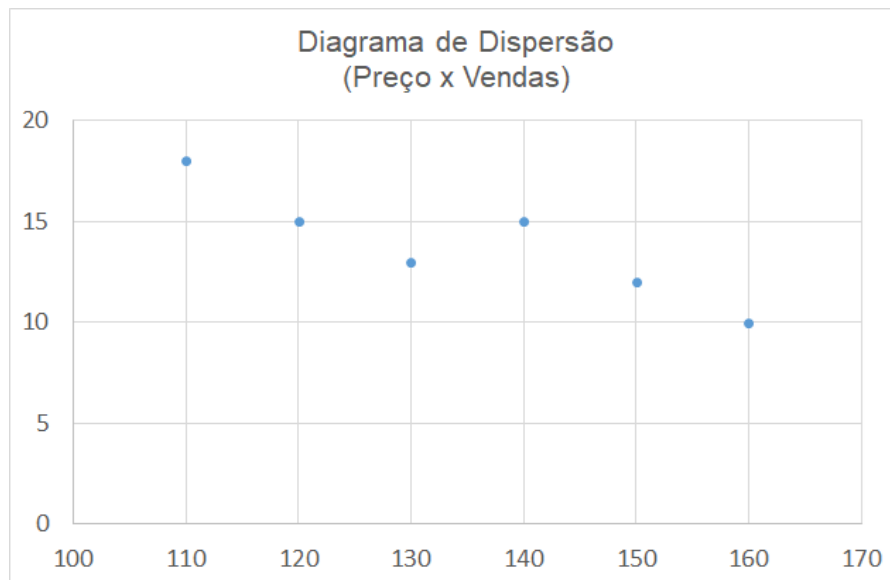


Diagrama de Dispersão

- Ex. 02: Vamos ver como seria o diagrama de dispersão para a amostra resumida na tabela a seguir (correlação negativa e forte):

x (preço)	y (vendas)
110	18
120	15
130	13
140	15
150	12
160	10

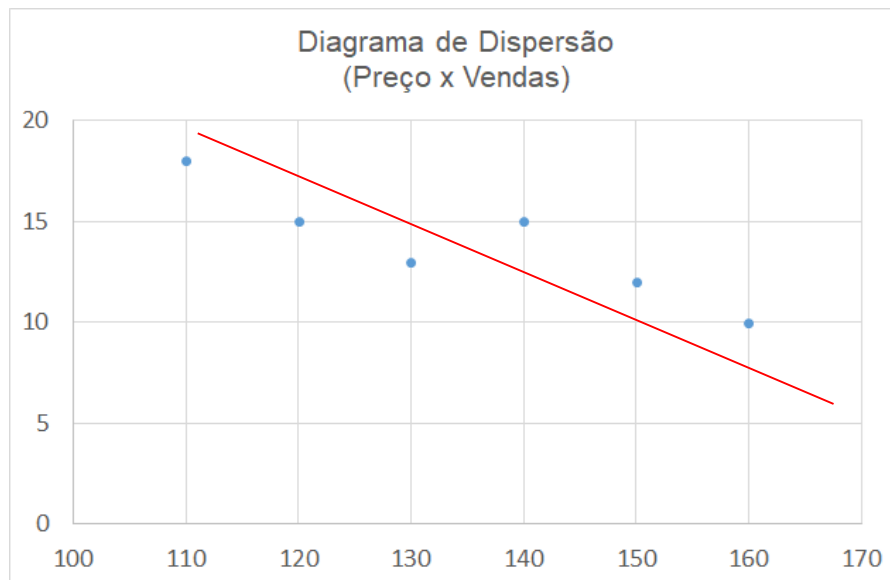
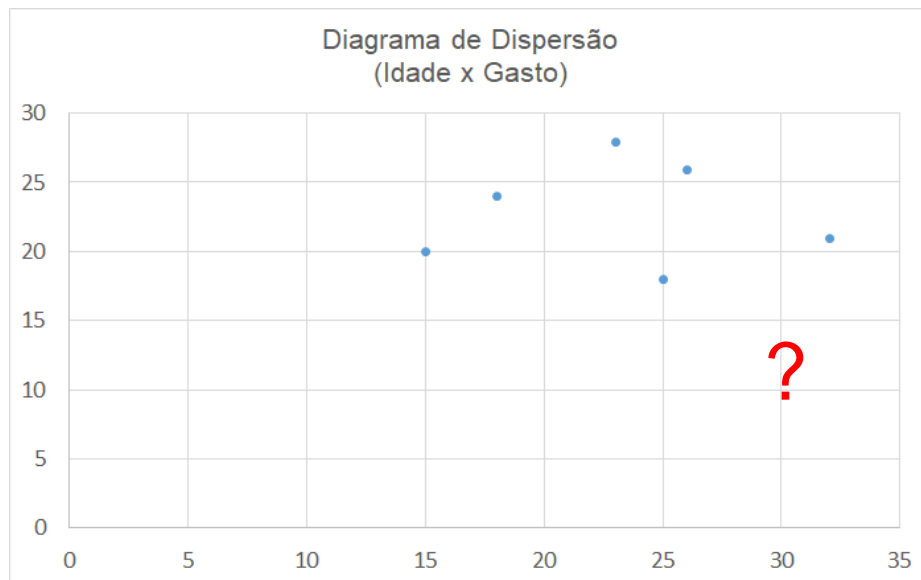


Diagrama de Dispersão

- Ex. 03: Vamos ver como seria o diagrama de dispersão para a amostra resumida na tabela a seguir (correlação fraca ou inexistente):

x (idade)	y (gastos)
15	27
18	25
23	23
25	21
26	28
40	22



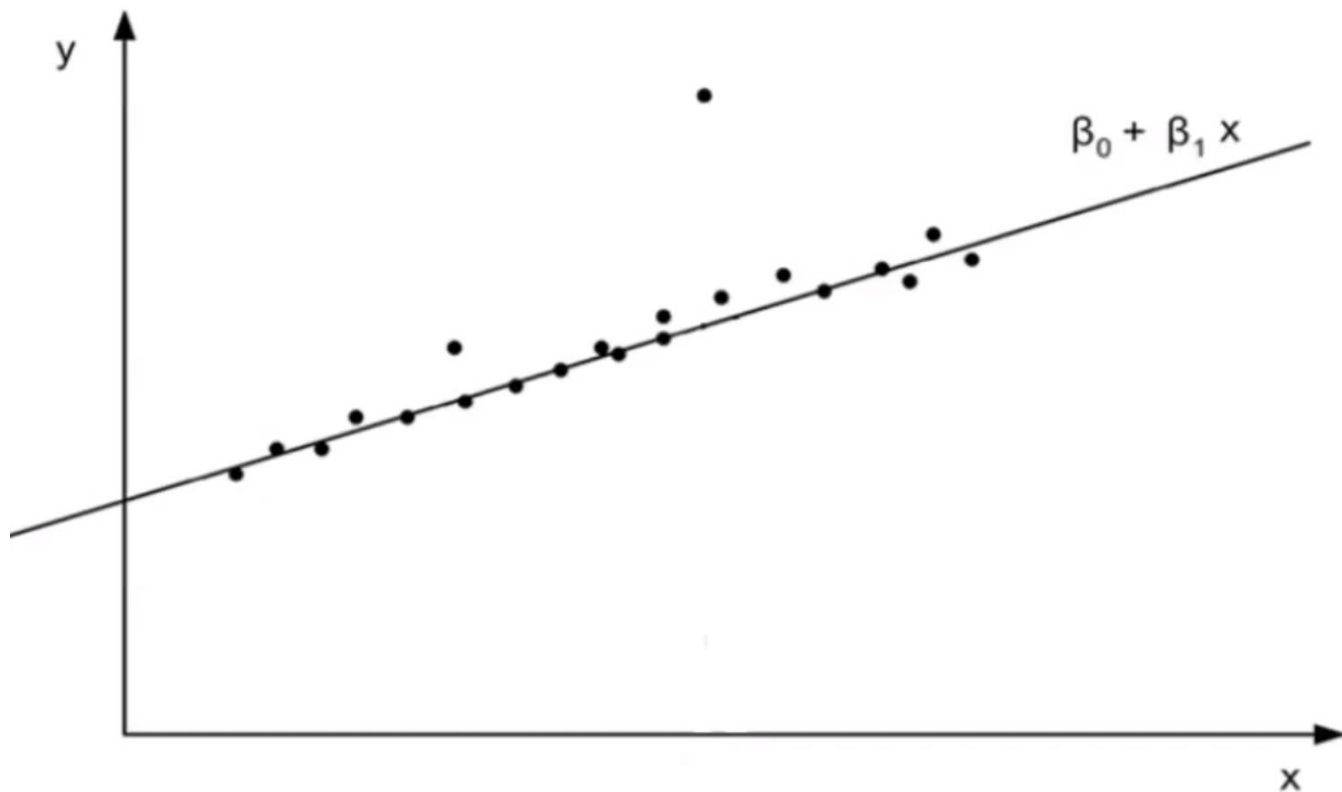
Visão Geral: Regressão Linear Simples

- Modela a relação entre duas variáveis
- Sendo essa relação linear, ela é matematicamente expressa por:

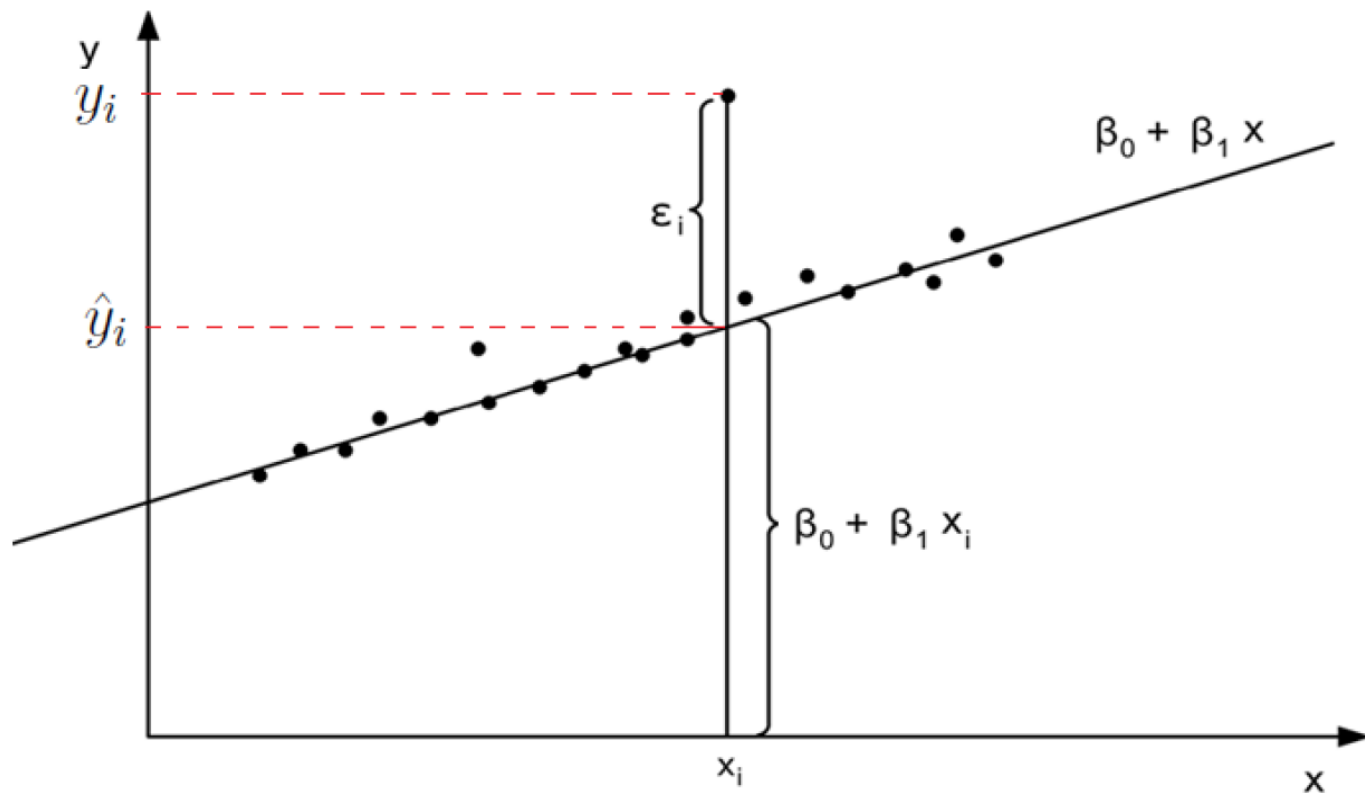
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- **y** é a variável **dependente** ou variável **resposta**
- **x** é a variável **independente**, **regressora**, **explicativa** ou **previsora**
- **ε** é o termo de **erro**. É a flutuação aleatória que ocorre ao tentar explicar a variável **y** por **x**.
Seja por imperfeições do modelo, erros de medida, ou outras variáveis fora de controle.

Visão Geral: Regressão Linear Simples



Visão Geral: Regressão Linear Simples



Visão Geral: Regressão Linear Simples

- Usaremos uma amostra $\{(x_1, y_1), \dots, (x_n, y_n)\}$ para estimar os parâmetros do modelo β_0 e β_1
- Métodos de estimação:
 - Mínimos quadrados ordinários (MQO)
 - Máxima verossimilhança (MV)
 - Método dos momentos (MM)
 - Melhor estimador não-enviesado (BLUE)
- A seguir, faremos inferências acerca dos parâmetros β_0 e β_1
 - ex.: propriedades dos estimadores, intervalos de confiança, testes de hipótese

Visão Geral: Regressão Linear Múltipla

- Modelo linear:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

- Matematicamente, dizemos que as derivadas parciais de y em relação aos coeficientes de regressão não dependem desses coeficientes

$$\frac{\partial y}{\partial \beta_0} = 1; \frac{\partial y}{\partial \beta_1} = x_1; \dots; \frac{\partial y}{\partial \beta_p} = x_p$$

- Contra exemplo:

$$y_i = \beta_0 + e^{\beta_2 x_i} + \varepsilon_i$$

Visão Geral: Tipos de Regressão

- Modelo linear simples:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Modelo linear múltiplo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

- Modelo não-linear:

$$y_i = \beta_0 + e^{\beta_2 x_i} + \varepsilon_i$$

Visão Geral: Linearização de Modelos

- Transformar modelos não-lineares em modelos lineares
 - É menos complexo trabalhar com as propriedades de modelos lineares
- Dada a função com erros na forma multiplicativa:

$$y_i = \alpha e^{\beta x_i} \varepsilon_i, \quad i = 1, 2, \dots, n$$

- Usando logaritmo em ambos os lados, temos:

$$\ln y = \ln \alpha e^{\beta t} \varepsilon$$

$$\ln y = \ln \alpha + \beta t + \ln \varepsilon$$

- Fazendo as substituições $z = \ln y$, $a = \ln \alpha$:

$$z = a + \beta t + \ln \varepsilon$$

Objetivos da Regressão

- Previsão: influência das variáveis independentes \mathbf{x} na variável resposta \mathbf{y}
- Descrição dos dados ou explanação: usar modelos para sumarizar ou descrever dados
- Seleção de variáveis ou triagem: determinar a importância de cada variável independente \mathbf{x} na determinação de \mathbf{y} . Quanto menor a contribuição de determinada variável, maior a possibilidade de sua exclusão do modelo
- Controle da saída: modelo estimado pode ser usado para controlar a saída \mathbf{y} . É possível encontrar um modelo ótimo para a variável de saída

Objetivos da Regressão

- Ex. 04: Utilizando um ajuste linear para estimar a tendência da série de consumo mensal de energia elétrica no período de janeiro de 2004 a dezembro de 2005, obteve-se a equação

$$T_t = 68,445 + 4,242t.$$

Sabendo-se que o valor observado em fevereiro de 2006 foi 185,8, calcule o erro absoluto de previsão associado à estimativa obtida para o mês de fevereiro de 2006, usando a equação apresentada, e assinale a opção correta.

Temos a seguinte equação de previsão:

$$T_t = 68,445 + 4,242t, \quad t = 1, 2, \dots, 24,$$

Para fevereiro de 2006, $t = 26$, com isso temos:

$$\hat{T}_{26} = 68,445 + 4,242 \cdot 26 \approx 178,7 \Rightarrow e = T_{26} - \hat{T}_{26} = 178,7 - 185,8 = 7,1$$

Objetivos da Regressão

- Ex. 04: Utilizando um ajuste linear para estimar a tendência da série de consumo mensal de energia elétrica no período de janeiro de 2004 a dezembro de 2005, obteve-se a equação

$$T_t = 68,445 + 4,242t.$$

Sabendo-se que o valor observado em fevereiro de 2006 foi 185,8, calcule o erro absoluto de previsão associado à estimativa obtida para o mês de fevereiro de 2006, usando a equação apresentada, e assinale a opção correta.


Temos a seguinte equação de previsão:

$$T_t = 68,445 + 4,242t, \quad t = 1, 2, \dots, 24,$$

Para fevereiro de 2006, $t = 26$, com isso temos:

$$\hat{T}_{26} = 68,445 + 4,242 \cdot 26 \approx 178,7 \Rightarrow e = T_{26} - \hat{T}_{26} = 178,7 - 185,8 = 7,1$$

Resíduo:

$$e_i = y_i - \hat{y}$$


Regressão Linear Simples

- Pressupostos:
 - Aquilo que se supõe antecipadamente, que se deseja alcançar
 - Em resumo, assumimos algumas inferências estatísticas para facilitar os cálculos, geramos um modelo e vemos se o modelo se adequa ao comportamento das amostras
 - Ex.: Dado $y_i = \alpha + \beta x_i + \varepsilon_i$, pode-se pressupor que:

$$E[\varepsilon_i | x_i] = 0;$$

$$\text{Var}[\varepsilon_i | x_i] = \sigma^2, \forall i \in 1, \dots, n \Rightarrow \text{homocedasticidade (variância constante)}.$$

$$\text{Cov}[\varepsilon_i, \varepsilon_j] = 0, i \neq j.$$

Regressão Linear Simples

- Pressupostos:

Opcionalmente $\varepsilon_i \sim N(0, \sigma^2)$.

$$E(y) = E[\alpha + \beta x_i + \varepsilon_i]$$

$$E(y) = E[\alpha + \beta x_i] + E[\varepsilon_i]$$

$$E(y) = \alpha + \beta x_i + \cancel{E[\varepsilon_i]} \overset{0}{\rightarrow}$$

$$E(y) = \alpha + \beta x_i$$

$$\text{Var}(y) = \text{Var}(\alpha + \beta x_i + \varepsilon_i) = \text{Var}(\varepsilon_i) = \sigma^2$$

$$\varepsilon_i \sim N(0, \sigma^2) \Rightarrow y_i | x_i \sim N(\alpha + \beta x_i, \sigma^2)$$

Portanto, no modelo vamos estimar a $\widehat{E(y)}$. Por preguiça, chamamos essa estimativa de \hat{y} .

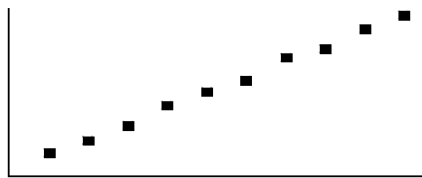
Teste Visual de Pressupostos

- Observação de gráficos é fundamental para verificar e formular pressupostos
- Alguns testes importantes:
 - Teste visual de linearidade
 - Teste visual de independência dos erros
 - Teste visual de erros normais
 - Teste visual de homocedasticidade (variância constante)
- Os testes podem ser realizados observando diagramas de dispersão

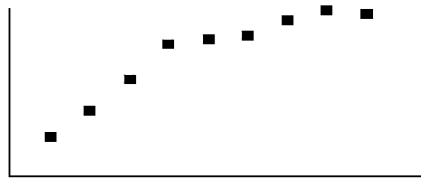
Teste Visual de Pressupostos

- Ex. (linearidade):

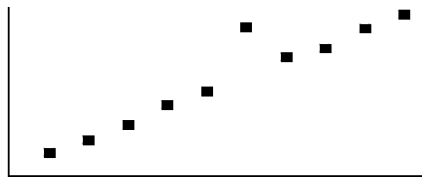
Gráficos de pontos x vs. y para ver o tipo básico da curva



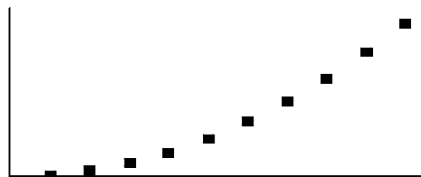
Linear



Linear por partes



Outlier/Exceção

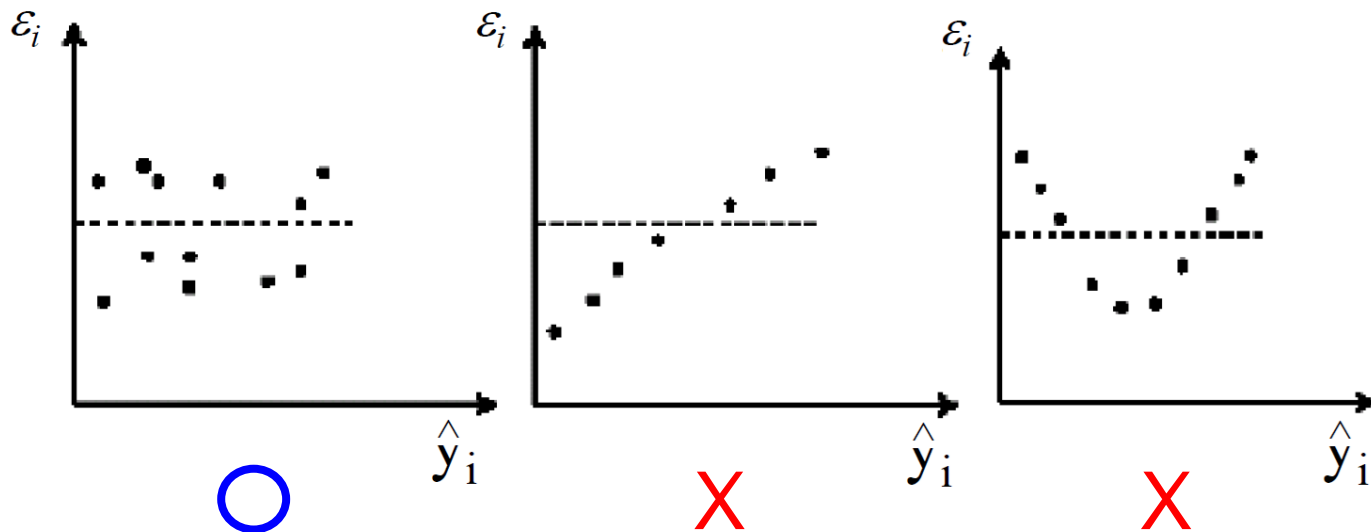


Não linear (Função de Potência)

Teste Visual de Pressupostos

- Ex. (independência dos erros):

Gráfico de pontos ε_i versus \hat{y}_i
Não deve haver tendência visível



Teste Visual de Pressupostos

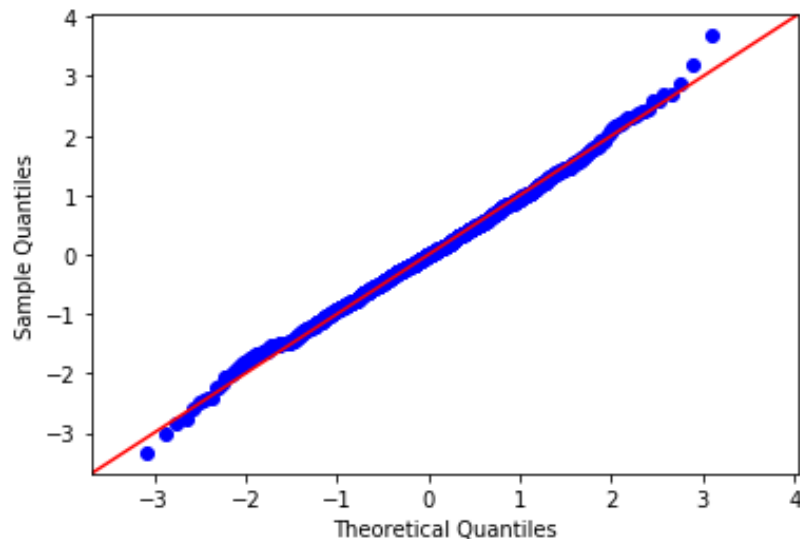
- Ex. (erros normais):

Analizar o Q-Q plot (gráfico quantil-quantil)

Caso os pontos plotados se aproximem da reta da normal, diz-se que os erros são normais

```
1 import numpy as np
2 import statsmodels.api as sm
3 import pylab
4
5 sample = np.random.normal(0,1, 1000)
6
7 sm.qqplot(sample, line='45')
8 pylab.show()
```

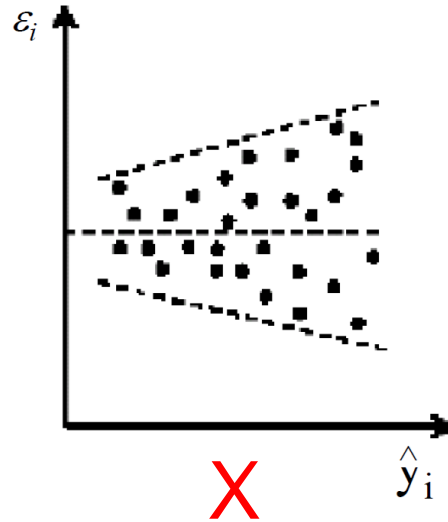
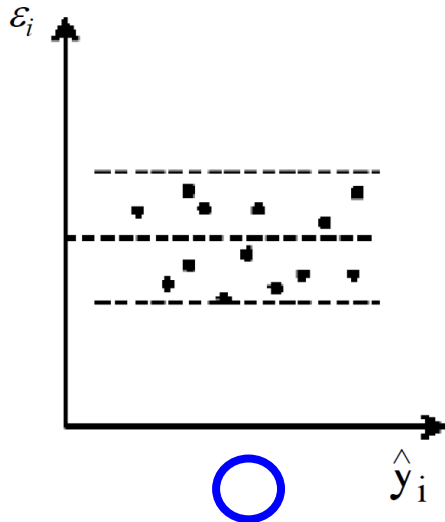
<https://www.statsmodels.org/stable/generated/statsmodels.graphics.gofplots.qqplot.html>



Teste Visual de Pressupostos

- Ex. (homocedasticidade):

Gráfico de pontos ε_i versus \hat{y}_i
Verificar tendência no espalhamento



**Caso haja tendência
ao espalhamento,
usar regressão
não-linear ou
linearização**

Método dos Mínimos Quadrados (MQO)

- Considerando o modelo: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Estimado por: $\hat{y}_i = b_0 + b_1 x$
 - **Ao inserir pressupostos no modelo, transformamos parâmetros em estimadores**
 - Os parâmetros de \hat{y}_i são calculados a partir de uma amostra
- Tomamos o erro aleatório (resíduo): $e_i = y_i - \hat{y}$
- Método dos mínimos quadrados:

$$f(\beta_0, \beta_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y})^2 = 0$$

Método dos Mínimos Quadrados (MQO)

- Os melhores parâmetros para regressão (que levam ao menor erro) são:

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} \quad b_0 = \bar{y} - b_1\bar{x}$$

- onde:

$$\bar{x} = \frac{1}{n} \sum x_i \quad \bar{y} = \frac{1}{n} \sum y_i$$

$$\sum xy = \sum x_i y_i \quad \sum x^2 = \sum x_i^2$$

Método dos Mínimos Quadrados (MQO)

- Ex. 05: Tempo de execução de um query para várias palavras:

x	Palavras	3	5	7	9	10
y	Tempo	1.19	1.73	2.53	2.89	3.26

$$\bar{x} = 6.8, \quad \bar{y} = 2.32, \quad \Sigma xy = 88.54, \quad \Sigma x^2 = 264$$

$$b_1 = \frac{88.54 - (5)(6.8)(2.32)}{264 - (5)(6.8)^2} = 0.29$$

$$b_0 = 2.32 - (0.29)(6.8) = 0.35$$

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

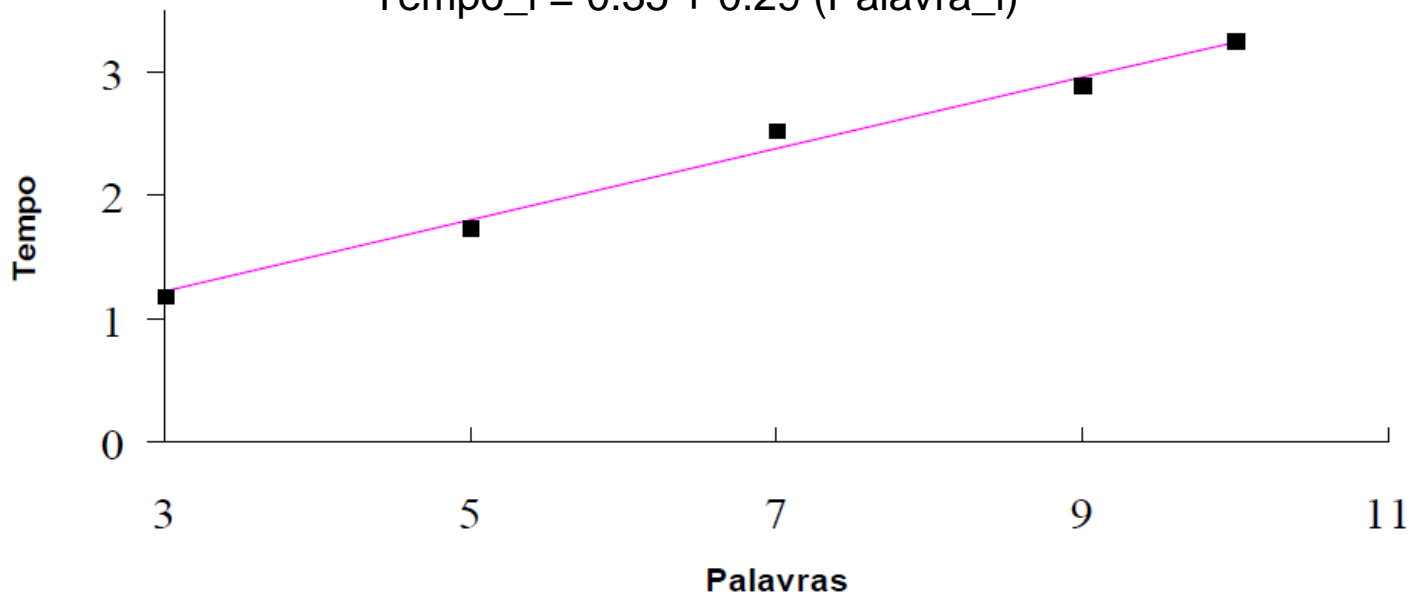
$$b_0 = \bar{y} - b_1 \bar{x}$$

Método dos Mínimos Quadrados (MQO)

- Ex. 05:

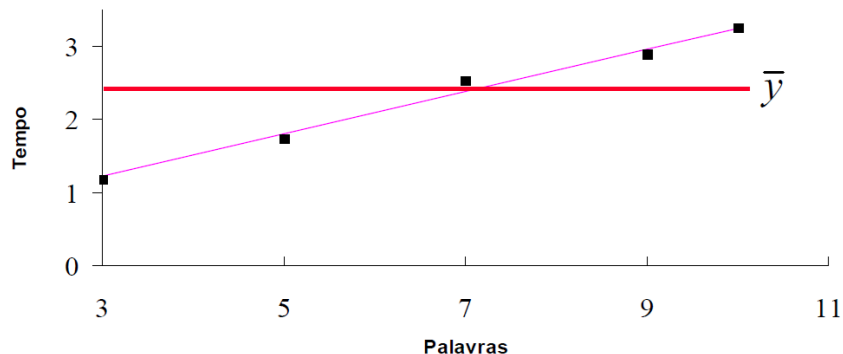
$$\hat{y}_i = 0.35 + 0.29 x_i$$

$$\text{Tempo}_i = 0.35 + 0.29 (\text{Palavra}_i)$$



Qualidade da Regressão

- Vale lembrar que a média acompanhada do desvio padrão é uma boa estimativa de uma amostra
- Regressão provê uma melhor estimativa, mas ainda existem erros
 - Ex.: Gráfico da regressão (linha inclinada) e da média (linha horizontal)



- Podemos avaliar a qualidade da regressão pela alocação das fontes de erros

Qualidade da Regressão

- Algumas notações importantes:

- SSE – Sum of Squared Errors (soma dos quadrados residuais, com regressão)

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

- SST – Total Sum of Squares (soma dos quadrados residuais, sem regressão)

$$\text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2 = \left(\sum_{i=1}^n y_i^2 \right) - n\bar{y}^2 = \text{SSY} - \text{SS0}$$

- SSY – Sum of Squares of y
- SS0 – Sum of Squares of \bar{y}
- SSR – Sum of Squares explained by Regression (SSR = SST - SSE)

Qualidade da Regressão

- Para avaliar a qualidade da regressão:

1. Calcule SST

2. Calcule SSE

3. Calcule o coeficiente de determinação (valor entre 0 e 1):

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

- Quanto maior o coeficiente de determinação, melhor a regressão

Qualidade da Regressão

Ex. 06 (do ex. 05): Tempo de execução de um query para várias palavras:

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

$$SST = \left(\sum_{i=1}^n y_i^2 \right) - n\bar{y}^2$$

$$SSR = SST - SSE$$

$$R^2 = \frac{SST - SSE}{SST}$$

x	Palavras	3	5	7	9	10
y	Tempo	1.19	1.73	2.53	2.89	3.26

$$n\bar{y}^2 = (5)(2.32)^2 = 26.9$$

$$\Sigma y = 11.60, \Sigma y^2 = 29.79, \Sigma xy = 88.54,$$

$$- SSE = 29.79 - (0.35)(11.60) - (0.29)(88.54) = 0.05$$

$$- SST = 29.79 - 26.9 = 2.89$$

$$- SSR = 2.89 - 0.05 = 2.84$$

$$\bullet R^2 = (2.89 - 0.05) / 2.89 = 0.98$$

Inferências sobre β_0 e β_1

- Relembrando a equação do modelo de Regressão Linear e seu estimador:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \hat{y}_i = b_0 + b_1 x$$

- Pressupostos:

- b_0 e b_1 seguem distribuição normal
- b_0 e b_1 possuem média dada por: $E(b_0) = \beta_0$ e $E(b_1) = \beta_1$
- Dado que $\text{Var}(b_0)$ e $\text{Var}(b_1)$ são variâncias, seus cálculos são dados por:

$$\text{Var}(b_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] \quad \text{Var}(b_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Inferências sobre β_0 e β_1

- Estimador da variância σ^2 (QME - quadrado médio residual):
 - Relembrando SSE (soma dos quadrados residuais):

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

$$QME = \frac{SSE}{n-2}$$

- Logo, o desvio padrão dos erros é dado por: $s_e = \sqrt{QME}$
- Usando QME (também conhecido como MSE) no pressuposto sobre as variâncias, temos:

$$S^2(b_0) = QME \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]$$

$$S^2(b_1) = \frac{QME}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Inferências sobre β_0 e β_1

- Como vimos, o desvio padrão dos erros é dado por: $s_e = \sqrt{QME}$
- O cálculo do desvio padrão para b_0 e b_1 é dado por:

$$s_{b_0} = s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum x^2 - n\bar{x}^2}}$$

$$s_{b_1} = \frac{s_e}{\sqrt{\sum x^2 - n\bar{x}^2}}$$

- O intervalo de confiança é obtido, usando a tabela-T, por:

$$b_i \pm t_{[1-\alpha, n-2]} s_{b_i}$$

Inferências sobre β_0 e β_1

- Ex. 07 (do ex. 05):

Tempo de execução de um query para várias palavras:

x	Palavras	3	5	7	9	10
y	Tempo	1.19	1.73	2.53	2.89	3.26

SSE = 0.05 (do ex. 06)

então: $MSE = 0.05/(5-2) = 0.05/3 = 0.017$

$$s_e = \sqrt{MSE} = 0.13$$

$MSE = QME = \frac{SSE}{n-2}$

Observe a alta qualidade da regressão do exemplo:

– $R^2 = 0.98$

– $s_e = 0.13$

Inferências sobre β_0 e β_1

- Ex. 07 (do ex. 05) continuação:

Lembre que $s_e = 0.13$, $n = 5$, $\sum x^2 = 264$, $\bar{x} = 6.8$

Assim

$$s_{b_0} = s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum x^2 - n\bar{x}^2}}$$

$$s_{b_1} = \frac{s_e}{\sqrt{\sum x^2 - n\bar{x}^2}}$$

$$s_{b_0} = 0.13 \sqrt{\frac{1}{5} + \frac{(6.8)^2}{264 - 5(6.8)^2}} = 0.16$$

$$s_{b_1} = \frac{0.13}{\sqrt{264 - 5(6.8)^2}} = 0.004$$

$$b_i \pm t_{[1-\alpha, n-2]} s_{b_i}$$

Usando um intervalo de confiança de 90%:

$$t_{0.90;3} = 2.353$$

	80%	90%	2-Tail Confidence Level
df	0.20	0.10	2-Tail Alpha
1	3.0777	6.3138	
2	1.8856	2.9200	
3	1.6377	2.3534	

Assim, o intervalo b_0

$$0.35 \pm 2.353(0.16) = (-0.03, 0.73)$$

b_1 é

$$0.29 \pm 2.353(0.004) = (0.28, 0.30)$$

Predições Baseadas em ***m*** Amostras

- O estimador da regressão linear da predição é dado por: $y_p = b_0 + b_1 x_p$
- Desvio padrão para a média de ***m*** amostras, com predição, é dado por:

$$S_{y_{mp}} = s_e \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum x^2 - n\bar{x}^2}}$$

- Quanto maior o ***m***, menor o desvio padrão
- Cálculo da predição com IC: $y_p \pm t_{[1-\alpha, n-2]} S_{y_{mp}}$

Predições Baseadas em *m* Amostras

- Ex. 08 (do ex. 05): Usando modelo desenvolvido, qual é o tempo previsto para uma execução com 8 palavras?

Tempo de execução de um query para várias palavras:

x	Palavras	3	5	7	9	10
y	Tempo	1.19	1.73	2.53	2.89	3.26

$$\bar{x} = 6.8, \quad \bar{y} = 2.32, \quad \Sigma xy = 88.54, \quad \Sigma x^2 = 264$$

$$b_1 = \frac{88.54 - (5)(6.8)(2.32)}{264 - (5)(6.8)^2} = 0.29$$

$$b_0 = 2.32 - (0.29)(6.8) = 0.35$$

$$\text{Tempo} = 0.35 + 0.29(8) = 2.67$$

$$\text{Desvio padrão de erros } s_e = 0.13$$

$$s_{y_p} = 0.13 \sqrt{1 + \frac{1}{5} + \frac{(8 - 6.8)^2}{264 - 5(6.8)^2}} = 0.14$$

90% do intervalo é então

$$2.67 \pm 2.353(0.14) = (2.34, 3.00)$$

Revisão

- Ex. 09: The number of disk I/O's and processing times of seven programs were measured as: (14,2), (16,5), (27,7), (42,9), (39,10), (50,13), (83,20).

For this data:

$$n=7, \sum xy=3375, \sum x=271, \sum x^2=13,855, \\ \sum y=66, \sum y^2=828, \bar{x} = 38.71, \bar{y} = 9.43.$$

Therefore,

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2} = 0.2438 \qquad b_0 = \bar{y} - b_1\bar{x} = -0.0083$$

Modelo linear : CPU time = - 0.0083 + 0.2438 (#Disk I/Os)

Revisão

- Ex. 09:

Disk I/O's	CPU Time	Estimate	Error	Error ²
x_i	y_i	$\hat{y}_i = b_0 + b_1 x_i$	$e_i = y_i - \hat{y}_i$	e_i^2
14	2	3.4043	-1.4043	1.9721
16	5	3.8918	1.1082	1.2281
27	7	6.5731	0.4269	0.1822
42	9	10.2295	-1.2295	1.5116
39	10	9.4982	0.5018	0.2518
50	13	12.1795	0.8205	0.6732
83	20	20.2235	-0.2235	0.0500
Σ	271	66.0000	0.00	5.8690

Revisão

- Ex. 09: Verificação da qualidade do estimador

$$\begin{aligned}\text{SSE} &= \Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy \\ &= 828 + 0.0083 \times 66 - 0.2438 \times 3375 = 5.87\end{aligned}$$

$$\begin{aligned}\text{SST} &= \text{SSY} - \text{SS0} = \Sigma y^2 - n(\bar{y})^2 \\ &= 828 - 7 \times (9.43)^2 = 205.71\end{aligned}$$

$$\text{SSR} = \text{SST} - \text{SSE} = 205.71 - 5.87 = 199.84$$

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{199.84}{205.71} = 0.9715$$

Modelo explica 97% da variacao: MUITO BOM!!!

Revisão

- Ex. 09: Desvio padrão dos erros e dos parâmetros

- The mean squared error is:

$$QME = \frac{SSE}{n-2} = \frac{5.87}{5} = 1.17$$

- The standard deviation of errors is:

$$s_e = \sqrt{QME} = \sqrt{1.17} = 1.0834$$

$$s_{b_0} = s_e \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum x^2 - n\bar{x}^2} \right]^{1/2} = 1.0834 \left[\frac{1}{7} + \frac{(38.71)^2}{13,855 - 7 \times 38.71 \times 38.71} \right]^{1/2} = 0.8311$$

$$s_{b_1} = \frac{s_e}{[\sum x^2 - n\bar{x}^2]^{1/2}} = \frac{1.0834}{[13,855 - 7 \times 38.71 \times 38.71]^{1/2}} = 0.0187$$

Revisão

$$b_i \pm t_{[1-\alpha, n-2]} s_{b_i}$$

- Ex. 09: \Rightarrow 90% confidence interval for b_0 is:

$$\begin{aligned} -0.0083 \mp (2.015)(0.8311) &= -0.0083 \mp 1.6747 \\ &= (-1.6830, 1.6663) \end{aligned}$$

Since, the confidence interval includes zero, the hypothesis that this parameter is zero cannot be rejected at 0.10 significance level. $\Rightarrow b_0$ is essentially zero.

90% Confidence Interval for b_1 is:

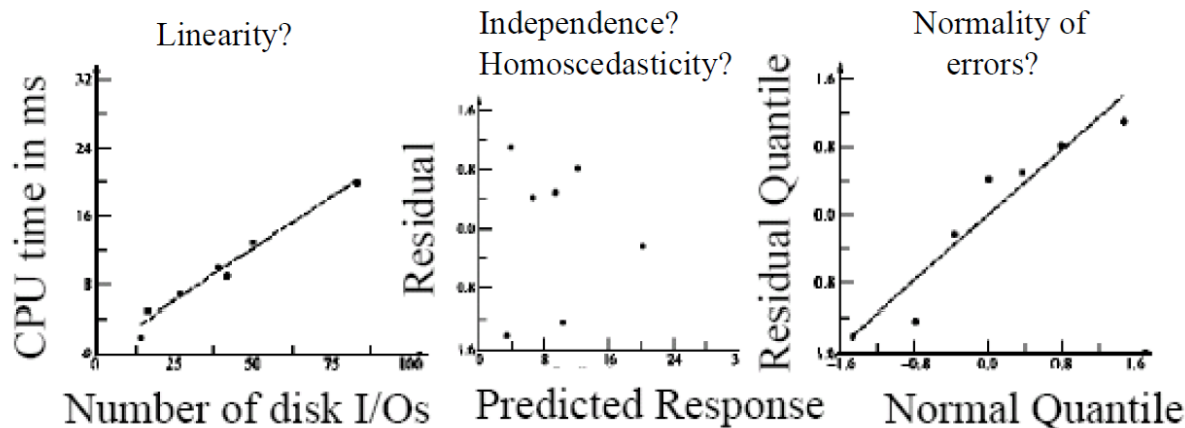
$$\begin{aligned} 0.2438 \mp (2.015)(0.0187) &= 0.2438 \mp 0.0376 \\ &= (0.2061, 0.2814) \end{aligned}$$

Since the confidence interval does not include zero, the slope b_1 is significantly different from zero at this confidence level.

Revisão

Ex. 09: Testes visuais

Fazer gráficos separados para esses 2 testes



1. Relationship is linear
2. No trend in residuals \Rightarrow Seem independent
3. Linear normal quantile-quantile plot \Rightarrow Larger deviations at lower values but all values are small

AAG05 Tarefa em Dupla

- Desenvolva em Python exemplo tão completo quanto o exemplo “Revisão”:
 1. Escolher/Criar uma amostra bivariada para o exemplo
 2. Calcule o coeficiente de correlação e só vá para “3” caso a amostra bivariada tenha correlação forte (positiva ou negativa)
 3. Estimar parâmetros, verificar a qualidade, calcular os erros
 4. Calcular desvio padrão dos erros e dos parâmetros
 5. Calcular intervalo de confiança dos parâmetros para níveis de confiança de 90%, 95% e 99%
 6. Testar linearidade, independência de erros, erros normais, homocedasticidade (com gráficos)
- Regras:
 1. Funções prontas de bibliotecas python DEVEM ser usadas ao máximo possível
 2. Código e resultados devem ser explicados em Markdown com comandos LaTeX
 3. Os formatos de entrega devem ser .pdf e .ipynb (código fonte+markdowns)
 4. Os dados devem ser entregues em anexo