Introdução à Probalidade e Estatística

Regressão Linear Simples

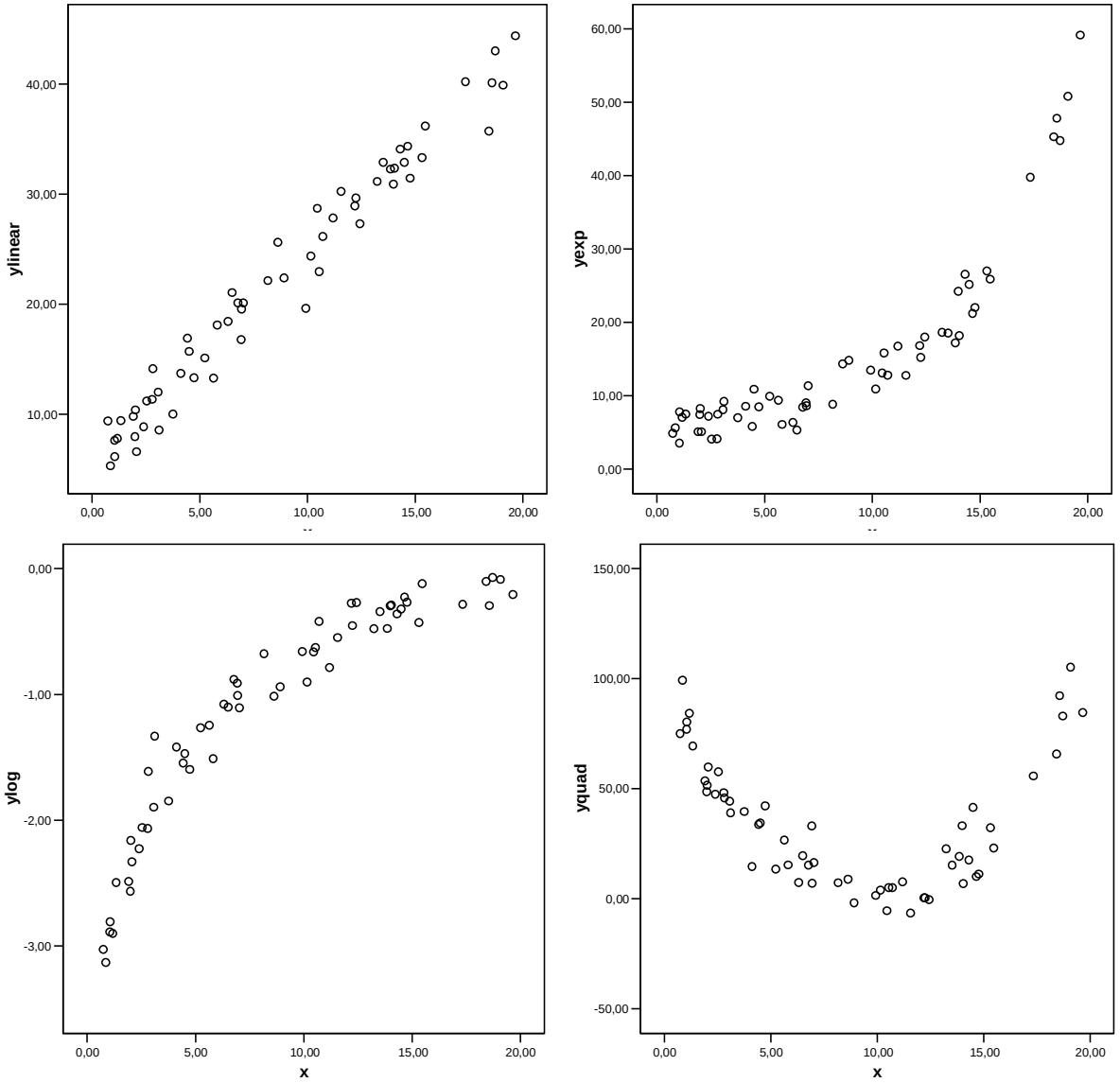
Departamento de Matemática
Universidade de Évora

Associação entre duas variáveis

Questões de interesse:

Será que duas variáveis são independentes ou pelo contrário dependentes? E se forem dependentes, qual o tipo e grau de dependência?

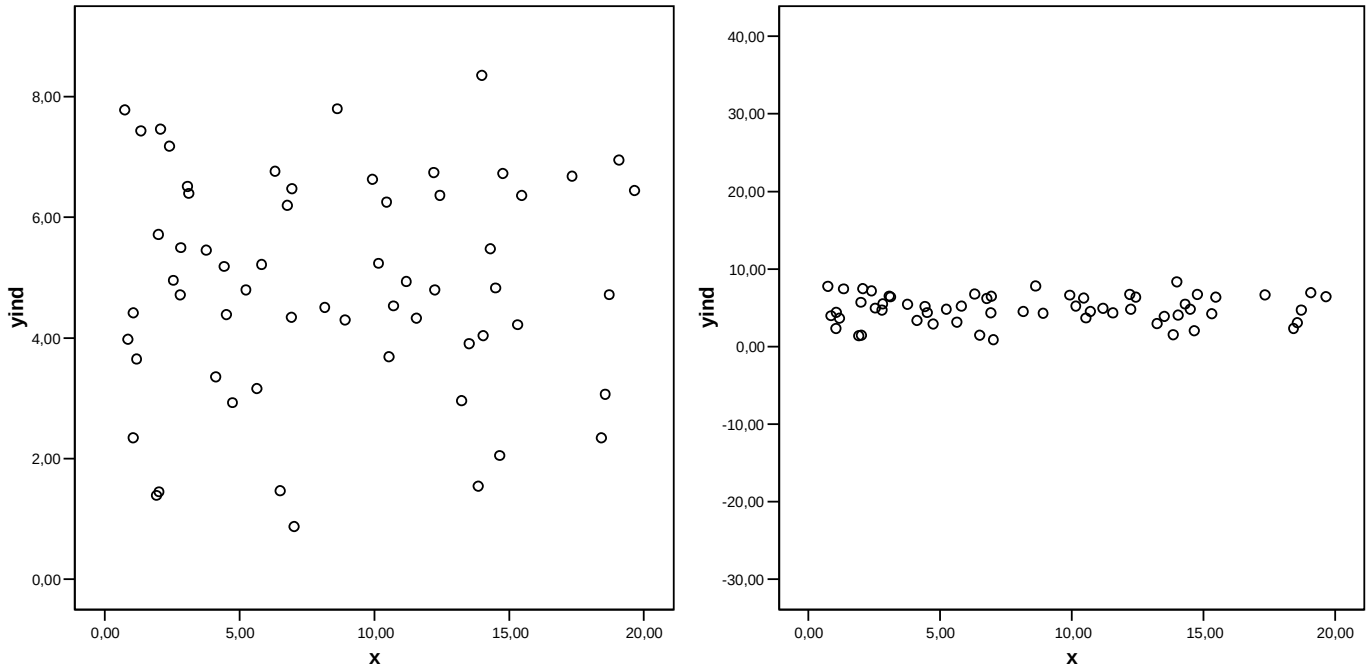
Existem diversas formas de associação entre variáveis numéricas. Por exemplo, podemos ter relações lineares, exponenciais, logarítmicas ou quadráticas.


ylinear
40,00
30,00
20,00
10,00
0,00
5,00
10,00
15,00
20,00
yexp
60,00
50,00
40,00
30,00
20,00
10,00
0,00
ylog
0,00
-1,00
-2,00
-3,00
x
yquad
150,00
100,00
50,00
0,00
-50,00

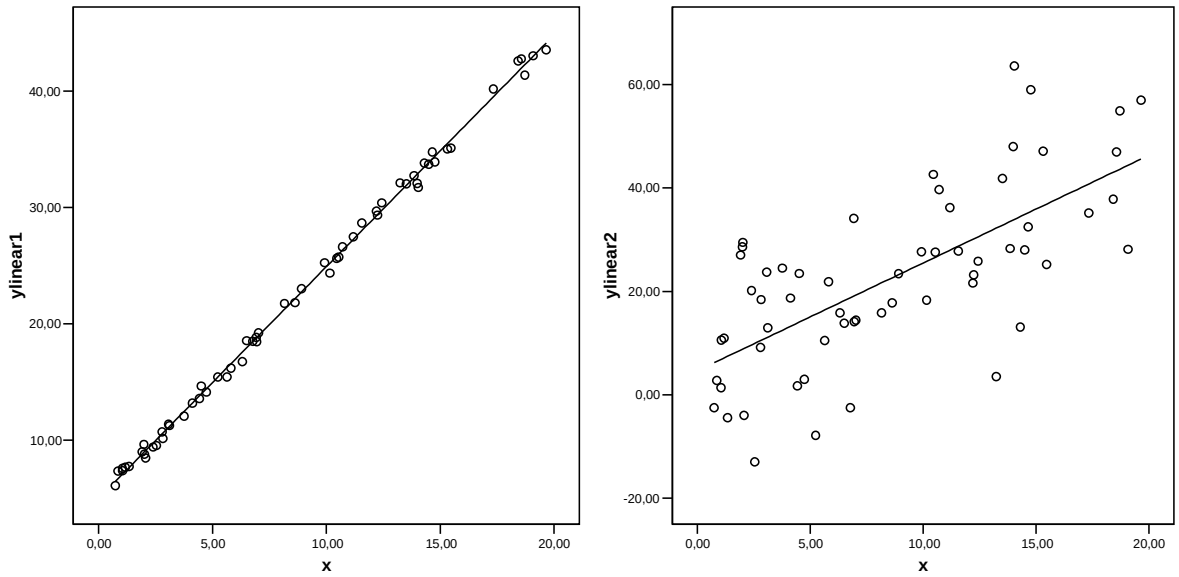
Como analisar a associação entre 2 variáveis numéricas

Primeiro passo: construção de diagramas de dispersão.

Quando duas variáveis são independentes, o diagrama de dispersão respectivo apresenta uma mancha de pontos aleatória (ou quando muito) um conjunto de pontos dispostos sobre uma recta horizontal.



Se a relação entre duas variáveis for linear, ao confrontarmos duas amostras num diagrama de dispersão devemos esperar observar um conjunto de pontos que se dispõem aproximadamente sobre uma recta. Por vezes os desvios em relação à recta são mínimos, mas noutras os pontos apresentam bastante dispersão tornando difícil a identificação da dita relação linear.



Segundo passo: calcular medidas de associação ou efectuar uma análise de regressão caso a relação seja linear.

Regressão Linear Simples

A equação $y = b_0 + b_1 x$ define uma recta no plano x, y. b_0 representa a ordenada na origem e b_1 o declive. Se um ponto (x_1, y_1) estiver sobre a recta então satisfaz a relação $y_1 = b_0 + b_1 x_1$.

Se o valor de y_1 estiver afectado de um erro aleatório, ϵ, passamos a ter $y_1 = b_0 + b_1 x_1 + \epsilon$.

Muitas vezes temos dados estatísticos que correspondem exactamente a pares de observações, (x_i, y_i), $i = 1, \ldots, n$, que têm subjacentes uma relação linear, mas que estão afectados de erros.

$$y_i = b_0 + b_1 x_i + \epsilon_i, \quad i = 1, \ldots n.$$

A análise de regressão é uma técnica estatística para modelar e investigar a relação entre variáveis. No modelo de regressão linear simples temos

- valores determinados x_i provenientes de uma variável independente também denominada regressor.

- valores aleatórios Y_i provenientes de uma variável dependente.

- um modelo probabilístico que relaciona Y_i com x_i

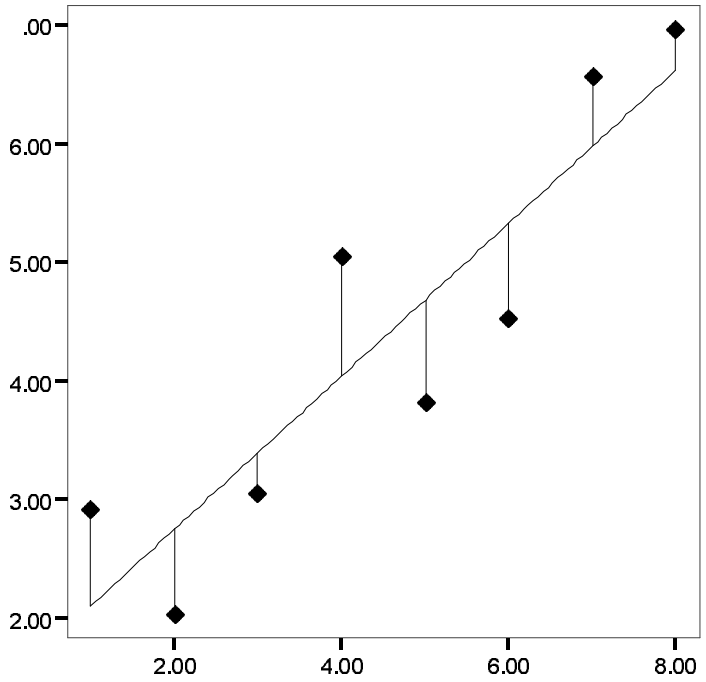
$$Y_i = b_0 + b_1 x_i + \epsilon_i, \quad \epsilon_i \text{ - erro,}$$

  b_0 e b_1 são designados coeficientes de regressão ou parâmetros de regressão.

- os erros devem ser independentes e identicamente distribuídos, $\epsilon_i \frown N(0, \sigma^2)$. Desta forma existe uma relação linear entre o valor esperado de Y_i e a variável independente x_i,

$$E[Y_i|x_i] = b_0 + b_1 x_i.$$

Graficamente, um exemplo de um modelo de regressão linear simples tem a seguinte forma:



Método dos mínimos quadrados e a recta de regressão

Como as observações estão afectadas de erros não é possível saber o valor exacto dos coeficientes b_0 e b_1. No entanto é possível estimá-los. O método que conduz aos melhores resultados (nas condições acima descritas) é o método dos mínimos quadrados

Este método conduz aos seguintes estimadores

$$\begin{cases} \hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{x} \\ \hat{b}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{cases}$$

A recta de regressão é então dada por

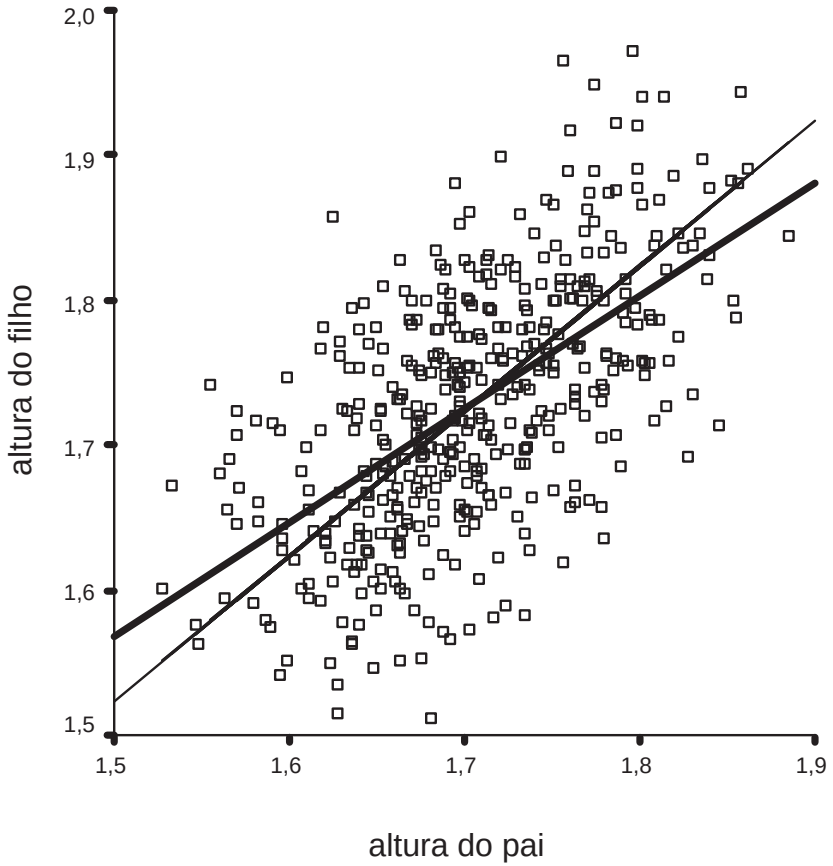
$$y = \hat{b}_0 + \hat{b}_1 x.$$

Chamamos valores preditos a

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i,$$

que são as nossas melhores estimativas para os pontos sobre a recta (desconhecida).

Exemplo: alturas dos filhos versus alturas dos pais. A equação da recta de regressão é dada por $y = 0.392 + 0.784x$ (traço grosso). A recta de traço mais fino tem declive unitário.



Testes e IC's para os coeficientes de regressão

Com base nos resultados anteriores podemos construir intervalos de confiança e efectuar testes de hipóteses aos parâmetros do modelo de regressão. Para tal é necessário utilizar as seguintes relações:

$$\frac{\hat{b}_0 - b_0}{\sqrt{\frac{SS_E}{(n-2)}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \frown t_{n-2}$$

$$\frac{\hat{b}_1 - b_1}{\sqrt{\frac{SS_E}{(n-2)S_{xx}}}} \frown t_{n-2}$$

Tem muito interesse testar se o declive da recta é nulo, ou seja, se Y não depende de x:

$$H_0: \quad b_1 = 0 \quad vs \quad H_1: b_1 \neq 0$$

Também pode ter interesse testar se a ordenada na origem é nula:

$$H_0: \quad b_0 = 0 \quad vs \quad H_1: b_0 \neq 0$$

Estatísticas de teste

Para a ordenada na origem:

$$T_0 = \frac{\hat{b}_0}{\sqrt{\frac{SS_E}{(n-2)}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} = \frac{\hat{b}_0}{\hat{\sigma}_{b_0}} \underset{sob\ H_0}{\frown} t_{n-2}$$

Para o declive:

$$T_1 = \frac{\hat{b}_1}{\sqrt{\frac{SS_E}{(n-2)S_{xx}}}} = \frac{\hat{b}_1}{\hat{\sigma}_{b_1}} \underset{sob\ H_0}{\frown} t_{n-2}$$

Tabela de regressão

A tabela de regressão contém, além de outras coisas, os valores das estimativas dos parâmetros de regressão e os p-values dos testes referidos anteriormente.

	Coeficientes não-estandardizados		Coeficientes estandardizados		
Coeficiente	b	Erro padrão	β	t	$p-value$
Ord. na origem	$\hat{b}_0$	$\hat{\sigma}_{b_0}$		t_{0obs}	$(\cdot)$
declive	$\hat{b}_1$	$\hat{\sigma}_{b_1}$	$\hat{\beta}_1$	t_{1obs}	$(\cdot)$

O exemplo dos pais e filhos no SPSS:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,392	,085		4,592	,000
	PAI	,784	,050	,598	15,665	,000

a. Dependent Variable: FILHO

Coefficients[a]

Model		95% Confidence Interval for B	
		Lower Bound	Upper Bound
1	(Constant)	,224	,560
	PAI	,686	,882

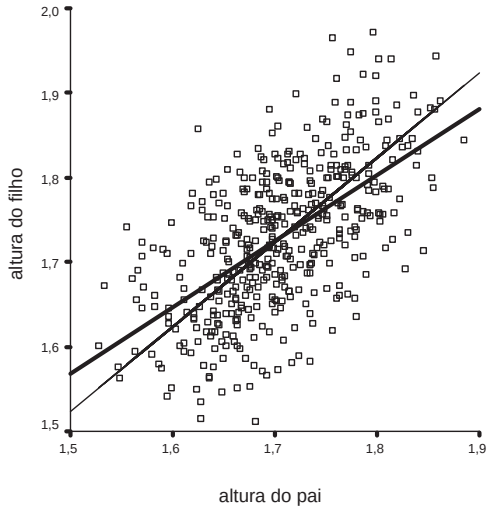
a. Dependent Variable: FILHO

A análise de regressão linear simples pode ser feita no SPSS utilizando o menu `Analyze / Regression / Linear`. Para obter os intervalos de confiança para os coeficientes é necessário seleccionar `Confidence Intervals` no botão `Statistics`.

Avaliação da qualidade e significado da regressão

1. **Análise gráfica:**

   Gráfico de dispersão de Y_i versus x_i: deve evidenciar uma relação linear; deve ter os pontos pouco dispersos para a regressão ter boa qualidade.



   Neste exemplo existe muita dispersão pelo que a regressão não terá muita qualidade.

2. Valor do coeficiente de determinação

$$R^2 = \frac{S_{xY}^2}{S_{xx}S_{YY}} = \frac{SS_R}{S_{YY}} = 1 - \frac{SS_E}{S_{YY}}$$

O coeficiente deve assumir valores próximos de 1 (superior a 0.9) se a relação entre Y e x for bem modelada por uma regressão linear simples. R^2 mede a proporção de variabilidade de Y explicada por x.

Por vezes utiliza-se o **coeficiente de determinação ajustado** que introduz uma correcção no coeficiente de determinação. Em geral os valores destes coeficientes são muito próximos.

$$R_a^2 = 1 - \frac{SS_E/(n-2)}{S_{YY}(n-1)}.$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,598[a]	,358	,357	,06968

a. Predictors: (Constant), PAI

3. **Teste ao declive**

Será que Y depende mesmo de x? Podemos responder a esta questão através do teste ao declive da tabela de regressão

$$H_0 : \ b_1 = 0 \quad vs \quad H_1 : \ b_1 \neq 0.$$

Validação dos pressupostos da regressão – análise de resíduos

Para avaliar se os erros se podem considerar como sendo provenientes de uma população com distribuição Normal:

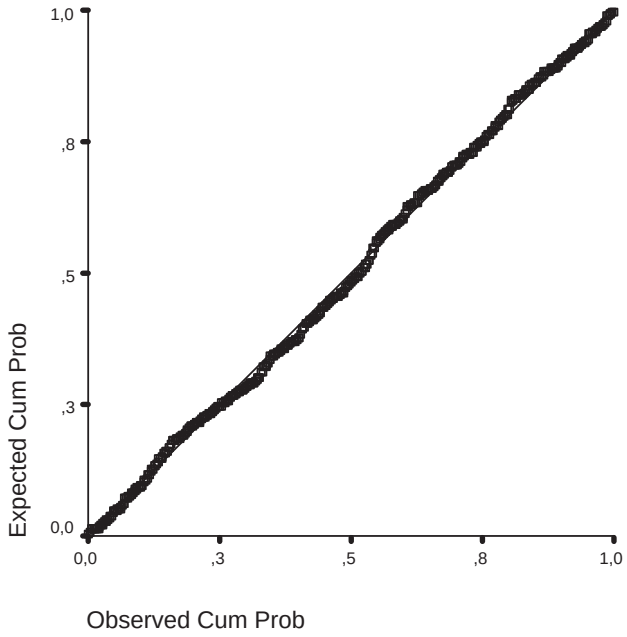
QQ-plot aos resíduos.

Chama-se resíduo a

$$e_i = y_i - \hat{b}_0 - \hat{b}_1 x_i = y_i - \hat{y}_i$$

que é a estimativa do erro ϵ_i.

Exemplo das alturas dos pais e filhos:

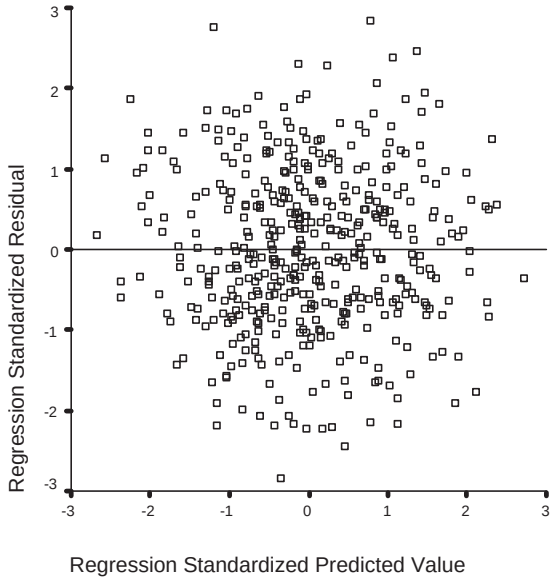


No SPSS pode-se obter o QQ-plot dos resíduos seleccionando a opção `Normal probability plot` no botão `Plots` do menu da regressão linear.

Também se podem fazer um teste de ajustamento à Normal.

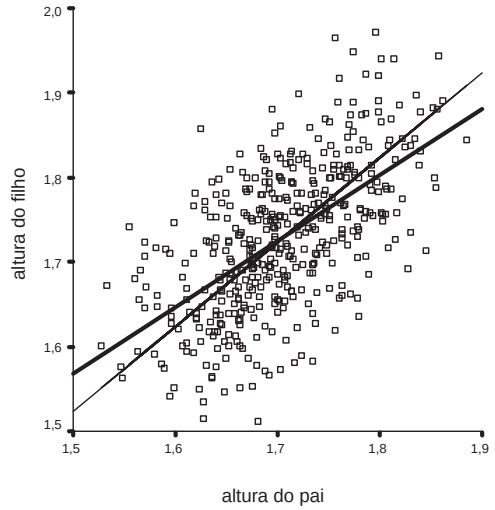
Para avaliar se os erros são independentes:

Gráficos de resíduos versus valores preditos $\hat{Y}_i$ (ou valores observados, ou regressores) que deve apresentar uma mancha de pontos aleatórios com o mesmo tipo de dispersão em torno do eixo do xx.



No SPSS pode-se obter este gráfico através do menu fornecido no botão `Plots` do menu da regressão linear.

Para avaliar se o modelo é correcto deve-se observar o gráfico de dispersão Y_i versus x_i:



Este gráfico deve apresentar uma relação linear e os pontos devem distribuir-se aleatoriamente em torno da recta com variabilidade constante.

Os gráficos de resíduos também podem ajudar a detectar que o modelo não é adequado em situações que o gráfico de dispersão não é claro.