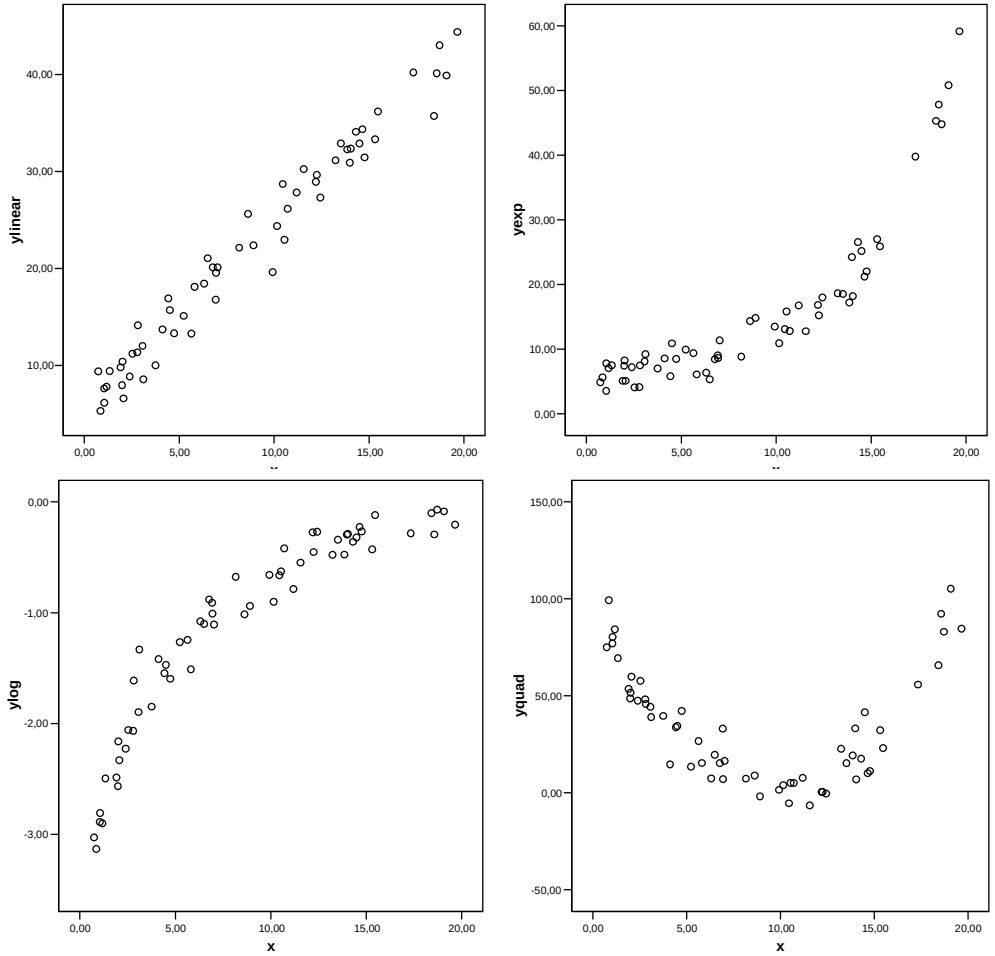
Associação entre duas variáveis

Questões de interesse:

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

Medir o grau de dependência é mais ambicioso do que simplesmente testar a existência de alguma associação entre variáveis. É obviamente de interesse poder medir o grau de associação entre dois conjuntos de observações obtidos a partir de um dado conjunto de unidades experimentais (indivíduos por exemplo). Mas, em muitas circunstâncias estamos apenas interessados em saber se uma certa associação observada nos dados indica ou não uma associação na população de onde foram retirados.

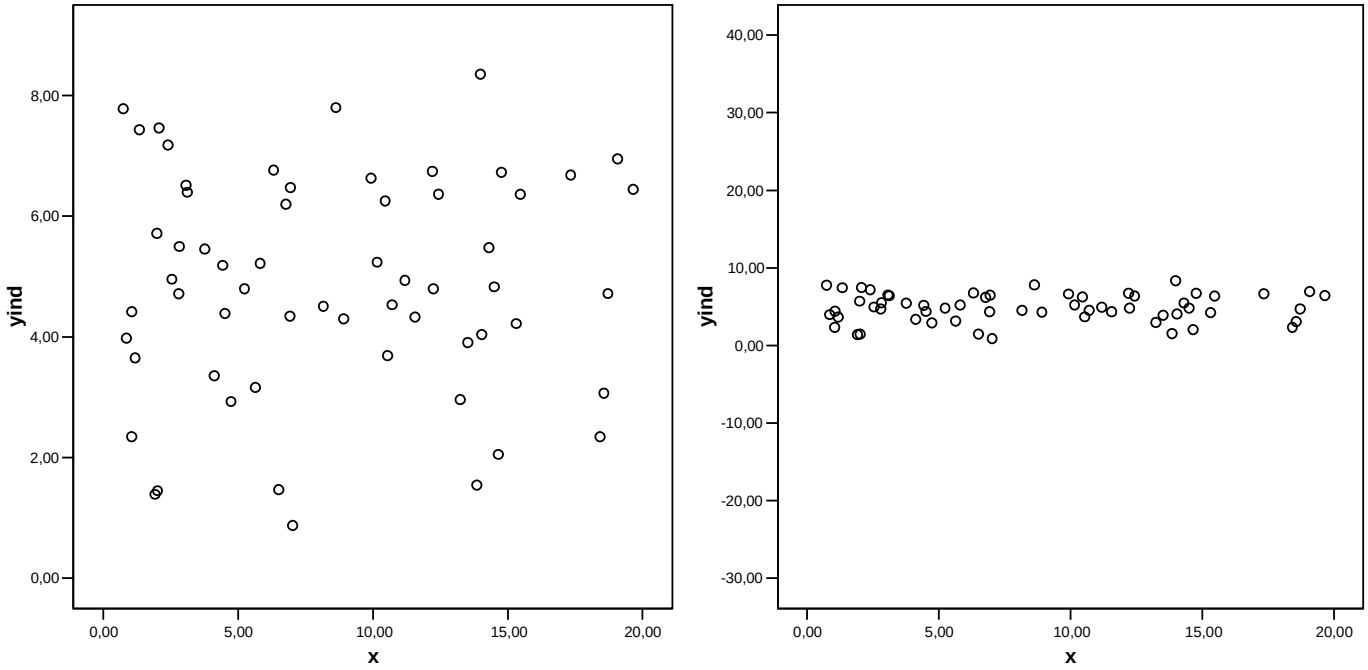
Existem diversas formas de associação entre variáveis numéricas. Por exemplo, podemos ter relações lineares, exponenciais, logarítmicas ou quadráticas.



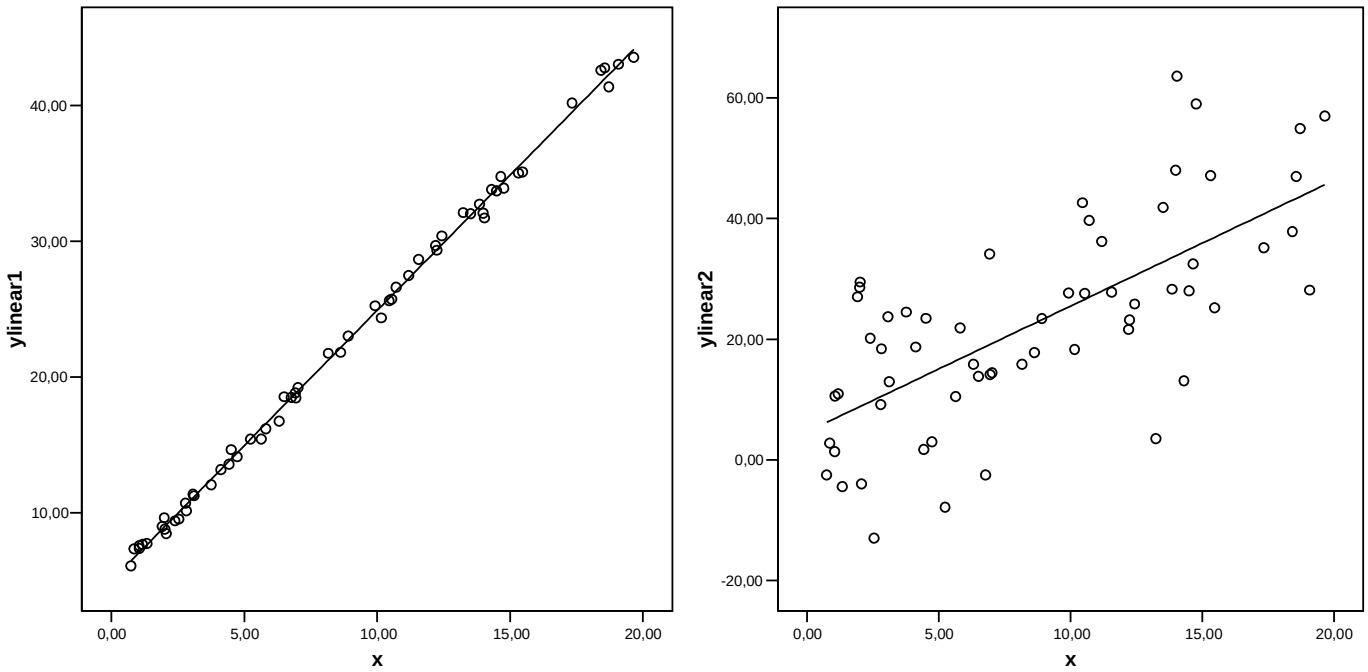
Como analisar a associação entre 2 variáveis numéricas

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 (coeficientes de correlação).

Último passo: realizar um teste de hipóteses para averiguar se os valores das medidas de associação observados nos dados são significativos, ou seja, se podemos estatisticamente concluir a favor de uma associação na população.

Coeficientes de correlação mais utilizados

Existem várias medidas de associação, quer para dados quantitativos, quer para dados qualitativos. Iremos apenas referir medidas de associação para dados quantitativos, que se designam habitualmente por coeficientes de correlação.

Os coeficientes de correlação mais utilizados são o de Pearson (em contexto paramétrico), o de Spearman e o de Kendall (em contexto não paramétrico).

No SPSS os coeficientes de associação (correlação) para dados numéricos ou ordinais podem ser obtido através do menu `Analyse / Correlate / Bivariate`.

Neste menu podem-se seleccionar mais do que duas variáveis, caso em que o SPSS fornece uma tabela de correlações para todas as combinações de pares de variáveis. O SPSS fornece também o p-value dos testes ao significado dos coeficientes, para cada par de variáveis.

1 - O coeficiente de correlação de Pearson (*Pearson product-moment correlation coefficient*)

Dadas duas amostras de observações medidas numa escala de intervalos ou razões, podemos medir o grau de associação linear através da estatística

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

r pertence ao intervalo $[-1, 1]$. Se $r = 1$ temos uma recta perfeita com declive positivo. Se $r = -1$ temos uma recta perfeita com declive negativo. Se as variáveis são independentes $r \simeq 0$.

Uma interpretação usual: r^2 mede a percentagem de variabilidade de uma das variáveis explicada pela outra.

Teste ao significado do coeficiente de Pearson (PARAMÉTRICO)

Podemos testar se duas variáveis são correlacionadas através das hipóteses:

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

onde ρ representa o coeficiente de correlação da população onde foram retirados os dados.

Estas hipóteses são equivalentes a

H_0 : As variáveis são independentes $\quad vs$
H_1 : As variáveis são (linearmente) dependentes.

Pressupostos do teste

1. os dados constituem duas amostras aleatórias emparelhadas,

2. ambas as populações de onde foram retirados as amostras têm distribuição Normal,

3. a relação entre as variáveis é de forma linear, caso exista.

2 - O coeficiente de correlação de Spearman (*Spearman rank-order coefficient*)

Aplica-se a duas variáveis medidas pelo menos numa escala ordinal, ou que apresentam uma relação não necessariamente linear mas monótona (se uma aumenta a outra tem sempre tendência a aumentar (ou a diminuir)). Aplica-se ainda quando não são satisfeitos os requisitos to teste ao coeficiente de Pearson (variáveis não Normais).

Dadas duas amostras de observação ordenáveis, substitui-se cada um dos seus valores pela sua ordem de ordenação, em inglês *rank*. O coeficiente de Spearman não é mais do que o coeficiente de Pearson aplicado aos *ranks*.

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n}$$

onde d_i representa a diferença de *ranks* correspondentes a cada par de observações x_i, y_i.

Teste ao significado do coeficiente de Spearman (NÃO PARAMÉTRICO)

Tal como no caso do coeficiente de Pearson é possível testar se a correlação é significativa:

H_0 : As variáveis não são correlacionadas $\quad vs$
H_1 : As variáveis são correlacionadas.

Pressupostos do teste

1. os dados constituem duas amostras aleatórias emparelhadas,
2. as variáveis foram medidas numa escala pelo menos ordinal,
3. as populações de onde foram retirados as amostras têm distribuição contínua,
4. a relação entre as variáveis deve ser monótona (não necessariamente linear).

3- O coeficiente de correlação de Kendall

Uma alternativa ao coeficiente de Spearman é o coeficiente de Kendall (*Kendall's tau coeffficient*) que se aplica nas mesmas condições.

Uma diferença muito importante entre os dois coeficientes (Kendall e Spearman) reside na sua interpretação e na impossibilidade de comparar directamente valores provenientes de ambos. Embora o objectivo comum seja o de medir associação, a forma de o fazer é distinta.

O coeficiente de Kendall é muitas vezes descrito como uma medida de concordância entre dois conjuntos de classificações relativas a um conjunto de objectos ou experiências.

$$T = \frac{\#\text{concordâncias} - \#\text{discordâncias}}{\text{número total de pares}}$$

Teste ao significado do coeficiente de Kendall (NÃO PARAMÉTRICO)

Tal como para os coeficientes de Pearson e Spearman é possível efectuar um teste de hipóteses para averiguar se a correlação é significativa.

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

onde τ representa o coeficiente na população.

Pressupostos do teste

1. os dados constituem duas amostras aleatórias emparelhadas,
2. as variáveis foram medidas numa escala pelo menos ordinal,
3. as populações de onde foram retirados as amostras têm distribuição contínua.

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

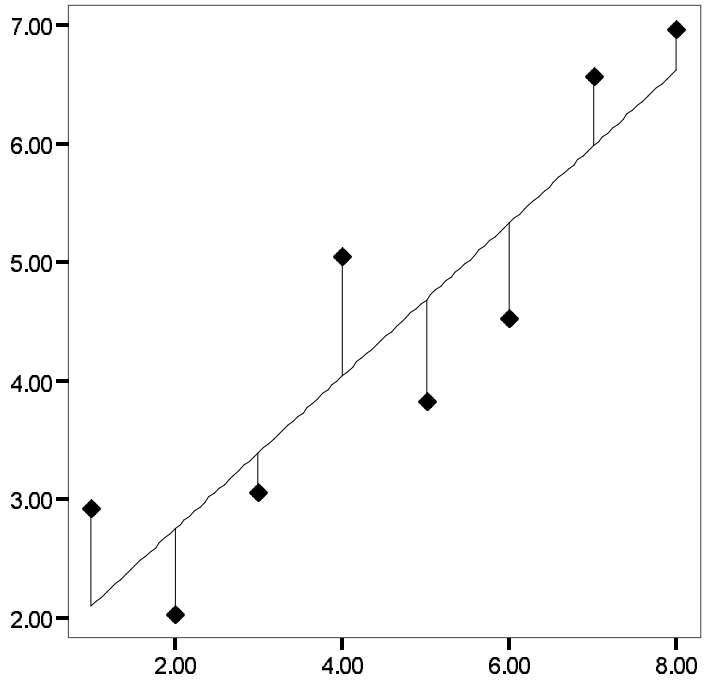
$$Y_i = b_0 + b_1 x_i + \epsilon_i, \quad \epsilon_i \text{ - erro,}$$

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

- os erros devem ser independentes e identicamente distribuídos, $\epsilon_i \frown N(0, \sigma)$. 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

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

Para simplificar a notação iremos adoptar as seguintes convenções habituais:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad S_{xY} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})$$

$$SS_E = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad SS_R = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

Os estimardoes de mínimos quadrados dos coeficientes da recta de regressão são dados por

$$\begin{cases} \hat{b}_0 = \bar{Y} - \hat{b}_1\bar{x} \\ \hat{b}_1 = \dfrac{S_{xY}}{S_{xx}} \end{cases} .$$

A recta de regressão é então dada por

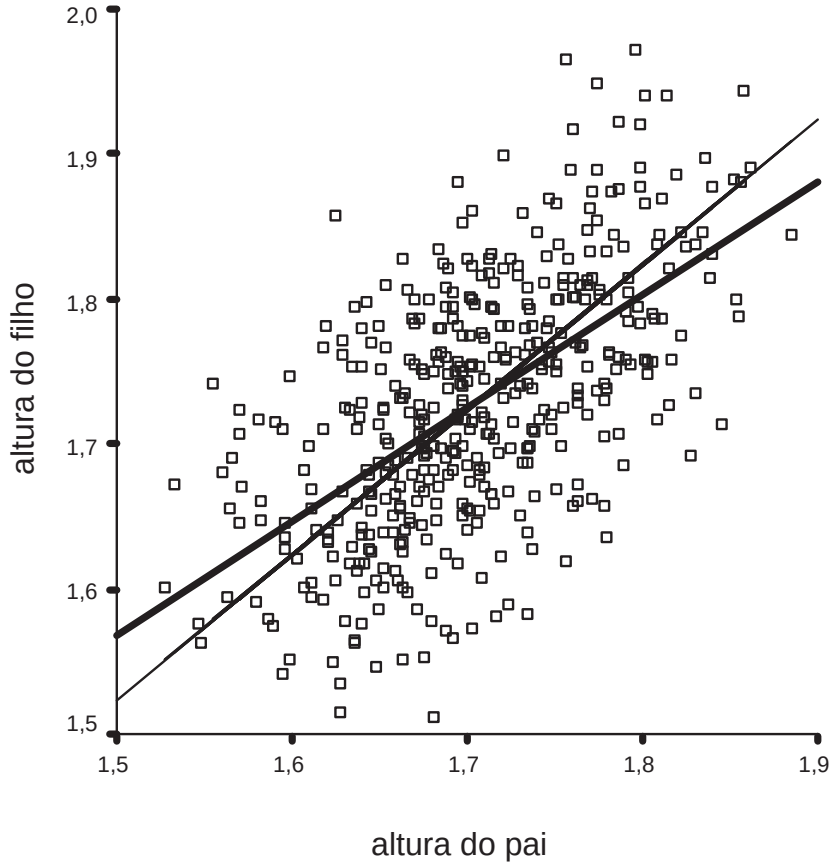
$$y = \hat{b}_0 + \hat{b}_1 x.$$

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.



Propriedades dos estimadores

Com base nos pressupostos do modelo de regressão linear simple podemos calcular a esperança e a variância dos estimadores $\hat{b}_0$ e $\hat{b}_1$.

$$E[\hat{b}_1] = b_1 \quad Var[\hat{b}_1] = \frac{\sigma^2}{S_{xx}}$$

$$E[\hat{b}_0] = b_0 \quad Var[\hat{b}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$

Uma vez que os erros têm distribuição Normal, deduz-se que

$$\hat{b}_1 \frown N\left(b_1, \frac{\sigma^2}{S_{xx}}\right)$$

$$\hat{b}_0 \frown N\left(b_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)\right)$$

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:

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

Para o declive:

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

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`.

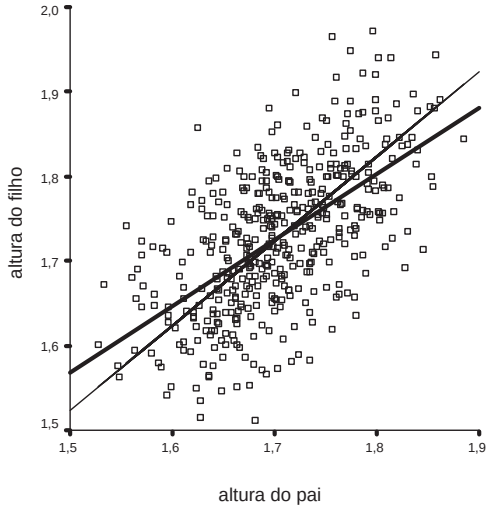
ANOVA da regressão

Em geral o software estatístico efectua uma ANOVA sobre a análise de regressão. No caso da regressão linear simples a ANOVA vai apenas repetir (indirectamente) o teste ao declive e não fornece informação adicional. (Reparar que o p-value da tabela de ANOVA tem o mesmo valor do p-value da tabela de regressão respeitante ao declive.) Só no caso de regressões múltiplas é que a ANOVA produz informação adicional. Por esta razão não iremos descrever a ANOVA da regressão.

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:

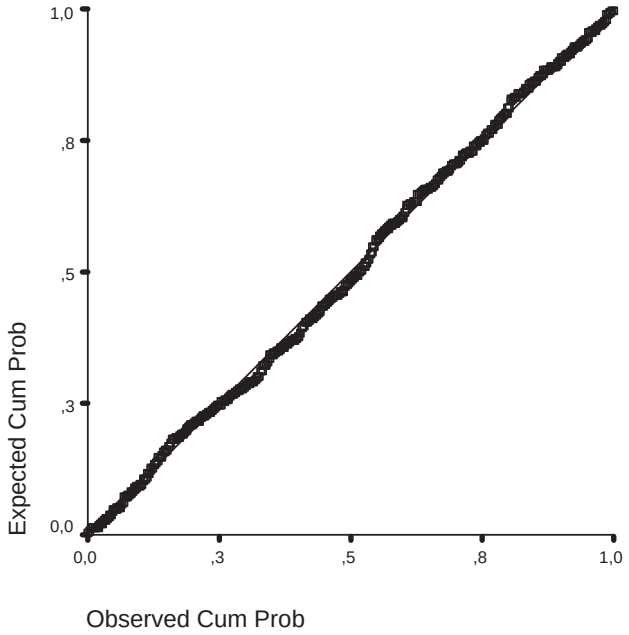
QQ-plot aos resíduos.

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:

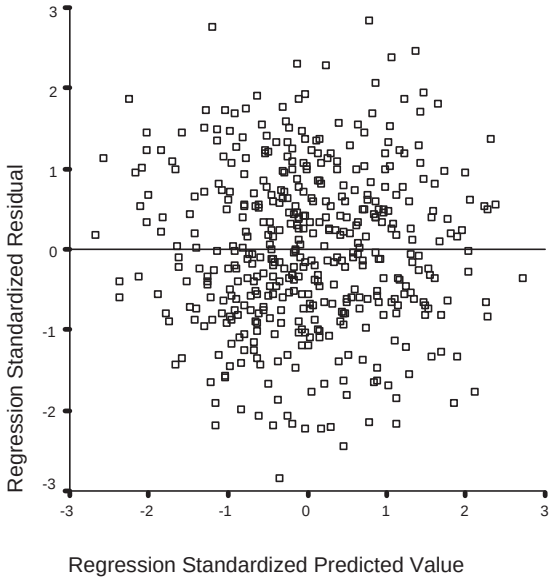


No SPSS pode-se obter o QQ-plot dos resíduos seleccionando a opção `Normal probability plot` no botão `Plots` do menu da regressão linear.

Também se podem fazer um teste de ajustamento à Normal.

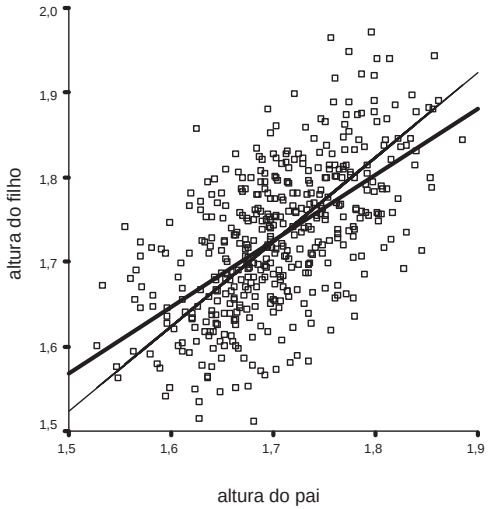
Para avaliar se os erros são independentes:

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:



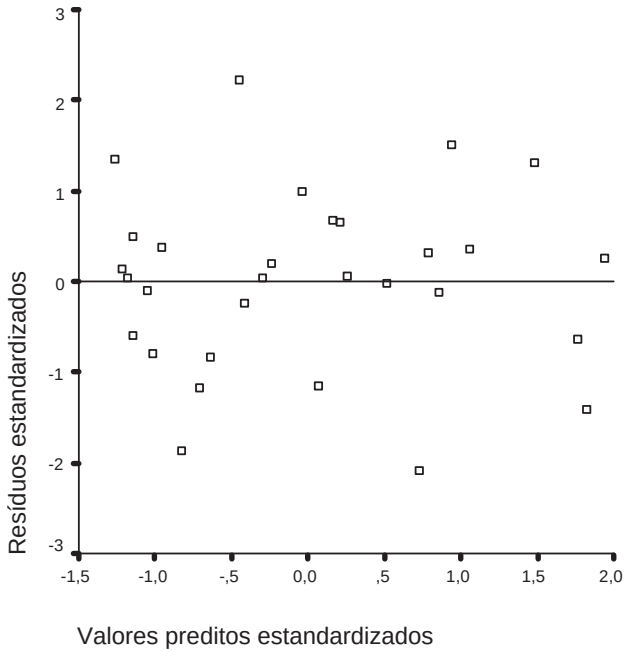
Este gráfico deve apresentar uma relação linear e os pontos devem distribuir-se aleatoriamente em torno da recta com variabilidade constante.

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.

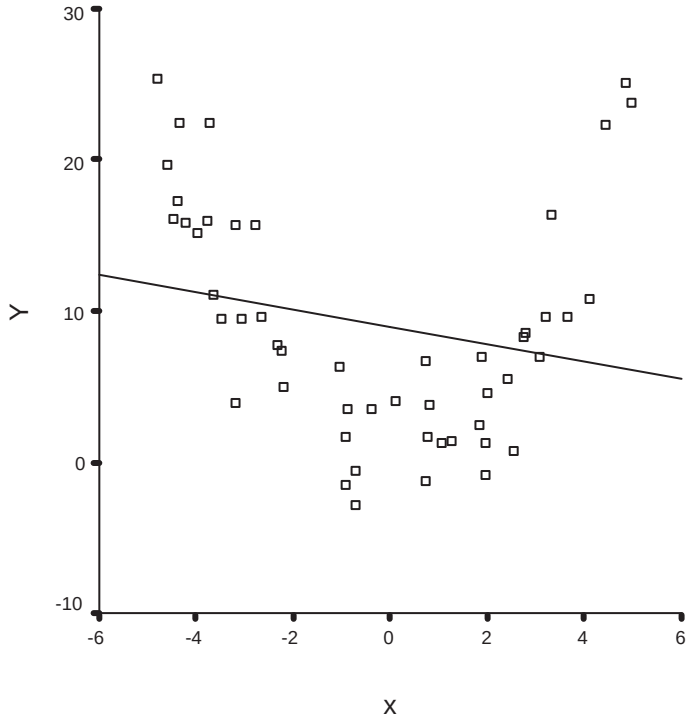
Outras formas de identificar uma possível não-adequação do modelo

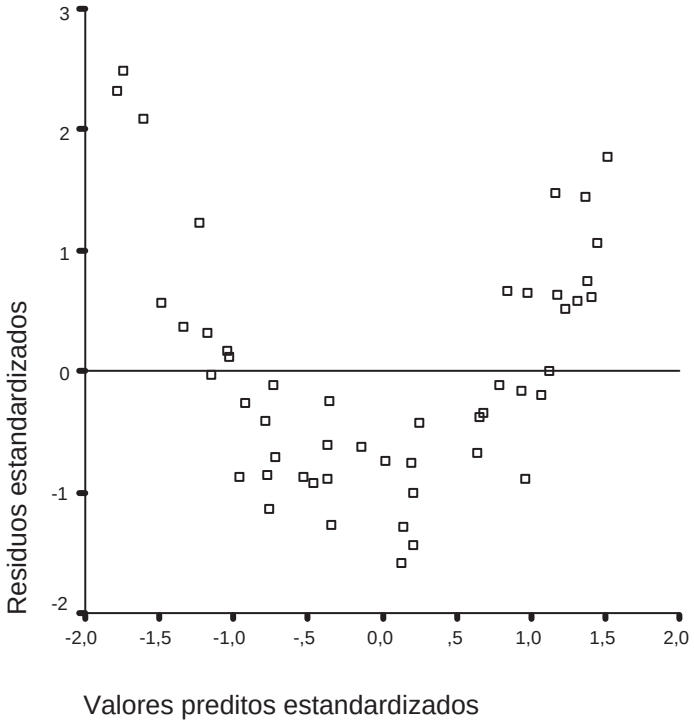
Os gráficos de resíduos podem sugerir não-linearidades na relação entre as variáveis ou alterações na variância dos erros.

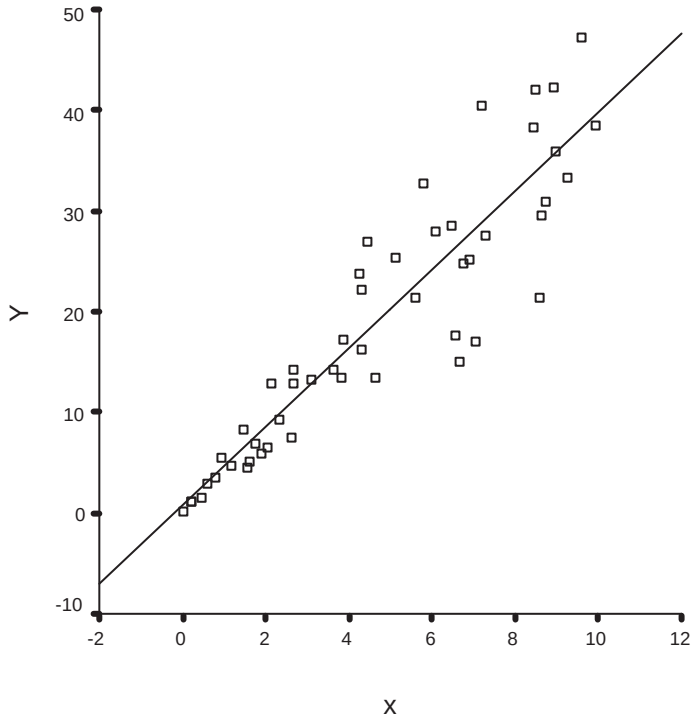
Gráfico de resíduos típico quando são válidos os pressupostos do moedelo:

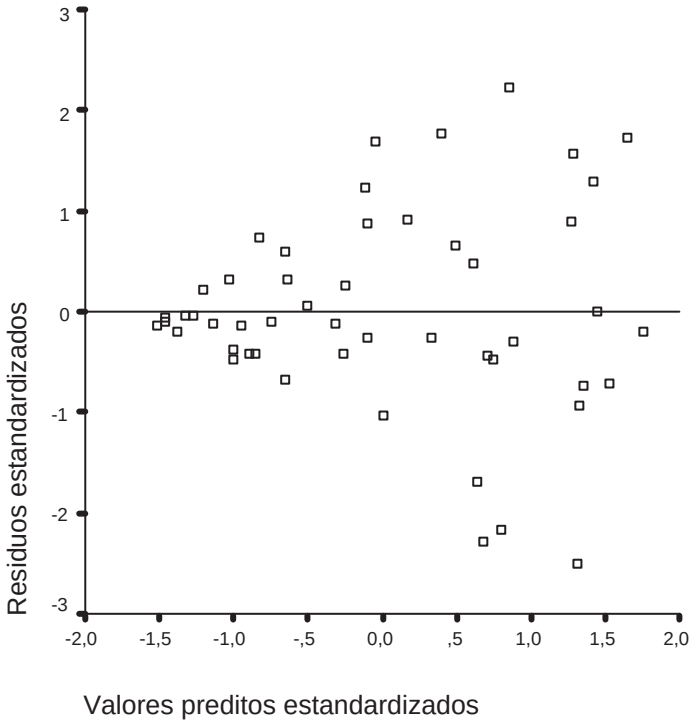


Exemplos de gráficos quando não são válidos os pressupostos do moedelo:









Quando há suspeitas de não linearidades no modelo deve-se transformar os dados por forma a obter um modelo linear (quando possível).

Transformações de variáveis

Quando um conjunto de dados não permite validar os pressupostos de aplicabilidade de uma determinada técnica estatística podemos procurar técnicas alternativas ou então tentar transformar os dados de forma a obter novas amostras em condições de validar os ditos pressupostos. Exemplos típicos destas situações são os seguintes:

1. Uma amostra evidencia bastante assimetria e não se pode considerar como sendo proveniente de uma população Normal (o QQ-plot não é linear e os testes de ajustamento rejeitam a hipótese de normalidade).

2. As amostras envolvidas numa ANOVA apresentam variâncias tão diferentes que se rejeita a hipótese de homogeneidade de variâncias.

3. Um gráfico de dispersão entre duas variáveis indicia existir uma relação entre as variáveis mas essa relação é claramente não-linear. Neste caso não é possível efectuar uma análise de regressão linear nem se pode fazer um teste de correlação utilizando o coeficiente de correlação de Pearson.

De entre as transformações possíveis as mais utilizadas são as seguintes:

- Transformação logarítmica:

$$X' = \ln X \ , X > 0 \quad (\text{ou } \ X' = \ln(X + a), \ a \in \mathbb{R})$$

  Esta transformação é útil para tornar mais simétrica uma distribuição que apresente assimetria positiva.
  Também é útil para diminuir a variabilidade nos valores mais elevados e aumentar a variabilidade nos valores próximos de 0.
  Quando um gráfico de dispersão apresenta um crescimento de tipo exponencial, uma transformação logarítmica aos valores de y tornam o gráfico linear.

- Raiz quadrada:

$$X' = \sqrt{X}, \ X > 0$$

  Tem uma função semelhante à transformação logarítmica mas a transformação não é tão acentuada.

- Transformação potência:

$$X' = X^b, b > 0$$

  Quando $b > 1$ esta transformação faz o contrário da transformação logarítmica, i.e.: pode tornar mais simétricas distribuições com assimetria negativa; pode diminuir a variabilidade de valores próximos de 0 e aumentar a variabilidade de valores elevados; pode tornar mais linear um gráfico de dispersão que apresente uma relação do tipo $y = \sqrt[a]{x},\ a > 1$. Quando $b < 1$ as consequências são semelhantes às da transformação logarítmica.

- Transformação inversa:

$$X' = 1/X$$

Cuidados a ter na transformações de variáveis

Atenção que quando se transformam variáveis, os resultados a que se chega para as variáveis transformadas não se podem converter facilmente para as variáveis originais. Por isso, as conclusões a retirar são relativas às variáveis transformadas e isso deve ficar explícito nos textos a elaborar.

Por exemplo, se construirmos um intervalo de confiança para a média de uma variável $X' = \ln X$, não se pode transformar o intervalo obtido num intervalo para a média de X por aplicação da transformação inversa (exponencial) aos extremos do intervalo. Isto acontece porque a média de X', $\mu' = E[X']$, não é o logaritmo da média de X, $\mu = E[X]$. ($\mu' \neq \ln \mu$!)