Metodologia de Diagnóstico e Elaboração de Relatório – FASHT

Correlação e Regressão linear simples

Prof. Cesaltina Pires
cpires@uevora.pt

Plano da Apresentação

- **Correlação linear**
  - Diagrama de dispersão
  - Covariância
  - Coeficiente de correlação de Pearson
  - Teste de correlação de Spearman
- **Regressão linear simples**
  - A recta de regressão
  - O método dos mínimos quadrados
  - Poder explicativo da regressão

Associação entre hábitos leitura e escolaridade

Anos de Escolaridade	Nº de livros lidos por ano
9	6
12	16
10	11
6	12
8	7
17	15
10	9
11	13
7	10

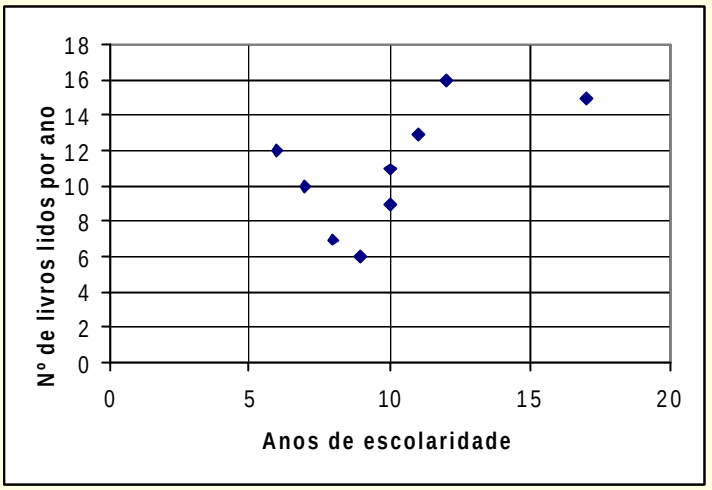
Escolaridade média = 10
Nº médio de livros lidos = 11

Em geral, quanto maior é a escolaridade, maior é o nº de livros lidos. Mas a relação não é exacta.

Correlação linear

- A análise de correlação é usada para medir o grau de associação (linear) entre variáveis quantitativas
- Queremos estudar «relação» entre variáveis. Esta relação não é uma relação matemática exacta, é uma relação estatística.
- Exemplo: relação entre nº de livros lidos por ano e anos de escolaridade. Em geral, quanto maior o nº de anos de escolaridade, maior é o nº de livros lidos (mas esta relação não é exacta, há muitas excepções)
- Diagrama de dispersão pode ajudar a visualizar o grau de associação

Diagrama de dispersão



Covariância

Anos de Escolaridade	Nº livros lidos	Desvio (anos esc)	Desvio (nº livros)	Produto desvios
9	6	-1	-5	5
12	16	2	5	10
10	11	0	0	0
6	12	-4	1	-4
8	7	-2	-4	8
17	15	7	4	28
10	9	0	-2	0
11	13	1	2	2
7	10	-3	-1	3

$$\text{cov} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n}$$

$$\text{Covariância} = \frac{5+10+0-4+8+28+0+2+3}{9}$$

6

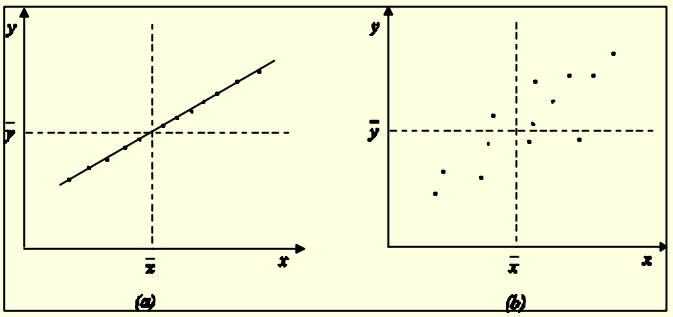
Coeficiente de correlação de Pearson

- A covariância é sensível às unidades de medida
- O coeficiente de correlação também mede o grau de associação entre as variáveis mas não é sensível às unidades de medida
- O coeficiente de correlação entre as variáveis *x* e *y* obtêm-se dividindo a covariância entre *x* e *y* pelos seus desvios padrões:

$$\rho_{x,y} = \operatorname{corr}(x,y) = \frac{\operatorname{cov}(x,y)}{s_x s_y}$$

O coeficiente de correlação pode tomar valores entre -1 e +1.

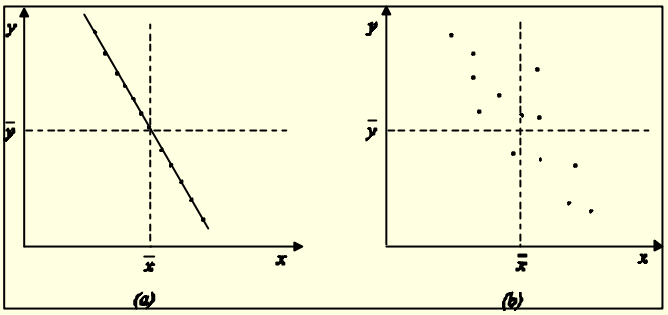
Diagrama de dispersão e correlação



Correlação positiva perfeita
? ?? ??

Correlação positiva
? ?= ? ?= 1

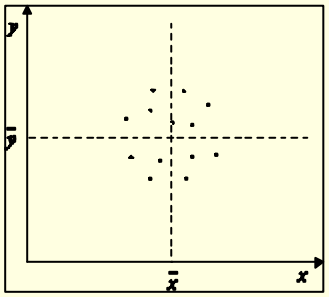
Diagrama de dispersão e correlação



Correlação negativa perfeita
? ?? ?? ?

Correlação negativa
? ?≔ ?≔ 0

Diagrama de dispersão e correlação



Correlação linear nula
? ?? ??

Quando ? = 0 as variáveis não Estão linearmente relacionadas. Mas pode existir outro tipo de relação (não linear) entre as variáveis.

Coeficiente de correlação de Spearman

- O coeficiente de correlação de Pearson é bastante sensível á presença de outliers
- Testes de hipóteses da correlação são baseados na hipótese da normalidade da população
- Por isso, pode ser conveniente usar medidas que sejam válidas seja qual for a distribuição na população e menos sensíveis á presença de outliers.
- Pode também ser interessante ter medidas de correlação mesmo que as variáveis sejam qualitativas (mas com escala ordinal)

Coeficiente de correlação de Spearman

- As observações são ordenadas por ordem crescente da variável *x* e são também ordenadas por ordem crescente da variável *y*.
- Ficamos assim a saber a ordem (o rank) de cada observação relativamente à variável x e relativamente à variável y.
- O coeficiente de Correlação de Spearman é o coeficiente de correlação entre a rank da variável x e o rank da variável y.

Coeficiente de correlação de Spearman

Anos de Escolaridade	Rank escolaridade	Nº livros lidos	Rank nº de Livros lidos
9	4	6	1
12	8	16	9
10	5	11	5
6	1	12	6
8	3	7	2
17	9	15	8
10	5	9	3
11	7	13	7
7	2	10	4

Que significa a existência de 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 (Ex: idade do marido, idade da esposa)
- As duas variáveis podem estar correlacionadas porque ambas são influenciadas por uma terceira variável e é o facto de ambas «responderem» a variações nessa variável que explica a correlação (Ex: nº de insolações e produção de trigo)

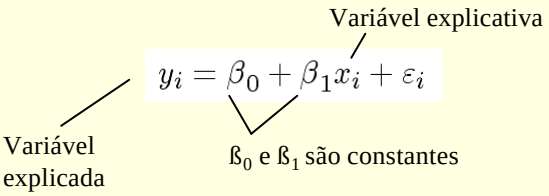
→ A existência de correlação não implica causalidade

Regressão linear simples

- Quando medimos correlação linear estamos a medir o grau de associação linear entre as variáveis. Tanto faz falar de correlação entre *x* e *y*, como correlação entre *y* e *x*.
- Quando fazemos regressão linear também queremos estudar relação entre variáveis, mas queremos estudar se uma das variáveis depende da outra.
- Na regressão linear simples há uma variável explicativa (ou independente) e uma variável explicada (ou dependente). O que queremos saber é se a variável explicativa ajuda (ou não) a explicar o comportamento da variável explicada.

Regressão linear simples

A relação entre *y* e *x* é uma relação linear:



$ß_0$ é a intersecção na origem da recta e $ß_1$ é o declive da recta.
O termo e_i é um termo aleatório que capta a influência de outros pequenos factores que influenciam *y*, para além do *x*. A média dos e_i é zero.

Na regressão linear simples há só uma variável explicativa.
Se houver várias variáveis explicativas temos regressão linear múltipla

Regressão linear simples

A relação entre *y* e *x* é uma relação linear:

Variável explicativa

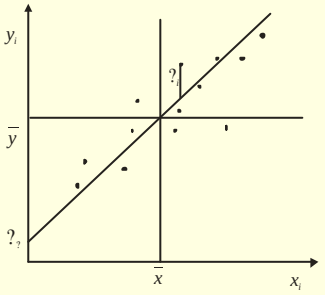
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Variável explicada

ß$_0$ e ß$_1$ são constantes

ß$_0$ diz-nos qual é o valor médio de *y* quando *x* toma o valor 0.
ß$_1$ diz-nos quão sensível é *y* às variações de *x*. Mais concretamente, se *x* aumentar de 1 unidade, o valor de y aumenta ß$_1$ unidades.

Regressão linear simples



Se a relação entre *y* e *x* fosse exacta todas as observações estariam na recta. Mas a relação não é exacta, há outros factores aleatórios que influenciam *y*, para além de *x*. Há pontos acima da recta (desvios positivos) e pontos abaixo da recta (desvios negativos).

Regressão linear simples

- O que vamos ter é uma amostra de observações, cada uma das quais com determinados valores de *x* e *y*.
- Se representamos o diagrama de dispersão ficamos com uma «nuvem de pontos» no espaço (*x*,*y*).
- Com base nessa amostra queremos estimar a relação entre *y* e *x*. Qual é a recta que melhor se ajusta à nuvem de pontos? Qual é a intersecção na origem e qual é o declive dessa recta (quanto são $ß_0$ e $ß_1$)?

Método dos mínimos quadrados

- É um método para estimar os parâmetros $ß_0$ e $ß_1$, com base na informação de uma amostra
- Para uma dada recta podemos calcular os desvios em relação à recta (desvios positivos e negativos compensam-se). Podemos também calcular a soma dos desvios ao quadrado.
- A «melhor recta» é aquela para o qual a soma dos desvios ao quadrado é menor.
- Usando este método ficamos com estimativas dos parâmetros $ß_0$ e $ß_1$.
- O Excel estima regressão, só é preciso saber interpretar.

Exemplo – o que explica o salário?

Consideremos o seguinte modelo:

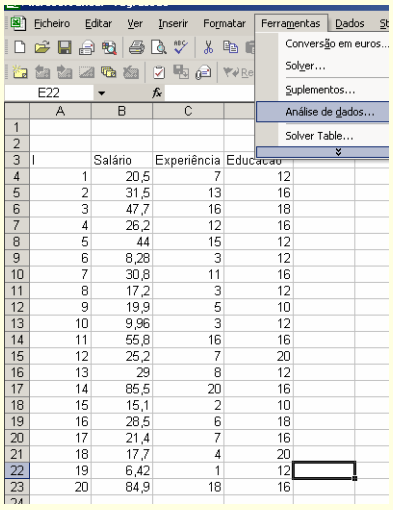
$S_i = ß_0 + ß_1 Exp_i + e_{i,}$

Onde S_i – salário do indivíduo i
Exp_i – anos de experiência do indivíduo i

$ß_0$ diz-nos qual é o valor médio do salário para trabalhadores sem experiência.
$ß_1$ diz-nos quanto varia o salário por cada ano adicional de experiência.

Será que este modelo é bom? Não haverá outros factores importantes que influenciam o salário? Será que a relação é linear?...

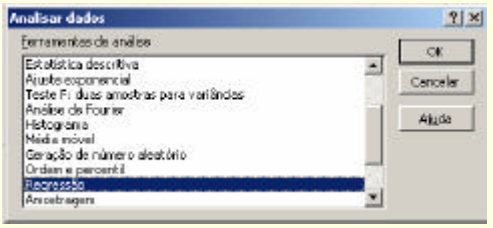
Estimação da regressão no Excel



	A	B	C	D
1				
2				
3	I	Salário	Experiência	Educação
4	1	20,5	7	12
5	2	31,5	13	16
6	3	47,7	16	18
7	4	26,2	12	16
8	5	44	15	12
9	6	8,28	3	12
10	7	30,8	11	16
11	8	17,2	3	12
12	9	19,9	5	10
13	10	9,96	3	12
14	11	55,8	16	16
15	12	25,2	7	20
16	13	29	8	12
17	14	85,5	20	16
18	15	15,1	2	10
19	16	28,5	6	18
20	17	21,4	7	16
21	18	17,7	4	20
22	19	6,42	1	12
23	20	84,9	18	16

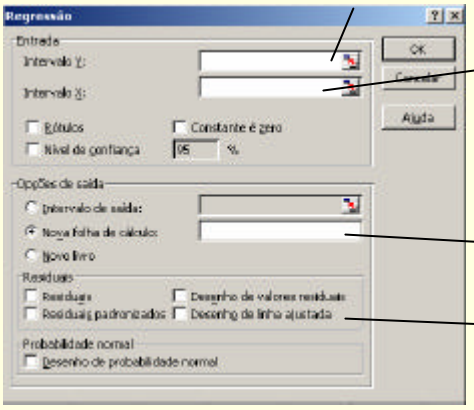
Amostra com 20 observações (é pouco!)
Variáveis Salário e experiência

No menu das Ferramentas escolher Análise de Dados, depois escolher regressão.



Estimação da regressão no Excel

onde estão dados da variável explicada



onde estão dados da variável explicativa

onde queremos por os resultados

opções

Resultados da regressão

medidas da qualidade do ajustamento

Estimadores dos parâmetros $ß_0$ e $ß_1$.

	A	B	C	D	E	F	G	H	I
1	SUMÁRIO DOS RESULTADOS								
2									
3	*Estatística de regressão*								
4	R múltiplo	0,912544							
5	Quadrado de R	0,832736							
6	Quadrado de R ajustado	0,823443							
7	Erro-padrão	9,40202							
8	Observações	20							
9									
10	ANOVA								
11		*gl*	*SQ*	*MQ*	*F*	*e significância*			
12	Regressão	1	7921,717	7921,717	89,61424	2,06E-08			
13	Residual	18	1591,164	88,39797					
14	Total	19	9512,881						
15									
16		*Coeficiente*	*Erro-padrão*	*Stat t*	*valor P*	*95% inferior*	*5% superior*	*inferior 95,0%*	*perior 95,0%*
17	Interceptar	0,347908	3,885271	0,089545	0,929637	-7,81475	8,510566	-7,81475	8,510566
18	Variável X 1	3,494926	0,36919	9,46648	2,06E-08	2,719287	4,270565	2,719287	4,270565

A recta estimada é:

$S_i = 0{,}348 + 3{,}495 Exp_i$,

Qualidade do ajustamento – R^2

Pode mostrar-se que a variação total da variável explicada se pode decompor em:

Variação explicada pelo modelo
(y depende de x, e com x varia y também varia)
+
variação residual
(não explicada pelo modelo)

$$R^2 = \frac{\text{Variação explicada}}{\text{Variação total}}$$

Qualidade do ajustamento – R^2

R^2 diz-nos qual é a percentagem da variação total da variável dependente que é explicada pelo modelo

No nosso exemplo $R^2 = 0,83$, o que significa que 83% da variação total no salário é explicada pelo nosso modelo.