UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS
STT – DEPARTAMENTO DE ENGENHARIA DE TRANSPORTES
STT0405 - Planejamento e Análise de Sistemas de Transportes

CORRELAÇÃO E REGRESSÃO

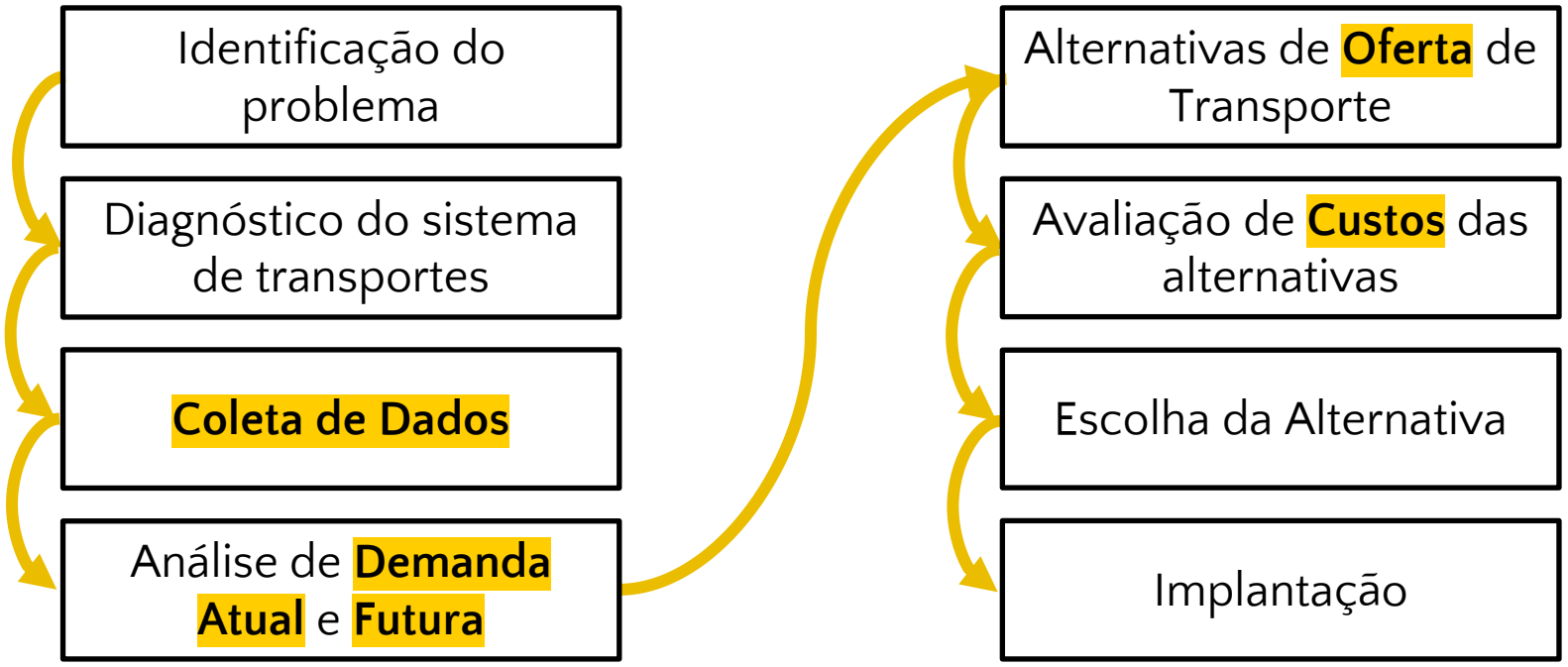
Professora:

Cira Souza Pitombo

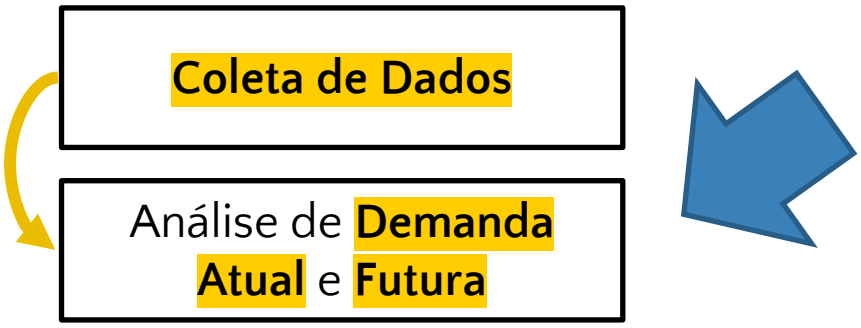

K!

Por qual razão devo aprender Regressão Linear?

Etapas gerais de um projeto de **mobilidade urbana**



Etapas gerais de um projeto de mobilidade urbana



EXEMPLOS

Quantidade de passageiros	População economicamente ativa
20.000	5.200
34.000	8.400
39.000	9.000
55.000	11.000
63.000	19.000
70.000	23.000
97.000	35.000

Quantidade de viagens motorizadas no domicílio	Quantidade de automóveis no domicílio
2	1
2	1
4	2
4	1
7	3
8	4
12	5

Quantidade de viagens originadas na zona de tráfego	População residente na zona de tráfego
25	13
45	22
50	25
55	26
58	27
61	30
70	35

CORRELAÇÃO

VARIÁVEIS CORRELACIONADAS

ASSOCIAÇÃO ENTRE DUAS VARIÁVEIS

QUANTITATIVAS – correlação

VARIÁVEIS CORRELACIONADAS

X e Y positivamente correlacionadas

Quando?

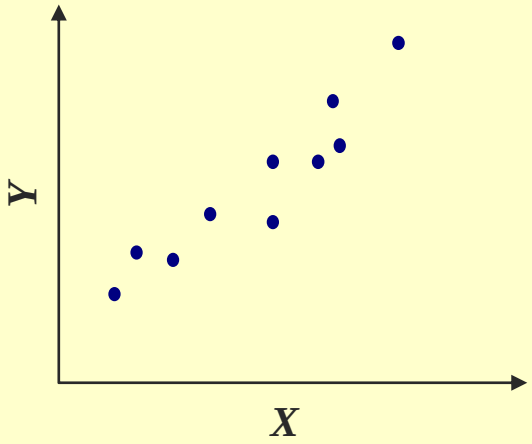
Exemplos?

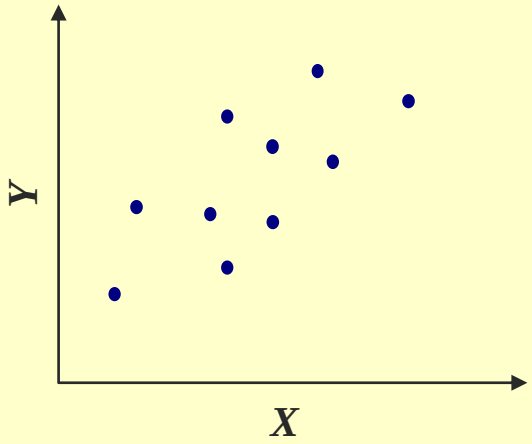
X e Y negativamente correlacionadas

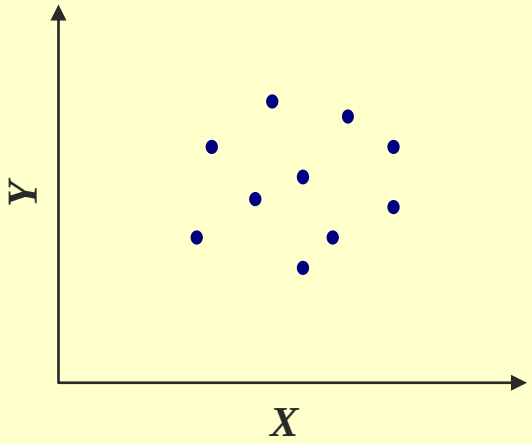
Quando?

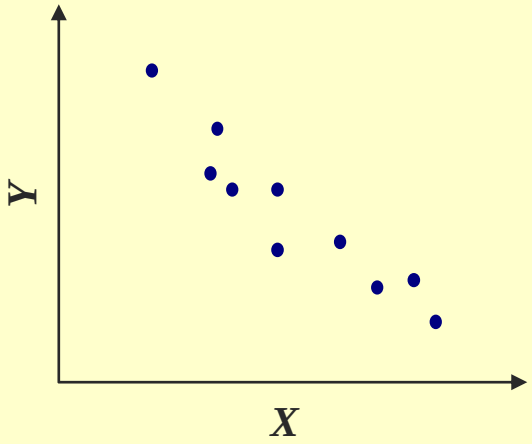
Exemplos?

Diagramas de dispersão









Diagramas de dispersão

DistCap: distância à capital da respectiva Unidade da Federação.

EspVida: esperança de vida ao nascer

MortInf: mortalidade (número médio de mortes em 1.000) até um ano de idade.

Alfab: taxa de alfabetização (percentagem da população adulta alfabetizada).

Renda: renda per capita do município (R$)

Diagramas de dispersão

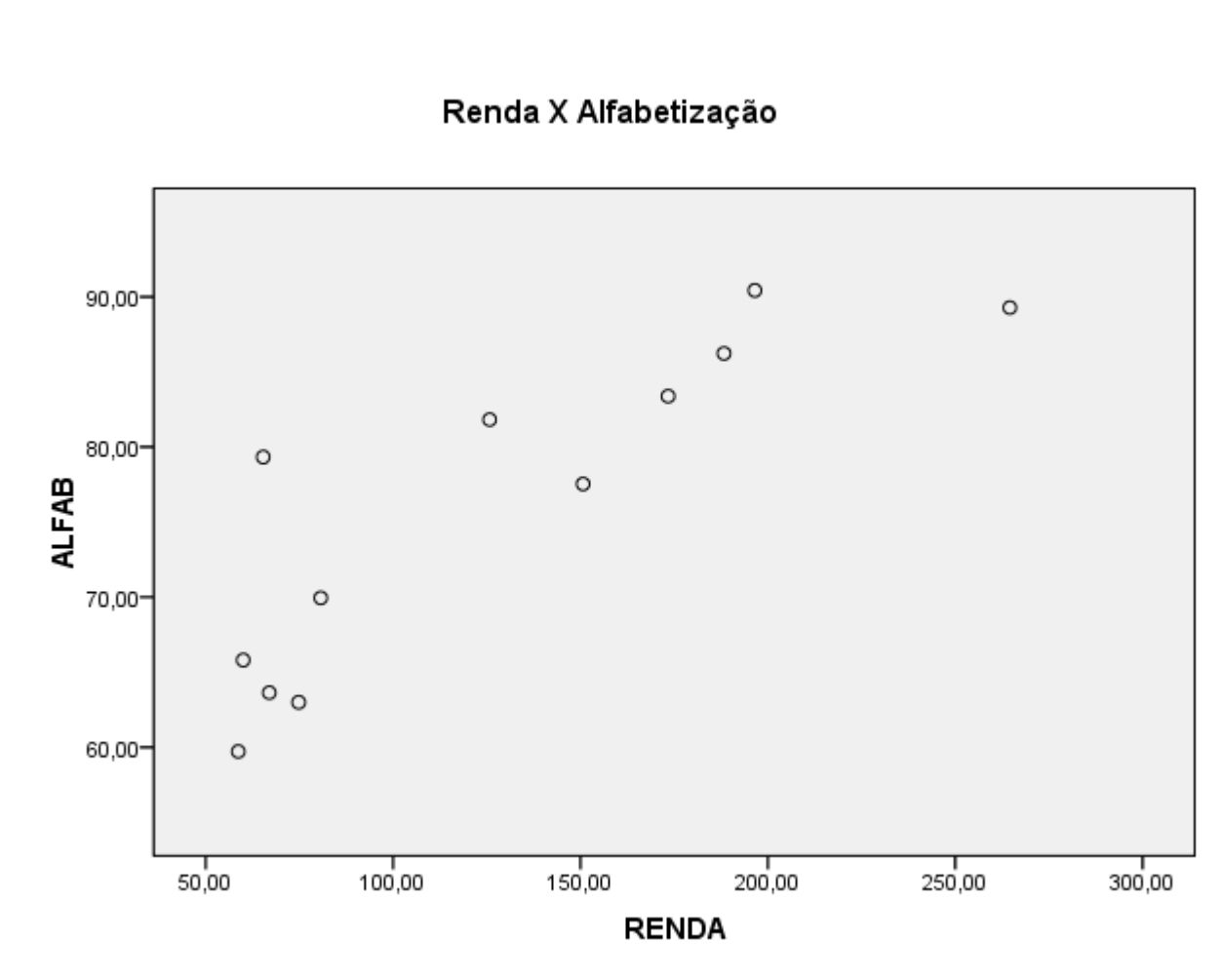
ID	Município	DISTCAP	ESPVIDA	MORTINF	ALFAB	RENDA
1	**Araruna**	365	67,99	23,19	86,23	188,29
2	**Nova Redenção**	278	61,19	56,56	63	74,79
3	**Monção**	150	59,58	63,32	63,64	66,96
4	**Porto Rico do Maranhão**	78	58,96	66,05	79,33	65,34
5	**Campo Erê**	468	68,1	31,71	83,38	173,38
6	**Lagoa do Piauí**	40	63,65	47,08	65,81	60
7	**São José das Palmeiras**	486	71,01	16,62	77,54	150,67
8	**Paraíba do Sul**	83	71,36	15,69	89,28	264,55
9	**Malhada dos Bois**	65	64,46	44,18	69,95	80,69
10	**Jandaíra**	175	62,45	51,57	59,72	58,68
11	**Vespasiano**	14	68,68	32,81	90,43	196,51
12	**Ipaba**	167	67,42	37,04	81,82	125,75

Diagramas de dispersão

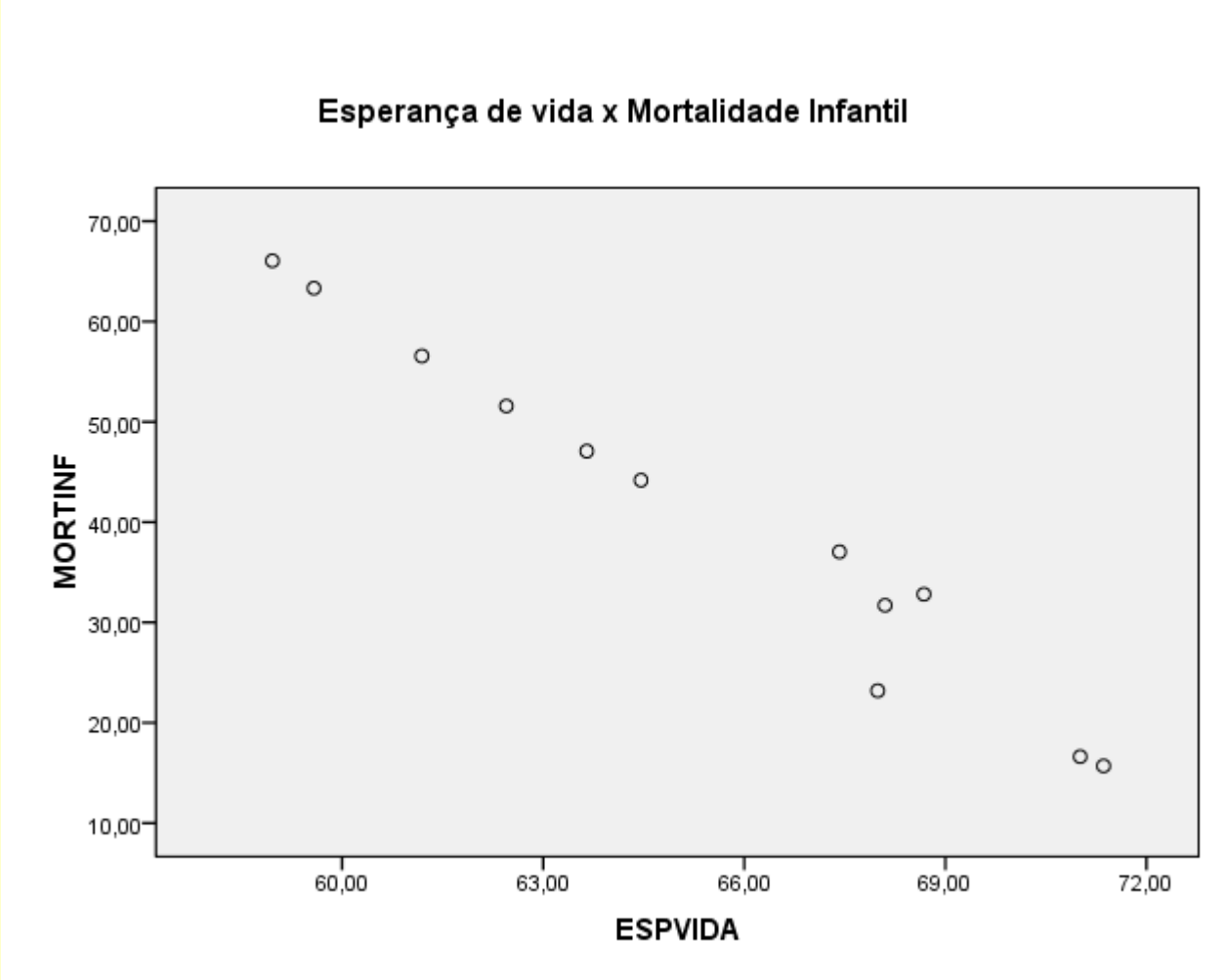
Construa o gráfico de dispersão Renda (eixo x) x Taxa de alfabetização (eixo y)

Qual a relação esperada?

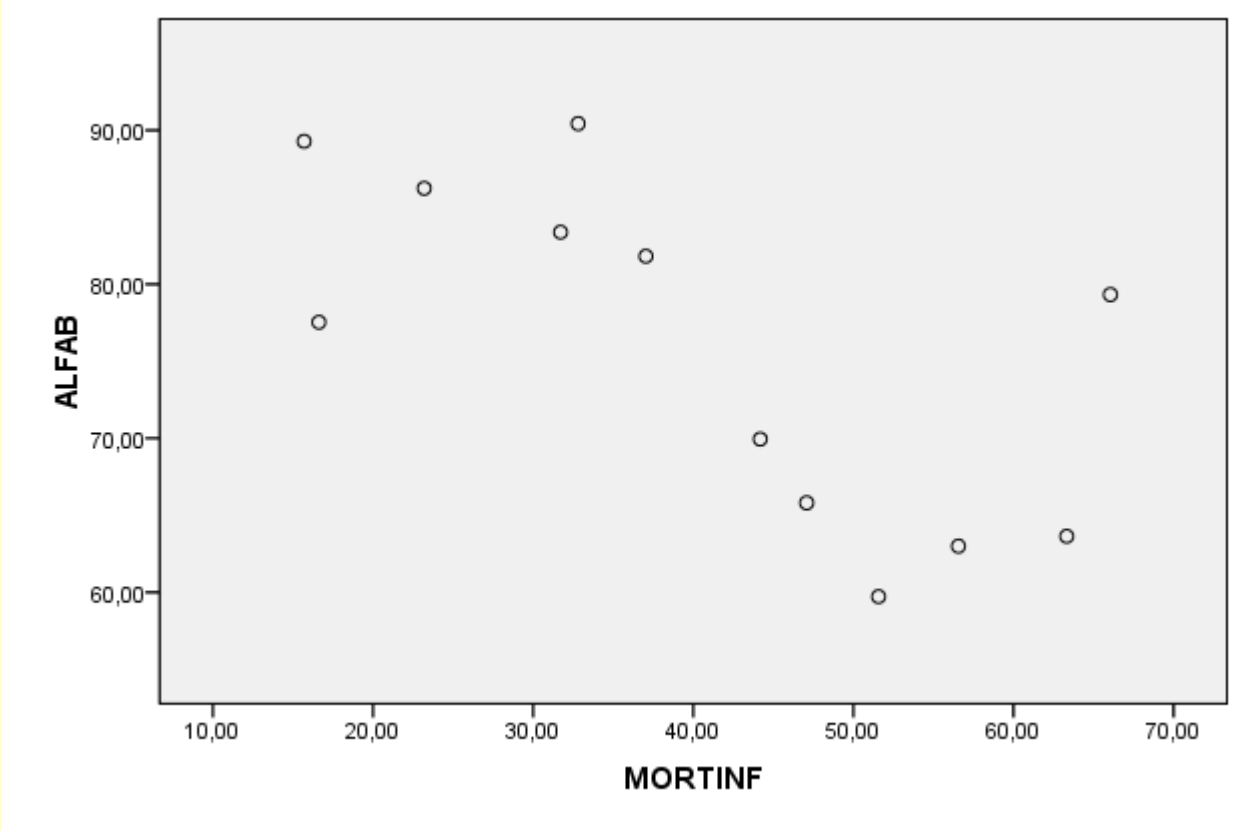
Diagramas de dispersão



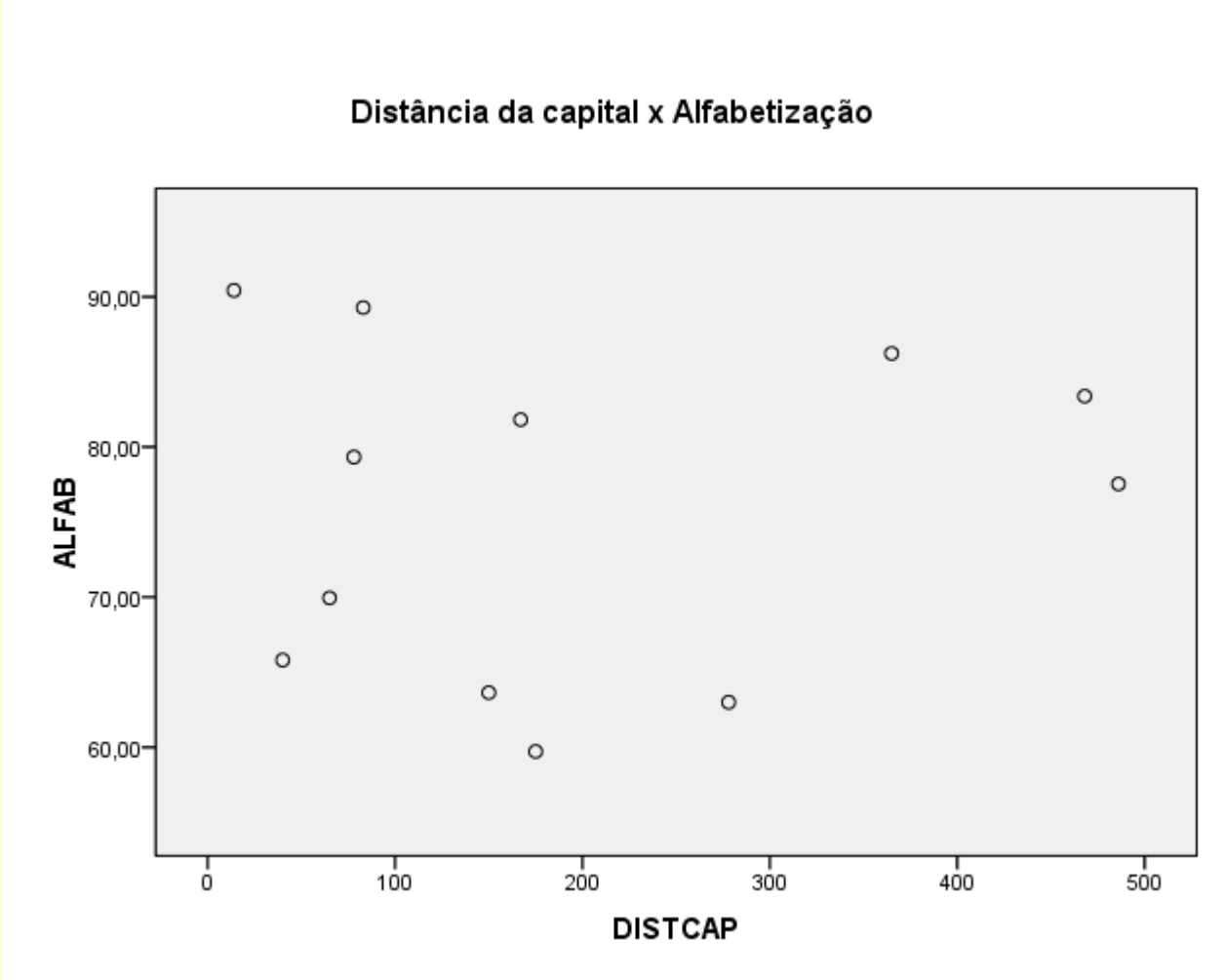
Diagramas de dispersão



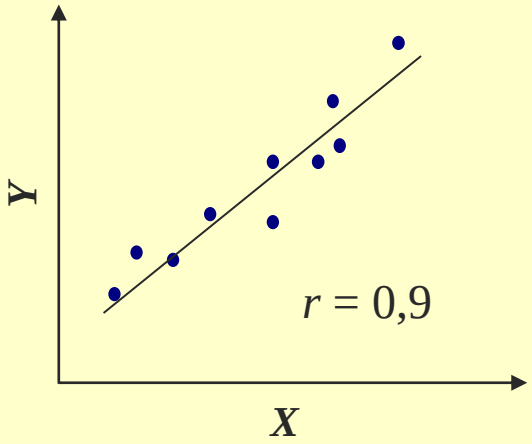
Diagramas de dispersão

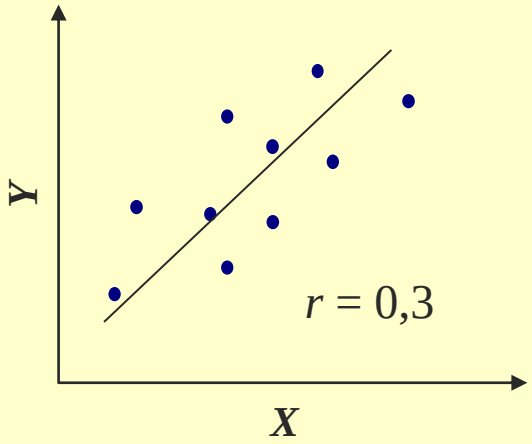


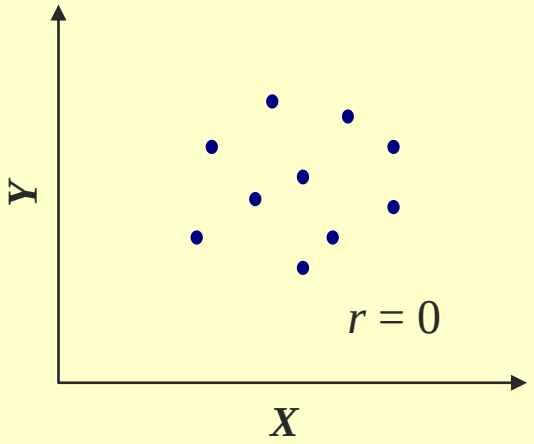
Diagramas de dispersão

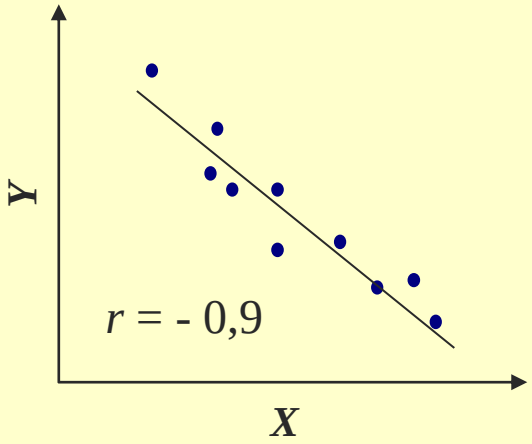


Coeficiente de correlação









Coeficiente de Correlação (de Pearson)
mede o grau de relação linear entre X e Y

Coeficiente de correlação

Os dados devem ser padronizados, X´ (valor padronizado da variável x) e Y´ (valor padronizado da variável y)

Como padronizar?

$$r = \frac{\sum(x' . y')}{n - 1}$$

Coeficiente de correlação

COMO SERIA UM TESTE DE HIPÓTESE??????

QUAL A HIPÓTESE NULA E ALTERNATIVA??????

Coeficiente de correlação

Teste de significância sobre R

Ho: As variáveis X e Y não são correlacionadas
H1: As variáveis X e Y são correlacionadas

P ≥ α – aceitar Ho
P < α – Rejeitar Ho em favor de H1

	Nível de significância, α, num teste bilateral					
n	**0,200**	**0,100**	**0,050**	**0,020**	**0,010**	**0,002**
5	0,687	0,805	0,878	0,934	0,959	0,986
6	0,608	0,729	0,811	0,882	0,917	0,963
7	0,551	0,669	0,754	0,833	0,875	0,935
8	0,507	0,621	0,707	0,789	0,834	0,905
9	0,472	0,582	0,666	0,750	0,798	0,875
10	0,443	0,549	0,632	0,715	0,765	0,847
11	0,419	0,521	0,602	0,685	0,735	0,820
12	0,398	0,497	0,576	0,658	0,708	0,795
13	0,380	0,476	0,553	0,634	0,684	0,772
14	0,365	0,458	0,532	0,612	0,661	0,750
15	0,351	0,441	0,514	0,592	0,641	0,730
16	0,338	0,426	0,497	0,574	0,623	0,711
17	0,327	0,412	0,482	0,558	0,606	0,694
18	0,317	0,400	0,468	0,543	0,590	0,678
19	0,308	0,389	0,456	0,529	0,575	0,662
20	0,299	0,378	0,444	0,516	0,561	0,648
21	0,291	0,369	0,433	0,503	0,549	0,635
22	0,284	0,360	0,423	0,492	0,537	0,622
23	0,277	0,352	0,413	0,482	0,526	0,610
24	0,271	0,344	0,404	0,472	0,515	0,599
25	0,265	0,337	0,396	0,462	0,505	0,588
26	0,260	0,330	0,388	0,453	0,496	0,578

Coeficiente de correlação – SPSS

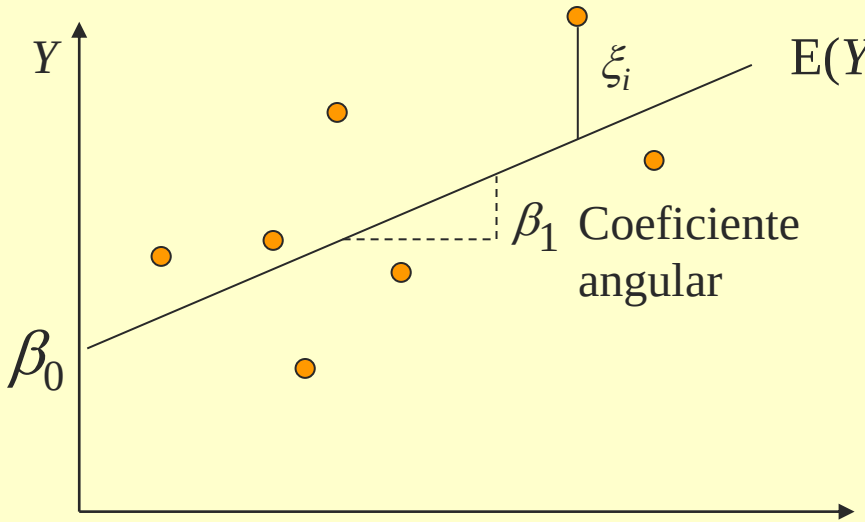
Correlations

		ESPVIDA	MORTINF	DISTCAP	ALFAB	RENDA
ESPVIDA	Pearson Correlation	1	-,983**	,337	,718**	,865**
	Sig. (2-tailed)		,000	,284	,009	,000
	N	12	12	12	12	12
MORTINF	Pearson Correlation	-,983**	1	-,400	-,684*	-,860**
	Sig. (2-tailed)	,000		,198	,014	,000
	N	12	12	12	12	12
DISTCAP	Pearson Correlation	,337	-,400	1	,087	,205
	Sig. (2-tailed)	,284	,198		,788	,523
	N	12	12	12	12	12
ALFAB	Pearson Correlation	,718**	-,684*	,087	1	,863**
	Sig. (2-tailed)	,009	,014	,788		,000
	N	12	12	12	12	12
RENDA	Pearson Correlation	,865**	-,860**	,205	,863**	1
	Sig. (2-tailed)	,000	,000	,523	,000	
	N	12	12	12	12	12

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Análise de Regressão Linear Simples



$E(Y) = \beta_0 + \beta_1 X$

Inclinação

Intercepto

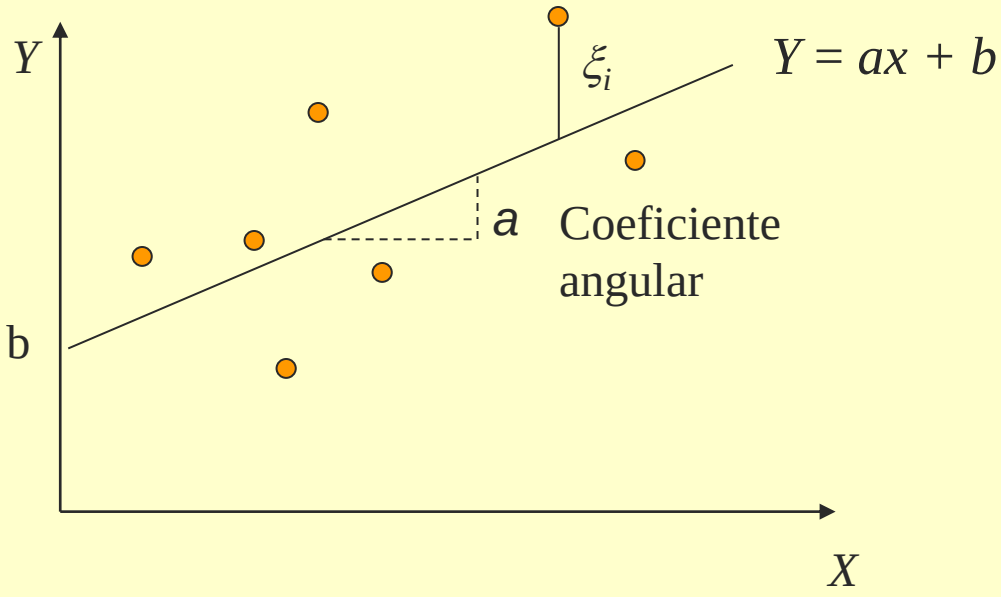
Erro Aleatório

$$Y_i = \beta_0 + \beta_1 X_i + \xi_i$$

Variável Dependente

Variável Independente

Estimativas dos parâmetros



$$a = \frac{n.\sum(X.Y) - (\sum X).(\sum Y)}{n.\sum x^2 - ((\sum X)^2}$$

$$b = \frac{\sum Y - a.\sum X}{n}$$

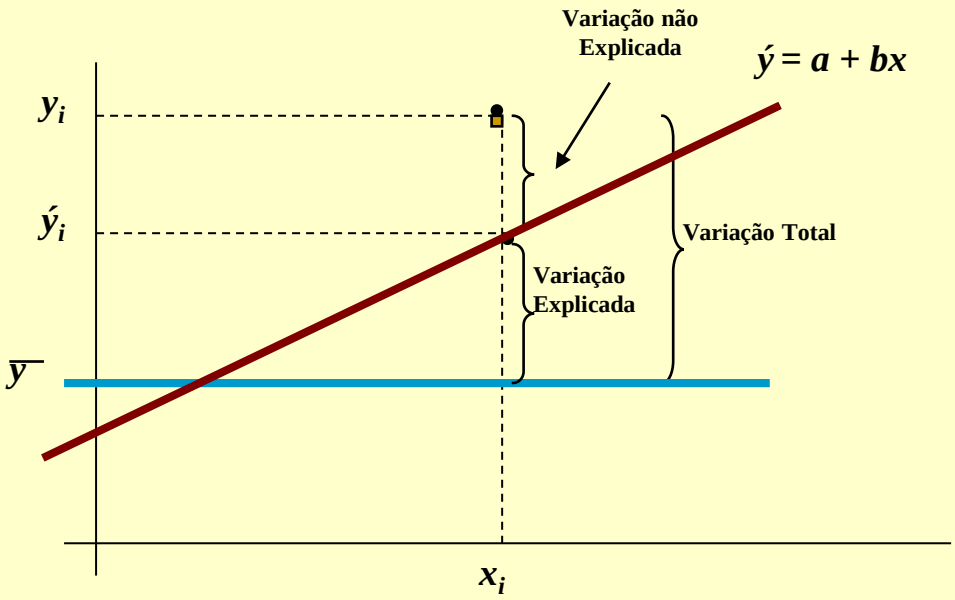
ANÁLISE DE REGRESSÃO

Coeficiente de determinação (R^2)

O coeficiente de determinação deve ser interpretado como a proporção de variação total da variável dependente que é explicada pela variação da variável independente X. R^2 igual a 0,7385 significa que 73,85 % das variações de Y são explicadas pela variação de X.

ANÁLISE DE REGRESSÃO

Poder de Explicação de r^2



$$r^2 = \frac{\text{variação explicada}}{\text{variação total}} = \frac{\sum\left(y_c - \bar{y}\right)^2}{\sum\left(y_i - \bar{y}\right)^2}$$

Para Yc = valor previsto

E Yi = valor observado

➢**A percentagem de variação explicada, r2, é a razão da variação explicada sobre a variação total.**

Regressão Linear Simples

Exercício:

Município	DISTCAP	ESPVIDA	MORTINF	ALFAB	RENDA
Araruna	365	67,99	23,19	86,23	188,29
Nova Redenção	278	61,19	56,56	63	74,79
Monção	150	59,58	63,32	63,64	66,96
Porto Rico do Maranhão	78	58,96	66,05	79,33	65,34
Campo Erê	468	68,1	31,71	83,38	173,38
Lagoa do Piauí	40	63,65	47,08	65,81	60
São José das Palmeiras	486	71,01	16,62	77,54	150,67
Paraíba do Sul	83	71,36	15,69	89,28	264,55
Malhada dos Bois	65	64,46	44,18	69,95	80,69
Jandaíra	175	62,45	51,57	59,72	58,68
Vespasiano	14	68,68	32,81	90,43	196,51
Ipaba	167	67,42	37,04	81,82	125,75

Regressão Linear Simples

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	MORTINF[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: ESPVIDA

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,983[a]	,967	,964	,81590

a. Predictors: (Constant), MORTINF

b. Dependent Variable: ESPVIDA

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	75,326	,626		120,349	,000
	MORTINF	-,245	,014	-,983	-17,110	,000

a. Dependent Variable: ESPVIDA

ANOVA - REGRESSÃO

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	685,986	1	685,986	31,348	,000[a]
	Residual	350,132	16	21,883		
	Total	1036,118	17			

a. Predictors: (Constant), Consumo médio de Vegetais(gr/pessoa/dia)

b. Dependent Variable: Média das taxas de mortalidade padronizadas por cancro do estômago para o sexo masculino de 1994, 95 e 96 por 100000 habitantes

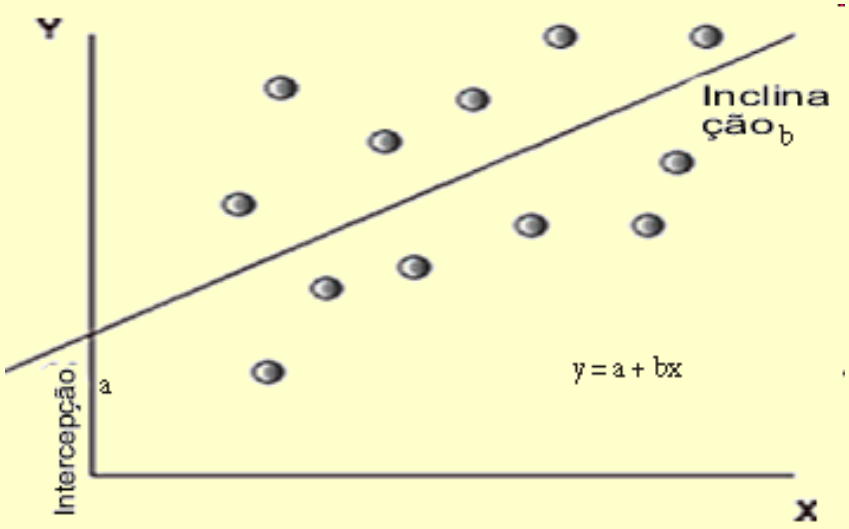
Quadro 2. Análise da variância (ANOVA) do modelo de regressão linear múltipla.

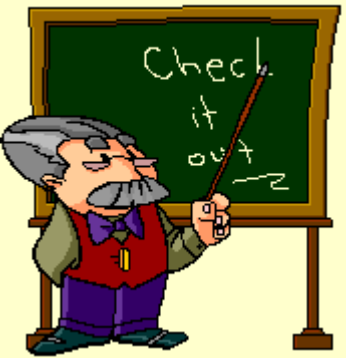
Fonte	[a] *df*	[b] *SS*	[c] *MS*	
Regressão	k	$\sum\left(\hat{Y}_i - \overline{Y}\right)^2$	SS_{Reg}/df_{Reg}	$F = MS_{Reg}/s^2$
Resíduo	$n - k - 1$	$\sum\left(Y_i - \hat{Y}_i\right)^2$	$s^2 = SS_{Res}/df_{Res}$	$R^2 = SS_{Reg}/SS_{Tot}$
Total	$n - 1$	$\sum\left(Y_i - \overline{Y}\right)^2$	SS_{Tot}/df_{Tot}	

[a]*df* = Graus de liberdade (*degrees of freedom*); [b]*SS* = Soma dos quadrados (*sum of squares*); [c]*MS* = Média da soma dos quadrados (*mean square*);

ANÁLISE DE REGRESSÃO MÚLTIPLA

O que é análise de Regressão Múltipla?





ANÁLISE DE REGRESSÃO MÚLTIPLA

TÉCNICA ESTATÍSTICA GERAL USADA PARA ANALISAR A RELAÇÃO ENTRE UMA ÚNICA VARIÁVEL DEPENDENTE E DIVERSAS VARIÁVEIS INDEPENDENTES.

$$Y_1 = X_1 + X_2 + \ldots + X_n$$

Métrica — **Métricas**

ANÁLISE DE REGRESSÃO MÚLTIPLA

Relação esta supostamente linear

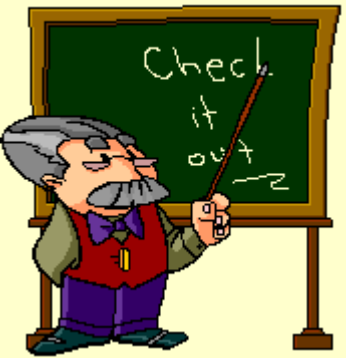
A RLM é uma extensão lógica dos princípios da Regressão Linear Simples (RLS)

Desta vez, há um coeficiente para cada uma das variáveis independentes

Assim, a variável dependente é prevista a partir da combinação de todas as variáveis independentes multiplicadas por seus respectivos coeficientes adicionada a um termo que representa o resíduo

ANÁLISE DE REGRESSÃO MÚLTIPLA

Qual a finalidade?



ANÁLISE DE REGRESSÃO MÚLTIPLA

Uma combinação linear das variáveis independentes que melhor prevê a variável dependente

Combinação linear das variáveis independentes - máxima correlação com a variável dependente.

$$Yi = b_o + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n + \varepsilon_i$$

y = variável dependente

b_o= constante

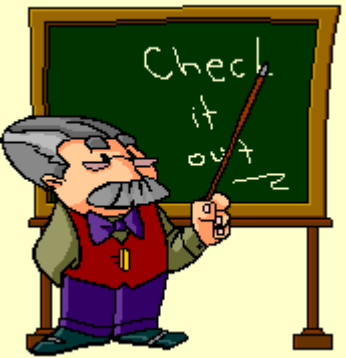
b_1 = coeficiente da primeira variável independente x_1

b_n = coeficiente da enésima variável independente x_n

ε_i = a diferença entre o valor previsto de y e o valor observado considerando o indivíduo/objeto i.

ANÁLISE DE REGRESSÃO MÚLTIPLA

Um exemplo



ANÁLISE DE REGRESSÃO MÚLTIPLA

Um exemplo de aplicação de RLM seria a previsão do número de cartões de crédito utilizados no domicílio em função do tamanho da família e da sua renda.

O modelo resultante, calcula os valores dos coeficientes para as variáveis independentes, assim como a constante.

$$Ncartõesdecrédito = b_0 + b_1 tamanhodafamília + b_2 rendafamiliar + \varepsilon$$

Ncartõesdecrédito = número de cartões de crédito no domicílio

b_0= constante

b_1 = coeficiente da variável tamanho da família

b_2 = coeficiente da variável renda familiar

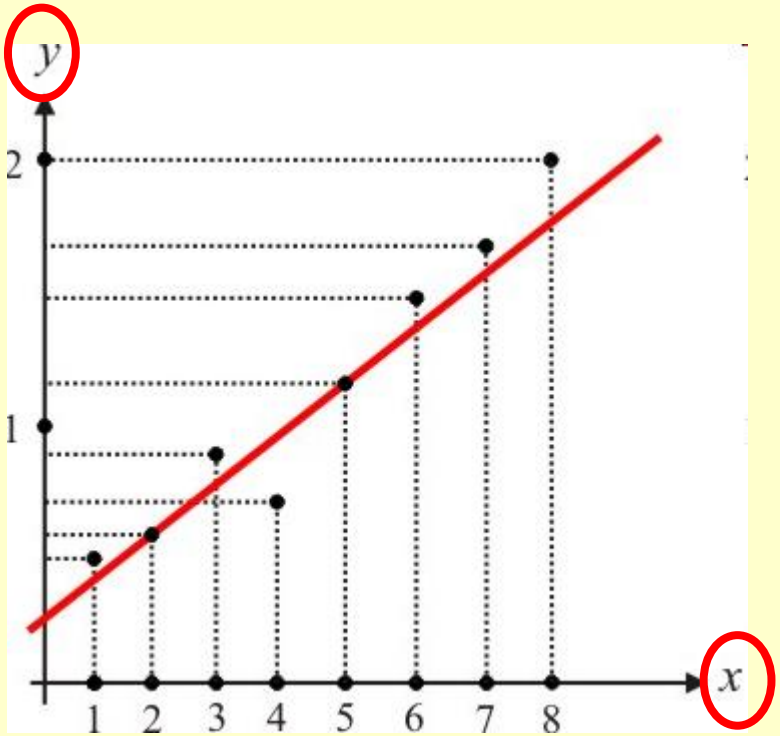
tamanhodafamília = número de pessoas na família

rendafamiliar = renda familiar

ε = termo residual

ANÁLISE DE REGRESSÃO MÚLTIPLA

Representação gráfica – Regressão Linear Simples – 2 dimensões

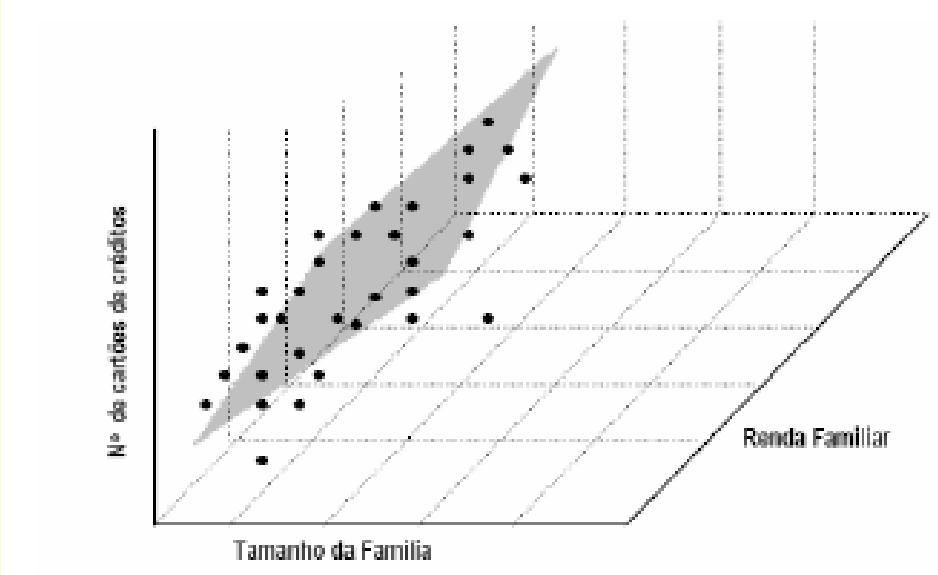


ANÁLISE DE REGRESSÃO MÚLTIPLA

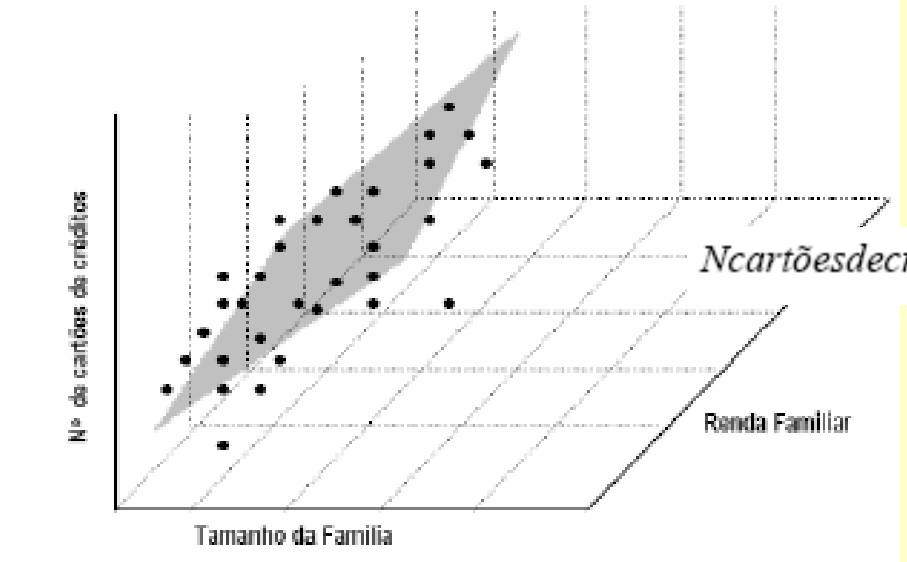
Um exemplo de aplicação de RLM seria a previsão do número de cartões de crédito utilizados no domicílio em função do tamanho da família e da sua renda.

Representação gráfica – 1 variável dependente, 2 variáveis independentes – 3 dimensões

$$Ncartõesdecrédito = b_0 + b_1 tamanhodafamília + b_2 rendafamiliar + \varepsilon$$



ANÁLISE DE REGRESSÃO MÚLTIPLA



A Equação descreve o plano cinza no gráfico e os pontos representam os valores observados

$$Ncartõesdecrédito = b_0 + b_1 tamanhodafamília + b_2 rendafamiliar + \varepsilon$$

O plano é ajustado com a finalidade de prever da melhor forma os dados observados.

No entanto, quando se trata de múltiplas variáveis, embora não se possa visualizar graficamente o modelo, deve-se aplicar os mesmos princípios da RLS aos cenários mais complexos.

Regressão Linear Múltipla

Exercício:

a) **Calcule, com auxílio do SPSS, o modelo de regressão para previsão da variável esperança de vida a partir das demais variáveis do banco de dados**

Regressão Linear Múltipla

Exercício 7:

Município	DISTCAP	ESPVIDA	MORTINF	ALFAB	RENDA
Araruna	365	67,99	23,19	86,23	188,29
Nova Redenção	278	61,19	56,56	63	74,79
Monção	150	59,58	63,32	63,64	66,96
Porto Rico do Maranhão	78	58,96	66,05	79,33	65,34
Campo Erê	468	68,1	31,71	83,38	173,38
Lagoa do Piauí	40	63,65	47,08	65,81	60
São José das Palmeiras	486	71,01	16,62	77,54	150,67
Paraíba do Sul	83	71,36	15,69	89,28	264,55
Malhada dos Bois	65	64,46	44,18	69,95	80,69
Jandaíra	175	62,45	51,57	59,72	58,68
Vespasiano	14	68,68	32,81	90,43	196,51
Ipaba	167	67,42	37,04	81,82	125,75

Regressão Linear Múltipla

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,987[a]	,973	,958	,87793

a. Predictors: (Constant), DISTCAP, ALFAB, MORTINF, RENDA

b. Dependent Variable: ESPVIDA

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	73,293	3,534		20,742	,000
	MORTINF	-,246	,034	-,986	-7,186	,000
	ALFAB	,036	,049	,092	,734	,487
	RENDA	-,003	,011	-,051	-,282	,786
	DISTCAP	-,001	,002	-,055	-,772	,466

a. Dependent Variable: ESPVIDA

Recapitulando...

a) **Como mensurar correlação entre duas variáveis**
b) **Defina Regressão Linear Simples**
c) **Defina Coeficiente de Determinação**
d) **Defina Regressão Linear Múltipla**

Kahoot!

PIN do jogo

Inserir

Quantidade de passageiros	População economicamente ativa
20.000	5.200
34.000	8.400
39.000	9.000
55.000	11.000
63.000	19.000
70.000	23.000
97.000	35.000

RESUMO DOS RESULTADOS

Estatística de regressão	
R múltiplo	0,96816
R-Quadrado	0,93734
R-quadrado ajustado	0,9248
Erro padrão	7051,9
Observações	7

ANOVA

	gl	*SQ*	*MQ*	*F*	*F de significação*
Regressão	1	3719353400	3719353400	74,79	0,000341492
Resíduo	5	248646600	49729320,1		
Total	6	3968000000			

	Coeficiente	*Erro padrão*	*Stat t*	*valor-P*	*95% inferiores*	*95% superiores*	*Inferior 95,0%*	*Superior 95,0%*
Interseção	16693,1	5070,81607	3,29200043	0,022	3658,18098	29728,07637	3658,18098	29728,07637
Variável X 1	2,36119	0,27302617	8,6482346	3E-04	1,659358269	3,063030506	1,659358269	3,063030506

Quantidade de viagens motorizadas no domicílio	Quantidade de automóveis no domicílio
2	1
2	1
4	2
4	1
7	3
8	4
12	5

RESUMO DOS RESULTADOS									
Estatística de regressão									
R múltiplo	0,968718959								
R-Quadrado	0,938416422								
R-quadrado ajustado	0,926099707								
Erro padrão	0,990867389								
Observações	7								
ANOVA									
	gl	*SQ*	*MQ*	*F*	*F de significação*				
Regressão	1	74,80519	74,80519	76,19047619	0,000326844				
Resíduo	5	4,909091	0,981818						
Total	6	79,71429							
	Coeficientes	*rro padrã*	*Stat t*	*valor-P*	*95% inferiores*	*95% superiores*	*Inferior 95,0%*	*perior 95,0%*	
Interseção	0,272727273	0,713274	0,38236	0,717907791	-1,560802873	2,106257419	-1,560802873	2,106257	
Variável X 1	2,181818182	0,249959	8,728716	0,000326844	1,539278954	2,82435741	1,539278954	2,824357	

RESUMO DOS RESULTADOS									
Estatística de regressão									
R múltiplo	0,991457								
R-Quadrado	0,982988								
R-quadrado ajustado	0,816321								
Erro padrão	0,917663								
Observações	7								
ANOVA									
	gl	*SQ*	*MQ*	*F*	*F de significação*				
Regressão	1	291,9473684	291,9473684	346,6875	8,22488E-06				
Resíduo	6	5,052631579	0,842105263						
Total	7	297							
	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*	*95% inferiores*	*95% superiores*	*Inferior 95,0%*	*Superior 95,0%*	
Interseção	0	#N/D	#N/D	#N/D	#N/D	#N/D	#N/D	#N/D	
Variável X 1	2,263158	0,121547425	18,61954618	1,54854E-06	1,96574206	2,56057373	1,96574206	2,56057373	

Quantidade de viagens originadas na zona de tráfego	População residente na zona de tráfego
25	13
45	22
50	25
55	26
58	27
61	30
70	35

RESUMO DOS RESULTADOS								
Estatística de regressão								
R múltiplo	0,992992748							
R-Quadrado	0,986034598							
R-quadrado ajustado	0,983241518							
Erro padrão	1,855013491							
Observações	7							
ANOVA								
	gl	*SQ*	*MQ*	*F*	*F de significação*			
Regressão	1	1214,795	1214,795	353,027646	7,86486E-06			
Resíduo	5	17,20538	3,441075					
Total	6	1232						
	Coeficientes	*rro padrã*	*Stat t*	*valor-P*	*95% inferiores*	*95% superiores*	*Inferior 95,0%*	*Superior 95,0%*
Interseção	-0,804259635	2,896515	-0,27766	0,792384839	-8,24998959	6,64147032	-8,24998959	6,64147032
Variável X 1	2,076572008	0,11052	18,78903	7,86486E-06	1,792470152	2,360673864	1,792470152	2,360673864

RESUMO DOS RESULTADOS								
Estatística de regressão								
R múltiplo	0,999566606							
R-Quadrado	0,999133399							
R-quadrado ajustado	0,832466732							
Erro padrão	1,706393559							
Observações	7							
ANOVA								
	gl	*SQ*	*MQ*	*F*	*F de significação*			
Regressão	1	20142,52933	20142,53	6917,602	4,76149E-09			
Resíduo	6	17,47067388	2,911779					
Total	7	20160						
	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*	*95% inferiores*	*95% superiores*	*Inferior 95,0%*	*Superior 95,0%*
Interseção	0	#N/D	#N/D	#N/D	#N/D	#N/D	#N/D	#N/D
Variável X 1	2,046797005	0,02460917	83,17213	2,03E-10	1,986580534	2,107013476	1,986580534	2,107013476