

Multivariada de dados aplicada à administração

PROFA. DRA. LUCIANA ORANGES CEZARINO

-AULA PPGAO/FEARP 201711

AGENDA

26/7	14h: Apresentações 14:15h – 15:45h: Técnicas da pesquisa quantitativa : AF e Regressão 16h – 18h: <i>Start</i> dos artigos
02/08	14h: Apresentações 14:15h – 15:45h: Análise de métodos quantitativos 16h – 18h: Utilização dos softwares
	14h às 18h: Apresentações de seminários

Quem sou?

Docente Fagen – UFU, desde 2013

Pós-doutora pelo Politécnico de Milão – 2016/2017

Mestre e Doutora pela FEA – USP (2005 e 2013)

Economista pela UEL;

Coordenadora Capítulo Brasileiro *Golden for Sustainability*;

Coordenadora do PRME Fagen UFU;

Coordenadora Scor;

Ex-consultora COHROS, Ex- gerente da Vivo, ex-consultora Sebrae, ex-professora substituta da FEARP-USP, ex-professora UNAERP (2005-2010);

Autora de artigos e consultora.

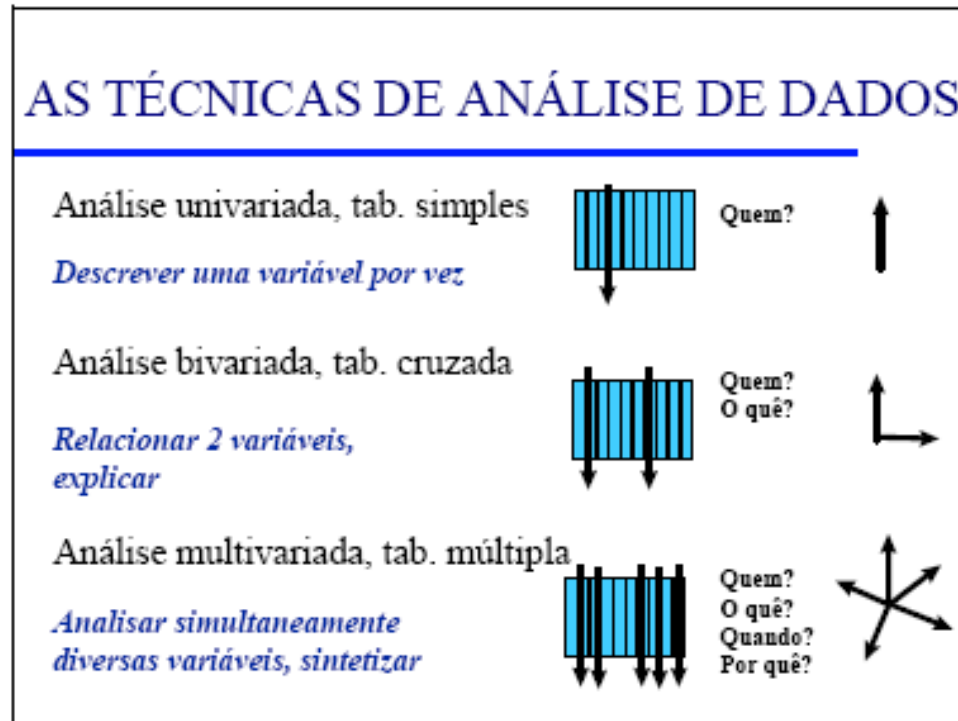


AMD

“O truque da estatística multivariada, se existe, não está nos cálculos, fácil e rapidamente feitos num computador com software adequado instalado. O truque consiste em escolher o método apropriado ao tipo de dado, usá-lo corretamente, saber interpretar os resultados e retirar deles as conclusões corretas”(REIS, 2001)

Introdução

1. Análise multivariada: conjunto de métodos estatísticos que possibilita a análise simultânea de medidas múltiplas para cada dado ou sequencia de dados;
2. Aplicado a duas ou mais variáveis;
3. Formam um único relacionamento



Freitas & Moscarola, 2017

Figura 6 - As técnicas de análise de dados

Tipos de variáveis

QUANTITATIVAS

- DISCRETAS: VALORES INTEIROS;
 - EX: número de empresas, funcionários, clientes, fornecedores
- CONTÍNUAS: FRACIONADAS
 - Tempo de produção, idade, índice de rentabilidade, preços.

QUALITATIVAS

- NOMINAIS OU CATEGÓRICAS: não apresentam ordenação: sexo, estado civil, sim ou não.
- ORDINAIS: possuem ordenação,
 - Salários, escolaridade, tempo (em meses), cargos, classe social.

Relações entre variáveis

VARIÁVEL INDEPENDENTE

É aquela que é fator determinante para que ocorra um determinado resultado; é a condição ou causa para um determinado efeito ou consequência; é o estímulo que condiciona uma resposta.

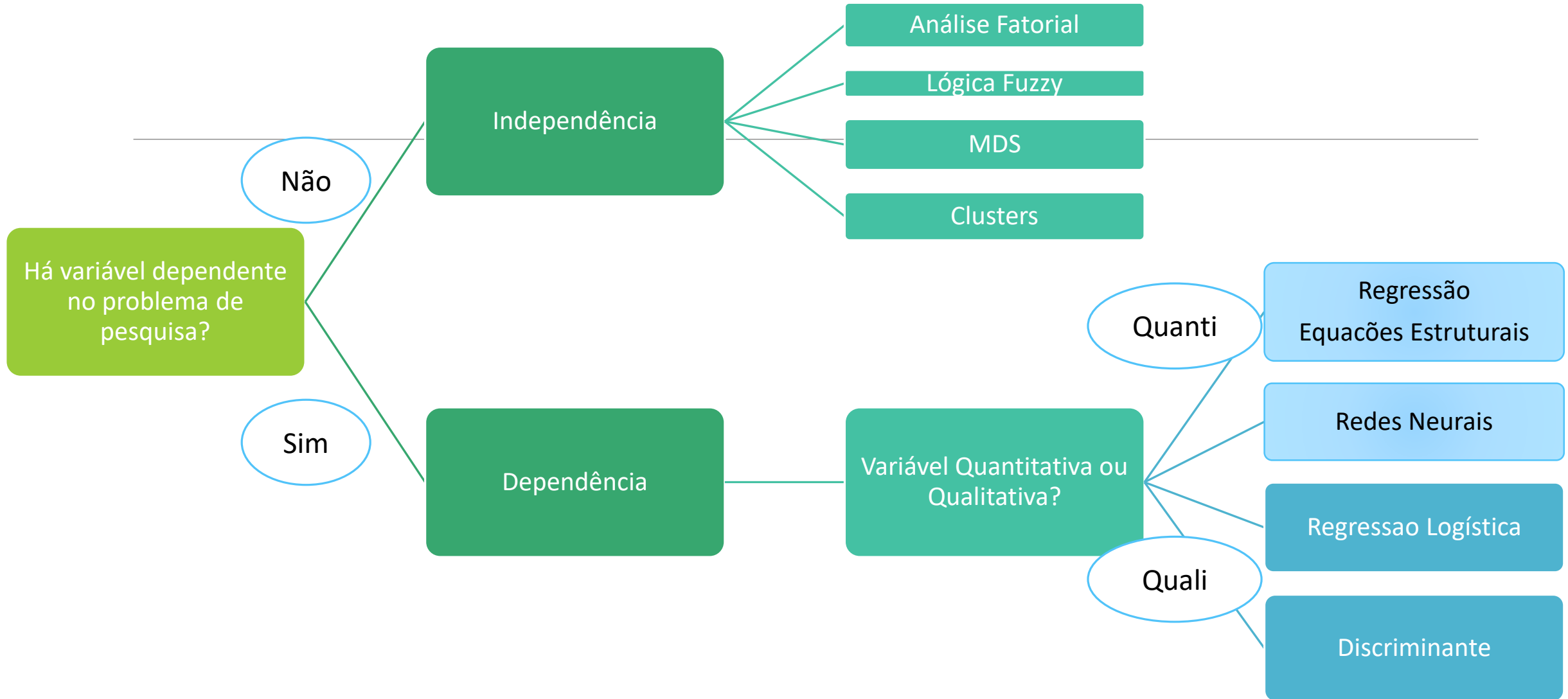
VARIÁVEL DEPENDENTE

É aquele fator ou propriedade que é efeito, resultado, consequência ou resposta de algo que foi estimulado; não é manipulada, mas é o efeito observado como resultado da manipulação da variável independente.

Hipóteses

- Suposição que pode ser colocada à prova para determinar sua validade; pode parecer contrária, ou de acordo com o senso comum, pode ainda ser correta ou errada.
- Conduz a uma verificação empírica, independentemente do resultado; ela é uma questão proposta de tal maneira que uma resposta de algum tipo pode estar próxima a aparecer.
- Karl Popper e a falseabilidade das hipóteses;
 - H_0 (hipótese nula) e $H_1, 2, 3, 4...$ (alternativas)
 - Hipótese alternativa é que o pesquisador quer comprovar
 - Erro do tipo I: decisão de rejeitar H_0 quando de fato H_0 é verdadeira.
 - Erro do tipo II: decisão de não rejeitar H_0 quando de fato H_0 é falsa.

Técnicas mais utilizadas na administração



Estatística Descritiva

N	Mínimo	Máximo	Média	Desvio-padrão
100				

N – número de observações;
Mínimo – Menor valor da base;
Máximo – maior valor da base;
Média – aritmética;
Desvio-padrão – dispersão da média.

Estatística Descritiva por gráfico

- Duas opções:
 - Histograma

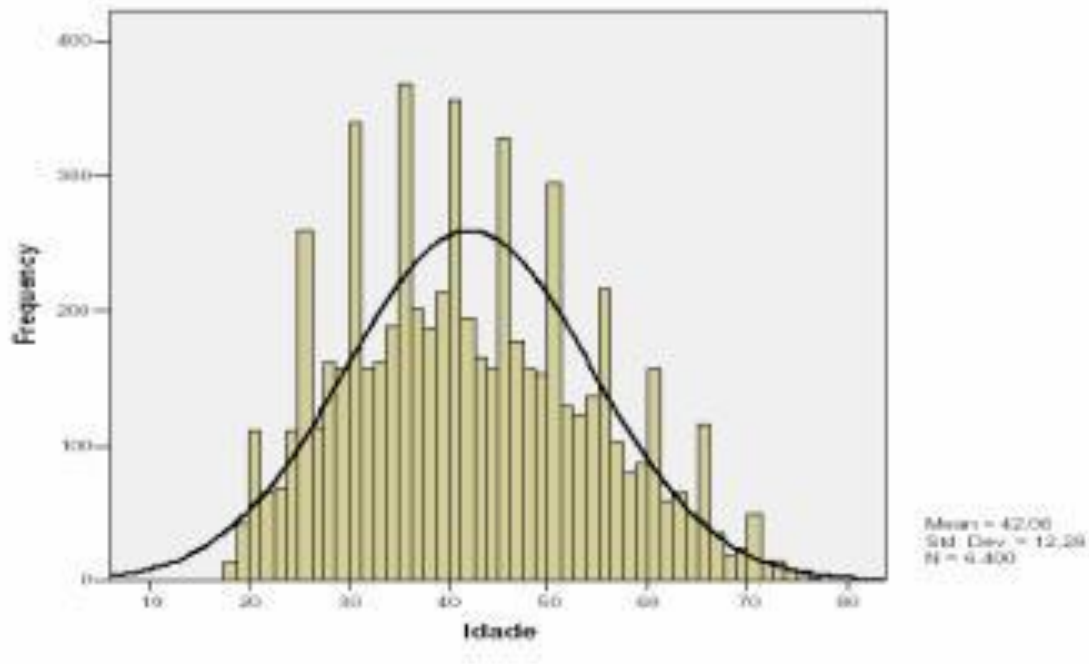
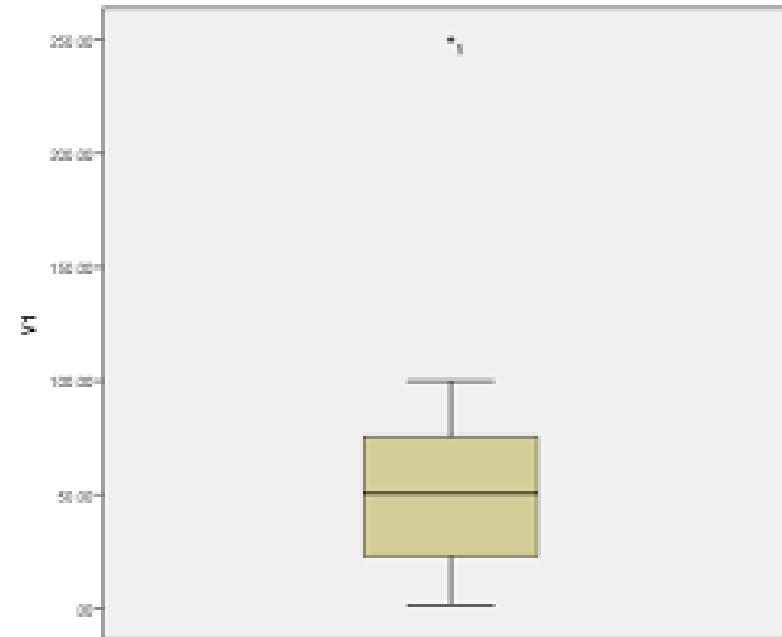


Gráfico 6: Histograma com as idades dos inquiridos

- Boxplot



Outliers

São observações atípicas, inconsistentes com o resto da base de dados.

- Causas
 - Erro de medição;
 - Variabilidade inerente da amostra
- Devem ser retirados?

Dados perdidos

- *Missing Values*
- São dados ausentes causados pela falta de respostas a um questionário ou pela ausência de informações de uma base de dados.
- Podem ser ignorados ou uma série deve ser eliminada.
- Ações corretivas:
 - Incluir atribuições medias (*replace missing values*)

Premissas da AMD

1) Normalidade

2) Homocedasticidade

3) Linearidade

Normalidade

- Significa que a base segue uma distribuição normal de dados.
- É a forma (física) dos dados da amostra.
- Normalidade pode ser individual ou multivariada
- Para a utilização de testes paramétricos é preciso que a amostra de dados seja normal.
- Para Hair Jr (1995) se a variação dos dados da normalidade é grande então os testes serão inválidos, sendo necessários os teste t e f (regressão, ANOVA, Variância etc)
- Teste não paramétrico mais utilizado é o Mann-Whitney
- Segundo Paes (2003) tem-se Wilcoxon como alternativa ao teste t pareado e Kruskal-Wallis como alternativa à ANOVA com um fator. Apesar de parecer uma solução relativamente simples, a grande desvantagem dos testes não-paramétricos é que eles tendem a ser bem menos poderosos que os paramétricos

Testes de normalidade

- Histograma com faixa de probabilidade normal;
- Teste KS (Kolmogorov-Smirnov) qualquer resultado abaixo de 0,005% de significância revela a não normalidade de dados
 - Outros testes: Lilliefors (LF) – que é apenas uma correção para o teste KS), Shapiro-Wilk (SW) e Shapiro-Francia (SF)
- Fazer testes diferentes ou eliminar variáveis não normais;
- Corrigir a normalidade univariada por meio da variável identidade, logarítima ou de raiz quadrada.

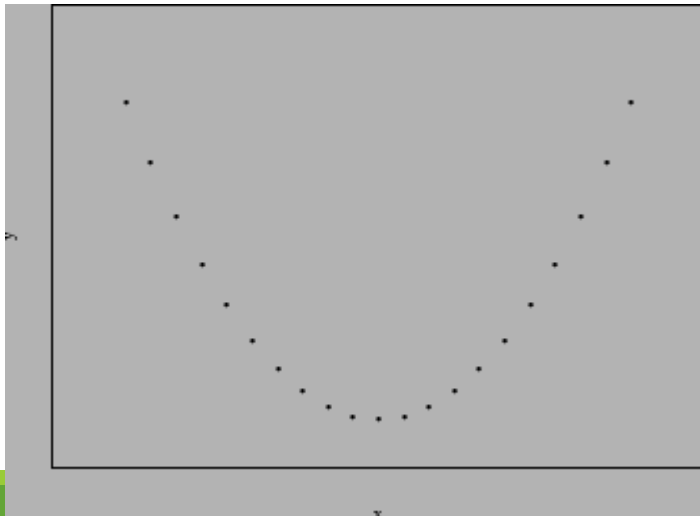
Homoscedasticidade

- correto: homocedasticidade

- É a suposição que há relacionamento entre as variáveis dependentes iguais, ou seja, a variância entre as variáveis dependentes é a mesma sob o domínio das independentes.
- Variância dos resíduos erros (ϵ) é constante.
- Teste mais conhecido para a detecção da heteroscedasticidade é o teste de Park

Linearidade

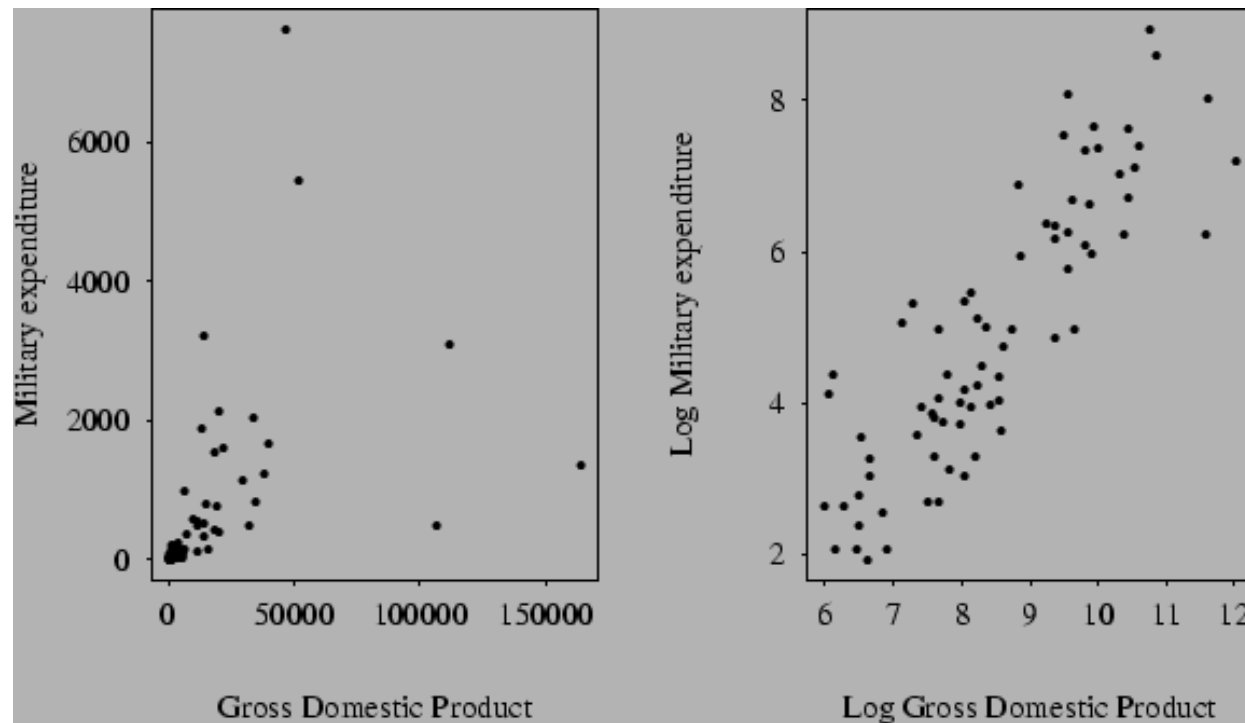
- A linearidade mede a variação da tendência para diferentes valores de referência na faixa de interesse;
- A linearidade é avaliada pela inclinação da reta formada pelos diferentes valores de referência em relação a respectiva tendência. Quanto menos inclinada a reta, melhor será a qualidade do sistema de medição.
- Somente relações lineares são detectadas pelo coeficiente de correlação



Dados não lineares

Linearidade

Pode ser corrigido, mudando y por x ou vice-versa



Análise Fatorial

Objetivo

- Simplificar os dados, agrupar as variáveis em fatores altamente correlacionados;
- Criar índices com variáveis que medem dimensões conceituais similares ou complementares;
- Tipos:
 - Exploratória: quando não temos uma idéia pré-definida da estrutura e de quantas dimensões estão presentes em um conjunto de variáveis. `factor var1 var2 var3 ... varn`
 - Confirmatória: quando queremos testar hipóteses específicas sobre a estrutura de um número de dimensões subjacentes a um conjunto de variáveis. Por exemplo, pensamos que nossos dados possuem duas dimensões e queremos verificar isto. `factor var1 var2 var3 ... varn, factors(#)`

Eigenvalues e factors

- Cargas fatoriais (factors) são as correlações entre as variáveis originais e os fatores.
- Esse é um dos pontos principais da análise fatorial, quanto maior a carga fatorial maior será a correlação com determinado fator. –
- Um valor negativo indica um impacto inverso no fator.
- A quantidade de cargas fatoriais é automaticamente calculada pelo Stata com base nos autovalores (iguais ou superiores a uma unidade - critério de Kaiser).
- Por sua vez, as cargas fatoriais relevantes são aquelas com valores maiores que 0,5.

Rotacão

- Rotação fatorial (rotação fatorial (rotation rotation) é o processo de é o processo de manipulação ou de ajuste dos eixos fatoriais para conseguir uma solução fatorial mais simples e pragmaticamente mais significativa, cujos fatores sejam mais facilmente interpretáveis
- Varimax e ortogonal são as mais utilizadas
- Teste:
 - O teste de Kaiser-Meyer-Olkin (KMO) varia entre 0 e 1.
Quanto mais perto de 1, melhor.
 - Friel (2009) sugere a seguinte escala para interpretar o valor da estatística KMO:
 - * Entre 0,90 e 1: excelente. * Entre 0,80 e 0,89: bom. * Entre 0,70 e 0,79: mediano. * Entre 0,60 e 0,69: medíocre. * Entre 0,50 e 0,59: ruim. * Entre 0 e 0,49: inadequado.
 - Pallant (2007) sugere 0,60 como um limite razoável.
 - Hair et al. (2006) sugerem 0,50 como patamar aceitável.

Resultados da AF

	EASY	FLEXIBLE	POWERFUL	RELIABLE	SUITABLE	TECHSUPP
EASY	1.000	.341	.277	.224	.252	.335
FLEXIBLE	.341	1.000	.595	.165	.539	.018
POWERFUL	.277	.595	1.000	-.023	.644	.071
RELIABLE	.224	.165	-.023	1.000	-.086	.318
SUITABLE	.252	.539	.644	-.086	1.000	-.007
TECHSUPP	.335	.018	.071	.318	-.007	1.000

Matriz de Correlação são as possíveis correlações de Pearson entre as variáveis

Os valores da diagonal principal é igual a um, devido à perfeita correlação entre as mesmas variáveis

	EASY	FLEXIBLE	POWERFUL	RELIABLE	SUITABLE	TECHSUPP
EASY	1.000	.341	.277	.224	.252	.335
FLEXIBLE	.341	1.000	.595	.165	.539	.018
POWERFUL	.277	.595	1.000	-.023	.644	.071
RELIABLE	.224	.165	-.023	1.000	-.086	.318
SUITABLE	.252	.539	.644	-.086	1.000	-.007
TECHSUPP	.335	.018	.071	.318	-.007	1.000

Resultados da AF

	EASY	FLEXIBLE	POWERFUL	RELIABLE	SUITABLE	TECHSUPP
EASY	1.000	.341	.277	.224	.252	.335
FLEXIBLE	.341	1.000	.595	.165	.539	.018
POWERFUL	.277	.595	1.000	-.023	.644	.071
RELIABLE	.224	.165	-.023	1.000	-.086	.318
SUITABLE	.252	.539	.644	-.086	1.000	-.007
TECHSUPP	.335	.018	.071	.318	-.007	1.000

Na apresentação as correções são espelhadas

	EASY	FLEXIBLE	POWERFUL	RELIABLE	SUITABLE	TECHSUPP
EASY	1.000	.341	.277	.224	.252	.335
FLEXIBLE	.341	1.000	.595	.165	.539	.018
POWERFUL	.277	.595	1.000	-.023	.644	.071
RELIABLE	.224	.165	-.023	1.000	-.086	.318
SUITABLE	.252	.539	.644	-.086	1.000	-.007
TECHSUPP	.335	.018	.071	.318	-.007	1.000

As correlações nestes casos são relativamente fracas, próximas de Zero. Como nestes casos, você deve reconsiderar o uso do método de análise fatorial com os seus dados.

		EASY	FLEXIBLE	POWERFUL	RELIABLE	SUITABLE	TECHSUPP
Correlation	EASY	1.000	.341	.277	.224	.252	.335
	FLEXIBLE	.341	1.000	.595	.165	.539	.018
	POWERFUL	.277	.595	1.000	-.023	.644	.071
	RELIABLE	.224	.165	-.023	1.000	-.086	.318
	SUITABLE	.252	.539	.644	-.086	1.000	-.007
	TECHSUPP	.335	.018	.071	.318	-.007	1.000
Sig. (1-tailed)	EASY		.000	.003	.013	.006	.000
	FLEXIBLE	.000		.000	.051	.000	.429
	POWERFUL	.003	.000		.411	.000	.242
	RELIABLE	.013	.051	.411		.197	.001
	SUITABLE	.006	.000	.000	.197		.471
	TECHSUPP	.000	.429	.242	.001	.471	

Estes valores representam a significância do teste de correlação de Pearson (quanto ficou de fora). Estes p-values da testagem indicam quais são as correspondências são diferente de zero.

Muitos deste valores devem ser pequenos para o emprego do método de análise fatorial.

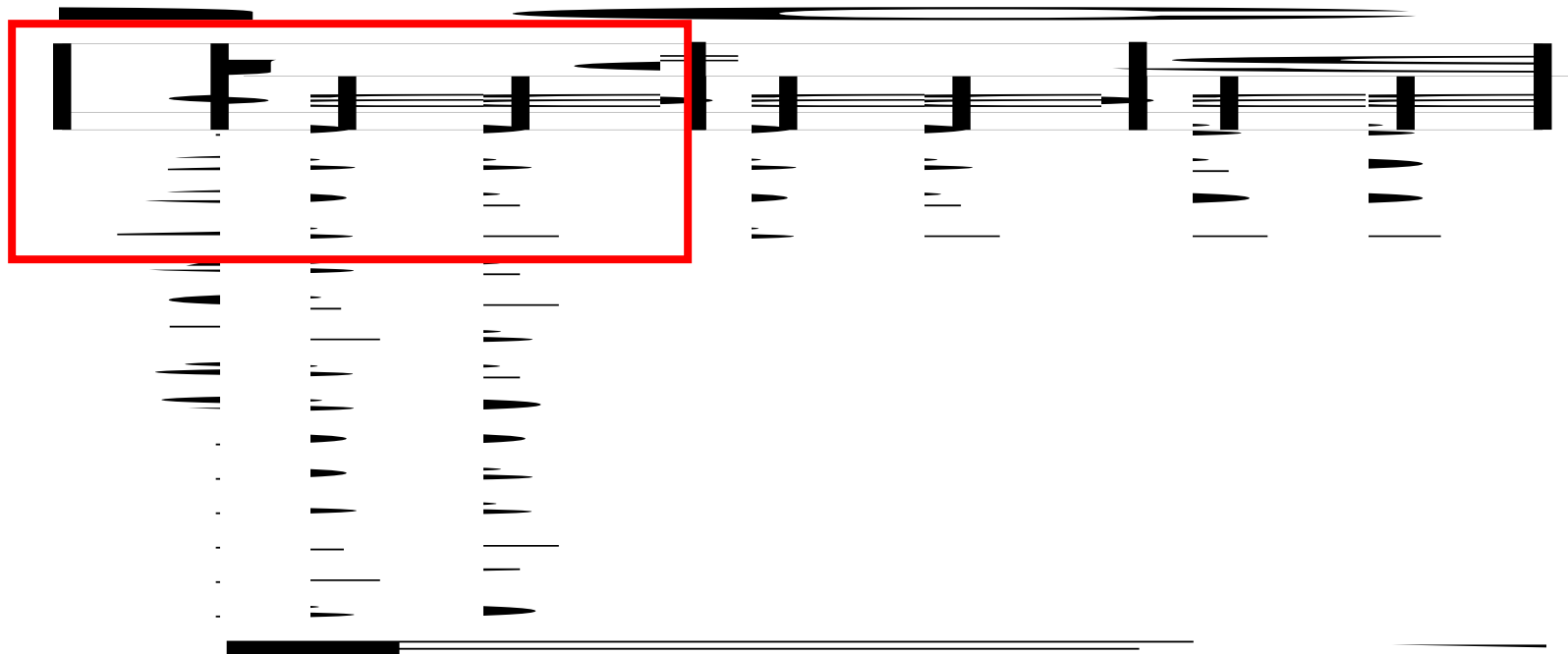
		EASY	FLEXIBLE	POWERFUL	RELIABLE	SUITABLE	TECHSUPP
Correlation	EASY	1.000	.341	.277	.224	.252	.335
	FLEXIBLE	.341	1.000	.595	.165	.539	.018
	POWERFUL	.277	.595	1.000	-.023	.644	.071
	RELIABLE	.224	.165	-.023	1.000	-.086	.318
	SUITABLE	.252	.539	.644	-.086	1.000	-.007
	TECHSUPP	.335	.018	.071	.318	-.007	1.000
Sig. (1-tailed)	EASY		.000	.003	.013	.006	.000
	FLEXIBLE	.000		.000	.051	.000	.429
	POWERFUL	.003	.000		.411	.000	.242
	RELIABLE	.013	.051	.411		.197	.001
	SUITABLE	.006	.000	.000	.197		.471
	TECHSUPP	.000	.429	.242	.001	.471	

	EASY	FLEXIBLE	POWERFUL	RELIABLE	SUITABLE	TECHSUPP
EASY	1.000	.341	.277	.224	.252	.335
FLEXIBLE	.341	1.000	.595	.165	.539	.018
POWERFUL	.277	.595	1.000	-.023	.644	.071
RELIABLE	.224	.165	-.023	1.000	-.086	.318
SUITABLE	.252	.539	.644	-.086	1.000	-.007
TECHSUPP	.335	.018	.071	.318	-.007	1.000

a. Determinant= .219

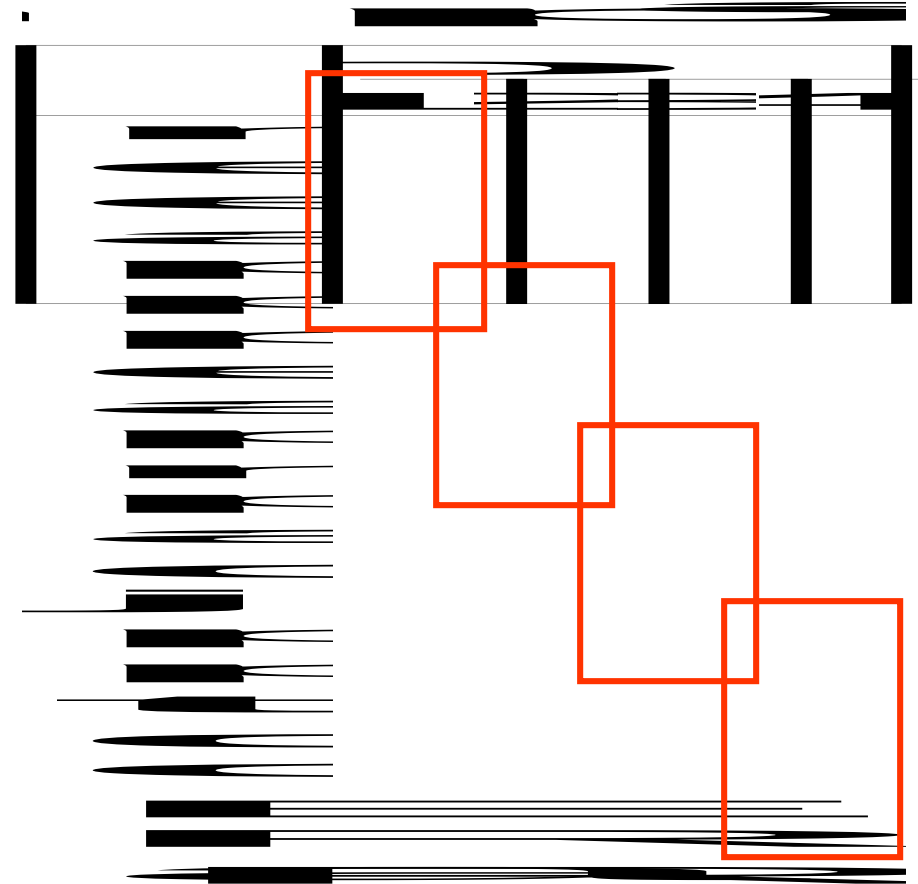
Deve-se analisar o valor do determinante da matriz de correlação. Este indica a possibilidade de inversão da matriz. Se o valor do determinante é zero, a matriz de correlação não pode ser invertida e certamente os métodos de extração de análise fatorial serão impossíveis de serem computados.

Análise de Variância



Matriz de rotação

VARIMAX



Conclusões

O resultado mostra que as variáveis anteriores podem ser estatisticamente agrupadas em 4 fatores ou dimensões

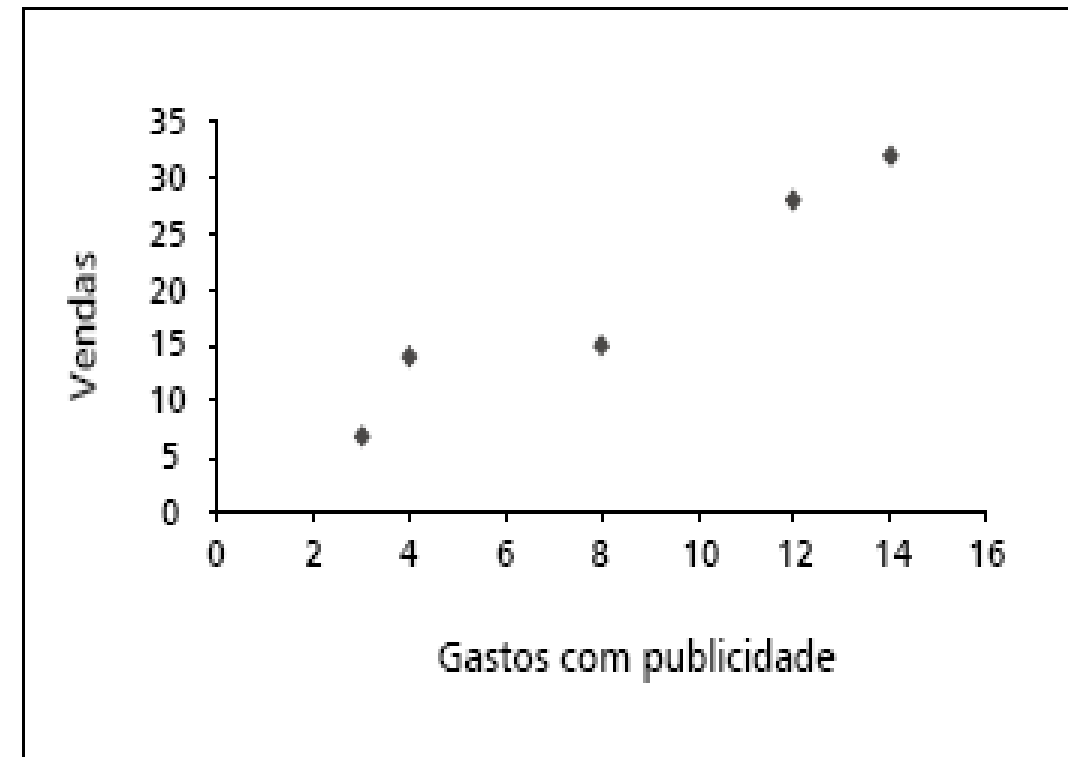
Análise de Regressão

Variáveis

Y : independente

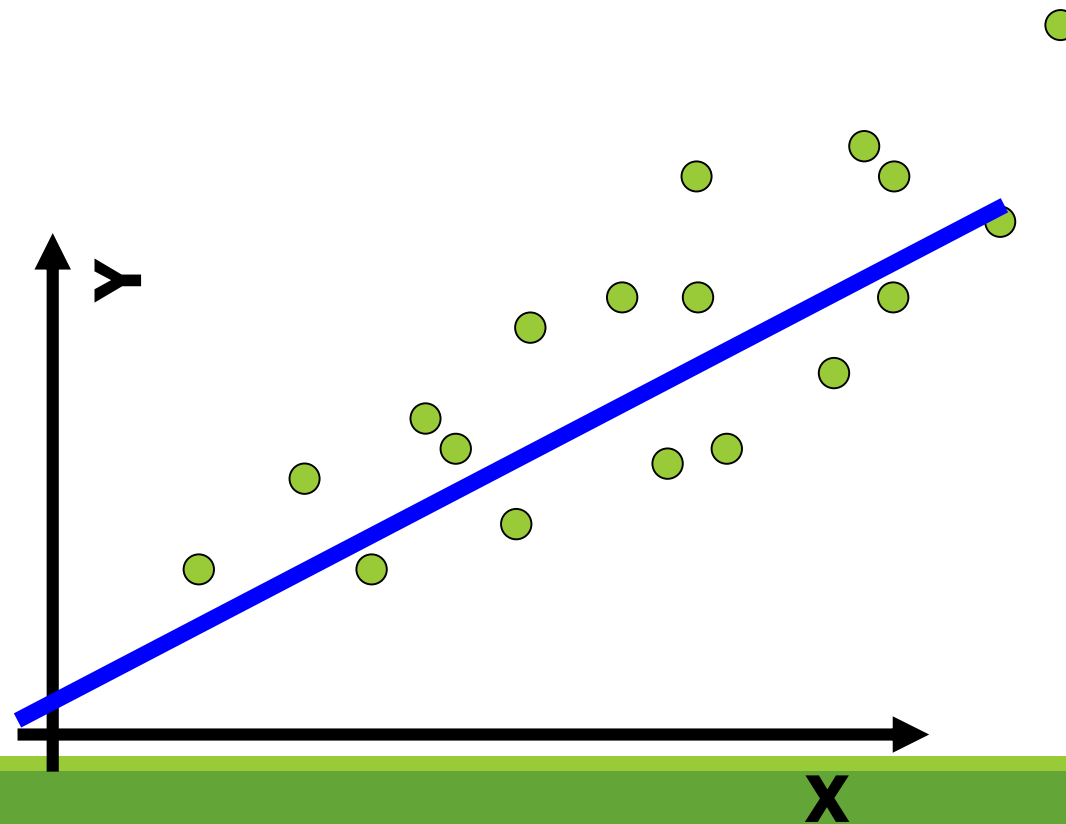
X1, X2 etc : dependentes

Gastos com publicidade (em \$ mil)	3	4	8	12	14
Vendas (em \$ mil)	7	14	15	28	32

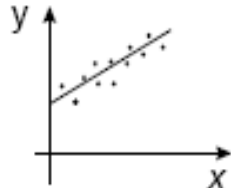
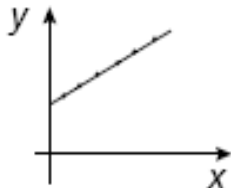
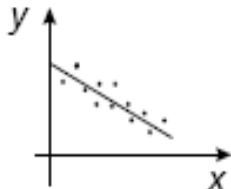
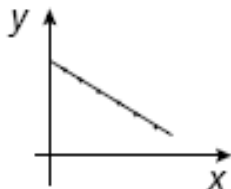
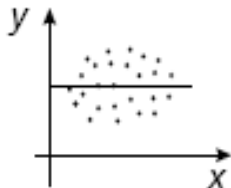


Objetivo

Verificar se há associação entre os dados, se as variáveis dependentes explicam a variação da independente.



Equação
 $Y = a + b.X$

Correlação	Digrama de dispersão	Descrição
Linear Positiva		A correlação é positiva se os valores crescentes ou decrescentes x e y estiverem ligados. Ou seja, quando y cresce, x cresce também. Quando y decresce, x decresce também e vice-versa. Nos modelos de correlação linear positiva, o valor do coeficiente de correlação de Pearson, r , é positivo: $0 < r < 1$.
Linear Perfeita Positiva		A correlação linear perfeita positiva apenas ocorre quando os valores de x e y estão perfeitamente alinhados. Nestas situações, o valor do coeficiente de correlação de Pearson, r , é igual à unidade: $r = 1$.
Linear Negativa		A correlação negativa é percebida quando valores crescentes de x ou y estão associados a valores decrescentes de y ou x , respectivamente. Ou seja, quando y cresce, x decresce e vice-versa. O valor do coeficiente de correlação de Pearson, r , é negativo: $-1 < r < 0$.
Linear Perfeita Negativa		A correlação é considerada perfeita negativa quando os valores de x e y estiverem perfeitamente alinhados, mas em sentido contrário. Nesta situação, o valor do coeficiente de correlação de Pearson, r , é igual a menos um: $r = -1$.
Nula		A correlação nula é percebida quando não há relação entre x e y . As variáveis ocorrem independentemente. Nestas situações, o valor do coeficiente de correlação de Pearson, r , é nulo: $r = 0$.

Resultados

Regression

[DataSet1] D:\dados\slides\estatistica\copia_de_bases\bases_spss\carros.sav

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	Cilindradas em polegadas cúbicas ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: Potência do motor em HPs

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,903 ^a	,816	,815	18,614

a. Predictors: (Constant), Cilindradas em polegadas cúbicas

Resultados

- ◆ O coeficiente apresentado na tabela refere-se ao coeficiente parcial da regressão e difere do da regressão simples considerando a relação de cada variável independente em separado.
- ◆ O coeficiente expressa o aumento médio em Y dado um aumento de 1 unidade de X, sem considerar o efeito de qualquer outra variável independente (mantendo todos os outros fatores constantes).
- ◆ Para um aumento de 1 unidade na altura, há um aumento médio de 0.724 no peso, para crianças de mesma idade.

- ◆ O coeficiente apresentado na tabela refere-se ao coeficiente parcial da regressão e difere do da regressão simples considerando a relação de cada variável independente em separado.
- ◆ O coeficiente expressa o aumento médio em Y dado um aumento de 1 unidade de X, sem considerar o efeito de qualquer outra variável independente (mantendo todos os outros fatores constantes).
- ◆ Para um aumento de 1 unidade na altura, há um aumento médio de 0.724 no peso, para crianças de mesma idade.

Referência

Básica:

- CORRAR, Luiz J., PAULO, Edilson, DIAS FILHO, José Maria. Análise Multivariada para os cursos de Administração, Ciências Contábeis e Economia. São Paulo: Atlas, 2007.
- FREITAS, Henrique; MOSCAROLA, Jean. Da observação à decisão: métodos de pesquisa e de análise quantitativa e qualitativa de dados. **RAE electron.**, São Paulo , v. 1, n. 1, p. 1-30, June 2002 . Available from <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1676-56482002000100006&lng=en&nrm=iso>. access on 25 July 2017. <http://dx.doi.org/10.1590/S1676-56482002000100006>