Capítulo 1 (final) + capítulo 2

Análise Multivariada

Data: 24/09/2021

DIRETRIZES PARA ANÁLISES MULTIVARIADAS E INTERPRETAÇÃO

1. Estabelecer significância prática, bem como significância estatística.
2. Reconhecer que tamanhos de amostras afetam resultados.
3. Conhecer seus dados.
4. Esforçar-se por modelos parcimoniosos.
5. Examinar seus erros.
6. Validar seus resultados.

Etapas para a construção de modelo multivariada.

1. Definir o problema de pesquisa, os objetivos e a técnica multivariada a ser usada.
2. Desenvolver o plano de análise.
3. Avaliar as suposições.
4. Estimar o modelo multivariado e avaliar o ajuste.
5. Interpretar as variáveis estatísticas.
6. Validar o modelo multivariado.

EXAME GRÁFICO DOS DADOS

Antes de aplicar qualquer técnica estatística multivariada, é necessário uma visão crítica dos dados

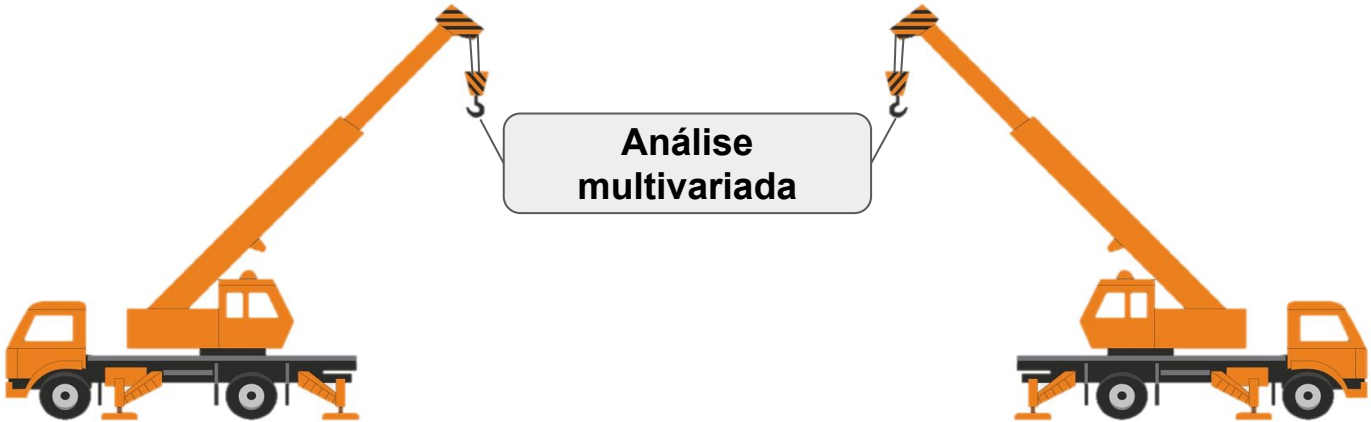
Teoria

"Constitui-se de princípios, categorias e conceitos, formando sistematicamente um conjunto logicamente coerente, dentro do qual o trabalho do pesquisador se fundamenta e se desenvolve"

(SEVERINO, 2004)

Estatística

- Verificar como as variáveis se distribuem e como se relacionam
- Outliers, dados atípicos, faltantes...

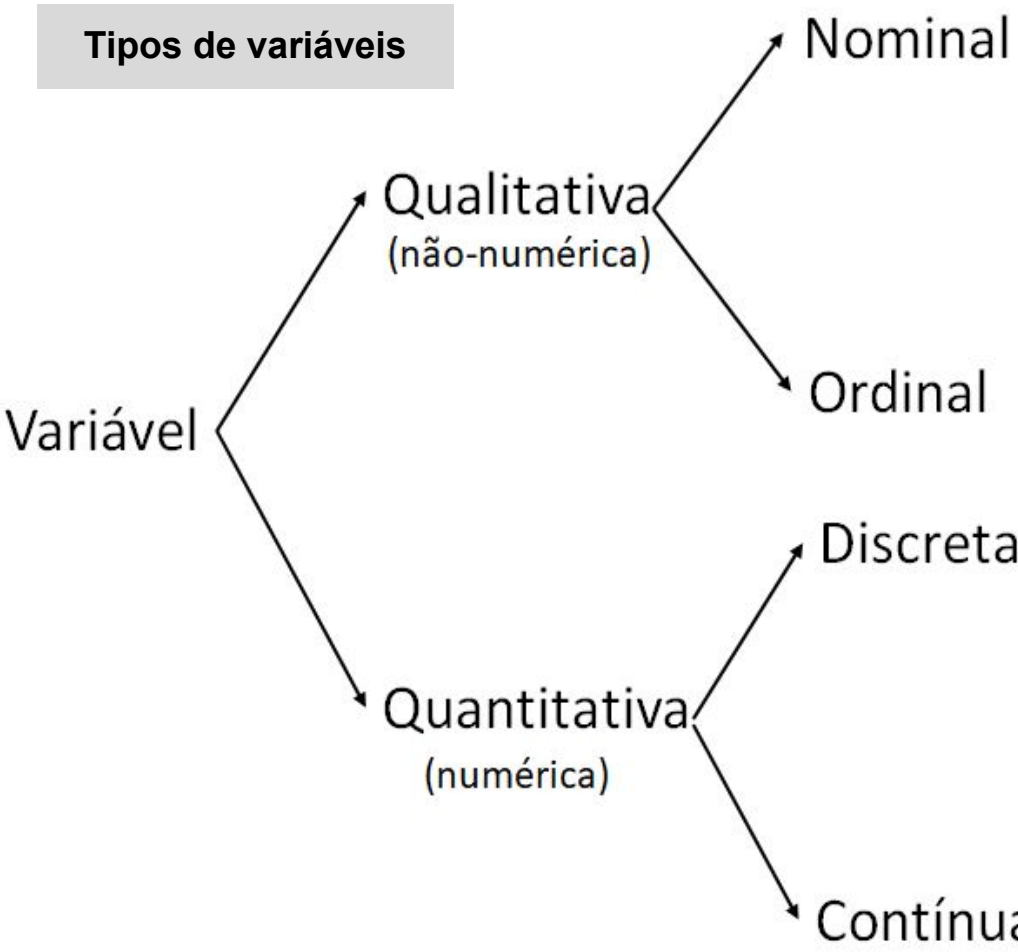


"O conhecimento das inter-relações de variáveis pode ajudar incrivelmente na especificação e no refinamento do modelo multivariado, bem como fornecer uma perspectiva racional para a interpretação dos resultados."

(HAIR, 2009)

SEVERINO, Antônio Joaquim. metodologia do trabalho científico, 22. ed. São Paulo: Cortez, 2004.

Tipos de variáveis



Consistem apenas em nomes, rótulos ou categorias. Os dados não podem ser dispostos segundo um esquema ordenado.
Ex: Masculino, Feminino.

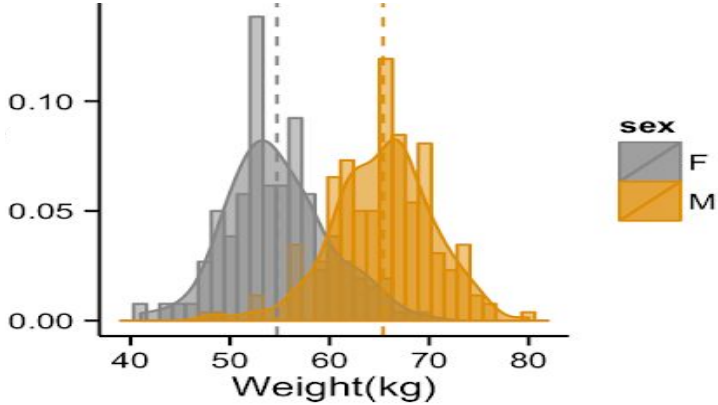
Envolve dados que podem ser dispostos em alguma ordem, mas as diferenças entre os valores dos dados não podem ser determinadas ou não tem sentido. Ex: Ens. Fundamental, médio e superior.

Assume valores pertencentes a um conjunto finito ou enumerável. Geralmente, seus valores são resultados de um processo de contagem, razão pela qual seus valores são expressos através de números inteiros não-negativos. Ex: Quantidade de membros por família.

Assume qualquer valor pertencente a um determinado intervalo do conjunto dos Reais. Pode-se dizer que a variável contínua resulta normalmente de mensurações. Ex: Nota, Altura, Peso.

Perfil univariado

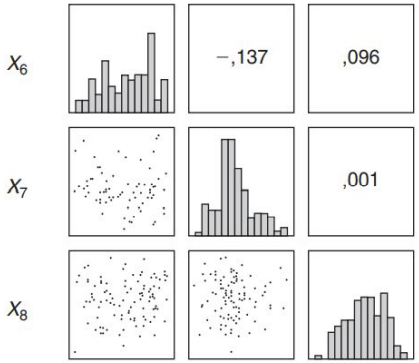
Histograma: é uma representação gráfica (barras verticais ou barras horizontais) da distribuição de frequências de um conjunto de dados quantitativos continuous.



O histograma permite termos uma ideia da média, moda, mediana, assimetria e curtose dos dados!

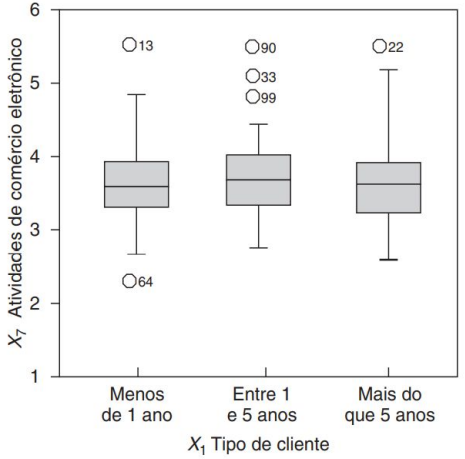
Perfil bivariado

Diagrama de dispersão: é um gráfico de pontos baseado em duas variáveis. O padrão de pontos representa a relação entre variáveis.



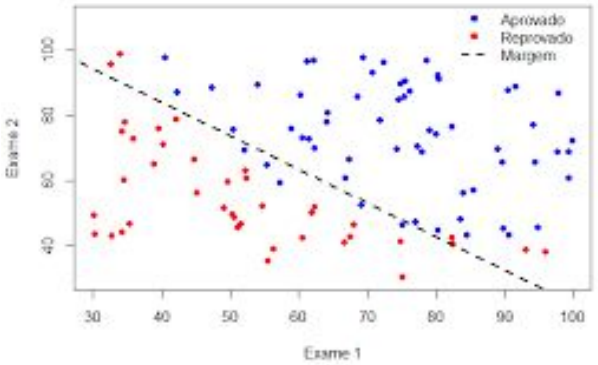
Uma forte organização de pontos ao longo de uma linha reta caracteriza uma relação linear ou correlação.

Gráfico de caixa (boxplot): representação da distribuição de dados de uma variável numérica para cada grupo de uma variável não-métrica

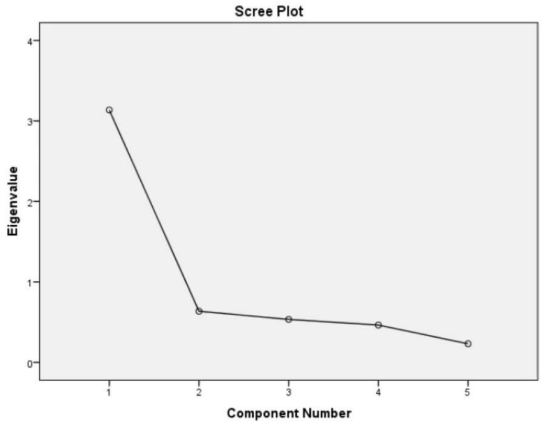


Perfis multivariados

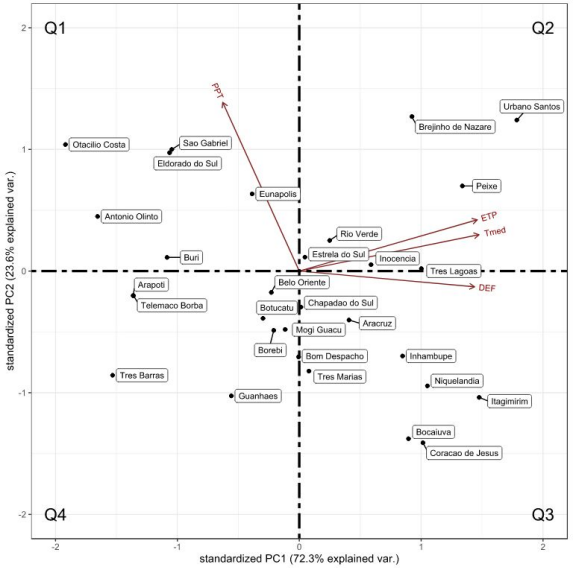
O objetivo dos perfis multivariados é retratar os dados de uma maneira que permita a identificação de diferenças e similaridades.



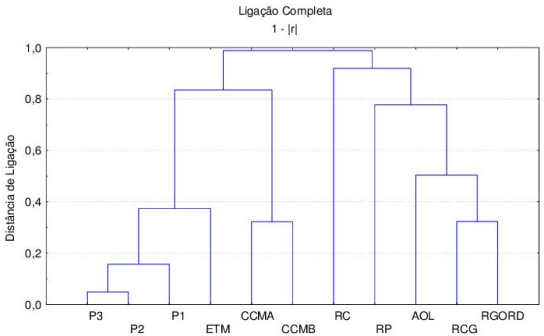
Exemplo 1 – Gráfico de análise discriminante



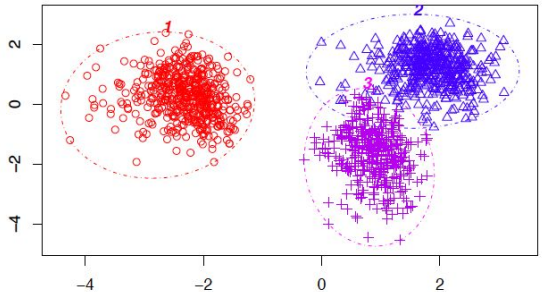
Exemplo 2 – Gráfico de Scree plot (gráfico de cotovelo)



Exemplo 3 – Gráfico de componentes principais



Exemplo 4 – Dendograma



Exemplo 5 – Gráfico de clusters (via k-means)

DADOS PERDIDOS

Dados Perdidos

Definição: valores válidos sobre uma ou mais variáveis que não estão disponíveis para análise devido a:

- eventos sistemáticos externos ao respondente (erros de entrada, problemas de coleta);
- ação por parte do respondente (recusa a responder).

O que fazer? Abordar questões geradas que podem afetar a generalidade dos resultados:

(1) Os dados perdidos estão distribuídos ao acaso ou há padrões e relações inerentes? -> Tratar pode ajudar a manter tanto quanto possível a distribuição original de valores e evitar viés.

(2) Qual é a freqüência dos dados perdidos? -> Mais determinante para o tipo de ação corretiva, portanto é uma questão secundária

Passo 1: Determinar o tipo de dados perdidos

Ignoráveis

- São esperados, fazem parte do planejamento;
- Valores observados são amostra aleatória do conjunto total de valores -> correção desnecessária

 Amostra x população -> dados perdidos são as observações em uma população que não estão incluídas numa amostra probabilística;

 Delineamento do instrumento admite “saltos” sobre seções de questões que não aplicáveis;

 Dados censurados - observações incompletas devido a seu estágio no processo de perda de dados.

Não ignoráveis

 Processos conhecidos - Ex.: entrada de dados que criam códigos inválidos, restrições de desfecho, falha para completar um questionário, morte do respondente -> pesquisador tem pouco controle sobre os processos de perda, mas algumas ações corretivas podem ser aplicadas se perda for aleatória;

 Processos desconhecidos - Diretamente relacionados com o respondente: recusa em responder questões de natureza sensível ou ausência de opinião ou conhecimento para responder -> menos facilmente identificados e acomodados, portanto, o pesquisador deve prever e minimizar no planejamento e coleta de dados.

TABELA 2-1 Exemplo hipotético de dados perdidos

Identificação do caso	V_1	V_2	V_3	V_4	V_5	Dados perdidos por caso: Número	Dados perdidos por caso: Percentual
1	1,3	9,9	6,7	3,0	2,6	0	0
2	4,1	5,7			2,9	2	40
3		9,9		3,0		3	60
4	0,9	8,6		2,1	1,8	1	20
5	0,4	8,3		1,2	1,7	1	20
6	1,5	6,7	4,8		2,5	1	20
7	0,2	8,8	4,5	3,0	2,4	0	0
8	2,1	8,0	3,0	3,8	1,4	0	0
9	1,8	7,6		3,2	2,5	1	20
10	4,5	8,0		3,3	2,2	1	20
11	2,5	9,2		3,3	3,9	1	20
12	4,5	6,4	5,3	3,0	2,5	0	9
13					2,7	4	80
14	2,8	6,1	6,4		3,8	1	20
15	3,7			3,0		3	60
16	1,6	6,4	5,0		2,1	1	20
17	0,5	9,2		3,3	2,8	1	20
18	2,8	5,2	5,0		2,7	1	20
19	2,2	6,7		2,6	2,9	1	20
20	1,8	9,0	5,0	2,2	3,0	0	0
Dados perdidos por variável						Valores perdidos totais	
Número	2	2	11	6	2	Número: 23	
Percentual	10	10	55	30	10	Percentual: 23	

Dados perdidos reduzem o tamanho amostral

Processo não-aleatório de perda = características sistematicamente relacionadas à perda -> Viés
Ex.: indivíduos com maior renda

Passo 2: Determinar a extensão de dados perdidos

Extensão ou quantia de dados perdidos baixa o suficiente para não afetar os resultados (viés) mesmo que opere de modo não-aleatório

SIM: Qualquer das abordagens corretivas (técnicas de atribuição) (passo 4) pode ser aplicada sem criar vieses;

NÃO: Determinar aleatoriedade antes de escolher uma ação corretiva (passo 3).

Avaliação da extensão e padrões de perda de dados

Tabulação: (1) do percentual de variáveis com dados perdidos para cada caso; (2) do número de casos com dados perdidos para cada variável;

-> Amostra disponível sem ações corretivas (casos sem perdas em qualquer variável);

-> Se perdas estão concentradas, exclusão de casos e/ou variáveis reduz a extensão dos dados perdidos;

-> Se um padrão não-aleatório está presente, esta solução pode ser a mais eficiente.

Passo 3: Diagnosticar aleatoriedade dos processos

Missing at random (MAR) = Perdidos ao acaso

Probabilidade de perdas em uma variável está relacionada a alguma outra variável, mas não ao valor da própria variável com valores perdidos

=

Valores perdidos de Y dependem de X, mas não de Y

Não pode ser confirmado, porque não pode ser testado se a probabilidade de perdas em uma variável apenas é função de outras variáveis medidas.

Missing completely at random (MCAR) = Perdidos completamente ao acaso

Casos com dados perdidos são indistinguíveis daqueles com dados completos, pois perda não está relacionada a nenhuma outra variável

=

Valores observados de Y são uma amostra aleatória de todos os valores Y e também não dependem de nenhum X

Único mecanismo verificável. É preferível porque acomoda qualquer tipo de ação corretiva.

Passo 3: Diagnosticar aleatoriedade dos processos

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	
26	
29	
30	
30	
31	
44	118
46	93
48	141
51	104
51	116
54	97

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	
26	121
29	91
30	
30	110
31	
44	118
46	93
48	
51	
51	116
54	

- Perda em uma única variável: Dicotomizar perda em Y e testar diferenças entre os dois grupos em outras variáveis de interesse -> SIM = MAR; NÃO = MCAR

- Teste geral de aleatoriedade: Padrão de perda em todas as variáveis X Padrão esperado para um processo aleatório -> SIM = MAR; NÃO = MCAR

Passo 4: Selecionar o método de atribuição

 Atribuição é o processo de estimação de valor perdido baseado em valores válidos de outras variáveis e/ou casos na amostra;

 Variáveis não-métricas não são tratáveis com atribuição.

Missing at random (MAR) = Perdidos ao acaso

(1) Técnicas de estimação de máxima verossimilhança: métodos iterativos de modelagem dos processos inerentes aos dados perdidos para estimar valores mais precisos e razoáveis, bem como os parâmetros após substituição (médias, desvios padrão ou correlações) (Ex.: abordagem EM);

(2) Inclusão das perdas na análise: Tratar observações com dados perdidos como subconjunto da amostra = dicotomizar as perdas e substituir valores ausentes pela média. Na regressão, o coeficiente dessa variável dicotômica avalia a significância estatística da diferença para a variável dependente entre observações com dados perdidos e aquelas com dados válidos.

Para pensar: Técnicas mantêm o tamanho amostral, mas adicionam complexidade, dificultam interpretabilidade e diminuem a heterogeneidade dos dados.

Passo 4: Selecionar o método de atribuição

Missing completely at random (MCAR) = Perdidos completamente ao acaso

Atribuição usando apenas dados válidos = representar amostra inteira com observações com dados válidos

- **Abordagem de caso completo ou Listwise deletion:** casos com perdas em qualquer variável são eliminados
  - Diminui tamanho e poder amostral e é muito afetada por perda não aleatória;
  - Adequada se perda de dados é pequena (<5%); amostra é grande e relações nos dados são fortes.

- **Uso de dados totalmente disponíveis ou Pairwise deletion:** atribuição das características de distribuição (p. ex., médias ou dp) ou de relação (p. ex., correlações) a partir de cada valor válido. Ou seja, assume-se que correlações obtidas nos casos com dados válidos são representativos da amostra inteira.

Passo 4: Selecionar o método de atribuição

Missing completely at random (MCAR) = Perdidos completamente ao acaso

Atribuição usando valores de substituição: valores perdidos substituídos por valores estimados com base em outras informações disponíveis

- **Uso de valores conhecidos de substituição:** valor conhecido substitui dados perdidos em caso similar.
  - **Atribuição por carta marcada** = valor vem de outra observação na amostra considerada semelhante, ou externa à amostra, se informação mais válida;
  - **Substituição por um caso** = observações com dados perdidos totalmente substituídas por outra observação escolhida fora da amostra (Ex.: substituir uma família da amostra, que não pode ser contactada ou que tem extensos dados perdidos, por outra família que não esteja na amostra).

- **Cálculo de valores de substituição:** valor de substituição calculado a partir de um conjunto de observações com dados válidos na amostra.
  - **Substituição pela média =** troca valores perdidos em uma variável pelo valor médio dessa variável, com base em todas as respostas válidas;
  - **Atribuição por regressão =** equação prevê valores perdidos de uma variável com base em sua relação com outras variáveis.

OUTLIERS

Outliers

Definição: observações atípicas, que destoam dos demais valores observados. Podem influenciar indevidamente nos resultados.

O que fazer? Retirar? Manter? Avaliar o contexto!

O evento extraordinário se ajusta aos propósitos da pesquisa?

É erro de procedimento? Retirar, se tornará um dado perdido.

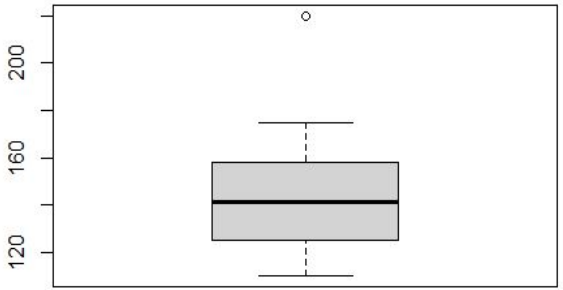
O valor atípico representa um segmento da população?

Outliers

Como identificar?

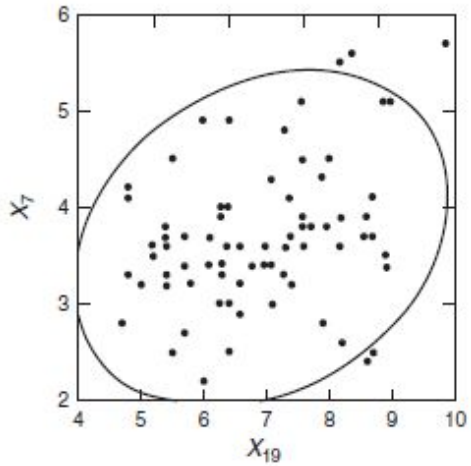
Detecção univariada

Boxplot



Detecção bivariada

Diagrama de dispersão



Detecção multivariada

1. Métodos bivariados são inadequados
2. Medir a posição multidimensional de cada variável em relação a um valor comum

Outliers

Como identificar?

Detecção multivariada

D^2 de Mahalanobis: mede a distância de cada observação em um espaço multidimensional a partir do centro médio de todas as observações. **Valores elevados indicam observações afastadas.**

D^2/nº de variáveis: valores maiores do que 2,5 (em amostras pequenas, n= 80 ou menos) e que 3 ou 4 (em amostras grandes) indicam possíveis outliers.

TESTE DAS SUPOSIÇÕES DA ANÁLISE MULTIVARIADA

NORMALIDADE

Como detectá-la?
Análises gráficas de normalidade: Histograma, Gráfico de probabilidade normal, Diagrama de dispersão, Gráfico de densidade, Boxplot

Testes estatísticos: Shapiro-Wilks, Kolmogorov-Smirnov, Anderson-darling, Cramer-von Mises

Quais as consequências da sua violação?
Forma da distribuição → Nível de Curtose e Assimetria
Tamanho da amostra

E como corrigir?
Transformação dos dados

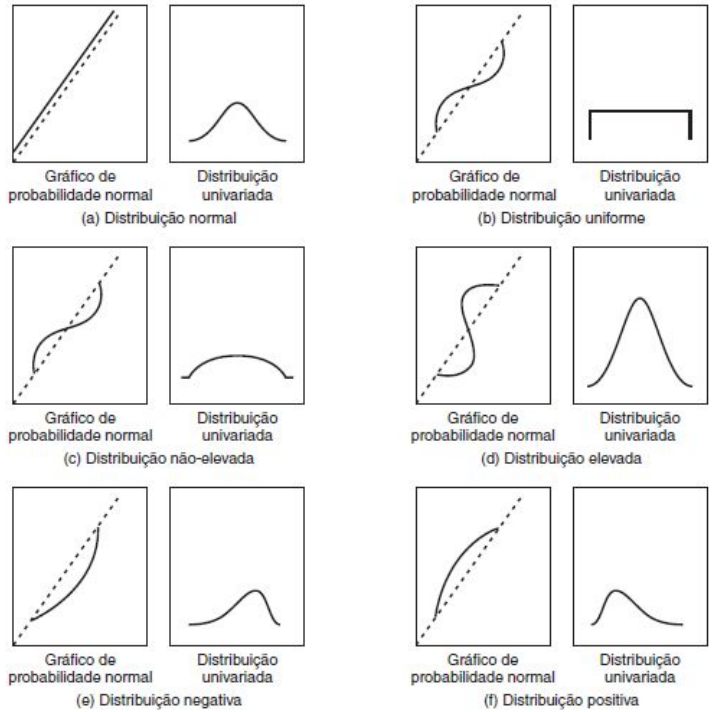


FIGURA 2-8 Gráficos de probabilidade normal e distribuições univariadas correspondentes.

HOMOCEDASTICIDADE

Como detectá-la?
Análise gráfica: Diagrama de dispersão, Boxplot

Testes estatísticos: Levene, M de Box, Bartlett, Breusch Pagan

Quais as consequências da sua violação?
Previsão melhor em alguns níveis do que em outros
Tamanho da amostra

E como corrigir?
Transformação dos dados

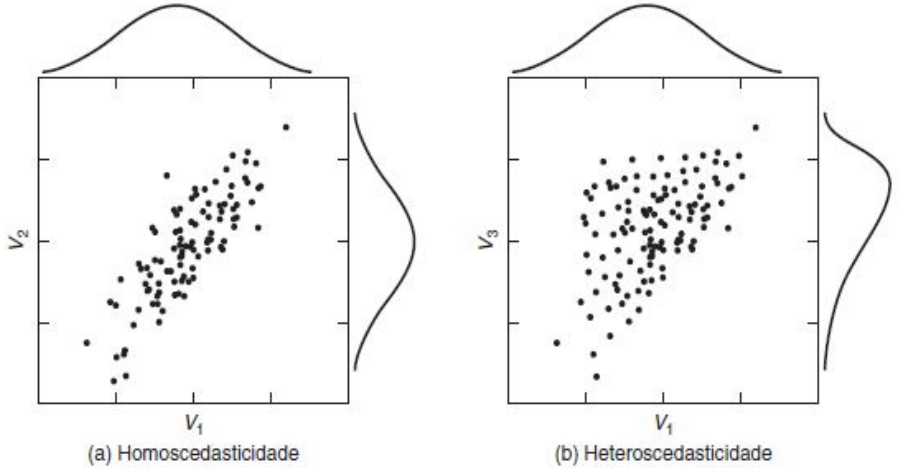


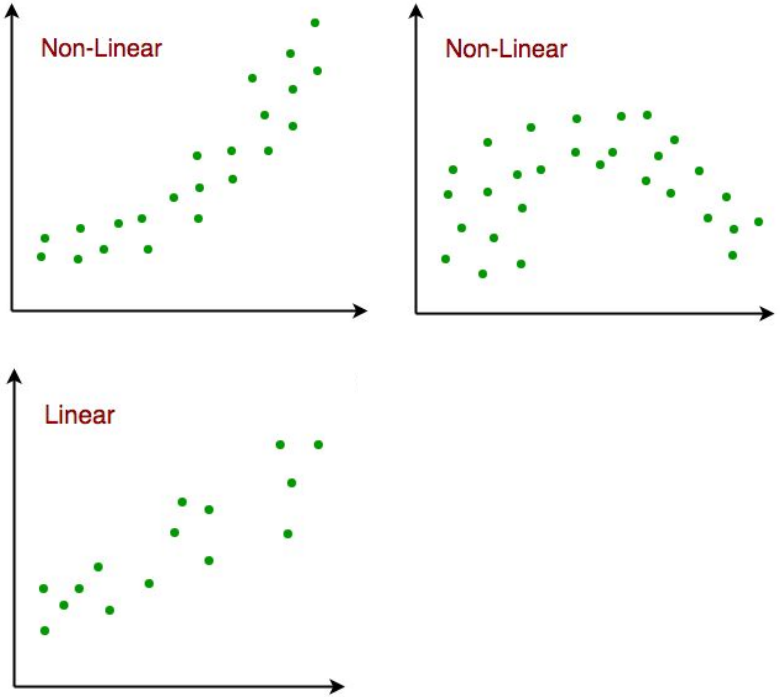
FIGURA 2-9 Diagramas de dispersão de relações homoscedásticas e heteroscedásticas.

LINEARIDADE

Como detectá-la?
Análise gráfica: Diagrama de dispersão
Regressão simples (e examinar os resíduos)
Modelar uma relação não linear

Quais as consequências da sua violação?
Subestimação da força real da relação

E como corrigir?
Transformação dos dados
Criar novas variáveis

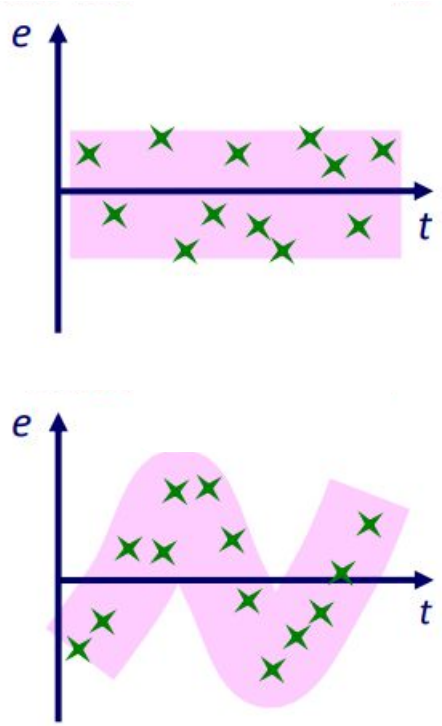


INDEPENDÊNCIA ou AUSÊNCIA DE ERROS CORRELACIONADOS

Como detectá-la?
Identificar possíveis causas: Coleta de dados e Dados em séries temporais

Quais as consequências da sua violação?
Resultado com viés

E como corrigir?
Inclusão do fator causal omitido



TRANSFORMAÇÃO DE VARIÁVEIS

As transformações devem ser aplicadas nas variáveis independentes, exceto no caso de heterocedasticidade.

Tipo de distribuição	Tipo de transformação
achatada	inversa
negativamente assimétricas	de quadrados ou cubos
positivamente assimétricas	raiz quadrada, logaritmos, ou transformação inversa

Gráficos de probabilidade normal

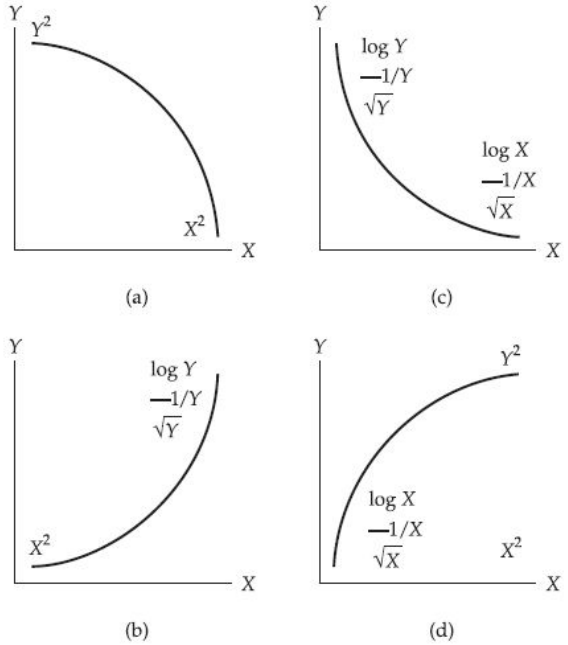


FIGURA 2-10 Seleção de transformações para atingir linearidade.
Fonte: F. Mosteller and J. W. Tukey, Data Analysis and Regression. Reading, MA: Addison-Wesley, 1977.

A fim de executar a maioria das análises multivariadas, não é necessário atender a todas as suposições de normalidade, homocedasticidade, linearidade e independência.

Efeito da violação de certas suposições → **Robustez da técnica**