Introdução à análise multivariada e análise fatorial de dados

Este conteúdo aborda os fundamentos da análise multivariada, um conjunto de técnicas estatísticas utilizadas para explorar, compreender e modelar relações entre múltiplas variáveis simultaneamente.

Profa. Manoela Gonçalves Cabo

Objetivos

- Compreender o que é a análise multivariada de dados e *clustering*, suas principais aplicações, classificações e técnicas envolvidas.
- Aprender a realizar a análise preliminar dos dados, utilizando recursos gráficos, tratamento de dados perdidos, testes das suposições e integração de variáveis não-métricas.
- Entender os conceitos, objetivos e etapas da análise fatorial, bem como sua aplicação prática para redução de variáveis e identificação de padrões latentes.

Introdução

Vivemos na era dos dados. Com o aumento exponencial de informações disponíveis, surge a necessidade de ferramentas que permitam interpretar e extrair sentido de grandes conjuntos de variáveis. A análise multivariada atende a essa demanda ao permitir que pesquisadores e profissionais entendam as relações complexas entre múltiplos fatores de forma simultânea.

O curso inicia com os conceitos básicos de análise multivariada e de agrupamento (*clustering*), avançando para o exame criterioso dos dados e finalizando com a análise fatorial, uma das principais técnicas utilizadas para sintetizar informações e detectar estruturas ocultas nos dados. Os conteúdos combinam teoria e prática, utilizando o *software* R como ferramenta principal para execução dos métodos.

Ao final, o aluno será capaz de aplicar métodos multivariados para resolução de problemas reais nas áreas de estatística, marketing, finanças, ciências sociais e demais campos que lidam com múltiplas variáveis.

Definição de análise multivariada de dados e clustering

Em muitas análises de dados não podemos usar apenas uma variável para tirar inferências, fazer previsões ou simplesmente entender o comportamento dos dados. Em muitos casos, necessitamos de várias variáveis para isso.

A seguir, assista ao vídeo e compreenda o que é análise multivariada e *cluster*.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Exemplo

Por exemplo, não podemos simplesmente usar apenas a ação da Petrobras para fazer previsões do IBOVESPA (Índice BOVESPA), que é composto por uma carteira teórica de ações, com os papéis de maior volume financeiro da Bolsa de Valores do Brasil por um determinado período. Parte de nossas análises não pode ser feita baseando-se apenas em um fator ou variável, nesse caso a Petrobras.

Ao analisar o Índice BOVESPA, os analistas verificam diversos dados como: Crescimento econômico, comportamento de outras ações da carteira, mercado nacional e internacional etc.

> As análises multivariadas consistem em um conjunto de métodos estatísticos utilizados em situações nas quais as variáveis são medidas simultaneamente, em cada elemento amostral. Sendo assim, quando utilizamos várias informações ou dados, realizamos uma análise multivariada de dados.

Análises multivariadas

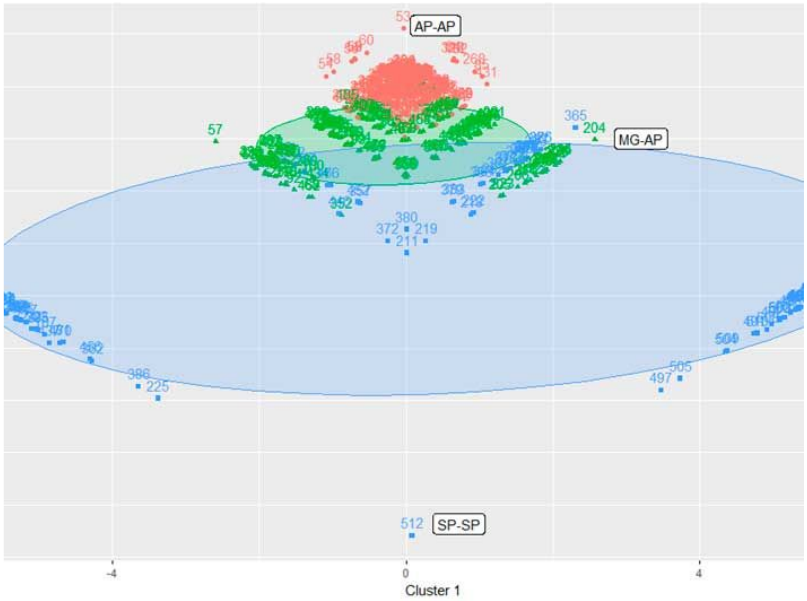
As análises multivariadas consistem em um conjunto de métodos estatísticos utilizados em situações nas quais as variáveis são medidas simultaneamente, em cada elemento amostral. Sendo assim, quando utilizamos várias informações ou dados, realizamos uma análise multivariada de dados.

Análise de clusters

A análise de *clusters*, *clustering* ou análise de agrupamento é o conjunto de metodologias de prospecção de dados (data mining) que utiliza variáveis com o objetivo de fazer agrupamentos segundo seu grau de semelhança e minimizar as diferenças.

Observe a seguir um gráfico, como exemplo, do resultado de uma clusterização, utilizando a técnica fuzzy.

Nesse exemplo abaixo, a partir de um conjunto de informações sobre as cidades brasileiras relacionadas ao transporte aéreo, após a aplicação de técnicas de agrupamento de dados, o método criou três grupos, ou *clusters*, contendo as cidades com maior similaridade e distanciando as cidades com maior variabilidade.



Agrupamento fuzzy cidades brasileiras.

Sendo assim, a análise multivariada e *clustering* são técnicas analíticas que usam informações de várias fontes, ao mesmo tempo, para obter uma ideia melhor, mais completa e mais otimizada do objeto a ser estudado. Ou seja, a análise multivariada é uma ferramenta que acha padrões e relações entre muitas variáveis, nos permitindo prever efeitos e alterações que uma variável terá sobre a outra.

As análises multivariadas podem ser aplicadas em diversos escopos, ainda nos eventos em que não se dispõe previamente de um modelo rigorosamente estruturado a respeito das relações entre esses dados. O objetivo de sua aplicação poderia ser diminuir dados ou simplificação, classificação e agrupamento, tentar verificar a dependência entre variáveis, para previsão e formar hipóteses e testar se são válidas.

Exemplo

Por exemplo, quando solicitamos um cartão de crédito, o banco ou financeira, analisa uma série de informações, como renda, tipo de trabalho, imóveis etc., com o intuito de prever se o solicitante possuirá crédito para honrar os compromissos. Na nossa linguagem, ele faz uma análise das variáveis para predizer ou testar hipóteses sobre o comportamento creditício da pessoa que está solicitando o cartão.

As Análises multivariadas podem ser divididas em duas classificações: A primeira, consistindo em técnicas exploratórias de sintetização ou simplificação da estrutura de variabilidade dos dados; a outra, em métodos de inferência estatística.

Primeira classificação

A primeira classificação consiste em técnicas como análises das componentes principais, análise fatorial, análise de correlações, análises de agrupamentos ou *clustering*, análises discriminantes e análises de correspondência.

Essas técnicas são muito utilizadas porque grande parte independe de conhecimento matemático de distribuições de probabilidade que gera o processo ou dados amostrais.

Segunda classificação

A segunda classificação consiste na utilização de técnicas estatísticas, ou seja, em encontrar os métodos de estimação de parâmetros, análise de variância e covariância (MANOVA) e de vários tipos de regressões multivariadas.

Uma definição muito importante para trabalhar com informações multivariadas é o conceito de vetor aleatório:

Atividade

Seja X um vetor contendo p componentes, em que cada componente é uma variável aleatória (ou dado, informação, por exemplo, a renda citada anteriormente), então, Xi é um dado ou informação ou variável aleatória (renda) para todo i = 1, 2, 3, ..., p. Assim X é chamado de vetor aleatório e é denominado por:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \cdots \\ x_p \end{pmatrix}$$

Um exemplo é o modelo de Credit Score desenvolvido pelos bancos: Considere o assunto abordado acima, seja uma empresa de cartão de crédito que deseja fazer análise creditícia de uma proposta para obtenção de um cartão de crédito.

Algumas informações importantes para essa empresa seriam a renda (para verificar capacidade de pagamento), o tipo de trabalho (se é informal, assalariado, servidor público, sem emprego); o tipo de moradia (aluguel ou próprio); entre outras variáveis.

Cada variável relacionada acima corresponde a um Xi, do vetor aleatório X, sendo que p=3, nesse exemplo. Essa empresa deseja fazer uma análise multivariada das informações dos proponentes de cartão de crédito. Assim, defina o vetor aleatório que essa empresa irá trabalhar:

Chave de resposta

O vetor aleatório será composto por três variáveis, $X_1 =$ Renda; $X_2 =$ Tipo de trabalho; $X_3 =$ Tipo de moradia.

O vetor aleatório será da seguinte forma:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

Sendo que:

$$x = \begin{pmatrix} x_1 = \text{Renda} \\ x_2 = \text{Tipo de trabalho} \\ x_3 = \text{Tipo de moradia} \end{pmatrix}$$

$X = [X_1 X_2 X_3]'$ é o vetor aleatório transposto, sendo $p = 3$, que são variáveis aleatórias.

Resumindo

Em resumo, os métodos de análise multivariada são utilizados com o propósito de simplificar ou facilitar a interpretação do acontecimento que estamos estudando ou trabalhando por meio de construções de índices ou variáveis alternativas que sintetizem a informação original dos dados; construção de conjunto de elementos que apresentem similaridade entre si, possibilitando a segmentação do conjunto de dados originais; investigação das relações de dependência entre as variáveis respostas associadas ao fenômeno e outros fatores com objetivo, muitas vezes, de se fazer previsões.

Para facilitar o entendimento, apresentaremos classificações das técnicas e alguns tipos de técnicas de análise multivariada que serão vistas ao longo do curso, com alguns exemplos de aplicação.

Classificação e tipos de técnicas de análise multivariada.

As técnicas de análise multivariada são divididas em dois grupos. Confira quais são no vídeo a seguir.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Primeiro conjunto

Conjunto que constitui técnicas exploratórias de sintetização ou simplificação da estrutura de variabilidade dos dados, em que são abordadas técnicas de análise de componentes principais, análise fatorial, análise de correlações canônicas, análise de agrupamentos (*clustering*), análise discriminante e análise de correspondência. Pela facilidade de implementação e por não precisar da distribuição de probabilidade, essas técnicas são muito utilizadas.

Segundo conjunto

Conjunto que constitui técnicas e métodos de inferência estatística que versam no emprego de técnicas estatísticas, análise de variância (MANOVA), análise de covariância e vários tipos de regressões multivariadas.

Existem diversos tipos de técnicas para realizar uma análise multivariada, sendo os mais utilizados:

Análise fatorial

Nesse tipo de técnica incluem-se as análises de componentes principais e as análises de fator. Ela é utilizada quando existe um número muito extenso de variáveis correlacionadas entre elas mesmas, com o objetivo de selecionar um número menor de outras variáveis alternativas, que não sejam correlacionadas e que sintetizem as informações básicas das variáveis originais achando os fatores ou variáveis latentes.

Ou seja, a Análise Fatorial tem como objetivo descrever a variabilidade de variáveis correlacionadas observadas em menos variáveis não observadas. No exemplo do cartão de crédito, um possível resultado da análise fatorial seria: Renda (0,7), Tipo de emprego (0,8) e Tipo de moradia (0,6); essas variáveis têm grandes cargas fatoriais positivas no fator 1, portanto, esse fator descreve uma adequação e potencial do proponente ao cartão de crédito (esse cálculo será feito em aulas posteriores).

Análise de regressão múltipla

Constitui um conjunto de técnicas estatísticas para construir modelos que descrevem de maneira plausível as relações entre variáveis explicativas de um determinado processo. A diferença entre a regressão linear simples e a múltipla é que na múltipla são tratadas duas ou mais variáveis explicativas, que é o caso de análise multivariada.

Sendo assim, é um método de fazer análises apropriadas quando existe uma única variável dependente relacionada a duas ou mais variáveis explicativas. O objetivo de se utilizar esse método é tentar fazer a previsão das alterações na variável dependente ou explicada, de acordo com as variações nas variáveis independentes ou explicativas.

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X$$

Em que a variável Yt é a variável dependente ou explicada e Xi são as p variáveis independentes ou explicativas.

Análise discriminante múltipla e regressão logística

Esse tipo de técnica ou análise é empregado quando os conjuntos são conhecidos antes de se fazer a análise, ou seja, a priori. Por exemplo, queremos formar três grupos de dados, o número de cluster é três. Na análise discriminante utilizam-se algumas ferramentas para distinguir grupos de populações e classificar as novas observações nos grupos determinados.

Ou seja, será utilizada para classificação de elementos de uma amostra ou população. Como exemplo, citamos os bancos e financeiras que podem utilizar análise discriminante para traçar o perfil dos clientes de cartão de crédito com o objetivo de julgar o risco de oferecer o crédito a um novo cliente.

Análise multivariada de variância

A análise multivariada de variância ou MANOVA (análise multivariada de variância) é utilizada para verificar a semelhança entre grupos de muitas variáveis, constatando ao mesmo tempo as relações entre diversas variáveis independentes ou explicativas e duas ou mais variáveis dependentes ou explicadas.

Esse tipo de análise ajuda a responder se as mudanças na variável independente, ou nas variáveis independentes, têm efeitos significantes nas variáveis dependentes, ou para conhecer as relações entre as variáveis dependentes e até as relações entre as variáveis independentes.

Análise conjunta

A análise conjunta é um método de dependência muito aplicado na utilização da avaliação de objetos, por exemplo, no lançamento de um produto novo, um serviço ou ideias novas e inovadoras. A pergunta que a análise conjunta apresenta seria um conjunto de cenários possíveis para os consumidores e pede que eles tomem uma decisão sobre a qual escolheriam.

Por exemplo, atributos (cor, tamanho, preço) que são detalhados por um conjunto de níveis (rosa, verde, pequeno, 15 reais, 50 reais etc.).

Análise de agrupamento ou *Clustering*

Análise de agrupamento ou *clustering* é o nome dado para o grupo de técnicas computacionais cujo propósito consiste em separar objetos em grupos ou *clusters*, baseando-se nas características que estes objetos possuem.

Ou seja, é uma análise que identifica grupos em objetos de dados multivariados. O objetivo é formar grupos (*clusters*) com propriedades homogêneas de amostras heterogêneas grandes.

Deve-se buscar grupos mais homogêneos possíveis e que as diferenças entre eles sejam as maiores possíveis. Existem vários métodos para esse tipo de análise, como método hierárquico de escolher os parâmetros distância (Euclidiana, Manhattan ou Gower) e os outros métodos (*ward, single, complete, average, median ou centroid*).

Também existem os métodos K-médias, em que há apenas a possibilidade de definir o número de grupos e padronizar os dados, definindo-se o número de *clusters*.

Escalonamento multidimensional e Análise de correspondência

Escalonamento multidimensional e análise de correspondência são técnicas que podem ser utilizadas quando se pretende modificar o julgamento de algum consumidor sobre similaridade ou preferência em distâncias representadas em um espaço multidimensional.

Existem outros tipos de técnicas, além das técnicas básicas relacionadas acima, como as técnicas de Modelagem de Equações Estruturais (SEM), de análise fatorial confirmatória e o teste de um modelo estrutural.

Diretrizes para análises multivariadas e interpretação

Existem diretrizes para análises multivariadas e interpretação. Apresentaremos a seguir alguns exemplos gerais de aplicação, como a construção de índices; classificação e discriminação; associação entre variáveis categórica; e inferência estatística.

Na construção de índices coletam-se diversas variáveis que descrevem um fenômeno, com intuito de construir um índice específico relativo à sua quantificação. A função básica do índice é sintetizar em uma única variável a informação de todas as variáveis que formam medidas sobre o fenômeno, sendo que seus valores podem ser analisados por métodos de estatísticas univariadas. Como exemplo, tem-se a inflação, o risco Brasil, o IDH (índice de desenvolvimento humano), o índice de desemprego etc.

Na classificação e discriminação, muitas são as situações em que há um conjunto de dados e se busca uma divisão desses dados em grupos, de modo que os grupos tenham coesão interna e sejam heterogêneos entre si (muito utilizada em *data mining*).

A seguir, assista ao vídeo e conheça alguns exemplos gerais de aplicação das diretrizes.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Verificando o aprendizado

Questão 1

Uma empresa, produtora de bens de consumo, deseja fazer uma análise multivariada da situação econômica do mercado do Rio de Janeiro, no ano de 2020, com o objetivo de identificar se deve aumentar a sua produção. Algumas informações importantes socioeconômicas que a empresa pretende analisar seriam o PIB (Produto Interno Bruto) do Estado, seu crescimento populacional e a taxa de desemprego do Rio de Janeiro. Defina o vetor aleatório que essa empresa irá trabalhar.

Chave de resposta

Cada variável relacionada acima corresponde a um X_i, do vetor aleatório X, sendo que p = 3. Essa empresa deseja fazer uma análise multivariada das informações socioeconômicas do Estado do Rio de Janeiro.

O vetor aleatório será composto por três variáveis, X_1 = PIB; X_2 = Crescimento populacional; X_3 = taxa de desemprego.

O vetor aleatório será da seguinte forma:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

Sendo que:

$$x = \begin{pmatrix} x_1 = \text{ PIB} \\ x_2 = \text{ Crescimento populacional} \\ x_3 = \text{ Taxa de desemprego} \end{pmatrix}$$

Questão 2

Os gestores de uma Universidade desejam formar quatro grupos de alunos considerando variáveis como renda, idade, localidade e tipo de emprego, com o objetivo de criar programas de incentivo financeiro. Eles pretendem formar esses grupos para separá-los de acordo com as similaridades entre os alunos. Qual técnica de análise multivariada eles devem utilizar?

Chave de resposta

Os gestores devem usar a técnica de análise de agrupamento ou *clustering*, cujo propósito consiste em separar objetos em grupos ou *clusters*, baseando-se nas características que esses objetos possuem.

Ou seja, é uma análise que identifica grupos em objetos de dados multivariados. O objetivo é formar grupos (*clusters*) com propriedades homogêneas de amostras heterogêneas grandes. Deve-se buscar grupos mais homogêneos possíveis, e as diferenças entre eles devem ser as maiores possíveis.

Questão 3

Uma interpretação e aplicação da análise multivariada é a construção do IDH (Índice de Desenvolvimento Humano). A estatística é composta a partir de dados de expectativa de vida ao nascer, educação e PIB per capita (como um indicador do padrão de vida) recolhidos em nível nacional. Qual diretriz do tipo de aplicação foi utilizada para construir o IDH?

Chave de resposta

A diretriz do tipo de aplicação utilizada para construir o IDH foi a construção de índice, em que são coletadas diversas variáveis. Nesse caso, expectativa de vida ao nascer, educação e PIB per capita (como um indicador do padrão de vida), recolhidos em nível nacional, descrevem um fenômeno, o IDH, com o intuito de construir um índice específico relativo à sua quantificação.

A função básica do IDH é sintetizar em uma única variável a informação de todas as variáveis que formam medidas sobre o fenômeno, sendo que seus valores podem ser analisados por métodos de estatísticas univariadas.

Análise dos dados ou variáveis multivariadas

Para começarmos, assista ao vídeo conheça o que são análise dos dados ou variáveis multivariadas.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

A análise dos dados ou variáveis multivariadas é o primeiro passo da nossa disciplina. Demanda conhecimento e tempo, mas muitas vezes não damos tanta importância a essa etapa extremamente importante e crucial para o entendimento das informações.

É nesse momento que analisamos os dados ou variáveis multivariadas, fazemos o exame gráfico dos dados, verificamos e classificamos as técnicas e tipos de análise multivariada, avaliamos o impacto e tratamos os dados perdidos e atípicos (*outliers*), realizando testes das suposições e a incorporação de dados não-métricos.

> O alvo dessa atividade de análise de dados é muito mais no sentido de revelar o que não é aparente do que retratar os dados reais, pois esses fatores, muitas vezes, passam despercebidos.

Ao fazermos a análise dos dados, antes de aplicar quaisquer técnicas de análise multivariada, passamos a ter um aspecto mais crítico das características dos dados ou das variáveis multivariadas.

Devemos ter duas visões dessa análise, conforme detalhamos a seguir:

01

Primeiramente, conquistamos uma compreensão básica dos dados e das relações entre variáveis. As técnicas de análise multivariada possuem muitas exigências para nós entendermos, interpretarmos e verificarmos os resultados. O conhecimento dessas relações pode ajudar imensamente na especificação e no aprimoramento do modelo de análise multivariada, oferecendo uma perspectiva para a interpretação dos resultados.

02

Em seguida, garantimos que os dados e variáveis atendem a todas as exigências para aplicação de uma técnica multivariada. Essas técnicas demandam mais informações, ou seja, maiores conjuntos de dados e suposições mais complexas do que aquilo que se encontra na análise univariada.

Dados perdidos, observações atípicas e as características estatísticas dos dados são muito mais difíceis de avaliar em um contexto multivariado. Por isso, para garantir que essas exigências sejam atendidas, temos que empregar uma série de técnicas de exame de dados tão complexas quanto as próprias técnicas multivariadas.

As variáveis, ou dados, são colocadas em um vetor contendo p componentes, e cujo cada componente é uma variável aleatória. Então, Xi é um dado ou informação ou variável aleatória para todo i = 1, 2, 3, ..., p. Sendo que X é chamado de vetor aleatório:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \cdots \\ x_p \end{pmatrix}$$

Inicialmente, faremos um exame gráfico desse vetor aleatório, de forma univariada e bivariada. Para começar, vamos fazer um gráfico de apenas uma variável (Xi), que é simples; no entanto, ter uma visão multivariada é um pouco mais complexo.

Exame gráfico dos dados

Assista ao vídeo saiba como realizar o exame gráfico dos dados.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

> O exame gráfico é um dos métodos para analisar um conjunto de variáveis ou dados multivariados; pode ser de forma univariada (uma variável), bivariada (duas variáveis) ou multivariável (várias variáveis). O exame gráfico multivariado pode mostrar o que os gráficos unidimensionais são incapazes de alcançar.

Assim que nós começarmos a realizar análises multivariadas mais complexas, a necessidade e o nível de compreensão vão aumentar e nos demandar medidas diagnósticas empíricas ainda mais poderosas.

Conquistaremos maior entendimento sobre o significado dessas variáveis utilizando técnicas gráficas, retratando as características básicas de variáveis individuais e relações entre elas em uma “imagem” simples.

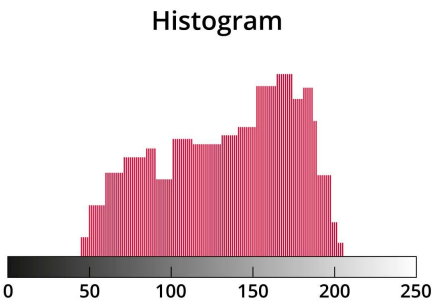
Os programas estatísticos nos dão acesso aos métodos gráficos. A maioria desses programas tem bibliotecas de técnicas gráficas para o exame de dados. No nosso curso utilizaremos o R como software padrão para esse tipo de análise por ser um programa livre e a que todos podem ter acesso.

Análise univariada – formato da distribuição

Para começar a análise, precisamos entender a distribuição ou forma das variáveis, se os dados são normalmente distribuídos, se são simétricos ou não, entre outras características. Um gráfico uni variado que poderá auxiliar-nos nesse processo é o histograma.

> O histograma, ou gráfico de distribuição de frequências, consiste em uma representação gráfica de dados divididos em classes, uma forma gráfica para distribuição de uma variável, um gráfico estatístico para a organização dos dados que exibe a frequência de uma determinada variável.

O histograma é uma variação do gráfico de barras. Enquanto o gráfico de barras descreve os dados em barras e categorias separadas, o histograma representa os dados da mesma categoria no intervalo analisado, por isso, sem espaço entre as barras.



Exemplo de histograma.

Ele consiste em uma distribuição de frequências cuja base de cada uma das barras representa uma classe, e, a altura, a quantidade ou frequência absoluta com que o valor da classe ocorre.

O histograma pode ser feito utilizando o próprio *software* Excel, pois ele tem um gráfico fácil de usar e montar.

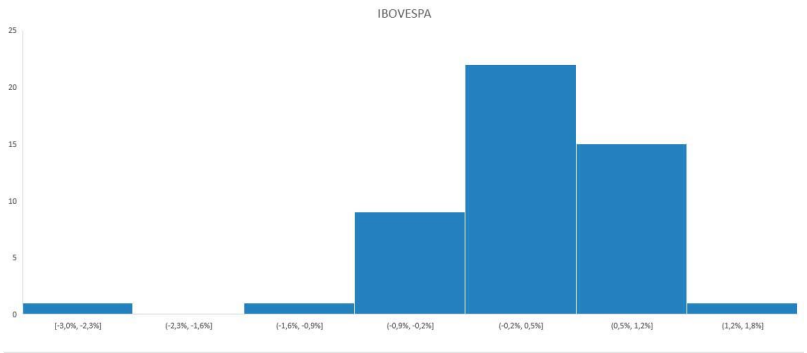
Dica

Primeiro, selecione todo o conjunto de dados. Vá para a guia “Inserir” e selecione “Gráficos recomendados” no grupo “Gráficos”. Clique na opção “Todos os Gráficos”, selecione “Histograma” e depois “OK”.

Não poderíamos usar apenas a ação da Petrobras para fazer previsões do IBOVESPA (Índice BOVESPA). Parte de nossas análises não poderia ser feita baseando-se apenas em um fator ou variável, nesse caso a Petrobras.

Ao analisar o Índice BOVESPA, os analistas verificam diversos dados como: Crescimento econômico, comportamento de outras ações da carteira, mercado nacional e internacional etc. Uma das variáveis Xi para análise multivariada seria o IBOVESPA.

Selecionamos as informações mensais do Índice de ações - Ibovespa - fechamento - (% a.m.), entre janeiro de 2017 e fevereiro de 2021, e montamos um histograma no Excel conforme a seguir:



Assista ao vídeo saiba como realizar a análise univariada.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Por meio do histograma, podemos interpretar a distribuição dos dados, e observamos que a distribuição é assimétrica. Notamos também que a distribuição difere da distribuição normal, pois parece ser mais achatada.

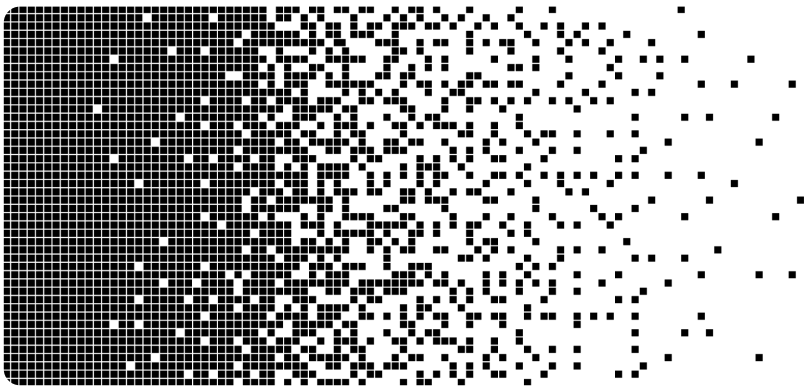
> Uma variante do histograma é o diagrama de ramos e folhas, que apresenta a mesma ilustração gráfica do histograma, mas fornece uma enumeração dos valores reais dos dados. O gráfico de ramos e folhas é um tipo de gráfico com dados na sua forma explícita.

O objetivo é separar os dados de maneira que, antes da barra, fique determinada a unidade de medida dos dados, por exemplo, as dezenas, e, depois da barra, fique a unidade de medida que falta.

Análise bivariada – relação entre variáveis

Muitas vezes, **analisar a distribuição de cada variável não é suficiente**; precisamos fazer a análise de duas variáveis para verificar o comportamento entre elas. O método mais popular para examinar relações bivariadas é o diagrama de dispersão, um gráfico de pontos baseado em duas variáveis.

> Os diagramas de dispersão ou gráficos de dispersão são representações de dados de duas (tipicamente) ou mais variáveis organizadas em um gráfico. O gráfico de dispersão utiliza coordenadas cartesianas para exibir valores de um conjunto de dados.



Os dados são exibidos como uma coleção de pontos, cada um com o valor de uma variável determinando a posição no eixo horizontal, e o valor da outra variável determinando a posição no eixo vertical (em caso de duas variáveis).

Para desenvolver o gráfico de dispersão, sugerimos a utilização do *software* R.

Para mostramos um exemplo, vamos usar um conjunto de dados padrão do R para esse tipo de análise, o conjunto Boston Housing, que contém 14 variáveis associadas com o mercado imobiliário americano na década de 1970.

Algumas dessas informações são a taxa de criminalidade per capita (X_1), proporção de área residencial (X_2), e a área destinada ao comércio (X_3).

Para acessar esse banco de dados no R, utilizaremos o código a seguir:

```
plain-text
>library(MASS)
>data(Boston)
```

Para fazer o diagrama de dispersão multivariado, usando quatro variáveis (rm, medv, lstat, age), utilizaremos o seguinte código:

```
plain-text
>pairs(Boston[,c("rm","medv","lstat","age")])
```

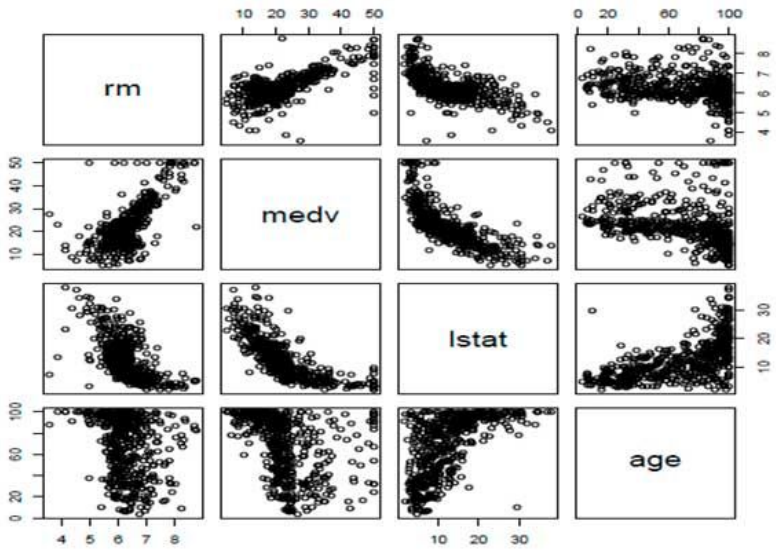


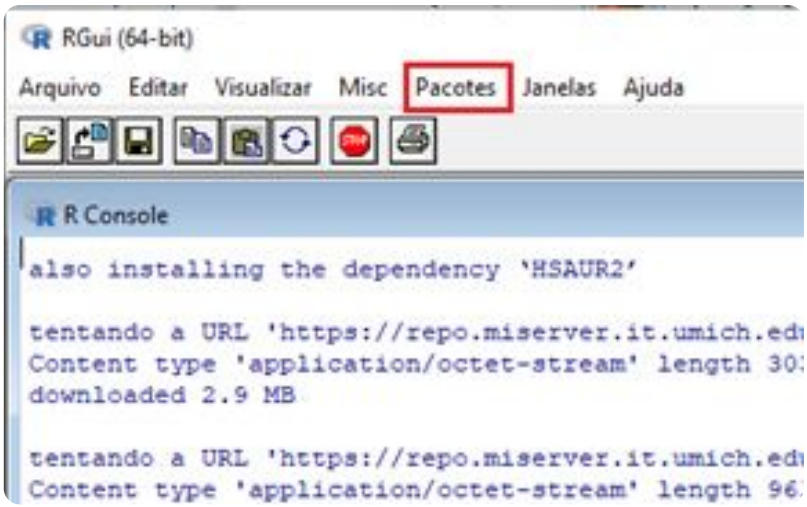
Diagrama de dispersão do R.

Além dos gráficos apresentados acima, existem outros gráficos como o boxplot bivariado. Para construirmos o boxplot bivariado para as variáveis, utilizamos o seguinte código:

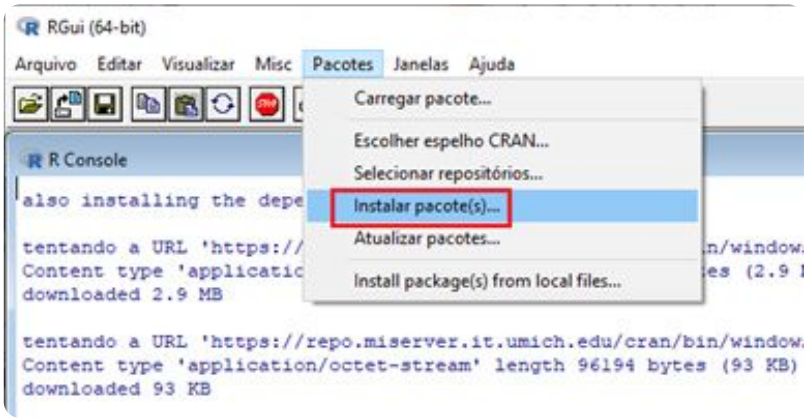
```
plain-text
> library(MVA)
> bvbox(cbind(Boston$rm, Boston$medv), xlab = "rm", ylab = "medv", method ="O")
```

Caso o pacote MVA não esteja instalado siga os seguintes passos.

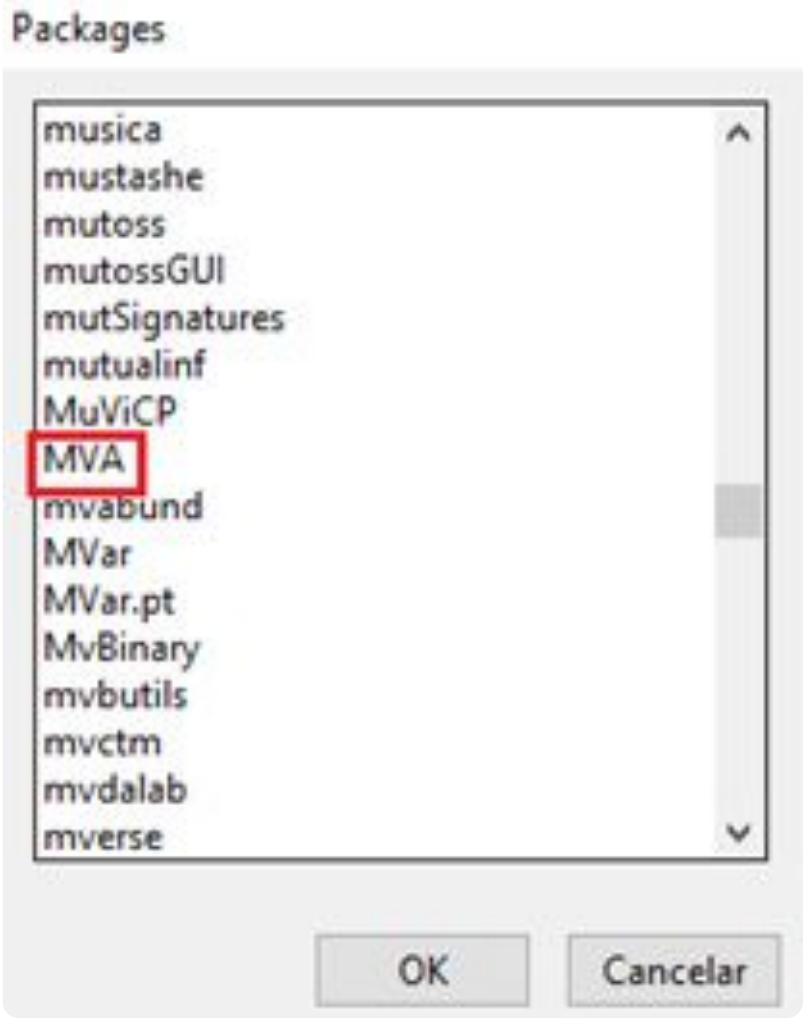
Passo 1: No R clique em Pacotes.

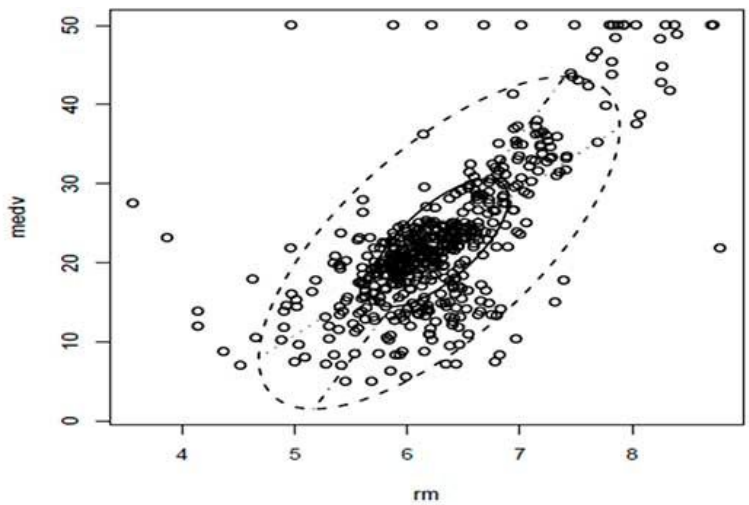


Passo 2: Clicar em Instalar pacotes.



Passo3: Buscar por MVA e clicar em ok.





Boxplot bivariado do R.

Dados perdidos, outliers e testes das suposições estatísticas.

Dados perdidos e outliers

Definimos os dados perdidos como sendo os valores válidos sobre uma ou mais variáveis que não estão disponíveis para análise. O nosso desafio é abordar as questões que podem ser geradas pelos dados perdidos que afetam os resultados da análise multivariada.

Para isso, temos que identificar o padrão e a relação intrínseca a essas variáveis, tentando chegar o mais próximo possível da distribuição de origem dos dados. Esses dados perdidos podem ter várias origens, como erros de entrada de dados, problemas de coleta de dados, o informante não responder a uma questão ou conjunto de questões.

Muitas vezes, tendemos a utilizar apenas os valores válidos, mas os dados perdidos nos levam à redução do tamanho de amostra para análise. E o resultado estatístico baseado em dados com um processo não aleatório de dados perdidos pode ser tendencioso.

Os três principais tipos de mecanismos estão descritos a seguir:

MCAR *(missing completely at random*)

A omissão não está relacionada às variáveis (dependentes ou independentes). Por exemplo, quando a falta da informação se deve à perda de acompanhamento do cliente por ele ter mudado de endereço por motivos alheios ao estudo;

MAR (*missing at random*)

A omissão pode depender do que é observado (variáveis dependentes ou independentes), mas ela não depende dos valores que estão faltando. Isso pode ter ocorrido no exemplo com falta de informação sobre o sexo do cliente, mas não dos valores que se deseja analisar;

MNAR (*missing not at random*)

A omissão depende também do que não é observado. Nesse caso, as variáveis observadas não explicam completamente a omissão dos dados. Por exemplo, faltam informações sobre a renda do cliente do sexo feminino

Uma possível solução para tratar dados perdidos pode ser eliminar casos ou variáveis; essa é uma solução prática. No entanto, se faz necessário um processo estruturado para identificar a presença de processos de dados perdidos, e então aplicar as soluções devidas.

O processo mais utilizado é o processo em quatro etapas para identificar os dados perdidos, conforme a seguir:

Etapa 01

Determinar o tipo de dados perdidos. Identificar se os dados perdidos são ignoráveis ou não.

Etapa 02

Determinar a extensão dos dados perdidos.

Etapa 03

Diagnosticar a aleatoriedade dos processos de perda de dados.

Etapa 04

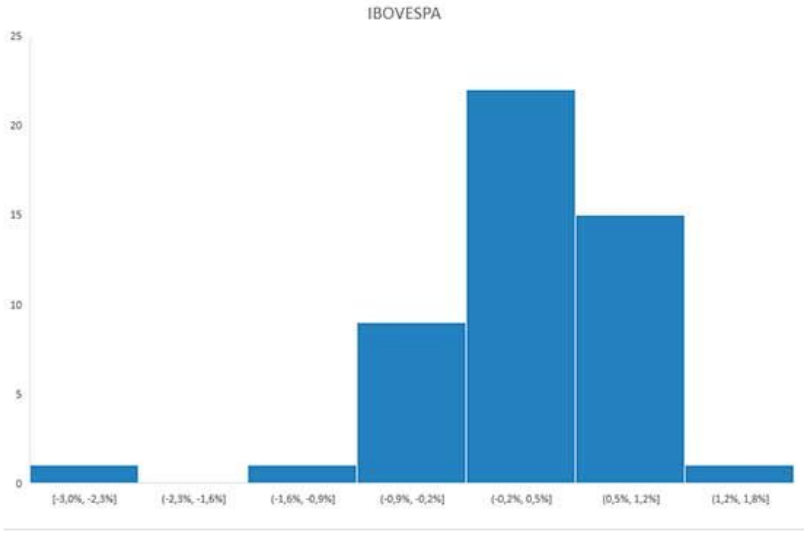
Selecionar o método de atribuição. Se deseja usar apenas casos com dados completos ou todos os dados válidos possíveis; se deseja usar valores conhecidos ou calcular valores de substituição a partir de dados válidos.

> A atribuição é a ação mais lógica; mesmo dando o benefício de eliminar os dados e as variáveis, estaremos nos prevenindo contra a solução simples de usar o método de caso completo, pois isso resulta em um tamanho de amostra inadequado.

Portanto, alguma forma de atribuição se faz necessária para manter um tamanho de amostra adequado para qualquer análise multivariada. As correlações atribuídas diferem de acordo com as técnicas. Múltiplos métodos para substituir os dados perdidos estão disponíveis e são apropriados.

Um outro ponto a que devemos estar atentos são as observações atípicas ou *outliers*, que se diferenciam das demais e levantam suspeitas de que aquela observação foi gerada por um mecanismo distinto. Têm uma combinação única de características identificáveis como sendo notavelmente diferentes das outras observações.

No exemplo do IBOVESPA, podemos identificar, em vermelho, possíveis dados atípicos ou *outliers*, conforme a figura a seguir:



Testes das suposições estatísticas

Em uma análise multivariada precisamos que as suposições pertinentes às técnicas estatísticas sejam testadas duas vezes:

- Primeiro, para as variáveis separadas, de modo semelhante aos testes para uma análise univariada.
- Segundo, para a variável estatística do modelo multivariado, a qual atua coletivamente para as variáveis na análise, e, assim, devem atender às mesmas suposições das variáveis individuais.

São realizados alguns testes para suposição estatística, sendo assim podemos evitar o risco de uma análise falha e com vieses.

Vários testes devem ser aplicados, dentre eles:

Teste de Normalidade

A normalidade pode ter sérios efeitos em pequenas amostras (com menos de 50 casos), mas o impacto diminui efetivamente quando a amostra atinge 200 casos ou mais.

Testar Heteroscedasticidade

A heterocedasticidade é o fenômeno estatístico que ocorre quando o modelo de hipótese matemático apresenta variâncias não iguais para todas as observações A maioria dos casos de heteroscedasticidade é resultado de não normalidade em uma ou mais variáveis; assim, corrigir normalidade pode não ser necessário devido ao tamanho de amostra, mas sim para igualar a variância.

Teste de Linearidade Relações não lineares

Relações não lineares podem ser bem definidas, mas seriamente subestimadas, a menos que os dados sejam transformados em um padrão linear ou componentes de modelo explícito sejam usados para representar a porção não linear da relação.

Teste de Erros Correlacionados.

Erros correlacionados surgem de um processo que deve ser tratado de forma muito parecida com a perda de dados; ou seja, o pesquisador deve primeiramente definir as causas entre variáveis como internas ou externas ao conjunto de dados; se não forem descobertas e remediadas, sérios vieses podem acontecer nos resultados, muitas vezes desconhecidos pelo pesquisador.

Verificando o aprendizado

Questão 1

O que é realizado na etapa de análise dos dados ou variáveis multivariadas?

Chave de resposta

É nesse momento que analisamos os dados ou variáveis multivariadas, fazemos o exame gráfico dos dados, verificamos e classificamos as técnicas e tipos de análise multivariada, avaliamos o impacto e tratamos os dados perdidos e atípicos (*outliers*), realizamos os testes das suposições e a incorporação de dados não-métricos.

Questão 2

O que é o diagrama de dispersão? Ele serve para fazer uma análise univariada ou bivariada?

Chave de resposta

Os diagramas de dispersão ou gráficos de dispersão são representações de dados de duas (tipicamente) ou mais variáveis organizadas em um gráfico. O gráfico de dispersão utiliza coordenadas cartesianas para exibir valores de um conjunto de dados.

Os dados são exibidos como uma coleção de pontos, cada um com o valor de uma variável determinando a posição no eixo horizontal e o valor da outra variável determinando a posição no eixo vertical (em caso de duas variáveis). Ele serve para fazer uma análise bivariada.

Questão 3

Quais são os três principais mecanismos geradores de dados perdidos ou *missing data*? E qual a sua definição?

Chave de resposta

Os três principais mecanismos geradores de dados perdidos são MCAR, MAR e MNAR. MCAR (*missing completely at random*): A omissão não está relacionada às variáveis (dependentes ou independentes). Por exemplo, quando a falta da informação se deve à perda de acompanhamento do cliente por ele ter mudado de endereço por motivos alheios ao estudo;

MAR (*missing at random*): A omissão pode depender do que é observado (variáveis dependentes ou independentes), mas ela não depende dos valores que estão faltando. Isso pode ter ocorrido no exemplo com falta de informação sobre o sexo do cliente, mas não dos valores que se deseja analisar;

MNAR (*missing not at random*): A omissão depende também do que não é observado. Nesse caso, as variáveis observadas não explicam completamente a omissão dos dados. Por exemplo, faltam informações sobre a renda do cliente do sexo feminino.

Conceito de análise fatorial

Assista ao vídeo e conheça o que se pretende alcançar com a análise multifatorial.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

> Quanto mais acesso à informação, maior é o número de variáveis a serem estudadas em todas as áreas, e são cada vez maiores as técnicas multivariadas utilizadas, sendo necessário o conhecimento da estrutura e das inter-relações das variáveis.

Nessa conteúdo, estudaremos a análise fatorial, uma técnica que nos ajuda a analisar os padrões de relações multidimensionais. Essa análise é empregada para examinar os padrões ou relações latentes para um grande número de informações e determinar se as variáveis podem ser condensadas ou resumidas a um conjunto menor de fatores ou componentes.

Relembrando

A análise fatorial é uma técnica de interdependência, que tem como objetivo definir a estrutura inerente entre as variáveis na análise. Ela analisa a estrutura das inter-relações (ou correlações) em um grande número de variáveis formando conjuntos de variáveis que são fortemente inter-relacionadas, que chamaremos de fatores.

Os fatores ou grupos de variáveis são altamente intercorrelacionados, e representam as dimensões dentro dos dados. Esses fatores podem ter por objetivo apenas a redução do número de variáveis, então as dimensões podem orientar a criação de novas medidas compostas.

Um outro objetivo é que os fatores da análise fatorial desempenhem um papel confirmatório, ou seja, que avaliem o grau em que os dados satisfazem a estrutura esperada.

Variabilidade original em termos de um número menor de variáveis aleatórias.

No vídeo, a especialista explica o que se pretende alcançar com a análise multifatorial.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Para descrever a variabilidade original em termos de menos variáveis aleatórias, com a aplicação da análise fatorial, devemos seguir os sete estágios listados a seguir:

1. Objetivos da análise fatorial.
2. Planejar uma análise fatorial, incluindo a seleção de variáveis e o tamanho da amostra.
3. Suposições da análise fatorial.
4. Obtenção de fatores e avaliação de ajuste geral, incluindo o modelo fatorial a ser usado e o número de fatores.
5. Rotação e interpretação de fatores.
6. Validação das soluções da análise fatorial.
7. Usos adicionais de resultados fatoriais, como seleção de variáveis substitutas, criação de escalas múltiplas ou cáclulo de escores fatoriais.

Estágio 1

No estágio 1, o objetivo da análise fatorial é encontrar um modo de resumir as informações contidas em diversas variáveis originais, transformando em um conjunto menor de novas dimensões compostas ou fatores, que são as variáveis estatísticas com uma perda mínima de informação.

Temos que buscar e definir as bases fundamentais ou dimensões assumidas como inerentes às variáveis originais. Ao atingir seus objetivos, a análise fatorial é ajustada com três questões:

- a) especificação da unidade de análise;
- b) obtenção do resumo de dados e/ou redução dos dados;
- c) seleção de variáveis e uso de resultados da análise fatorial com outras técnicas multivariadas

Estágio 2

No estágio 2, o planejamento de uma análise fatorial, é importante verificar se uma perspectiva exploratória ou confirmatória é assumida. Vamos confiar na técnica para fornecer uma visão sobre a estrutura dos dados, mas a estrutura revelada na análise depende de decisões em tópicos como variáveis incluídas, tamanho da amostra e assim por diante. Desse modo, envolve três decisões básicas:

a) cálculo dos dados de entrada (uma matriz de correlação) para atender aos objetivos especificados de agrupamento de variáveis ou respondentes;

b) planejamento do estudo em termos de número de variáveis, propriedades de medida das variáveis e tipos de variáveis admissíveis; e

c) o tamanho necessário para a amostra em termos absolutos e como função do número de variáveis na análise

A análise fatorial é executada geralmente sobre variáveis métricas, apesar de existirem métodos especializados para o emprego de variáveis dicotômicas; um número pequeno de "variáveis dicotômicas" pode ser incluído em um conjunto de variáveis métricas que são analisadas por fatores.

Se um estudo está sendo planejado para revelar estrutura fatorial, temos que nos esforçar para ter pelo menos cinco variáveis para cada fator proposto.

Para tamanho de amostra: A amostra deve ter mais observações do que variáveis; o menor tamanho absoluto de amostra deve ser de 50 observações; maximize o número de observações por variável, com um mínimo de 5 e, com sorte, com pelo menos 10 observações por variável.

Estágio 3

No estágio 3, as suposições na análise fatorial são mais conceituais do que estatísticas. Temos que estar preocupados em atender à exigência estatística para qualquer técnica multivariada, mas em análise fatorial as preocupações que se impõem concentram-se muito mais no caráter e na composição das variáveis incluídas na análise do que em suas qualidades estatísticas.

As questões conceituais estão relacionadas com o conjunto de variáveis selecionadas e a amostra escolhida. Afinal, existe uma estrutura subjacente a esse conjunto de dados. Devemos garantir que a amostra é homogênea com relação à estrutura fatorial.

Com relação às questões estatísticas, devemos verificar os desvios da normalidade, a homocedasticidade e a linearidade, pois eles diminuem as correlações observadas. No entanto, temos que garantir que as variáveis são suficientemente correlacionadas umas com as outras para produzir fatores representativos.

Uma das questões estatísticas é a realização dos testes das suposições da análise fatorial. O teste das suposições da análise fatorial deve seguir os seguintes critérios:

- Uma forte fundamentação conceitual é necessária para embasar a suposição de que existe uma estrutura antes que a análise fatorial seja realizada.
- Um teste de esfericidade de Bartlett estatisticamente significante (sign. < 0,05) indica que correlações suficientes existem entre as variáveis para se continuar a análise.
- Medidas de valores de adequação da amostra (MSA) devem exceder 0,50 tanto para o teste geral quanto para cada variável individual; variáveis com valores inferiores a 0,50 devem ser omitidas da análise fatorial uma por vez, sendo aquela com menor valor eliminada a cada vez.

Estágio 4

No estágio 4, de determinação de fatores e avaliação do ajuste geral, uma vez que as variáveis sejam especificadas e a matriz de correlação seja preparada, estaremos prontos para aplicar a análise fatorial para identificar a estrutura latente de relações. Nisso, as decisões devem ser tomadas com relação:

a) ao método de extração dos fatores (análise de fatores comuns versus análise de componentes)

b) ao número de fatores selecionados para explicar a estrutura latente dos dados

Estágio 5

No estágio 5, interpretação dos fatores, deve-se verificar a avaliação de cargas fatoriais que, apesar de cargas fatoriais de ± 0,30 a ± 0,40 serem minimamente aceitáveis, valores maiores do que ± 0,50 são geralmente considerados necessários para significância prática.

Também deve ser considerada significante uma carga menor com uma amostra maior ou um número maior de variáveis sob análise e uma carga maior faz-se necessária com uma solução fatorial com um número maior de fatores, especialmente na avaliação de cargas em fatores posteriores.

Os testes estatísticos de significância para cargas fatoriais são geralmente conservadores e devem ser considerados apenas como pontos de partida necessários para inclusão de uma variável para futura consideração.

Estágio 6

No estágio 6, a validação da análise fatorial, devemos envolver a avaliação do grau de generalidade dos resultados para a população e da influência potencial de casos ou respondentes individuais sobre os resultados gerais.

Estágio 7

Por último, no estágio 7, usos adicionais dos resultados da análise fatorial, dependendo dos objetivos da aplicação da análise fatorial, podemos parar com a interpretação fatorial ou utilizar um dos métodos para redução de dados.

Processo de decisão em análise fatorial

Assista ao vídeo e conheça o processo de decisão em análise fatorial.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Sendo assim, a análise fatorial tem como objetivo descrever a variabilidade original da variável aleatória X em termos de menos variáveis aleatórias, chamadas de fatores comuns, e que se relacionam com o vetor original X através de um modelo linear.

> Ela é uma técnica de sumarização e redução de dados que não considera variáveis dependentes ou independentes, mas sim a interdependência entre todas as variáveis que são consideradas simultaneamente.

Seja X um vetor contendo p variáveis aleatórias, $X = [\ X_1 \quad amp; X_2 \quad amp; X_3\ ]', \mu$ é a média e σ a variância dessas variáveis, o modelo de fatores pode ser apresentado da seguinte forma:

$$\begin{aligned}
\frac{X_1 - \mu_1}{\sigma_1} &= l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1 \\
\frac{X_2 - \mu_2}{\sigma_2} &= l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2 \\
&\ldots \\
\frac{X_p - \mu_p}{\sigma_p} &= l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \varepsilon_p
\end{aligned}$$

Podemos transformar a expressão acima em um modelo de fatores ortogonais com m fatores comuns:

$$\begin{aligned}
Z_1 &= l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1 \\
Z_2 &= l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2 \\
&\ldots \\
Z_p &= l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \varepsilon_p
\end{aligned}$$

Os vetores F e ε satisfazem as seguintes condições:

$Fe\varepsilon$ são independentes

$E(F) = 0$ e $\mathrm{Cov}(F) = I, I =$ identidade

$E(\varepsilon) = 0$ e $\mathrm{Cov}(\varepsilon) = \psi$, onde ψ é uma matriz diagonal

A suposição para o modelo de fatores ortogonais é:

$$\mathrm{Cov}(Z) = LL' + \psi$$

$$\widehat{L}_{pxm} = \left(\sqrt{\widehat{\lambda}_1\hat{e}_1}\sqrt{\widehat{\lambda}_2\hat{e}_2}\cdots\sqrt{\widehat{\lambda}_m\hat{e}_m}\right)$$

Sendo (λ_i) os autovalores e (e_i) os autovetores normalizados

O grande objetivo da análise fatorial é encontrar as matrizes L (comunalidades) e Ψ (variância) que possam representar a matriz $\mathrm{Cov}(Z)$ para m < p.

Como exemplo prático, utilizando o software R para o cálculo da análise fatorial, carregamos no R os retornos semanais de cinco ações negociadas em bolsa de valores.

Para encontrarmos a matriz de covariância (S) e a correlação (R) utilizamos os seguintes códigos:

```
plain-text
> # Matriz de Covariâncias
> S<-cov(X)
> S

> # Matriz de Correlações
> R<-cor(X)
> R
```

Para encontrarmos as matrizes L e Ψ vamos obter os autovalores e autovetores da matriz de covariâncias:

```
plain-text
> lambda<-eigen(S)$values
> lambda

[1] 0.0035953 0.0007921 0.0007364 0.0005086 0.0003437

> evec<-eigen(S)$vectors

> evec
                              [,1]                    [,2]                    [,
3]                    [,4]                    [,5]
[1,]           -0.5605914
0.73884565             -0.1260222              -0.28373183             -0.20846832
[2,]           -0.4698673              -0.09286987             -0.4675066
0.68793190             0.28069055
[3,]           -0.5473322              -0.65401929             -0.1140581              -0
.50045312              -0.09603973
[4,]           -0.2908932              -0.11267353             0.6099196
0.43808002             -0.58203935
[5,]           -0.2842017              0.07103332
0.6168831              -0.06227778             0.72784638
```

Para obtermos a matriz de cargas fatoriais L para o modelo, com o mesmo número de fatores igual ao número de variáveis, teremos:

```
plain-text

> L=sqrt(lambda)*t(evec)

> L<-t(L)

> LLT<-L%*%t(L)

> LLT

                      [,1]                 [,2]                 [,
3]                    [,4]                 [,5]
[1,]    0.0016299269    0.0008166676    0.0008100713    0.0004422405    0.0005139715
[2,]    0.0008166676    0.0012293759    0.0008276330    0.0003868550    0.0003109431
[3,]    0.0008100713    0.0008276330    0.0015560763    0.0004872816    0.0004624767
[4,]    0.0004422405    0.0003868550    0.0004872816    0.0008023323    0.0004084734
[5,]    0.0005139715    0.0003109431    0.0004624767    0.0004084734    0.0007587370
```

Para encontrarmos a proporção da variabilidade explicada por cada componente, fazemos:

```
plain-text

> lambda/sum(lambda)


[1] 0.60159252 0.13255027 0.12322412 0.08511218 0.05752091
```

Ou seja, o primeiro componente explica 60,159% e o segundo 13,255% da variabilidade do modelo; poderíamos utilizar apenas os dois primeiros no modelo.

Verificando o aprendizado

Questão 1

A análise fatorial é uma técnica que nos ajuda a analisar os padrões de relações multidimensionais encontradas. Defina análise fatorial e descreva para que essa técnica empregada.

Chave de resposta

Análise fatorial é uma técnica de interdependência que tem como objetivo definir a estrutura inerente entre as variáveis na análise. Ela analisa a estrutura das inter-relações (ou correlações) em um grande número de variáveis formando conjuntos de variáveis que são fortemente inter-relacionadas, que chamaremos de fatores. Essa análise é empregada para examinar os padrões ou relações latentes para um grande número de informações e determinar se as variáveis podem ser condensadas ou resumidas a um conjunto menor de fatores ou componentes.

Questão 2

Descreva o estágio 2, que é o planejamento de uma análise fatorial. Quais as três decisões básicas a se tomar nessa etapa?

Chave de resposta

No estágio 2, o planejamento de uma análise fatorial, é importante verificar se uma perspectiva exploratória ou confirmatória é assumida. Confiamos na técnica para fornecer uma visão sobre a estrutura dos dados, mas a estrutura revelada na análise depende de decisões em tópicos como variáveis incluídas, tamanho da amostra e assim por diante.

Desse modo, envolve três decisões básicas: Cálculo dos dados de entrada (uma matriz de correlação) para atender aos objetivos especificados de agrupamento de variáveis ou respondentes; planejamento do estudo em termos de número de variáveis, propriedades de medida das variáveis e tipos de variáveis admissíveis; e o tamanho necessário para a amostra em termos absolutos e como função do número de variáveis na análise.

Questão 3

Quais são as condições estatísticas que os vetores F (fatores) e ε (erros) devem satisfazer?

Chave de resposta

Os vetores F e ε satisfazem as seguintes condições:

Fe ε são independentes

$E(F) = 0$ e $\mathrm{Cov}(F) = 1, l =$ identidade

$E(\varepsilon) = 0$ e $\mathrm{Cov}(\varepsilon) = \psi$, onde ψ é uma matriz diagonal

A suposição para o modelo de fatores ortogonais é:

$$\mathrm{Cov}(Z) = LL' + \psi$$

Considerações finais

O que você aprendeu neste conteúdo?

- A importância da análise multivariada na interpretação de dados com múltiplas variáveis.
- A utilização de técnicas gráficas e estatísticas para diagnóstico e preparação dos dados.
- A relevância do tratamento de dados perdidos e das suposições estatísticas na qualidade das análises.
- A aplicação da análise fatorial para reduzir a complexidade dos dados.
- A identificação de padrões latentes e estrutura interna dos dados por meio de fatores.
- A integração entre técnicas exploratórias e confirmatórias no contexto multivariado.
- A contribuição da análise multivariada para a tomada de decisões em áreas como negócios, saúde e ciências sociais.

Expore +

- Leia o Manual de Análise Multivariada com exemplos práticos de aplicação
- http://www.portalaction.com.br/manual-analise-multivariada
- No site da UFMG existe o manual de introdução ao R com exemplos práticos de aplicação.

Referências

HAIR JR., J.F. *et al.* **Análise Multivariada de Dados**, 6. ed., Porto Alegre, Bookman, 2009.

MINGOTI, S. **Análise de dados através de métodos de estatística multivariada**. Belo Horizonte: Editora UFMG, 2013.