ANÁLISE MULTIVARIADA

DE DADOS COM R

Breno Cauã Rodrigues da Silva

Índice

Prefácio

Em um mundo cada vez mais impulsionado por dados, a capacidade de extrair *insights* significativos de conjuntos de dados complexos tornou-se uma habilidade indispensável. A análise multivariada, com sua gama de técnicas para explorar relações entre múltiplas variáveis, figura como uma ferramenta poderosa nesse cenário.

Este livro, "**Análise Multivariada de Dados com R**", foi concebido com o propósito de desmistificar e tornar acessíveis os conceitos e aplicações da análise multivariada, utilizando a linguagem R como nossa principal ferramenta. Acreditamos que, ao combinar a teoria com exemplos práticos e implementações em R, podemos capacitar estudantes, pesquisadores e profissionais a aplicar essas técnicas em suas próprias áreas de interesse.

Nosso objetivo foi criar um recurso que não apenas introduza os fundamentos da análise multivariada, mas que também sirva como um guia prático para a resolução de problemas reais. Esperamos que este material seja um valioso companheiro em sua jornada de exploração de dados, abrindo novas perspectivas e *insights*.

Este é um *Quarto book*. Para saber mais sobre *Quarto books*, visite https://quarto.org/docs/books/.

Part I

MULTIVARIADA I

Part II

MULTIVARIADA II

1 Unidade I – Introdução

A *análise multivariada* de dados refere-se a um conjunto de técnicas estatísticas que possibilita a análise simultânea de múltiplas medidas para indivíduos, objetos ou fenômenos diversos observados. De maneira geral, a análise multivariada é usada para:

1. **Redução de dados ou simplificação estrutural:** explora a correlação entre as variáveis originais para construir índices ou outro conjunto de variáveis que sintetizam as variáveis originais, sem perder a informação.
    - **Ex:** Análise Fatorial.
2. **Classificação e discriminação:** agrupa indivíduos ou objetos ou variáveis similares de acordo com as suas características, pode ser utilizada para dados com variável resposta (dados supervisionados), ex: análise discriminante, ou para dados sem variável resposta (dados não supervisionados).
    - **Ex:** Análise de Agrupamentos.
3. **Analisar a relação entre as variáveis:** avalia a relação de dependência entre uma variável e um conjunto de outras variáveis, ou a dependência mútua entre grupos de variáveis.
    - **Ex:** Modelos de Regressão ou Equações Estruturais.

Em análise multivariada o interesse da análise pode ser na estrutura das variáveis ou dos indivíduos.

1.1 Leitura de Dados Multivariados

A matriz de dados

$$\mathbf{X_{n\times p}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix},$$

onde x_{jk} representa o j-ésimo "indivíduo" para a k-ésima variável. Essa matriz pode ser simplificada por meio de um vetor aleatório, cujos elementos são as variáveis aleatórias: $\mathbf{X}^\top = \begin{bmatrix} \mathbf{X_1} & \mathbf{X_2} & \cdots & \mathbf{X_p} \end{bmatrix}$. Quando se tem um vetor aleatório, cada variável pode ser analisada separadamente (análise univariada), mas sempre é importante analisá-lo conjuntamente para avaliar as interrelações entre as variáveis. As variáveis podem ser quantitativas ou qualitativas.

1.2 Pré-processamento de Dados Multivariados

1.2.1 Familiarização com os Dados

Na primeira parte do pré-processamento é feita a análise do banco de dados original, para conhecer as variáveis, selecionar as variáveis necessárias e identificar eventuais problemas. Incluem os seguintes passos:

1. Identificar as variáveis necessárias ao estudo.
2. Verificar se o banco já contém todas as variáveis de estudo ou se precisam ser buscados outros bancos de dados para agregar informações. Verificar se o banco possui variável ou variáveis de identificação, que permitirá fazer a ligação com outros bancos de dados, caso seja necessário.
3. Definir e classificar todas as variáveis no banco de dados.
4. Realizar a análise univariada por tipo de variáveis: quantitativas e qualitativas.
5. Identificar variáveis com valores faltantes (missing values) em excesso, verificar possibilidade de imputação de dados.
6. Identificar variáveis com valores aberrantes (outliers), ou com dados inconsistentes.
7. Verificar se alguma variável tem características especiais. Por exemplo, uma variável quantitativa com excesso de zero, variáveis qualitativas com grande concentração em uma das categorias.

1.2.2 Limpeza de Dados

Se for verificada a necessidade de fazer alterações no banco de dados, isto deve ser feito antes que qualquer análise estatística mais aprofundada possa ser executada. Para lidar com esta situação, deve ser feita a limpeza de dados. Então aqui deve-se filtrar informações irrelevantes, questionar a fonte primária sobre inconsistências, confirmar a veracidade de valores aberrantes, verificar a necessidade de fazer a imputação de dados faltantes (Depende do tipo de dado e do percentual de dados faltantes em relação ao total).

1.2.3 Inserção de Dados/Variáveis

Dependendo do objetivo da análise pode ser necessária a inserção de dados de outras fontes ou bancos de dados. Nesse caso as variáveis que identificam cada caso de maneira única (variáveis de identificação, variáveis chave) são essenciais para que a ligação entre os bancos possa ser feita.

1.2.4 Transformação de Dados

Caso seja necessário, pode-se transformar os dados originais em formatos mais apropriados:

1. **Normalização/Padronização:** é comum em dados multivariados termos informação no banco de dados de natureza distinta, então é prática comum em muitas áreas fazer a normalização ou a padronização de todo o conjunto de variáveis. As técnicas de normalização e padronização têm o mesmo objetivo: transformar todas as variáveis para a mesma ordem de grandeza. E a diferença básica entre as duas técnicas é que padronizar as variáveis irá resultar em uma média igual a 0 e

um desvio padrão igual a 1. Já normalizar tem como objetivo colocar as variáveis dentro do intervalo de 0 e 1, e caso tenha resultado negativo -1 e 1.

2. **Criação de novas variáveis:** tais como variáveis indicadoras (dummies), ou uma categoria agregada, ou uma variável quantitativa criada a partir de outras variáveis existentes no banco de dados.

3. **Discretização/categorização:** É o processo de transformação de variáveis contínuas em discretas ou categóricas. Algumas técnicas só trabalham com entradas de valores discretos ou categóricos, então a solução é a discretização que cria um número limitado de possíveis estados ou a categorização que transforme de quantitativa para qualitativa.

1.2.5 Resumo de Dados (Análise Exploratória)

A análise exploratória de dados multivariados (AED) possibilita a detecção de erros e inconsistências, determinação do relacionamento entre as variáveis, verificação da similaridade ou dissimilaridade entre as observações (indivíduos ou casos) e indicar as técnicas multivariadas adequadas de acordo com o tipo de variável e objetivo do trabalho. A partir de uma matriz de dados quantitativos podem ser calculadas algumas estatísticas para início da análise:

1. O vetor de médias amostral: $\overline{\mathbf{X}}^\top = \begin{bmatrix}\overline{X}_1 & \overline{X}_2 & ... & \overline{X}_k & ... & \overline{X}_p\end{bmatrix}$, onde $\overline{X}_k = \frac{1}{1}\sum_{j=1}^{n} x_{jk}$.

2. A matriz de variância-covariância amostral: $\mathbf{S_{p\times p}} = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{x_j} - \overline{\mathbf{X}})(\mathbf{x_j} - \overline{\mathbf{X}})^\top = \begin{bmatrix} s_{11} & s_{12} & ... & s_{1p} \\ s_{21} & s_{22} & ... & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & ... & s_{pp}\end{bmatrix}$.

3. Matriz de correlação amostral: $\mathbf{R_{p\times p}} = \begin{bmatrix} 1 & r_{12} & ... & r_{1p} \\ r_{21} & 1 & ... & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & ... & 1\end{bmatrix}$, em que $r_{ik} = \frac{r_{ik}}{\sqrt{r_{ii}r_{kk}}}$.
   Lembrando que a matriz $\mathbf{R}$ capta apenas o relacionamento linear entre as variáveis. Relações não lineares geram covariância e correlação nulas.

4. Matriz de distâncias: $\mathbf{D_{n\times n}} = \begin{bmatrix} d_{11} & d_{12} & ... & d_{1n} \\ d_{21} & d_{22} & ... & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & ... & d_{nn}\end{bmatrix}$, onde $d_{11} = d_{22} = ... = d_{nn} = 0$. As distâncias entre os pares de indivíduos e são calculadas de diversas maneiras, de acordo com o tipo de variável. Para variáveis quantitativas, são usualmente utilizadas:

   - **Distância de Minkowski:** $d(x_i, x_j) = \left[\sum_{k=1}^{p}(x_{ik} - x_{jk})^n\right]^{1/n}$, que é a distância mais geral, tendo como casos particulares a distância Euclidiana, quando $n = 2$ e a distância de Manhattan quando $n = 1$.

- **Distância Euclidiana:** $d(x_i, x_j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$
- **Distância de Mahalanobis:** $d(x_i, x_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_j)$

Essas medidas podem ser apresentadas em tabelas agregando as diversas variáveis, ou em gráficos de duas ou três dimensões que possam representar o inter-relacionamento entre as variáveis como correlogramas, diagramas de dispersão (matrix plot), faces de Chernoff, etc.

Se os dados são qualitativos podem ser criadas tabelas de contingência, testes de hipóteses e análises específicas para esse tipo de variável como a análise de correspondência.

Se o banco de dados contém variáveis quantitativas e qualitativas e o objetivo do trabalho for identificar diferenças entre grupos representados pelas variáveis qualitativas, a representação deve ser desde o início comparativa entre os grupos, com estatísticas descritivas e testes de hipóteses tendo o cuidado de ressaltar o atendimento aos pressupostos dos testes. Os gráficos também devem seguir o padrão comparativo como os box-plots múltiplos.

Toda a AED multivariada deve ser feita com a finalidade de respaldar os objetivos do trabalho portanto deve-se evitar o excesso de tabelas e gráficos que não agregam informação ao estudo.

References