

**MBA
USP
ESALQ**

**UNSUPERVISED
MACHINE LEARNING:
Análise de
Correspondência
Simples e Múltipla**

Prof. Dr. Wilson Tarantin Junior

*A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados é do professor.

Proibida a reprodução, total ou parcial, sem autorização. Lei nº 9610/98

Contextualização

- Quando aplicar a análise de correspondência?
 - Técnica adequada para a análise de variáveis categóricas (qualitativas)
 - O objetivo é verificar se existe associação estatisticamente significativa entre as variáveis e suas categorias, criando o mapa perceptual para visualizar as associações
 - Caso exista uma variável quantitativa, é necessário que ela passe por um processo de categorização previamente
 - Exemplo: a idade é uma variável quantitativa (25, 42, 73, 81 anos) e poderia ser categorizada como: 0-30 anos é a categoria 1, 31-60 anos é categoria 2, 61-90 anos é categoria 3...

Contextualização

- Trata-se de técnica exploratória (não supervisionada)
 - Para avaliar a relação conjunta entre as variáveis (interdependência)
 - Não há modelos do tipo “ $y_i = x_{1i} + x_{2i} + \dots + u_i$ ”
 - Não são adequadas para fins de inferência
 - Se novas observações forem adicionadas ao banco de dados, é adequado refazer a análise

Contextualização

- Exemplos de aplicação
 - Faixa de renda e status na aprovação de crédito
 - Nível de escolaridade e cargo ocupado em empresas
 - Tipo de solo e cultura implementada
 - Gravidade dos sintomas da doença e comorbidades
 - Outros...

Contextualização

- Análise de variáveis geradas por escala Likert
 - Exemplos: concordo plenamente; concordo parcialmente; não concordo, nem discordo; discordo parcialmente; discordo plenamente
 - Evita o problema da ponderação arbitrária
 - Cada ponto da escala Likert torna-se uma categoria da variável na análise de correspondência simples ou múltipla

Implementação

Análise de Correspondência Simples

Contextualização

- Também conhecida como Anacor
 - Quando o objetivo é estudar a associação entre duas variáveis e suas categorias
 - É possível separar a Anacor em duas partes:
 - 1. Análise da significância estatística da associação entre as variáveis e suas categorias por meio do teste qui-quadrado (χ^2)**
 - 2. Elaboração e interpretação do mapa perceptual**

Teste qui-quadrado para associação

Análise de Correspondência Simples

1. Tabela de contingência

- Contém as frequências absolutas observadas para cada par de categorias das variáveis: trata-se de uma tabela de classificação cruzada (*cross-tabulation*)

		Variável B					
		Categoria 1	Categoria 2	Categoria 3	...	Categoria J	Total
Variável A	Categoria 1	n_{11}	n_{12}	n_{13}	...	n_{1J}	Σ_{L1}
	Categoria 2	n_{21}	n_{22}	n_{23}	...	n_{2J}	Σ_{L2}
	Categoria 3	n_{31}	n_{32}	n_{33}	...	n_{3J}	Σ_{L3}

	Categoria I	n_{I1}	n_{I2}	n_{I3}	...	n_{IJ}	Σ_{LI}
	Total	Σ_{C1}	Σ_{C2}	Σ_{C3}	...	Σ_{CJ}	N

Análise de Correspondência Simples

2. Tabela de frequências absolutas esperadas

- Para a célula referente às categorias 1 das duas variáveis, a frequência absoluta esperada é:

$$\text{Freq. absoluta esperada}_{11} = \frac{(\Sigma_{L1} \times \Sigma_{C1})}{N}$$

- Este mesmo cálculo deve ser realizado para cada par de categorias das variáveis, alterando-se apenas o numerador

Análise de Correspondência Simples

3. Tabela de resíduos

- Para a célula referente às categorias 1 das duas variáveis, o valor do resíduo é:

$$\text{Resíduo}_{11} = n_{11} - \frac{(\Sigma_{L1} \times \Sigma_{C1})}{N}$$

- Ou seja, resíduo = frequência absoluta observada – frequência absoluta esperada
- O mesmo cálculo é realizado para cada par de categorias

Análise de Correspondência Simples

4. Tabela com os valores χ^2

- Para a célula referente às categorias 1 das duas variáveis, o valor da estatística χ^2 é:

$$\chi^2_{11} = \frac{(\text{resíduo}_{11})^2}{(\text{freq. absoluta esperada}_{11})}$$

- O mesmo cálculo é realizado para cada par de categorias
- Em seguida, os valores de todas as células são somados para obter o valor χ^2 total

Análise de Correspondência Simples

4. Tabela com os valores χ^2

- O objetivo é verificar se há associação estatisticamente significativa entre as variáveis (utilizando a soma do χ^2)
 - H_0 : as variáveis se associam de forma aleatória.
 - H_1 : a associação entre as variáveis não se dá de forma aleatória.
- Dados o nível de significância e os graus de liberdade, se o valor da estatística χ^2 for maior do que seu valor crítico, há associação significativa entre as duas variáveis (H_1)
 - Graus de liberdade = $(I - 1) \times (J - 1)$

Análise de Correspondência Simples

Tabelas de resíduos padronizados e resíduos padronizados ajustados

- Enquanto a análise do χ^2 permite verificar se há ou não a dependência entre as duas variáveis, a análise de resíduos padronizados ajustados permite aprofundar a análise com foco nas categorias das variáveis
- Como as categorias de uma variável se relacionam com as categorias da outra variável?
 - Observa-se o excesso ou falta de ocorrências de casos nas categorias das duas variáveis

Análise de Correspondência Simples

5. Tabela de resíduos padronizados

- Para a célula referente às categorias 1 das duas variáveis, o valor do resíduo padronizado é:

$$\text{Resíduo Padronizado}_{11} = \frac{\text{resíduo}_{11}}{\sqrt{\text{freq. absoluta esperada}_{11}}}$$

- O mesmo cálculo é realizado para cada par de categorias

Análise de Correspondência Simples

6. Tabela de resíduos padronizados ajustados

- Para a célula referente às categorias 1 das duas variáveis, o valor do resíduo padronizado ajustado é:

$$\text{Resíduo Padronizado Ajustado}_{11} = \frac{\text{resíduo padronizado}_{11}}{\sqrt{\left[\left(1 - \frac{\Sigma_{C1}}{N} \right) \times \left(1 - \frac{\Sigma_{L1}}{N} \right) \right]}}$$

- O mesmo cálculo é realizado para cada par de categorias

Análise de Correspondência Simples

6. Tabela de resíduos padronizados ajustados

- Se o valor do resíduo padronizado ajustado em certa célula for maior do que 1,96, interpreta-se que existe associação significativa, ao nível de significância de 5%, entre as duas categorias que interagem na célula; se for menor do que 1,96, não há associação estatisticamente significativa
- 1,96 é o valor crítico da normal padrão para o nível de significância de 5%

Elaboração do mapa perceptual

Análise de Correspondência Simples

1. Determinar os autovalores (λ^2)

- A quantidade (m) de autovalores depende da quantidade de categorias nas variáveis: **$m = \min(I - 1, J - 1)$**
- Na Anacor, os autovalores referem-se às inércias principais parciais e são base para determinar a inércia principal total e o percentual da inércia principal total em cada dimensão do mapa perceptual

Análise de Correspondência Simples

1. Determinar os autovalores (λ^2)

- Como base para o cálculo dos autovalores, inicialmente, define-se uma matriz A
- Um modo de obter a matriz A, baseando-se nas etapas anteriores, é fazer para cada célula da matriz de resíduos padronizados o seguinte cálculo:

$$\text{Matriz } A_{11} = \frac{\text{resíduo padronizado}_{11}}{\sqrt{N}}$$

- Com base na matriz A, obtém-se a matriz **W**: $\mathbf{W} = \mathbf{A}' \cdot \mathbf{A}$

Análise de Correspondência Simples

1. Determinar os autovalores (λ^2)

- Identificando W , os autovalores são obtidos pela solução da seguinte expressão:
 $\det(\lambda^2 \cdot I - W) = 0$

$$\begin{vmatrix} \lambda^2 - w_{11} & -w_{12} & -w_{13} \\ -w_{21} & \lambda^2 - w_{22} & -w_{23} \\ -w_{31} & -w_{32} & \lambda^2 - w_{33} \end{vmatrix} = 0$$

- I é a matriz identidade

Análise de Correspondência Simples

1. Determinar os autovalores (λ^2)

- Com base nos autovalores (λ^2), encontra-se o percentual da inércia principal total de cada dimensão

$$\% \text{ da Inércia Principal Total} = \frac{\lambda^2_{\text{dimensão}}}{\lambda^2_{\text{total}}} \quad (\text{para cada dimensão})$$

- Quanto maior a inércia principal total (e o χ^2), mais forte será a associação entre as variáveis em análise

$$\text{Inércia Principal Total} = \frac{\chi^2}{N}$$

Análise de Correspondência Simples

2. Determinar as massas das variáveis em linha e em coluna

- As massas representam a influência que cada categoria exerce sobre as demais categorias de sua variável
- Com base nos “totais” da tabela de contingência, para a categoria 1 das variáveis, obtém-se as massas:

$$\text{Massa Linha}_{\text{Categ1}} = \frac{\Sigma_{L1}}{N} \quad \text{Massa Coluna}_{\text{Categ1}} = \frac{\Sigma_{C1}}{N}$$

- O mesmo cálculo é realizado para as demais categorias das variáveis

Análise de Correspondência Simples

3. Determinar os autovetores

- É possível encontrar os autovetores a partir dos autovalores (λ^2) calculados
 - Cada autovalor gera seus autovetores
- Para cada um dos m autovalores:
- Substitui o autovalor na matriz $(\lambda^2 \cdot \mathbf{I} - \mathbf{W})$, multiplica por um vetor coluna e iguala a zero. Resolvendo o sistema de equações gerado, é possível encontrar o autovetor da coluna (V). Com base nele, encontra-se o autovetor da linha (U):

$$u_k = \text{Matriz } A \cdot \begin{pmatrix} v_k \\ \lambda_k \end{pmatrix}$$

Análise de Correspondência Simples

4. Obter as coordenadas das categorias: **variável em linha na tabela**

- Coordenadas das abscissas (X)

$$X_I = \left(\lambda_1 \cdot \frac{1}{\sqrt{D_1}} \right) \cdot u_1$$

- Coordenadas das ordenadas (Y)

$$Y_I = \left(\lambda_2 \cdot \frac{1}{\sqrt{D_1}} \right) \cdot u_2$$

- Coordenadas da K-ésima dimensão (k = quantidade de autovalores)

$$Z_I = \left(\lambda_k \cdot \frac{1}{\sqrt{D_1}} \right) \cdot u_k$$

Análise de Correspondência Simples

4. Obter as coordenadas das categorias: **variável em coluna na tabela**

- Coordenadas das abscissas (X)

$$X_c = \left(\lambda_1 \cdot \frac{1}{\sqrt{D_c}} \right) \cdot v_1$$

- Coordenadas das ordenadas (Y)

$$Y_c = \left(\lambda_2 \cdot \frac{1}{\sqrt{D_c}} \right) \cdot v_2$$

- Coordenadas da K-ésima dimensão (k = quantidade de autovalores)

$$Z_c = \left(\lambda_k \cdot \frac{1}{\sqrt{D_c}} \right) \cdot v_k$$

Implementação

Análise de Correspondência Múltipla

Análise de Correspondência Múltipla

- A Análise de Correspondência Múltipla (ACM) tem o objetivo de analisar a associação entre mais de duas variáveis categóricas
 - Só participam da ACM as variáveis que apresentam associação estatisticamente significativa com pelo menos uma outra variável contida na análise
 - Antes de elaborar a ACM, é importante realizar um teste χ^2 para cada par de variáveis
 - Se alguma delas não apresentar associação com outras, então não é incluída na ACM
 - A ACM apresenta a mesma lógica de análise definida na Anacor

Análise de Correspondência Múltipla

- 1º método: **matriz binária Z** → **coordenadas-padrão**

- A matriz binária é obtida pela transformação das variáveis qualitativas em variáveis binárias, ou seja, valores 0 ou 1
- Supondo que a **matriz binária Z** seja a **tabela de contingência da Anacor**, é possível obter a inércia principal parcial das dimensões, autovalores, autovetores e, portanto, as coordenadas no mapa perceptual dessa matriz
 - Quantidade de dimensões (λ^2) = J – Q**, em que “J” é a quantidade total de categorias em todas as variáveis e “Q” a quantidade de variáveis

$$\text{Inércia Principal Total} = \frac{J - Q}{Q}$$

Análise de Correspondência Múltipla

- Matriz binária Z: exemplo com $n = 7$

ID	Variável A		Variável B			Variável C			
	Categ. 1	Categ. 2	Categ. 1	Categ. 2	Categ. 3	Categ. 1	Categ. 2	Categ. 3	Categ. 4
1	1	0	0	1	0	1	0	0	0
2	0	1	0	0	1	0	1	0	0
3	0	1	1	0	0	0	0	1	0
4	1	0	0	0	1	0	0	0	1
5	0	1	0	1	0	0	1	0	0
6	1	0	0	1	0	0	0	1	0
7	1	0	1	0	0	1	0	0	0

Análise de Correspondência Múltipla

- 2º método: **matriz de Burt** → **coordenadas principais**
 - A matriz de Burt é definida como: $\mathbf{B} = \mathbf{Z}' \cdot \mathbf{Z}$
 - É possível combinar em uma única matriz o cruzamento de todos os pares variáveis e suas categorias, obtendo, desta forma, uma matriz que contém as frequências absolutas observadas para todos os cruzamentos
 - Ao considerar a **matriz de Burt como uma tabela de contingência**, é possível realizar uma Anacor e obter as coordenadas das categorias das variáveis

Análise de Correspondência Múltipla

- **Matriz de Burt:** obtida com base na matriz Z do exemplo anterior

		Variável A		Variável B			Variável C			
		Categ. A1	Categ. A2	Categ. B1	Categ. B2	Categ. B3	Categ. C1	Categ. C2	Categ. C3	Categ. C4
Variável A	Categ. A1	4	0	1	2	1	2	0	1	1
	Categ. A2	0	3	1	1	1	0	2	1	0
Variável B	Categ. B1	1	1	2	0	0	1	0	1	0
	Categ. B2	2	1	0	3	0	1	1	1	0
	Categ. B3	1	1	0	0	2	0	1	0	1
Variável C	Categ. C1	2	0	1	1	0	2	0	0	0
	Categ. C2	0	2	0	1	1	0	2	0	0
	Categ. C3	1	1	1	1	0	0	0	2	0
	Categ. C4	1	0	0	0	1	0	0	0	1

Referência

Fávero, Luiz Paulo; Belfiore, Patrícia. (2024). Manual de análise de dados: estatística e machine learning com Excel®, SPSS®, Stata®, R® e Python®. 2 ed. Rio de Janeiro: LTC.

OBRIGADO!

[linkedin.com/in/wilson-tarantin-junior-359476190](https://www.linkedin.com/in/wilson-tarantin-junior-359476190)