Análise de Dados

Mestrado em Métodos Quantitativos para a Decisão Económica e Empresarial

EN
2/6/2021
Duração: 2 horas

Nome_____ **N^o**_____

Espço reservado a classificações

A utilização do telemóvel, em qualquer circunstância, é motivo suficiente para a anulação da prova.

Justifique todas as suas respostas.

1. Com o objectivo de construir um índice de qualidade de vida observaram-se as seguintes variáveis em 39 países: rendimento médio anual das famílias em milhares de dólares, `rendimento`, percentagem de emprego, `emprego`, média do número de anos de escolaridade da população, `escolaridade`, esperança média de vida, `EMV` e média de horas semanais dedicadas a actividades pessoais e de lazer, `lazer`. Os valores e vetores próprios da matriz de correlação das observações são os seguintes:

```
## eigen() decomposition
## $values
## [1] 2.3324672 0.9649517 0.7691424 0.4795849 0.4538538
##
## $vectors
##                    PC1         PC2        PC3         PC4        PC5
## rendimento   0.4278701  0.57424363 -0.3221510 -0.19361899 -0.5881342
## emprego      0.5040268  0.25747404  0.2555305  0.74882237  0.2315887
## escolaridade 0.3958341 -0.62699156  0.4016454 -0.02182017 -0.5370305
## EMV          0.5224219  0.07311534  0.2560890 -0.62393080  0.5165832
## lazer        0.3650682 -0.45330685 -0.7771887  0.10959463  0.2126143
```

(a) Qual a proporção da variância explicada pelo primeiro componente principal? (1.0)

(b) Qual a correlação entre a esperança média de vida e o primeiro componente principal? (0.5)

(c) Atendendo ao critério de Kaiser, diga, justificando, quantos componentes principais devem ser retidos para compor o índice da qualidade de vida? (0.5)

(d) Qual o valor do índice da qualidade de vida de um país em que, após estandardização, o rendimento médio das famílias é -0.02026039, a taxa de emprego é 0.3153712, o número médio de anos de escolaridade é -2.648652, a esperança média de vida é 0.06843028 e o número médio de horas por semana dedicadas ao lazer é -8.537931? (1.0)

(e) Como poderia usar a Analise de Componentes Principais para obter um ranking dos 39 países quanto à qualidade de vida. (1.0)

2. Os dados do exercício anterior foram usados para uma Análise Fatorial Exploratória, tendo-se obtido o seguinte *output* através do *software R*:

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = R)
## Overall MSA =  0.65
## MSA for each item =
##   rendimento      emprego escolaridade          EMV        lazer
##         0.58         0.67         0.57         0.70         0.70
##
## Principal Components Analysis
## Call: principal(r = R, nfactors = 2, rotate = "varimax", n.obs = 39)
## Standardized loadings (pattern matrix) based upon correlation matrix
##               PC1   PC2   h2   u2 com
## rendimento   0.86 -0.04 0.75 0.25 1.0
## emprego      0.76  0.27 0.66 0.34 1.3
## escolaridade 0.10  0.86 0.74 0.26 1.0
## EMV          0.67  0.43 0.64 0.36 1.7
## lazer        0.17  0.69 0.51 0.49 1.1
##
##                        PC1  PC2
## SS loadings           1.82 1.48
```

```
## Proportion Var        0.36 0.30
## Cumulative Var        0.36 0.66
## Proportion Explained  0.55 0.45
## Cumulative Proportion 0.55 1.00
##
## Mean item complexity =  1.2
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.14
##  with the empirical chi square  20.1  with prob <  7.3e-06
##
## Fit based upon off diagonal values = 0.83
```

(a) Explique a diferença entre a Análise de Componentes Principais realizada no exercício anterior e a Análise Fatorial Exploratória realizada neste exercício. (1.5)

(b) Diga o que representa a comunalidade de uma variável relativamente a um fator latente? Qual a comunalidade de cada uma das variáveis `escolaridade` e `rendimento` relativamente a cada um dos fatores latentes e que proporção das comunalidades totais essas comunalidades representam? Comente e dê uma interpretação possível para cada um dos fatores latentes. (2.0)

3. Aos mesmos dados, após a Análise Fatorial Exploratória, foi aplicada uma Análise Fatorial Confirmatória que resultou no seguinte *output*:

```
## lavaan 0.6-8 ended normally after 97 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        11
##
##   Number of observations                            39
##
## Model Test User Model:
##
##   Test statistic                                 3.153
##   Degrees of freedom                                 4
##   P-value (Chi-square)                           0.533
##
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   F1 =~
##     rendimento        1.000                               19.676    0.564
##     emprego           0.289    0.103    2.800    0.005    5.681    0.711
##     EMV               0.178    0.063    2.811    0.005    3.493    0.751
##   F2 =~
##     escolaridade      1.000                                0.775    0.626
##     lazer             0.501    0.278    1.798    0.072    0.388    0.509
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   F2 ~
##     F1                0.028    0.013    2.095    0.036    0.706    0.706
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##    .rendimento      830.588  223.216    3.721    0.000  830.588    0.682
##    .emprego          31.492   11.361    2.772    0.006   31.492    0.494
##    .EMV               9.445    3.929    2.404    0.016    9.445    0.436
##    .escolaridade      0.933    0.396    2.358    0.018    0.933    0.608
##    .lazer             0.431    0.129    3.351    0.001    0.431    0.741
##     F1              387.145  234.763    1.649    0.099    1.000    1.000
##    .F2                0.302    0.333    0.907    0.364    0.502    0.502
```

(a) Comente, usando um nível de significância de 5%, a qualidade do modelo dos pontos de vista do ajuste global e da significância estatística dos parâmetros. (1.0)

(b) Qual o loading do fator F_1 na variável `rendimento` dado pela solução completamente estandardizada e o valor da variância específica de `escolaridade`? Trata-se de um modelo ortogonal ou oblíquo? Justifique. (1.5)

(c) Obtenha, justificando, a matriz de variâncias induzida pelo modelo fatorial e confirme o número de graus de liberdade. (2.5)

4. Uma empresa investiu em publicidade na internet. Com o objetivo de prever a eficácia do investimento, a empresa recolheu dados de 800 utilizadores de internet com o intuito de prever se o utilizador clica ou não no anúncio. Assim, observou-se para cada utilizador a idade (`Age`, em anos), o tempo gasto em sites onde o anúncio foi colocado (`Daily_time_on_site`, em minutos) e o tempo diário gasto na internet (`Daily_interne_usage`, em minutos). O resultado de uma análise discriminante encontra-se no *output* seguinte.

```
##  Two Sample t-test
##
## data:  Daily_time_on_site by Clicked_on_add
## t = 32.177, df = 798, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  21.82302 24.65862
## sample estimates:
## mean in group 0 mean in group 1
##        76.85462        53.61380
##
##  Two Sample t-test
##
## data:  Age by Clicked_on_add
## t = -16.228, df = 798, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.661915 -7.576752
## sample estimates:
```

```
## mean in group 0 mean in group 1
##        31.68400        40.30333
##
##  Two Sample t-test
##
## data:  Daily_internet_usage by Clicked_on_add
## t = 37.304, df = 798, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  66.49736 73.88432
## sample estimates:
## mean in group 0 mean in group 1
##        214.5137        144.3229
##
##  One-way MANOVA (Bartlett Chi2)
##
## data:  x
## Wilks' Lambda = 0.1812, Chi2-Value = 1360.6, DF = 3.0, p-value < 2.2e-16
## sample estimates:
##   Daily_time_on_site      Age Daily_internet_usage
## 0           76.85462 31.68400             214.5137
## 1           53.61380 40.30333             144.3229
##
## Call:
## lda(Clicked_on_add ~ Daily_time_on_site + Age + Daily_internet_usage,
##     data = my_data, CV = F)
##
## Prior probabilities of groups:
##     0     1
## 0.625 0.375
##
## Group means:
##   Daily_time_on_site      Age Daily_internet_usage
## 0           76.85462 31.68400             214.5137
## 1           53.61380 40.30333             144.3229
##
## Coefficients of linear discriminants:
##                               LD1
## Daily_time_on_site   -0.07376243
## Age                   0.05461113
## Daily_internet_usage -0.03134917
```

(a) Escreva a equação da função discriminante estimada e diga qual o score de um indivíduo com 30 anos, que passa 180 minutos por dia na internet, dos quais 60 são em sites onde se econtra o anúncio da empresa e que vive numa zona em que o rendimento médio anual é de 30 000 euros. (1.5)

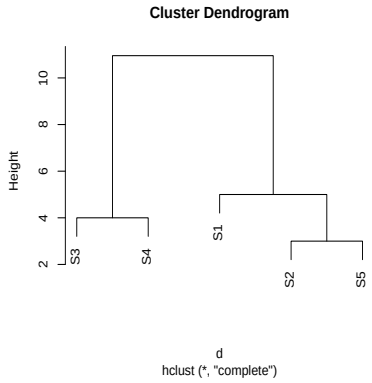
(b) Diga o que significam as `Prior probabilities of groups` e use-as para determinar o valor de corte, sabendo que os valores médios da função discriminante para os indíviduos que clicaram no anúncio e não clicaram no anúncio são, respetivamente, -10.663512 e -6.278077, e que na amostra 500 utilizadores clicaram no anúncio. Com base nos resultados obtidos, classificaria o indivíduo da alínea anterior como alguém que irá clicar no anúncio? (2.0)

(c) O que aconselharia os analistas da empresa a fazer se quisessem incluir na análise uma váriavel que indicasse o género do indíviduo? (1.0)

5. Uma empresa vende determinado produto em 5 cidades e o diretor de marketing está interessado em agrupar as principais 5 cidades onde o produto é consumido com base em características que considera importantes. Primeiro, analisou o quadrado da distância euclidiana entre as observações das 5 cidades:

```
##     S1  S2  S3  S4
## S2  10
## S3 120  88
## S4  68  60  16
## S5  25   9  51  35
```

Em seguida realizou uma Análise de Clusters hierárquica, com base na distância euclidiana, usando o método do vizinho mais afastado ou *complete linkage*, tendo obtido o seguinte dendograma:



(a) Diga, com base no dendograma, o número de passos usado no algorítmo e qual o agrupamento em cada passo. (1.5)

(b) Qual o quadrado a distância euclidiana entre o *cluster* formado pelas cidades $S2, S5$ e a cidade $S1$? (0.75)

(c) Quantos grupos de cidades, e quais as cidades em cada grupo, se devem considerar? Justifique. (0.75)