Métodos Multivariados de Análise de Dados*

1ª Lista de Exercícios

Alberson da Silva Miranda

3 de setembro de 2024

*Código disponível em https://github.com/albersonmiranda/analise_multivariada.

Índice

1 PRIMEIRA QUESTÃO

Primeiramente, verificamos a estrutura dos dados.

```
# estrutura dos dados
str(tabaco)
```

```
'data.frame':   2968 obs. of  7 variables:
 $ fuma       : 'labelled' chr  "0" "0" "0" "0" ...
  ..- attr(*, "label")= chr "0 = não; 1 = sim"
 $ fuma_freq  : 'labelled' num  0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "label")= chr "1 = Não fumo atualmente; 2 = Sim, menos que diariamente; 3 = Sim, diariament
 $ bebe_freq  : 'labelled' num  2 2 1 1 2 0 0 2 2 2 ...
  ..- attr(*, "label")= chr "1 = Não bebo nunca; 2 = Menos de uma vez por mês; 3 = Uma vez ou mais por mês"
 $ comodos    : 'labelled' num  5 4 4 5 5 6 6 5 5 5 ...
  ..- attr(*, "label")= chr " Número de cômodos na casa onde vive"
 $ sexo       : 'labelled' num  0 1 1 0 1 1 1 1 1 1 ...
  ..- attr(*, "label")= chr "0 = mulher; 1 = homem"
 $ idade      : num  31 20 67 64 35 32 63 58 27 38 ...
 $ educacional: 'labelled' chr  "2" "1" "1" "2" ...
  ..- attr(*, "label")= chr "1 = Fundamental; 2 = Médio; 3 = Superior"
```

Como as colunas categóricas estão no formato `labelled`, vamos convertê-las para o padrão `factor`.

```
# convertendo para factor
tabaco <- data.frame(
  lapply(tabaco, function(x) {
    if (inherits(x, "labelled")) as.factor(x) else x
  })
)
```

Agora, vamos verificar a estrutura dos dados após a conversão.

```r
# estrutura dos dados
str(tabaco)
```

```
'data.frame':   2968 obs. of  7 variables:
 $ fuma       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ fuma_freq  : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 1 1 ...
 $ bebe_freq  : Factor w/ 3 levels "0","1","2": 3 3 2 2 3 1 1 3 3 3 ...
 $ comodos    : Factor w/ 21 levels "1","2","3","4",..: 5 4 4 5 5 6 6 5 5 5 ...
 $ sexo       : Factor w/ 2 levels "0","1": 1 2 2 1 2 2 2 2 2 2 ...
 $ idade      : num  31 20 67 64 35 32 63 58 27 38 ...
 $ educacional: Factor w/ 3 levels "1","2","3": 2 1 1 2 2 2 1 1 2 2 ...
```

O próximo passo é verificar as estatísticas básicas:

```r
# estatísticas básicas
summary(tabaco)
```

```
 fuma     fuma_freq bebe_freq    comodos    sexo         idade
 0:1000   0:1000    0:1182    5      :824   0:1257   Min.   :16.00
 1:1968   1: 968    1: 475    6      :592   1:1711   1st Qu.:32.00
          2:1000    2:1311    4      :481            Median :44.00
                              7      :337            Mean   :44.77
                              3      :235            3rd Qu.:56.00
                              8      :193            Max.   :91.00
                              (Other):306
 educacional
 1:1521
 2:1010
 3: 437
```

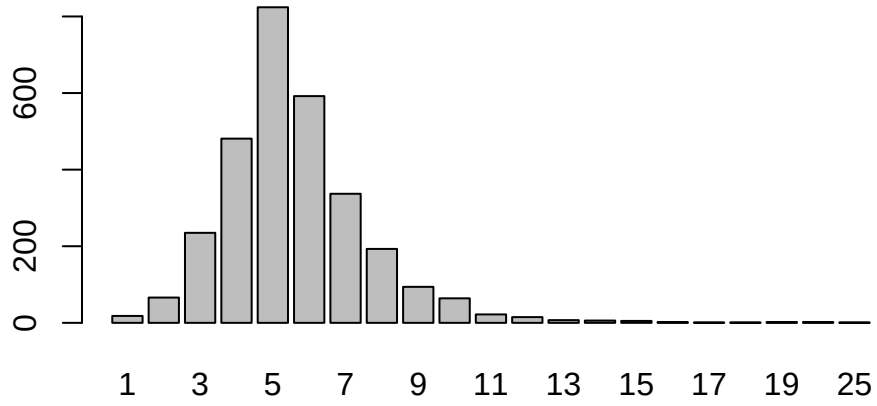
Podemos notar que a variável dependente `fuma` é binária, com 0 representando não fumante e 1 fumante. Todas as demais são categóricas, com exceção da idade. `fuma` também está desbalanceada, com 66,3% dos indivíduos fumantes. A variável `fuma_freq` é a única contínua, com média de 1,5 e desvio padrão de 1,2. A variável `fuma_freq` será utilizada como variável resposta para a questão 2.

```r
# proporção de fumantes
prop.table(table(tabaco$fuma))
```

```
        0         1
0.3369272 0.6630728
```

A variável `comodos` é assimetrica à esquerda, com uma estranha cauda longa à direita, implicando na presença de alguns bilionários detentores de residências de incríveis 25 cômodos. Será considerada um modelo com *binning* desse fator.

```r
# histograma de comodos
plot(tabaco$comodos)
```




```r
# agrupando maiores que 10 em um único grupo
tabaco$comodos_bin <- cut(
  as.numeric(tabaco$comodos),
  breaks = c(1:9, 25),
  labels = c(2:9, "⩾ 10"),
  include.lowest = TRUE
)
```

Agora, vamos verificar a correlação entre as variáveis. Fora da diagonal principal, os únicos índices de correlação extremos são entre `fuma` e `fuma_freq`, o que é esperado, pois a frequência de fumar é um indicador de fumante. `fuma_freq` será reservada como variável resposta para a questão 2.

```r
# correlação entre as variáveis
cor(sapply(tabaco, as.numeric), method = "spearman")
```

```
                  fuma    fuma_freq   bebe_freq     comodos        sexo
fuma         1.00000000  0.868369185 -0.04348923  0.05452729 -0.06566551
fuma_freq    0.86836918  1.000000000 -0.19655564  0.09828441 -0.12376863
bebe_freq   -0.04348923 -0.196555642  1.00000000  0.01835156  0.22205859
comodos      0.05452729  0.098284408  0.01835156  1.00000000 -0.04053535
sexo        -0.06566551 -0.123768628  0.22205859 -0.04053535  1.00000000
idade       -0.09772748  0.001278575 -0.26769966  0.16884965 -0.09782802
educacional  0.16346151  0.147430368  0.16903719  0.25218783 -0.06787551
comodos_bin  0.05436478  0.098087281  0.01815687  0.99995795 -0.04079725
                  idade educacional comodos_bin
fuma        -0.097727481  0.16346151  0.05436478
fuma_freq    0.001278575  0.14743037  0.09808728
bebe_freq   -0.267699661  0.16903719  0.01815687
comodos      0.168849651  0.25218783  0.99995795
sexo        -0.097828025 -0.06787551 -0.04079725
idade        1.000000000 -0.21885176  0.16867840
educacional -0.218851763  1.00000000  0.25192742
comodos_bin  0.168678401  0.25192742  1.00000000
```

Agora, a modelagem. Avaliaremos 4 pipelines:

1. Sem balanceamento de classes e sem *binning* de `comodos`.
2. Com balanceamento de classes e sem *binning* de `comodos`.
3. Sem balanceamento de classes e com *binning* de `comodos`.
4. Com balanceamento de classes e com *binning* de `comodos`.

A técnica para balanceamento de classes será *oversampling* e *undersampling* simultaneamente. Aumentaremos em 4/3 a classe minoritária e diminuiremos em 3/4 a classe majoritária.

```r
# realizando oversample
classe_minoritaria <- tabaco[tabaco$fuma == 0, ]
tabaco_ovrsmp <- classe_minoritaria[sample(nrow(classe_minoritaria), nrow(classe_minoritaria) * 4/3
nrow(tabaco_ovrsmp)
```

```
[1] 1333
```

```r
# realizando undersample
classe_majoritaria <- tabaco[tabaco$fuma == 1, ]
tabaco_undrsmp <- classe_majoritaria[sample(nrow(classe_majoritaria), nrow(classe_majoritaria) * 3,

nrow(tabaco_undrsmp)
```

```
[1] 1476
```