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
```

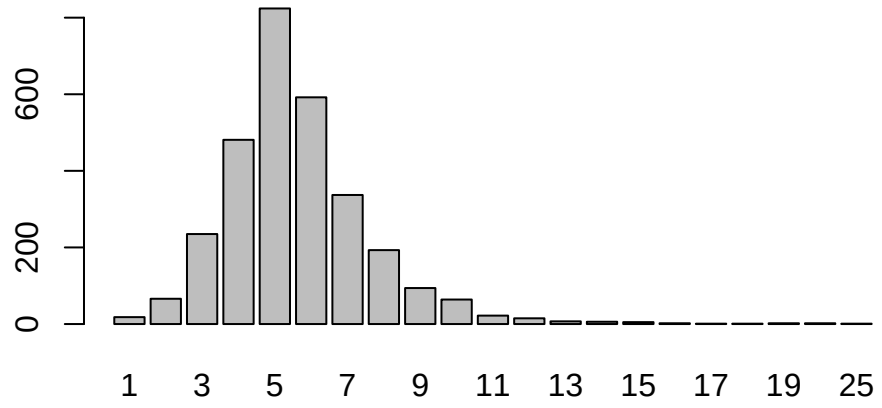
Podemos notar que a variável dependente `fuma` é binária, com 0 representando não fumante e 1 fumante. Todas as demais são categóricas, com exceção da idade. `fuma` também está desbalanceada, com 66,3% dos indivíduos fumantes. A variável `fuma_freq` é a única contínua, com média de 1,5 e desvio padrão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

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




```r
# transformando comodos em numérico
tabaco$comodos <- as.numeric(tabaco$comodos)
```

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. As correlações entre as variáveis `comodos` e `educacional` são positivas, indicando que quanto maior a renda e educação, maior a probabilidade de fumar, o que é contraintuitivo.

```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
                    idade educacional
fuma        -0.097727481  0.16346151
fuma_freq    0.001278575  0.14743037
bebe_freq   -0.267699661  0.16903719
comodos      0.168849651  0.25218783
sexo        -0.097828025 -0.06787551
idade        1.000000000 -0.21885176
educacional -0.218851763  1.00000000
```

Agora, a modelagem. Realizamos o *split* treino e teste com 70% e 30%, respectivamente.

```r
# split treino e teste
train_index <- sample(1:nrow(tabaco), nrow(tabaco) * 0.7)
tabaco_train <- tabaco[train_index, ]
tabaco_test <- tabaco[-train_index, ]
```

Avaliaremos 2 modelos:

1. Sem balanceamento de classes
2. Com balanceamento de classes

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_train[tabaco_train$fuma == 0, ]
tabaco_ovrsmp <- classe_minoritaria[sample(nrow(classe_minoritaria), nrow(classe_minoritaria) * 4/3

nrow(tabaco_ovrsmp)
```

```
[1] 925
```

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

nrow(tabaco_undrsmp)
```

```
[1] 1037
```

```r
# combinando
tabaco_train_bal <- rbind(tabaco_ovrsmp, tabaco_undrsmp)

# checando balanceamento
prop.table(table(tabaco_train_bal$fuma))
```

```
        0         1
0.4714577 0.5285423
```

Enfim, o ajuste dos modelos. Primeiro o modelo sem balanceamento. A variável `idade` foi transformada em `idade^2` para capturar a possível não linearidade. Por conta do desbalanceamento, o modelo é enviesado para a classe majoritária, com a maior parte das predições sendo fumantes. Em consequência, a acurácia de 65,5%, é superestimada. Deixaremos isso claro ao verificar as demais métricas.

```r
metricas <- function(matriz_confusao) {
  true_negative <- matriz_confusao[1,1]
  true_positive <- matriz_confusao[2,2]
  false_positive <- matriz_confusao[1,2]
  false_negative <- matriz_confusao[2,1]

  precisao <- true_positive / (true_positive + false_positive)
  recall <- false_positive / (false_positive + true_negative)
  f1 <- 2 * (precisao * recall) / (precisao + recall)
  acuracia <- (true_positive + true_negative) / (true_positive + true_negative + false_positive + f
  false_positive_rate <- false_positive / (false_positive + true_negative)
  false_negative_rate <- false_negative / (false_negative + true_positive)

  print(paste("Precisão: ", round(precisao, 2)))
  print(paste("Acurácia: ", round(acuracia, 2)))
  print(paste("Recall: ",round(recall,2)))
  print(paste("Taxa de falso positivo: ", round(false_positive_rate, 2)))
  print(paste("Taxa de falso negativo: ", round(false_negative_rate, 2)))
  print(paste("f1: ",round(f1, 2)))
}
```

```r
# ajuste dos modelos
modelo <- glm(
```

```r
  fuma ~ bebe_freq + comodos + sexo + idade + idade^2 + educacional,
  data = tabaco_train,
  family = binomial
)

# sumário
summary(modelo)
```

```
Call:
glm(formula = fuma ~ bebe_freq + comodos + sexo + idade + idade^2 +
    educacional, family = binomial, data = tabaco_train)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.003094   0.220779   4.543 5.53e-06 ***
bebe_freq1   -0.003121   0.147292  -0.021 0.983092
bebe_freq2   -0.379636   0.112199  -3.384 0.000715 ***
comodos       0.030955   0.025959   1.192 0.233078
sexo1        -0.223206   0.099779  -2.237 0.025286 *
idade        -0.010209   0.003335  -3.061 0.002205 **
educacional2  0.481405   0.112086   4.295 1.75e-05 ***
educacional3  0.955917   0.165987   5.759 8.46e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2646.4  on 2076  degrees of freedom
Residual deviance: 2559.5  on 2069  degrees of freedom
AIC: 2575.5

Number of Fisher Scoring iterations: 4
```

```r
# predict
pred <- predict(
  modelo,
  newdata = subset(
    tabaco_test,
    select = c(
      "bebe_freq",
```

```
      "comodos",
      "sexo",
      "idade",
      "educacional"
    )
  ),
  type = "response"
)

# adicionando threshold
pred <- ifelse(pred > 0.5, 1, 0)

# matriz de confusão
confusao <- table(tabaco_test$fuma, pred)

# métricas
metricas(confusao)
```

```
[1] "Precisão:  0.66"
[1] "Acurácia:  0.66"
[1] "Recall:  0.95"
[1] "Taxa de falso positivo:  0.95"
[1] "Taxa de falso negativo:  0.03"
[1] "f1:  0.78"
```

Utilizando classes balanceadas, verificamos que a acurácia é menor, mas, sem o viés para a classe majoritária, o modelo ganha em precisão e reduz a taxa de falso positivo de 95% para 49%, sendo um modelo muito mais confiável.

```
# ajuste dos modelos
modelo <- glm(
  fuma ~ bebe_freq + comodos + sexo + idade + idade^2 + educacional,
  data = tabaco_train_bal,
  family = binomial
)
# predict
pred <- predict(
  modelo,
  newdata = subset(
    tabaco_test,
    select = c(
```

```
      "bebe_freq",
      "comodos",
      "sexo",
      "idade",
      "educacional"
    )
  ),
  type = "response"
)

# adicionando threshold
pred <- ifelse(pred > 0.5, 1, 0)

# matriz de confusão
confusao <- table(tabaco_test$fuma, pred)

# métricas
metricas(confusao)
```

```
[1] "Precisão:  0.72"
[1] "Acurácia:  0.61"
[1] "Recall:  0.49"
[1] "Taxa de falso positivo:  0.49"
[1] "Taxa de falso negativo:  0.34"
[1] "f1:  0.58"
```

Interpretação direta, considerando apenas as direções, beber frequentemente reduz a probabilidade de fumar (-0.38 logodds), assim como ser do sexo masculino. Quanto maior a idade, menores as chances de ser fumante e quanto maior o nível educacional, maior as chances de ser fumante.

2 SEGUNDA QUESTÃO

Nesta questão, usaremos a variável multiclasse `fuma_freq`, adotando como *baseline* 0 = não fumo atualmente. Como as três classes já estão balanceadas, não precisaremos balancear as classes.

```
# ajuste dos modelos
modelo <- nnet::multinom(
  fuma_freq ~ bebe_freq + comodos + sexo + idade + idade^2 + educacional,
  data = tabaco_train
)
```

```
# weights:  27 (16 variable)
initial  value 2281.817724
iter  10 value 2131.846863
iter  20 value 2092.250643
final  value 2092.248510
converged
```

```
# sumário
summary(modelo)
```

```
Call:
nnet::multinom(formula = fuma_freq ~ bebe_freq + comodos + sexo +
    idade + idade^2 + educacional, data = tabaco_train)

Coefficients:
  (Intercept) bebe_freq1 bebe_freq2     comodos       sexo1         idade
1   0.4762081  0.7613984  0.5085071 -0.02162464 -0.06099632 -0.0207598204
2  -0.1330209 -0.5913979 -1.2087263  0.08208917 -0.36926726 -0.0003079305
  educacional2 educacional3
1    0.1925696     0.823909
2    0.8077368     1.133480

Std. Errors:
  (Intercept) bebe_freq1 bebe_freq2    comodos     sexo1       idade
```

```
1   0.2606990  0.1720757  0.1361059 0.03048098 0.1164684 0.003984939
2   0.2559384  0.1722865  0.1341999 0.02996268 0.1157627 0.003913266
  educacional2 educacional3
1    0.1304247    0.1858044
2    0.1349676    0.1917958

Residual Deviance: 4184.497
AIC: 4216.497
```

```r
# predict
pred <- predict(
  modelo,
  newdata = subset(
    tabaco_test,
    select = c(
      "bebe_freq",
      "comodos",
      "sexo",
      "idade",
      "educacional"
    )
  ),
  type = "class"
)

# matriz de confusão
confusao <- table(tabaco_test$fuma_freq, pred)

# acurácia
acuracia <- sum(diag(confusao) / sum(confusao))

# precisão
precisao <- diag(confusao) / colSums(confusao)

# recall
recall <- diag(confusao) / rowSums(confusao)

# métricas
confusao
```

```
   pred
      0   1   2
```

```
  0 109 104  93
  1  56 162  60
  2  56  72 179
```

```
acuracia
```

```
[1] 0.5050505
```

```
precisao
```

```
        0         1         2
0.4932127 0.4792899 0.5391566
```

```
recall
```

```
        0         1         2
0.3562092 0.5827338 0.5830619
```

Na interpretação direta, beber aumenta a probabilidade de fumar ocasionalmente, enquanto reduz a probabilidade de fumar diariamente. A *proxy* de renda, a quantidade de cômodos, influencia positivamente a probabilidade de fumar diariamente (para o nível 1 não foi significativo). Demais variáveis seguem a mesma direção e sentido do modelo biclasse.

```
# test z bicaudal
z <- summary(modelo)$coefficients / summary(modelo)$standard.errors

p <- (1 - pnorm(abs(z), 0, 1)) * 2
p
```

```
  (Intercept)   bebe_freq1   bebe_freq2     comodos       sexo1        idade
1  0.06775108 9.653752e-06 0.0001868866 0.478047079 0.600476207 1.892780e-07
2  0.60324622 5.977284e-04 0.0000000000 0.006149265 0.001423391 9.372801e-01
  educacional2 educacional3
1 1.398149e-01 9.237959e-06
2 2.168238e-09 3.424648e-09
```