Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Departamento de Estatística

Atividade 1 - Estatística Multivariada 1

Grupo 2: Antônio M. dos Santos Jr. - 744845

Crystiane Souza - 760955

Douglas Nestlehner - 752728

Eric Sato - 729739

Agosto, 2021

Sumário

Capítulo 1

Introdução

Este estudo tem como objetivo realizar análises multivariadas no banco de dados de Clientes de Cartão de Crédito, que foi retirado da plataforma Kaggle. O banco de dados contém 10127 observações e 21 variáveis.

Para a realização das primeiras análises, no momento, as variáveis categóricas do banco de dados foram separadas. Logo, serão analisadas 10 variáveis contínuas. A ferramenta de linguagem de programação utilizada será o software estatístico R.

No Capítulo 2 discorre-se sobre os resultados obtidos na análise e por fim, no Apêndice, está apresentado o código utilizado.

Variáveis analisadas nesse estudo:

$\mathbf{X}_1$ **:** Idade do cliente (em anos);
$\mathbf{X}_2$ **:** Período de relacionamento com banco (em meses);
$\mathbf{X}_3$ **:** Limite de crédito no cartão de crédito (U$);
$\mathbf{X}_4$ **:** Saldo rotativo total no cartão de crédito (U$);
$\mathbf{X}_5$ **:** Linha de crédito aberta para compra (média dos últimos 12 meses);
$\mathbf{X}_6$ **:** Mudança no valor da transação (Q4 sobre Q1);
$\mathbf{X}_7$ **:** Valor total da transação (últimos 12 meses);
$\mathbf{X}_8$ **:** Contagem total de transações (nos últimos 12 meses);
$\mathbf{X}_9$ **:** Mudança na contagem de transações (Q4 sobre Q1);
$\mathbf{X}_{10}$ **:** Taxa de utilização média do cartão.

Capítulo 2

Resultados

2.1 Vetor de Médias

Para realizar o cálculo do vetor de médias amostrais foi utilizada a expressão:

$$\bar{x}_p = \sum_{i=1}^{n} \frac{x_{ip}}{n},$$

e assim obteve-se o vetor de médias das 10 variáveis do banco de dados em estudo, sendo:

$$\bar{\mathbf{x}} = \begin{pmatrix} 46.326 \\ 35.928 \\ 8631.953 \\ 1162.814 \\ 7469.140 \\ 0.760 \\ 4404.086 \\ 64.859 \\ 0.712 \\ 0.275 \end{pmatrix}.$$

Esse vetor nos dá então, informações sobre a média de cada variável, por exemplo: A média de idade ($\bar{x}_1$) dos clientes é de pouco mais de 46 anos, a média do período de relacionamento com o banco ($\bar{x}_2$) é de quase 36 meses, a média da taxa de utilização média do cartão ($\bar{x}_{10}$) pelos clientes é de 0,275.

2.2 Matriz de Variâncias e Covariâncias (Var-Cov)

Para calcular a matriz de variâncias e covariâncias amostrais, utilizamos a seguinte expressão,

$$S_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{n},$$

assim, obtendo a matriz das 10 variáveis analisadas. Essa matriz é simétrica e serve para medir o grau de relacionamento linear entre duas variáveis.

$$S = \begin{bmatrix} 64.26 & 50.51 & 180.41 & 96.56 & 83.85 & -0.11 & -1264.81 & -12.62 & -0.02 & 0.02 \\ & 63.78 & 544.86 & 56.12 & 488.74 & -0.09 & -1046.89 & -9.34 & -0.03 & -0.02 \\ & & 82597704.01 & 314721.77 & 82282982.24 & 25.52 & 5301773.45 & 16196.42 & -4.37 & -1210.05 \\ & & & 664138.77 & -349417.00 & 10.39 & 178199.62 & 1072.32 & 17.43 & 140.19 \\ & & & & 82632399.24 & 15.13 & 5123573.83 & 15124.09 & -21.80 & -1350.24 \\ & & & & & 0.05 & 29.54 & 0.03 & 0.02 & 0.00 \\ & & & & & & 11539347.59 & 64358.62 & 69.21 & -77.76 \\ & & & & & & & 550.91 & 0.63 & 0.02 \\ & & & & & & & & 0.06 & 0.00 \\ & & & & & & & & & 0.08 \end{bmatrix}$$

Tal matriz traz muitas informações, em sua diagonal principal, carrega a variância de cada variável (de X_1 a X_{10}), e fora da diagonal principal apresenta as covariâncias entre tais variáveis.

Consegue-se por exemplos então, visualizar que a variância de idade dos clientes (X_1) é de 64,26 anos, que para essa circunstância pode ser considerada baixa, com desvios de aproximadamente 8 anos em torno da média anteriormente apresentada. Outro exemplo é o que acontece com X_3 (Limite de crédito no cartão do cliente), que apresenta uma variância de 80597704,01 resultando desvios de aproximadamente 9 mil dólares em torno da média (que é de 8.632), ou seja, uma alta variabilidade no limite de crédito.

2.3 Variância Total

A variância total é uma forma de sintetização de todas variâncias apresentadas anteriormente, já que esta é a soma das variâncias de todas as variáveis analisadas no banco de dados. Ou seja, é simplesmente a soma das variâncias (listadas na diagonal da matriz S), sendo então representada por:

$$\text{traço}(S) = tr(S) = S_{11} + S_{22} + \cdot + S_{pp}.$$

Nesse contexto, no banco de dados analisado, temos que a variância total é de:

$$tr(S) = 177451791.$$

A soma dessas variabilidades, por si só, no momento, como as variáveis em estudo possuem escalas distintas, não nos dá uma interpretação além da já descrita separadamente. Entretanto,

essa soma será útil para comparações futuras, onde pode-se haver interesses em diminuir essa variabilidade através de algum método multivariado.

2.4 Variância Generalizada

Outro método de sintetizar a variabilidade multivariada dos dados em um único valor numérico, é através da variância generalizada. Entretanto, essa, além de trazer informações apenas da variância (como a Total), também resume as covariâncias. Tal medida é definida como o determinante da matriz S. Sendo definida como:

$$\det(S) = \sum_{i=1}^{n} (-1)^{i+j} \cdot s_{ij} \cdot \det(S_{ij}),$$

onde $i, j = 1, 2, ..., 10$.

Assim, a partir da matriz Var-Cov gerada anteriormente, temos que:

$$\det(S) = 244049967398863$$

Da mesma forma que na Variância Total, por mais que seja um valor aparentemente grande, a Variância Generalizada, por si só, não nos acrescenta muitas informações além das já vistas, entretanto será útil para comparações futuras onde poderá haver interesses em diminuir o resultado desse determinante através de algum método multivariado.

2.5 Matriz de Desvios Padrão

Calculamos a matriz de desvio padrão utilizando as informações apresentadas em aula:

$$D^{1/2} = \begin{bmatrix} \sqrt{S_{11}} & 0 & ... & 0 \\ 0 & \sqrt{S_{22}} & & \\ ... & & ... & \\ 0 & ... & & \sqrt{S_{pp}} \end{bmatrix}$$

.

Implementando então para os dados que estão sendo utilizados, obtivemos os seguintes resultados:

$$D^{1/2} = \begin{bmatrix} 8.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & 7.98 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & 9088.32 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & 814.94 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & 9090.23.24 & 0 & 0 & 0 & 0 & 0 \\ & & & & & 0.21 & 0 & 0 & 0 & 0 \\ & & & & & & 3396.96 & 0 & 0 & 0 \\ & & & & & & & 23.47 & 0 & 0 \\ & & & & & & & & 0.23 & 0 \\ & & & & & & & & & 0.27 \end{bmatrix}$$

.

Tendo calculado a matriz de desvios padrão podemos encontrar outros resultados, que podem ser obtidos por meio das seguintes relações:

$$S = D^{1/2} R D^{1/2}$$

$$R = D^{-1/2} S D^{-1/2}$$

Ao realizar essas multiplicações de matrizes, conseguimos obter o mesmo resultado para a matriz de Var-Cov (S) já calculada anteriormente, e também para a matriz de correlação (R) que será calculada na próxima seção.

2.6 Matriz de Correlações

Para analisar o grau de relacionamento entre duas variáveis, pode ser calculada a correlação destas. A matriz de correlações nada mais é então que uma tabela que indica os coeficientes de conexão entre as variáveis, onde cada célula mostra a conexão entre duas variáveis respectivas da linha e coluna.

No caso, é utilizada a Correlação Linear de Pearson, dada por:

$$r_{jk} = \frac{S_{jk}}{\sqrt{S_{jj}}\sqrt{S_{kk}}} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x_k})}{\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}},$$

onde $j, k = 1, 2, ..., 10$.

Nesse contexto, no banco de dados analisado, a matriz de correlações é dada por:

	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$	$\mathbf{X}_5$	$\mathbf{X}_6$	$\mathbf{X}_7$	$\mathbf{X}_8$	$\mathbf{X}_9$	$\mathbf{X}_{10}$
$\mathbf{X}_1$	1.000	0.789	0.002	0.015	0.001	-0.062	-0.046	-0.067	-0.012	0.007
$\mathbf{X}_2$	0.789	1.000	0.008	0.009	0.007	-0.049	-0.039	-0.050	-0.014	-0.008
$\mathbf{X}_3$	0.002	0.008	1.000	0.042	0.996	0.013	0.172	0.076	-0.002	-0.483
$\mathbf{X}_4$	0.015	0.009	0.042	1.000	-0.047	0.058	0.064	0.056	0.090	0.624
$\mathbf{X}_5$	0.001	0.007	0.996	-0.047	1.000	0.008	0.166	0.071	-0.010	-0.539
$\mathbf{X}_6$	-0.062	-0.049	0.013	0.058	0.008	1.000	0.040	0.005	0.384	0.035
$\mathbf{X}_7$	-0.046	-0.039	0.172	0.064	0.166	0.040	1.000	0.807	0.086	-0.083
$\mathbf{X}_8$	-0.067	-0.050	0.076	0.056	0.071	0.005	0.807	1.000	0.112	0.003
$\mathbf{X}_9$	-0.012	-0.014	-0.002	0.090	-0.010	0.384	0.086	0.112	1.000	0.074
$\mathbf{X}_{10}$	0.007	-0.008	-0.483	0.624	-0.539	0.035	-0.083	0.003	0.074	1.000

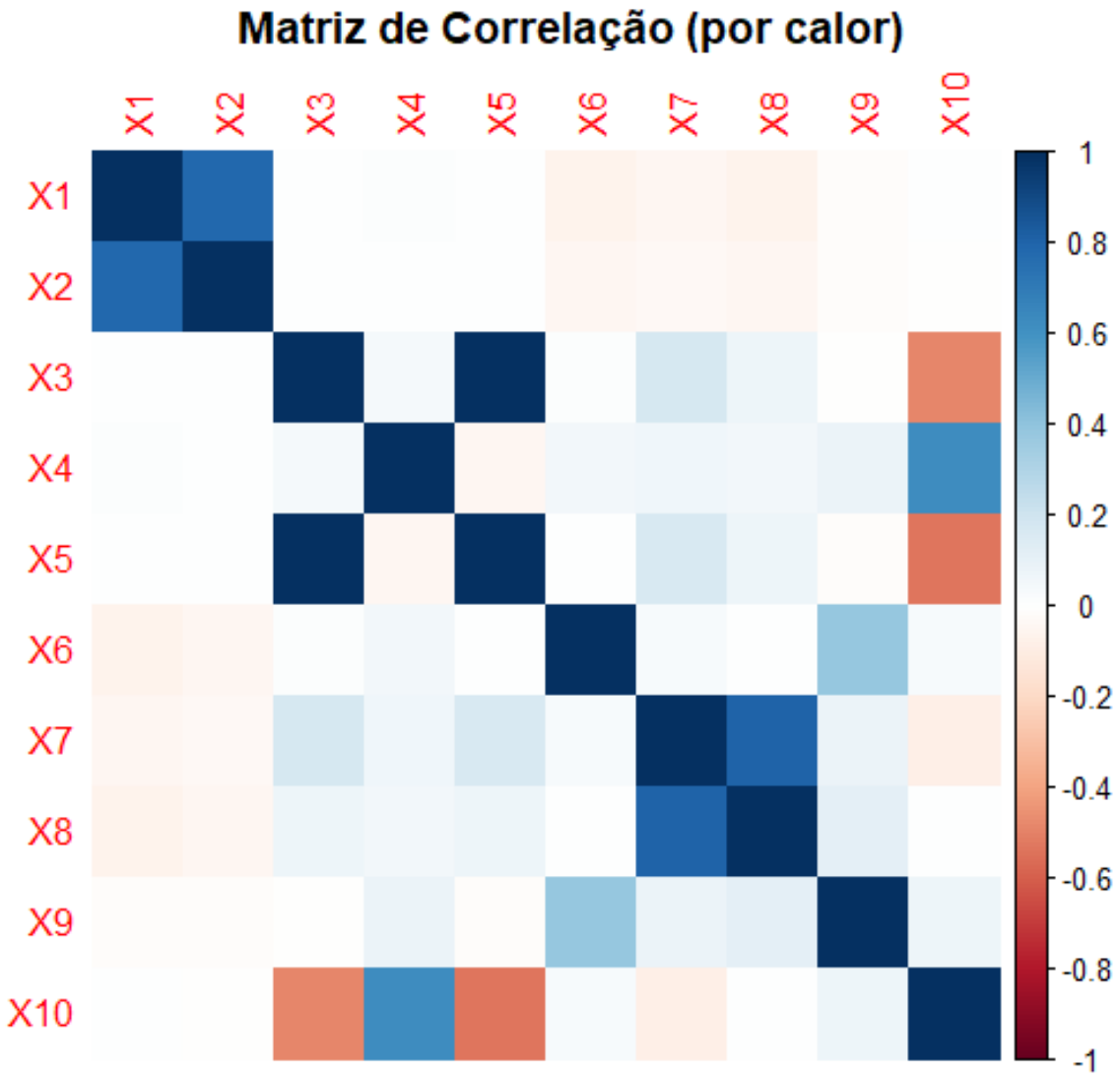

Figura 2.1: Matriz de Correlação - Método de calor

Para visualização mais clara, a matriz foi construída também em sistema de cores, onde o termômetro ao lado direito indica as correlações, sendo tonalidades vermelhas as correlações negativas, e azuis as positivas. Lembrando que a matriz é simétrica, percebe-se que além da diagonal principal, são poucas as cores fortes (correlações altas).

Algumas correlações fortes que podemos notar são:
$\rho(X_3, X_5) = 0{,}996$
$\rho(X_8, X_7) = 0{,}807$
$\rho(X_1, X_2) = 0{,}789$
$\rho(X_{10}, X_4) = 0{,}624$
$\rho(X_{10}, X_5) =$ -0,539
$\rho(X_{10}, X_3) =$ -0,483
$\rho(X_9, X_6) = 0{,}384$
O restante assume valores próximo a zero, ou seja, pouca correlação linear entre as variáveis.

Como exemplo, as duas correlações mais fortes são entre: (Limite do Cartão , Média de Linhas abertas para compra) e (Valor da Transação , Contagem de Transações). O que de certa forma já era esperado, dado que uma característica em um pré conceito aparenta ter grande possibilidade de impacto na outra.

No mais,
com essas análises, dar-se-á continuidade no estudo, aprofundando-as e detalhando caso necessário.

Apêndice A

Códigos

```
### Atividade 1 - Estatistica Multivariada 1 ###

## Importando a base de dados
original <- read.csv("C:/Users/Crys/Desktop/UFSCar/2021.1/Estatistica
   Multivariada 1/BankChurners.csv",
                     header = TRUE, sep = ",")


dados = original[,-c(22,23)]
nomes=c("Identificador","Atividade","Idade","Sexo","Dependentes","N vel
   Educacional","Estado Civil",
        "Renda Anual", "Tipo do Cart o", "Per odo de Relacionamento", "N  de
    Produtos Mantidos",
        "Meses inativos U.A","N  de Contatos U.A","Limite de Cr dito","Saldo
   Rotativo","M dia de Cr dito Aberto U.A",
        "Mudan a no Valor Transacional","Valor Total da Transa  o U.A","N
   de Transa  es U.A","Mudan a no N  Transacional", "Taxa de Utiliza  o
   M dia")

setnames(dados, nomes)
dados

dadosnum=dados[,c(3,10,14,15,16,17,18,19,20,21)]
dadoscat=dados[,-c(3,10,14,15,16,17,18,19,20,21)]

#Organizando base Numerica
nomes2=c("X1","X2","X3","X4","X5","X6","X7","X8","X9","X10")
library("data.table")
setnames(dadosnum, nomes2)
```

```
24
25 ## Encontrando o vetor de medias
26 vetor_medias <- matrix(data = NA, nrow=ncol(dadosnum), ncol=1)
27 for (i in 1:ncol(dadosnum)){
28   vetor_medias[i] <- sum(dadosnum[i])/nrow(dadosnum)
29 }
30
31 vetor_medias
32
33 ## Calculando a matriz VarCov
34 varcov <- matrix(data = NA, nrow=ncol(dadosnum), ncol=ncol(dadosnum))
35 for (j in 1:ncol(dadosnum)){
36   for (k in 1:ncol(dadosnum)){
37     varcov[j,k] <- sum((dadosnum[j]-vetor_medias[j])*(dadosnum[k]-vetor_medias[
     k]))/nrow(dadosnum)
38   }
39 }
40
41 round(varcov,2)
42
43 ## Calculando a variancia generalizada
44 determinate<- function(x=matrix()){
45   n <- nrow(x)
46   s <- 0
47   det <- function(x=matrix()){
48     a <- x[1,1]*x[2,2]
49     b <- x[1,2]*x[2,1]
50     d <- a-b
51     return(d)
52   }
53
54   if(n>2){
55     for(i in 1:(n-2)){
56       for(j in 1:n){
57         if(j%%2==0){p<- -1
58         }else{p<- 1}
59         s <- s + p*x[i,j]*determinate(x[-i,-j])
60       }
61     }
62   }else{
63     s<- det(x)
64   }
```

```
  return(s)
}

determinate(varcov)

## Calculando a variancia total
traco <- function(var_total){
  n <- dim(var_total)[1]
  tr <- 0
  for (i in 1:n){
    j <- var_total[i,i]
    tr <- tr + j
  }
  return(tr[[1]])
}

traco(varcov)

## Matriz de desvio padrao
D <- matrix(data = NA, nrow=ncol(dadosnum), ncol=ncol(dadosnum))
for (i in 1:ncol(dadosnum)){
  for (j in 1:ncol(dadosnum)){
    if(i == j){
      D[i,j] = varcov[i,j]^{1/2}
    }else{
      D[i,j] = 0
    }
  }
}
D
# D^{-1/2}
D_1 <- matrix(data = NA, nrow=ncol(dadosnum), ncol=ncol(dadosnum))
for (i in 1:ncol(dadosnum)){
  for (j in 1:ncol(dadosnum)){
    if(i == j){
      D_1[i,j] = 1/{varcov[i,j]^{1/2}}
    }else{
      D_1[i,j] = 0
    }
  }
}
D_1
```

```r

## Relacao
#D_1 = D^{-1/2}

R = (D_1%*%varcov)%*%D_1
R

S = (D%*%varcov)%*%D
S


## Matriz de correlacao
corr <- matrix(data = NA, nrow=ncol(dadosnum), ncol=ncol(dadosnum))
for (l in 1:ncol(dadosnum)){
  for (m in 1:ncol(dadosnum)){
    corr[l,m] <- varcov[l,m]/sqrt(varcov[l,l]*varcov[m,m])
  }
}

round(corr,2)

## Gr fico da Matriz de correlacao
library(corrplot)
corrplot(cor(dadosnum), method = "color", title = "Matriz de Correlacao (por
   calor)", mar=c(0,0,2,0))
```