**Universidade Estadual Do Oeste do Paraná**
**PGEAGRI**
***Programa de Pós-Graduação em Engenharia Agrícola***

**ANÁLISE MULTIVARIADA**
**Prof.: Dra. Luciana Pagliosa**

# Resolução da 2ª Lista de Análise Multivariada - 2018

Vanessa Mendes Pientosa
Willyan Goergen de Souza

Cascavel - PR
Dezembro de 2018

**2ª Lista de Análise Multivariada - 2018**

**QUESTÃO 1) No pacote MVar.pt do R existe um banco de dados chamado DataMix. Este banco de dados se refere a dados de 10 cooperativas/degustadores nas quais obtiveram-se as seguintes informações: Medias das notas dadas aos cafés analisados, Anos de trabalho como degustador, Degustador com formação técnica, Degustador com dedicação exclusiva, Frequência media dos cafés classificados como especiais, Frequência media dos cafés classificados como comerciais. Analise a associação entre os seguintes grupos de variáveis: GRUPO 1 (médias das notas dadas aos cafés analisados, Anos de trabalho como degustador) e GRUPO 2 (Frequência media dos cafés classificados como especiais, Frequência media dos cafés classificados como comerciais.**

**R:** Quando se realiza uma pesquisa, pode haver a necessidade de estudar a relação entre variáveis (ou grupos de variáveis), sendo elas relacionadas ou não.

A análise de correlação canônica (ACC) tem como objetivo principal medir a associação entre dois grupos de variáveis quantitativas. Para isso, obtêm-se combinações lineares das variáveis (U= a'X e V= b'Y) que resultem na melhor associação existente (correlação linear) entre esses dois grupos.

**Tabela 1:** Dados referentes as cooperativas/degustadores

| | **GRUPO 1** | | **GRUPO 2** | |
|---|---|---|---|---|
| | **X1** | **X2** | **X3** | **X4** |
| **A1** | 60.20 | 2.9 | 96 | 38 |
| **A2** | 72.20 | 3.1 | 98 | 86 |
| **A3** | 65.88 | 2.8 | 54 | 50 |
| **A4** | 54.98 | 2.1 | 59 | 19 |
| **A5** | 63.77 | 2.9 | 102 | 29 |
| **B1** | 77.20 | 6.2 | 105 | 342 |
| **B2** | 71.30 | 6.9 | 111 | 358 |
| **B3** | 82.22 | 7.3 | 93 | 587 |
| **B4** | 71.15 | 8.2 | 133 | 574 |
| **B5** | 63.44 | 6.2 | 124 | 453 |

Na tabela 1, encontram-se os dados referentes aos grupos 1 e 2, onde X1 é a "média das notas dadas aos cafés analisados", X2 refere-se aos "Anos de trabalho

como degustador", X3 é a "Frequência media dos cafés classificados como especiais" e X4 é a "Frequência media dos cafés classificados como comerciais", cujas são apenas as variáveis quantitativas de interesse, do banco de dados original (DataMix), para essa análise.

**Tabela 2:** Matriz de correlação linear entre os pares das variáveis consideradas no experimento

| | X1 | X2 | X3 | X4 |
|---|---|---|---|---|
| X1 | 1.0000 | | | |
| X2 | 0.6801 | 1.0000 | | |
| X3 | 0.3430 | 0.7158 | 1.0000 | |
| X4 | 0.6640 | 0.9719 | 0.6521 | 1.0000 |

Variáveis: X1 (media das notas dadas aos cafés analisados; X2 (anos de trabalho como degustadores); X3 (frequência média dos cafés classificados como especiais), X4 (frequência média dos cafés classificados como comerciais).

De acordo com a tabela 2, verifica-se a correlação linear de Pearson entre cada par das variáveis estudadas. Quanto mais próximo de 0, menor correlação existe entre as variáveis e quanto mais próximo de 1, maior é a correlação entre os dados. Segundo Callegari-Jacques (2003), para avaliar de maneira qualitativa o coeficiente de correlação de Pearson, podemos adotar o seguinte critério:

- se $0,00 < |r| < 0,30$, existe fraca correlação linear;
- se $0,30 \leq |r| < 0,60$, existe moderada correlação linear;
- se $0,60 \leq |r| < 0,90$, existe forte correlação linear;
- se $0,90 \leq |r| < 1,00$, existe correlação linear muito forte.

A maior relação existente neste estudo é entre as variáveis X4 e X2, ou seja, entre a frequência das medias dos cafés classificados como comerciais e os anos de trabalho como degustador, apresentando valor de 0,9719, representando assim, uma correlação muito forte. A segunda maior correlação linear se apresenta entre X3 e X2, ou seja, entre a frequência média dos cafés classificados como especiais e os anos de trabalho como degustadores, com valor correspondente a 0,7158, demostrando assim, uma correlação forte. X2 e X1, que representa a correlação entre a média das notas dadas aos cafés analisados e anos de trabalho como degustador, apresentam uma correlação forte (0,6801), e o mesmo acontece com as variáveis X1 e X4 (0.6640), ou seja, médias das

notas dadas as cafés analisados e frequência média dos cafés classificados como comerciais. As variáveis X4 e X3, que são frequência média dos cafés analisados como especiais e frequência média dos cafés classificados como comerciais, apresentaram ainda uma correlação linear forte, com valor de 0,6521, enquanto uma correlação fraca ocorreu entre as variáveis X1 e X3, que correspondem a medias das notas dadas aos cafés analisados e frequência média dos cafés classificados como especiais, com valor de 0,3430.

Calcula-se também os autovalores e autovetores da matriz A. A tabela 3 apresenta a matriz A que possui dimensões 2x2.

**Tabela 3:** Matriz A

| | X1 | X2 |
|---|---|---|
| X1 | 0.0462 | -0.0333 |
| X2 | 0.6010 | 0.9788 |

Em seguida, obtêm-se os autovalores e autovetores da Matriz A.

**Tabela 4:** Autovalores da Matriz A

| | |
|---|---|
| 0.9569 | 0.0682 |

**Tabela 5:** Autovetores da Matriz A

| | [,1] | [,2] |
|---|---|---|
| [1,] | 0.0365 | -0.8346 |
| [2,] | -0.9993 | 0.5508 |

A tabela 6 apresenta os coeficientes canônicos a serem aplicados nas variáveis do grupo 1 e na tabela 7 são apresentadas as estimativas das variáveis canônicas nos indivíduos do primeiro grupo de variáveis.

**Tabela 6**. Coeficientes canônicos para o grupo 1.

| | [,1] | [,2] |
|---|---|---|
| X1 | 0.0015 | -0.0557 |
| X2 | -0.1487 | 0.1306 |

**Tabela 7:** Variáveis canônicas do primeiro grupo de variáveis

| | U1 | U2 |
|---|---|---|
| **A1** | 0.8166 | 0.3525 |
| **A2** | 0.7833 | -0.8280 |
| **A3** | 0.8855 | -0.2529 |
| **A4** | 1.1412 | 0.6950 |
| **A5** | 0.8326 | -0.0130 |
| **B1** | -0.5427 | -0.5965 |
| **B2** | -0.8736 | 0.1752 |
| **B3** | -0.9987 | -0.8467 |
| **B4** | -1.4398 | 0.5023 |
| **B5** | -0.6043 | 0.8120 |

Em seguida, calcula-se os autovalores e os autovetores da matriz B.

A tabela 8 apresenta a matriz B. que possui dimensões 2x2.

**Tabela 8:** Matriz B

| | X3 | X4 |
|---|---|---|
| **X3** | 0.1701 | 0.1373 |
| **X4** | 0.5839 | 0.8550 |

Obtém-se os autovalores e autovetores da matriz B.

**Tabela 9:** Autovalores da Matriz B

| | |
|---|---|
| 0.9569 | 0.0682 |

**Tabela 10:** Autovetores da Matriz B

| | [,1] | [,2] |
|---|---|---|
| [1,] | -0.1719 | -0.8030 |
| [2,] | -0.9851 | 0.5960 |

Semelhante ao grupo anterior, a tabela 11 apresenta os coeficientes canônicos e a tabela 12 apresenta as variáveis canônicas do segundo grupo de variáveis.

**Tabela 11**. Coeficientes canônicos para o grupo 2.

| | [,1] | [,2] |
|---|---|---|
| X3 | -0.0021 | 0.0175 |
| X4 | -0.0013 | -0.0014 |

**Tabela 12:** Variáveis canônicas do segundo grupo de variáveis

| | V1 | V2 |
|---|---|---|
| **A1** | 0.9171 | -0.5003 |
| **A2** | 0.7014 | -0.4425 |
| **A3** | 1.1556 | 0.8800 |
| **A4** | 1.2516 | 0.6405 |
| **A5** | 0.9136 | -0.7161 |
| **B1** | -0.4234 | -0.0161 |
| **B2** | -0.5320 | -0.1683 |
| **B3** | -1.3712 | 0.9929 |
| **B4** | -1.5918 | -0.3257 |
| **B5** | -1.0209 | -0.3443 |

As combinações lineares U e V são chamadas de variáveis canônica, enquanto que a correlação entre os pares dos mesmos são nomeados de correlação canônica (Silva et al., 2014). Assim, a tabela 13 apresenta a correlação das variáveis canônicas par a par, sendo destacadas as de maior relevância.

**Tabela 13:** Correlação canônica entre U e V:

| | |
|---|---|
| **cor(u1,v1)** | **0.9782** |
| cor(u1,v2) | 1.21E-16 |
| cor(u2,v1) | 2.24E-17 |
| **cor(u2,v2)** | **-0.2611** |

Analisando a tabela 13, nota-se que há uma forte correlação entre U1 e V1, e também uma correlação relevante entre U2 e V2 (-0,2611). Enquanto que as demais correlações não apresentaram significância. Isto pode ser observado graficamente na figura 1.

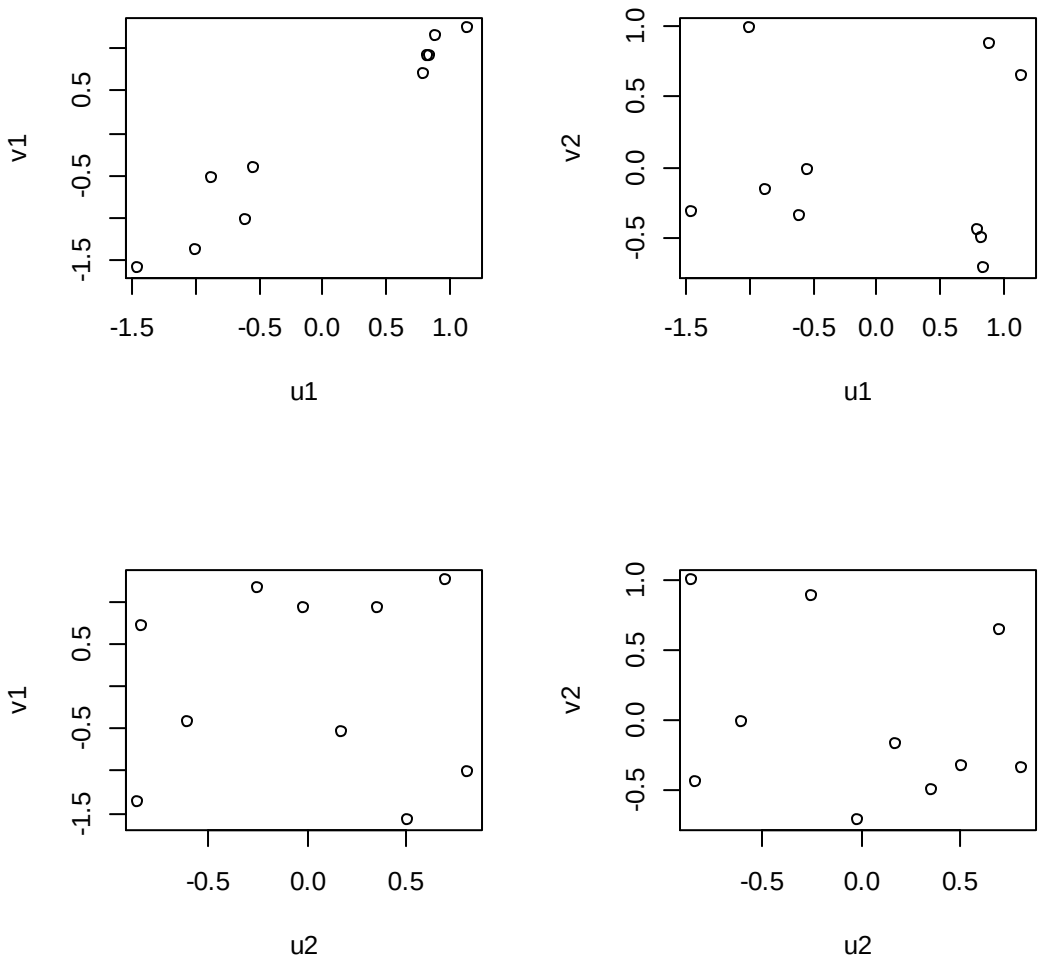

Figura 1: Gráfico das correlações canônicas.

Na figura 1 é possível observar a forte correlação que existe entre a primeira variável canônica do primeiro grupo e a primeira variável canônica do segundo grupo (U1, V1), onde o gráfico de molda como uma reta perfeita e ainda que demostrando que essa correlação é positiva. Também é possível observar que U1, V2, os pontos aparecem dispersos no gráfico, o que significa que não existe correlação entre essas duas variáveis. O mesmo se observa com relação as variáveis U2, V1, onde verifica-se também uma dispersão dos dados. Já com relação as variáveis U2, V2, constata-se uma tendência a uma reta, sendo essa reta mais alocado para a esquerda, o que demostra uma correlação negativa.

As tabelas 14 e 15 representam a correlação linear das variáveis canônicas U e V, respectivamente, com as variáveis originais.

**Tabela 14:** Correlação linear das variáveis U com variáveis originais:

| Correlação das variáveis canônicas U | Valor da correlação |
|---|---|
| cor(U1,X1) | -0.6597 |
| cor(U1,X2) | -0.9996 |
| cor(U2,X1) | -0.7515 |
| cor(U2,X2) | -0.0275 |
| cor(U1,X3) | -0.7209 |
| cor(U1,X4) | -0.9714 |
| cor(U2,X3) | 0.1765 |
| cor(U2,X4) | -0.0308 |

**Tabela 15:** Correlação linear das variáveis V com variáveis originais:

| Correlação das variáveis canônicas U | Valor da correlação |
|---|---|
| cor(V1,X3) | -0.7370 |
| cor(V1,X4) | -0.9930 |
| cor(V2,X1) | 0.1962 |
| cor(V2,X2) | 0.0072 |
| cor(V1,X1) | -0.6453 |
| cor(V1,X2) | -0.9778 |
| cor(V2,X3) | -0.6759 |
| cor(V2,X4) | 0.1180 |

Da Tabela 14, é possível afirmar que as variáveis originais tiveram maiores correlações com U1 que com U2, exceto X1, sendo as correlações mais fortes com o mesmo a de X2 e X4. Na Tabela 15, de forma análoga, as variáveis tiveram maior correlação com V1 que com V2, sendo as maiores também com X2 e X4.

Por fim, calcula-se a medida de qualidade do modelo, cujo resultado se apresenta na tabela 16 abaixo:

**Tabela 16:** Medida de qualidade do modelo

| | |
|---|---|
| U1 | 71.7238 % |
| V1 | 76.4614 % |

Na tabela 16, verifica-se os valores da medida de qualidade do modelo que é a proporção da variância total para cada grupo das variâncias originais, que é explicada pelas variáveis canônicas. Neste caso, observa-se que V1 apresentou maior valor, sendo este 76,4614 do que U1, que apresentou valor de 71,7238.

**<u>QUESTÃO 2</u>) Os dados da Tabela 1 são baseadas no exemplo descrito em Mingoti (2013). A tabela abaixo apresenta os resultados de uma pesquisa realizada com 1025 pessoas, a respeito da preferência entre quatro modelos de automóveis. As pessoas pesquisadas foram também classificadas quanto a variável (sexo/trabalho – considere como uma única variável).**

**Tabela 1 - Número de pessoas classificadas segundo a preferência pelo modelo do carro e sexo/trabalho.**

| | **Modelo 1** | **Modelo 2** | **Modelo 3** | **Modelo 4** |
|---|---|---|---|---|
| **Feminino/não trabalha** | **114** | **47** | **83** | **32** |
| **Feminino/ trabalha** | **148** | **92** | **128** | **49** |
| **Masculino/ trabalha** | **59** | **165** | **61** | **47** |

**<u>Alternativa a</u>) Interprete a tabela de dupla**

**Tabela 17:** Número de pessoas classificadas segundo a preferência pelo modelo do carro e sexo/trabalho.

| | Modelo 1 | Modelo 2 | Modelo 3 | Modelo 4 | **Total** |
|---|---|---|---|---|---|
| Feminino/não trabalha | 114 | 47 | 83 | 32 | **276** |
| Feminino/trabalha | 148 | 92 | 128 | 49 | **417** |
| Masculino/trabalha | 59 | 165 | 61 | 47 | **332** |
| **Total** | **321** | **304** | **272** | **128** | **1025** |

A tabela de dupla entrada (tabela 17) apresenta valores de frequência do requisito entre os quatro modelos de automóveis em relação ao sexo e trabalho. No modelo 1, observa-se uma predominante preferência deste pelo sexo feminino que trabalha. Já para o modelo 2, a preferência é pelo sexo masculino que trabalha. Em seguida, o modelo 3 teve a maior preferência pelo sexo feminino que trabalha. Já o modelo 4, teve uma preferência quase igualitária entre o grupo feminino que trabalha e o masculino que trabalha.

Analisando de forma geral, observa-se que o modelo 1 teve a preferência do sexo feminino, que trabalha e não trabalha e o modelo 2 a preferência ficou com o grupo masculino que trabalha.

**Alternativa b) Realize o teste do Qui-Quadrado (com 5% de significância) e a análise de correspondência. Interprete os resultados**.

O teste Qui-Quadrado é um teste não paramétrico, onde avalia a relação/associação entre variável A e variável B. É uma medida de padronização de frequências reais de células comparadas com frequências esperadas.

As hipóteses para o teste são:

*$H_0$: frequências observadas = frequências esperadas*

*Não há associação entre as variáveis*

*$H_1$: frequências observadas ≠ frequências esperadas*

*Há associação entre as variáveis*

Para as variáveis sexo e trabalho e modelo de automóveis o p-valor calculado pelo software R foi de < 2.2e-16, nota-se, portanto, que esse valor é menor que 5% de significancia, sendo possível concluir que ocorre associação entre as variáveis, portanto rejeita-se a hipotese H0.

Identificada essa associação, realizou-se a análise de correspondência (figura 2), que permite a visualização gráfica das relações existentes através da redução da dimensionalidade dos conjuntos de dados.

A análise de correspondência objetiva determinar o grau de associação global entre suas linhas e colunas, onde os níveis de linha e de coluna assumam posições nos gráficos de acordo com associação ou similaridade entre as mesmas (INFANTOSI et al., 2014).

**Tabela 18:** Variância das dimensões (eixos vetoriais)

| | **Dim.1** | **Dim.2** |
|---|---|---|
| **Variância** | 0.1080 | 0.0010 |
| **% da Var.** | 99.3340 | 0.6660 |
| **% Acumulada.** | **99.33** | **100.00** |

Na tabela 18 é possível verificar que pela proporção de variância a dimensão 1 explica 99,33 % dos dados analisados, enquanto a dimensão 2 explica apenas 0,66%, porém, totalizando assim representatividade de 100% da variância na análise de correspondência.

Foi realizado em seguida o gráfico de correspondência, representado pela Figura 2.

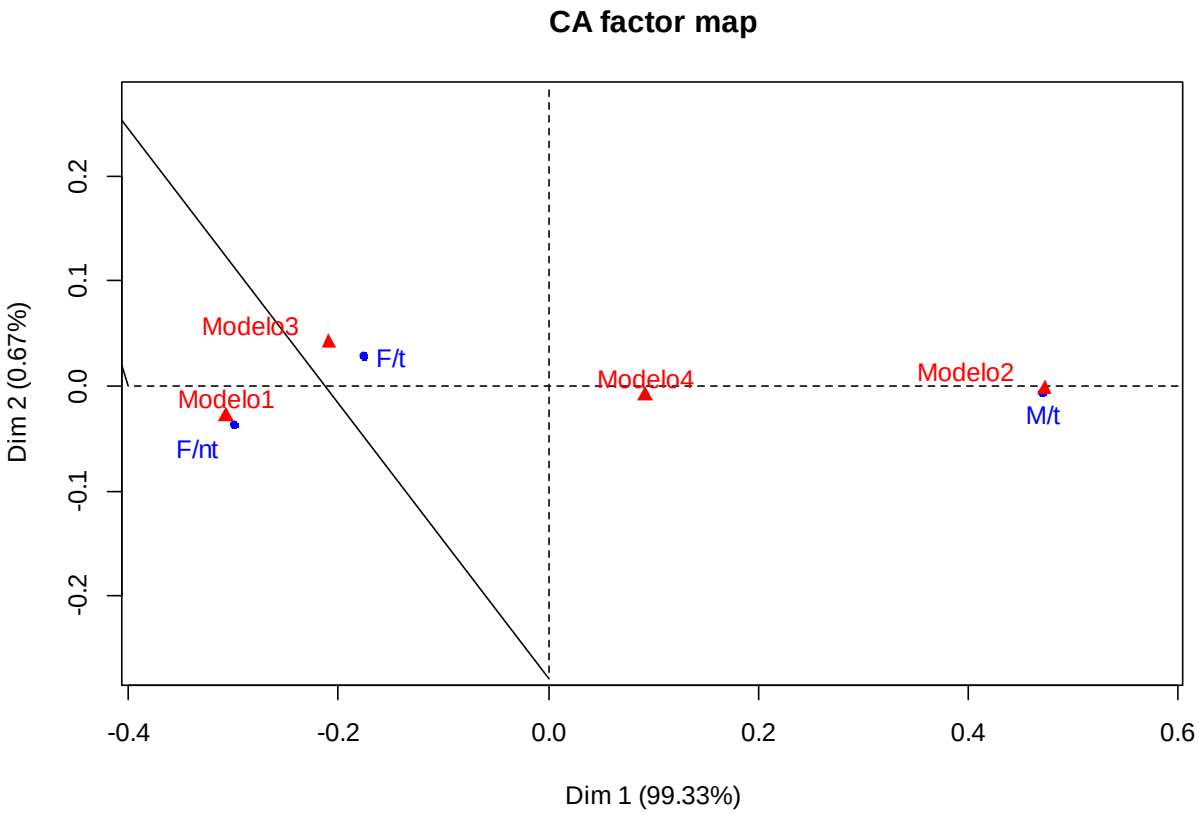


Figura 2: Gráfico da análise de correspondência

F/NT – Feminino/não trabalha; F/T – Feminino/trabalha; M/T – Masculino/trabalha

Analisando o gráfico da Figura 2, pode-se afirmar que:

i) Em relação aos modelos de veículos, o Modelo 1 e 3 apresentam aspectos de similaridade em relação às escolhas, sendo inversamente proporcional ao Modelo 2. Enquanto o Modelo 4 apresenta um menor valor, em módulo, na Dimensão 1, se apresenta mais centralizado entre os citados anteriormente;

ii) Quanto às classes “sexo/trabalho”, nota-se uma semelhança na Dimensão 1 entre as duas classes de sexo Feminino (trabalha e não trabalha), mostrando uma semelhança de escolhas em relação ao sexo. Já a classe do sexo Masculino/trabalha se apresenta distante, na porção positiva da Dimensão 1;

iii), ao ponderar-se sobre a correspondência das variáveis entre si (Modelo de automóvel *vs* pessoas classificadas em sexo/trabalho), nota-se uma estreita relação entre a classe Feminino/não trabalha com o automóvel Modelo 1, entre Feminino/trabalha com o Modelo 3, e entre o Modelo 2 com a classe Masculino/trabalha. O Modelo 4 não apresentou associação clara com nenhuma das 3 classes, ou seja, não foi preferência de nenhuma delas. Assim, considerando também os dois itens anteriores, verifica-se uma correspondência maior das duas classes Feminino com os Modelos 1 e 3, sendo inversamente proporcional à correspondência do Masculino com o Modelo 4;

iv) A Dimensão 2, por sua pequena significância, bem como o pequeno intervalo dos valores no seu eixo, não permite afirmações mais contundentes de semelhanças/diferenças de correspondência entre as variáveis.

**<u>QUESTÃO 3</u>) Um pesquisador está interessado em determinar a influência que alguns fatores químicos exercem em características produtivas do capim elefante. Para isso ele coletou como variáveis respostas que descrevem a característica produtiva da planta: altura e produção de massa seca $ha^{-1}$. Como variáveis independentes ele coletou as seguintes propriedades químicas do solo: Fosforo (P), Potássio (K) e Calcio (Ca). Os resultados na tabela 2 correspondem a 20 parcelas nas quais foram observadas as variáveis acima. Faça a análise de regressão linear multivariada e interprete os resultados obtidos.**

**Tabela 2: Altura, produção de massa seca $ha^{-1}$, e propriedades químicas do solo (Fósforo (P), Potássio (K) e Cálcio (Ca) em parcelas com cultivo de capim elefante.**

| Obs | Altura | Prod | P | K | Ca |
|---|---|---|---|---|---|
| 1 | 9,95 | 130 | 3 | 5 | 10 |
| 2 | 24,45 | 250 | 8 | 11 | 11 |
| 3 | 31,75 | 300 | 11 | 12 | 20 |
| 4 | 35 | 370 | 10 | 15 | 20 |
| 5 | 25,02 | 260 | 8 | 29 | 14 |
| 6 | 16,86 | 170 | 4 | 20 | 15 |
| 7 | 14,38 | 200 | 2 | 37 | 18 |
| 8 | 9,6 | 100 | 2 | 5 | 11 |
| 9 | 24,35 | 245 | 9 | 10 | 13 |
| 10 | 27,5 | 280 | 8 | 30 | 21 |
| 11 | 17,08 | 180 | 4 | 41 | 23 |
| 12 | 37 | 365 | 11 | 40 | 32 |
| 13 | 41,95 | 430 | 12 | 50 | 33 |
| 14 | 11,66 | 120 | 3 | 36 | 18 |
| 15 | 21,65 | 223 | 4 | 21 | 25 |
| 16 | 17,89 | 170 | 4 | 40 | 22 |
| 17 | 69 | 600 | 20 | 60 | 40 |
| 18 | 10,3 | 100 | 2 | 60 | 20 |
| 19 | 34,93 | 360 | 10 | 54 | 22 |
| 20 | 46,59 | 430 | 15 | 25 | 31 |

A análise de regressão linear multivariada tem como objetivo a obtenção de uma relação matemática entre uma das variáveis estudadas com o restante das variáveis. A variável que é estudada é chamada de dependente e as variáveis explicativas são chamadas de independente.

Na análise de regressão linear multivariada, cada variável resposta é explicada pelo seu próprio modelo. A equação básica para uma regressão linear multivariada pode ser expressa por:

$$Y_m = \beta_{0m} + \beta_{1m}x_1 + \beta_{2m}x_2 + \cdots + \beta_{rm}x_r + e_1$$

Onde:

- i=1,...,r representa a i-ésima variável preditora;
- j=1,...,m representa a j-ésima variável resposta;
- βij é o coeficiente associado a i-ésima variável preditora do modelo de regressão linear múltipla que explica a j-ésima variável resposta.

Na tabela 19 encontra-se a estatística descritiva das variáveis respostas realizado com o auxilio do Software R.

**Tabela 19:** Estatística descritiva para as variáveis respostas

| | **Altura** | **Prod** | **P** | **K** | **Ca** |
|---|---|---|---|---|---|
| **Min.** | 9.60 | 100.00 | 2.00 | 5.00 | 10.00 |
| **1st Qu.** | 16.24 | 170.00 | 3.75 | 14.25 | 14.75 |
| **Median** | 24.40 | 247.50 | 8.00 | 29.50 | 20.00 |
| **Mean** | 26.35 | 264.10 | 7.50 | 30.05 | 20.95 |
| **3rd Qu.** | 34.95 | 361.20 | 10.25 | 40.25 | 23.50 |
| **Max.** | 69.00 | 600.00 | 20.00 | 60.00 | 40.00 |

Observa-se que o maior valor de produção foi de 600,00, sendo também esse o valor máximo de, Fósforo, Potássio e Cálcio, apresentando valores de 20,00, 60,00 e 40,00 respectivamente. A maior média de produtividade foi de 264,10 e a mínima de 100,00. Os valores de altura para as 20 parcelas se encontram entre o intervalo de 9,60 a 69,00 cm, com média de 26,35 cm.

A correlação linear entre tais variáveis dependentes e independentes foi calculada a partir da correlação linear de Pearson e a representação gráfica pelo diagrama de dispersão. Segundo Callegari-Jacques (2003) para a interpretação dos coeficientes de correlação (r) é necessário considerar as seguintes magnitudes:

- 0 (Não existe associação entre as variáveis preditoras e descritoras);

- < 0,3 (a associação é muito fraca);
- 0,3 < x ≤ 0,6 (a associação é moderada);
- 0,6 < x ≤ 0,9 (a associação é forte);
- 0,9 < x < 1,0 (a associação é muito forte);
- 1 (a associação é perfeita)

A correlação linear entre as variáveis respostas Y1, Y2 foi de 0.98855, correlação essa considerada muito forte. Esta tendência dos dados pode ser comprovada no gráfico da Figura 3, onde pode-se notar nitidamente um alinhamento dos dados em torno de uma reta, observando que essa correlação é positiva.

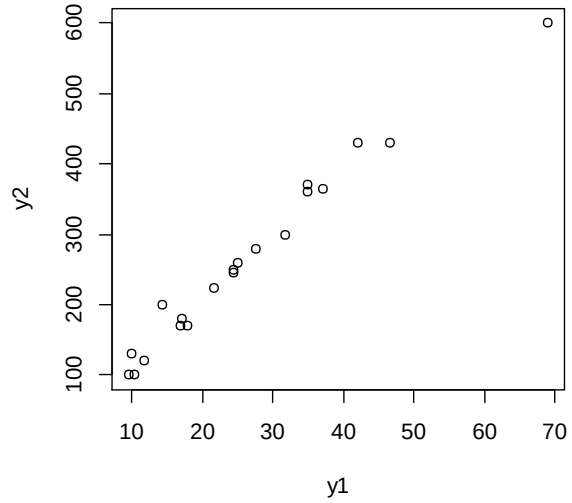


**Figura 3:** Gráfico da correlação linear entre as variáveis respostas

A tabela 20 apresenta a correlação linear entre as variáveis respostas e independentes.

**Tabela 20**: Correlação linear entre as variáveis respostas e independentes.

| | Altura | Prod |
|---|---|---|
| P | 0.9763 | 0.9619 |
| K | 0.3555 | 0.3408 |
| Ca | 0.7930 | 0.7720 |

A figura 4 apresenta os gráficos da correlação linear das variáveis independentes com a variável resposta altura.

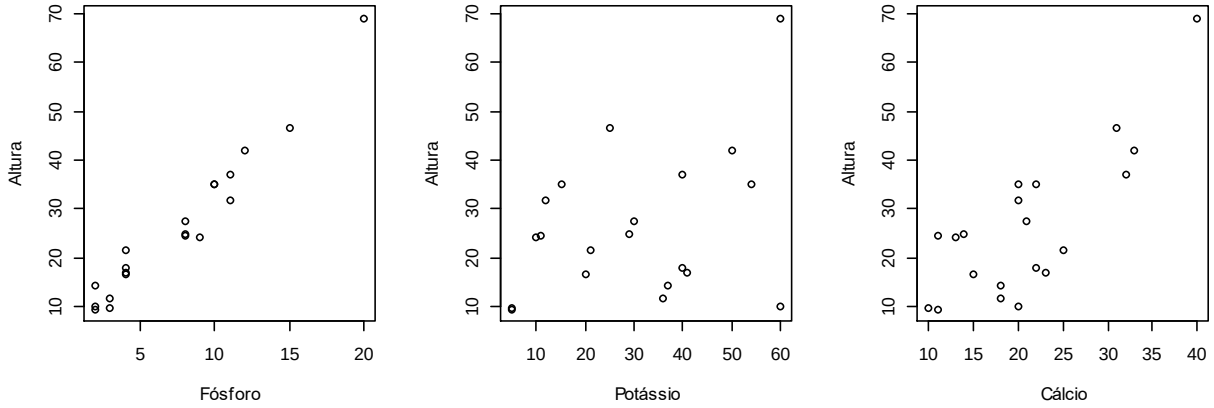


**Figura 4:** Correlação da variável dependente altura, com as variáveis independentes P, K, Ca.

De acordo com a Tabela 20 e com os gráficos acima, a variável resposta altura apresenta forte relação com o Fósforo, com o valor de 0,9763, sendo essa correlação positiva. Com relação da altura com o Potássio, a correlação é considerada fraca, com valor de 0,3555, sendo visualizado no gráfico de dispersão que não apresenta uma reta. Já a relação da variável Altura com o Cálcio, sua correlação linear é moderada, com valor de 0,7930, notando-se no gráfico uma tendência dos dados a uma reta.

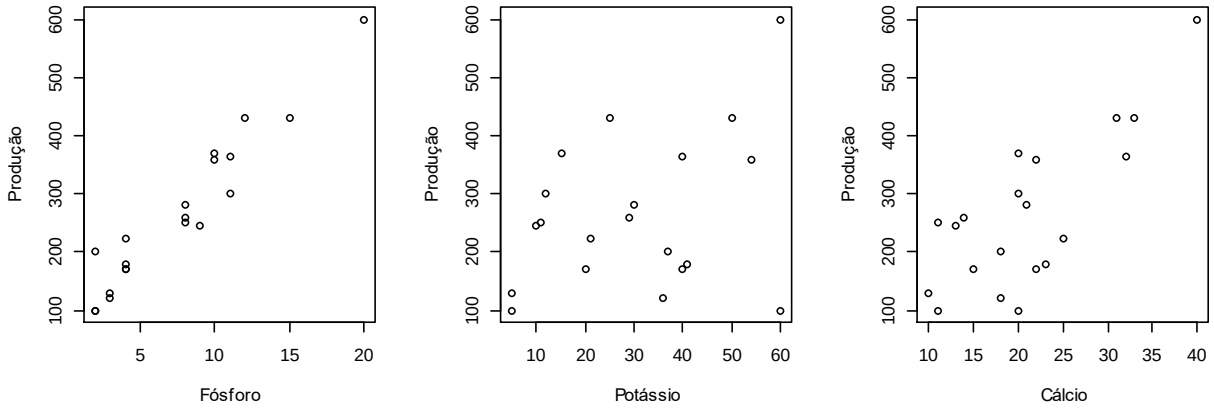


**Figura 5:** Correlação da variável dependente produção com as variáveis independentes P, K, Ca.

Ainda de acordo com a tabela 20, e com a Figura 5, a variável produção apresenta uma correlação muito forte com as variáveis independentes Fósforo com valor de 0,9619, resultado esse, observado na figura 5, onde nota-se uma reta se formando com os dados.

Já a produção apresenta uma correlação fraca com a variável independente Potássio, com valor de 0,3408, sendo notável esse valor no gráfico da figura 5, onde os pontos aparecem todos dispersos. Em relação ao Cálcio, a relação com a Produção apresenta uma correlação linear moderada, com valor 0,7720, onde se observa uma leve tendência dos pontos em formar uma reta.

**Tabela 21**: Correlação linear entre as variáveis independentes

| | P | K | Ca |
|---|---|---|---|
| P | 1.0000 | | |
| K | 0.2403 | 1.0000 | |
| Ca | 0.6908 | 0.6616 | 1.0000 |

Na tabela 21, apresenta a correlação linear entre as variáveis independentes. A correlação linear entre Fósforo e Potássio é uma correlação fraca, com valor de 0,2403. Já a correlação de Fósforo com Cálcio é uma correlação moderada, com valor de 0,6908. Já a relação entre Potássio e Cálcio também é uma correlação moderada de valor 0,6616.

Afim de verificar melhor o comportamento dos dados, realizou-se a verificação de normalidade dos mesmos, sendo empregado os testes de Shapiro-Wilk e Anderson-Darling a 5% de significância. O resultado do p-valor para os dois testes é apresentado na Tabela 22 e as hipóteses de teste são:

*H0 : Os dados possuem distribuição normal de probabilidade;*

*H1 : Os dados não possuem distribuição normal de probabilidade.*

**Tabela 22**: Resultado dos testes de normalidade.

| | **p-valor** | |
|---|---|---|
| **Teste** | **Shapiro-Wilk** | **Anderson-Darling** |
| **P** | 0.0445 | 0.0726 |
| **K** | 0.2992 | 0.5274 |
| **Ca** | 0.2018 | 0.2290 |

Rejeita-se H0 em caso do p-valor menor que 5% de significância. Sendo assim, Pelo teste de Shapiro-Wilk, K e Ca possuem distribuição normal de probabilidade, enquanto P não. Porém, como no teste de Anderson-Darling não há evidências estatísticas para rejeitar H0 ao nível de 5% de significância para as três variáveis independentes, isto é, apresentam p-valor > 0,05, considerou-se para essa análise que os dados possuem distribuição normal de probabilidade.

Parte-se então para o cálculo dos coeficientes de interesse da análise de regressão multivariada. A tabela 23 apresenta os resultados do teste de mínimos quadrados para o conjunto dos dados, onde nota-se que 94,79% da variação ocorre nas variáveis respostas. Esse resultado permite a explicação pelo modelo de regressão linear multivariada envolvendo as variáveis preditoras.

**Tabela 23**: Resultado do teste de mínimos quadrados.

| | N | DF | SSR | detRCov | OLS-R2 | McElroy-R2 |
|---|---|---|---|---|---|---|
| system | 40.00 | 32.00 | 17078.10 | 2225.73 | 0.9479 | 0.9637 |

Em seguida calculou-se valores dos coeficientes β, estimativas, erro padrão, valores t, valores de probabilidade, para as variáveis resposta Y1 (Tabela 24) e Y2 (Tabela 25), com as seguintes hipóteses:

$$H_0: X_1, X_2, X_3 = 0$$

$$H_1: Pelo\ menos\ um\ coeficiente\ \neq 0$$

**Tabela 24**: Modelo de regressão linear multivariada para a variável dependente Altura (Y1).

| | Estimativa | Desv. Pad. | t valor | Pr(>\|t\|) | Significância |
|---|---|---|---|---|---|
| **(Intercept)** | -1.1197 | 1.4656 | -0.7640 | 0.4560 | |
| **x1** | 2.5212 | 0.1617 | 15.5880 | 0.0000 | *** |
| **x2** | 0.0151 | 0.0430 | 0.3520 | 0.7298 | |
| **x3** | 0.3867 | 0.1266 | 3.0550 | 0.0076 | ** |

Níveis de significância: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

**Tabela 25**: Modelo de regressão linear multivariada para a variável dependente Produção (Y2).

| | Estimativa | Desv. Pad. | t valor | Pr(>\|t\|) | Significância |
|---|---|---|---|---|---|
| **(Intercept)** | 30.3913 | 20.9339 | 1.4520 | 0.17 | |
| **x1** | 22.1023 | 2.3102 | 9.5670 | 0.00 | *** |
| **x2** | 0.1217 | 0.6147 | 0.1980 | 0.85 | |
| **x3** | 3.0708 | 1.8082 | 1.6980 | 0.11 | |

Níveis de significância: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Nota-se que em ambas as tabelas acima, todas as estimativas dos coeficientes são diferentes de zero, logo rejeita-se H0. Os coeficientes X1 e X3 são significativos ao nível de 5% para a variável Y1, enquanto apenas X1 é significativo para a variável Y2.

**Tabela 26**: Resultado do teste de mínimos quadrados para as variáveis respostas

| | N | DF | SSR | MSE | RMSE | R2 | Adj R2 |
|---|---|---|---|---|---|---|---|
| **Altura** | 20.00 | 16.00 | 83.30 | 5.21 | 2.28 | 0.9802 | 0.9765 |
| **Prod** | 20.00 | 16.00 | 16994.81 | 1062.18 | 32.59 | 0.9475 | 0.9377 |

A Tabela 26, em complemento as duas anteriores, apresenta os valores do coeficiente de determinação ($R^2$ e $R^2$ ajustado) para cada uma das variáveis dependentes analisadas, onde verifica-se que esse modelo linear explica adequadamente acima de 97% das variações da variável Altura, bem como acima de 93% da variável Produção.

A Tabela 27 apresenta de forma resumida os coeficientes lineares calculados pelo método de regressão, possibilitando a construção das equações.

**Tabela 27**: Valores de coeficientes ajustados para os modelos de regressão multivariada pelo método de mínimos quadrados (OLS)

| | (Intercept) | x1 | x2 | x3 |
|---|---|---|---|---|
| | **β0** | **β1** | **β2** | **β3** |
| **Altura** | -1.1197 | 2.5212 | 0.0151 | 0.3867 |
| **Prod** | 30.3913 | 22.1023 | 0.1217 | 3.0708 |

O modelo de regressão linear (Y1) para a primeira variável dependente (Altura) fica ajustado então como:

$$\boldsymbol{Y_1} = -1{,}1197 + 2{,}5212 * \boldsymbol{X_1} + 0{,}0151 * \boldsymbol{X_2} + 0{,}3867 * \boldsymbol{X_3}$$

Analisando a equação acima, pode-se afirmar que a variável X1 (P, ou seja, Fósforo) apresenta maior interferência na variável resposta Altura, apresentando o maior coeficiente, cujo se traduz em aumento de 2,52 u.m. na altura para cada unidade de P.

De forma análoga, o modelo de regressão linear (Y2) para a segunda variável dependente (Prod) fica ajustado da seguinte maneira:

$$\boldsymbol{Y_2} = 30{,}3913 + 22{,}1013 * \boldsymbol{X_1} + 0{,}1217 * \boldsymbol{X_2} + 3{,}0708 * \boldsymbol{X_3}$$

Aqui, nessa equação, nota-se o comportamento semelhante aos coeficientes da equação Y1, o que se justifica uma vez que ambas as variáveis respostas Y1 e Y2 apresentam correlação muito forte e comportamento parecido. Assim, a variável que mais influencia na resposta Prod é X1, onde cada unidade de P eleva em 22,10 u.m. a Produção. A variável que menos influencia nas duas equações é X2 (K, ou seja, Potássio).

Tendo sido o modelo de regressão linear ajustado, calcula-se então os valores estimados das variáveis respostas para cada parcela, e compara-se com os valores originais para evidenciar-se o erro (resíduo), o que está apresentado na Tabela 28 a seguir.

**Tabela 28**: Valores reais e estimados para as variáveis respostas

| Parcela | Altura | Altura (estimado) | Erro | Prod | Prod (Estimado) | Erro |
|---|---|---|---|---|---|---|
| 1 | 9.95 | 10.39 | -0.44 | 130.00 | 128.02 | 1.99 |
| 2 | 24.45 | 23.47 | 0.98 | 250.00 | 242.33 | 7.67 |
| 3 | 31.75 | 34.53 | -2.78 | 300.00 | 336.39 | -36.39 |
| 4 | 35.00 | 32.05 | 2.95 | 370.00 | 314.66 | 55.34 |
| 5 | 25.02 | 24.90 | 0.12 | 260.00 | 253.73 | 6.27 |
| 6 | 16.86 | 15.07 | 1.79 | 170.00 | 167.30 | 2.70 |
| 7 | 14.38 | 11.44 | 2.94 | 200.00 | 134.37 | 65.63 |
| 8 | 9.60 | 8.25 | 1.35 | 100.00 | 108.98 | -8.98 |
| 9 | 24.35 | 26.75 | -2.40 | 245.00 | 270.45 | -25.45 |
| 10 | 27.50 | 27.62 | -0.12 | 280.00 | 275.35 | 4.65 |
| 11 | 17.08 | 18.48 | -1.40 | 180.00 | 194.42 | -14.42 |
| 12 | 37.00 | 39.59 | -2.59 | 365.00 | 376.65 | -11.65 |
| 13 | 41.95 | 42.65 | -0.70 | 430.00 | 403.04 | 26.96 |
| 14 | 11.66 | 13.95 | -2.29 | 120.00 | 156.36 | -36.36 |
| 15 | 21.65 | 18.95 | 2.70 | 223.00 | 198.13 | 24.87 |
| 16 | 17.89 | 18.08 | -0.19 | 170.00 | 191.23 | -21.23 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 17 | 69.00 | 65.68 | 3.32 | 600.00 | 602.57 | -2.57 |
| 18 | 10.30 | 12.56 | -2.26 | 100.00 | 143.32 | -43.32 |
| 19 | 34.93 | 33.42 | 1.51 | 360.00 | 325.55 | 34.45 |
| 20 | 46.59 | 49.06 | -2.47 | 430.00 | 460.16 | -30.16 |

A qualidade do ajuste do modelo pode ser analisada com base nos menores valores possíveis de erros (resíduos). Na tabela 28 é possível evidenciar alguns valores de erro relativamente altos para a variável Produção, sendo o maior valor de erro 65.63 u.m. (32,82%), enquanto o maior erro 3.32 u.m. (4,81%) para a variável Altura.

A tabela 29 apresenta os valores de intervalo de confiança para as estimativas dos coeficientes, calculados a partir do método de estimativa. Esses intervalos indicam que, com 95% de confiança, um valor aleatório das variáveis respostas esteja contido nos intervalos citados. Verifica-se que os valores dos coeficientes estimados estão dentro dos intervalos.

**Tabela 29**: Intervalos de confiança para as estimativas dos coeficientes.

| | | **2.50%** | **97.50%** |
|---|---|---|---|
| **Altura** | (Intercept) | -4.227 | 1.987 |
| | x1 | 2.178 | 2.864 |
| | x2 | -0.076 | 0.106 |
| | x3 | 0.118 | 0.655 |
| **Prod** | (Intercept) | -13.987 | 74.769 |
| | x1 | 17.205 | 27.000 |
| | x2 | -1.181 | 1.425 |
| | x3 | -0.762 | 6.904 |

Por fim, para verificar se as variáveis preditoras são significativas para explicar as variáveis respostas, ou seja, se exercem influência sobre as mesmas, aplicou-se o teste da razão de verossimilhança (Tabela 30), com critério de Lambda de Wilks, e com as seguintes hipóteses:

*H0 : $\beta_{(2)} = 0$, as variáveis independentes (P,K,Ca) não exercem influência sobre as variáveis dependentes;*

*H1 : $\beta_{(2)} \neq 0$, as variáveis independentes (P,K,Ca) exercem influência sobre as variáveis dependentes.*

**Tabela 30**: Teste de Verossimilhança.

| | #Df | LogLik | Df | Chisq | Pr(>Chisq) | Signif. |
|---|---|---|---|---|---|---|
| Modelo 1 | 9 | -129.370 | | | | |
| Modelo 2 | 3 | -169.350 | -6.000 | 79.9490 | 0.0000 | *** |

Níveis de significância: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

O resultado apresentado na Tabela 30 permite apontar que as variáveis independentes são significativas para explicar a produtividade (Altura e Produção) do capim elefante no presente estudo, uma vez que com 5% de significância (p-valor < 0,05), rejeita-se a H0 e aceita-se H1.

## REFERÊNCIAS BIBLIOGRÁFICAS


CALLEGARI-JACQUES, S.M. **Bioestatística – Princípios e Aplicações**. Porto Alegre: Artmed, 2003.

INFANTOSI, A. F. C.; COSTA, J. C. G. D.; ALMEIDA, R. M. V. R. **Análise de Correspondência: bases teóricas na interpretação de dados categóricos em Ciências da Saúde**. Cad. Saúde Pública, Rio de Janeiro, 30(3):473-486, mar, 2014.

SILVA et al. Análise de correlação canônica na descrição de potenciais de desenvolvimento nos municípios de Minas Gerais. **Revista de Estatística UFOP**. Volume III (3), 2014.

## ANEXO A – Script software R (Questão 1)

```
### QUESTÃO 1 ###

## Análise de correlação canônica

## Dados ##

require(MVar.pt)
help(DataMix)
data(DataMix)
data.frame(DataMix)

dados <- data.frame(cbind(DataMix[,2,],DataMix[,3,],DataMix[,6,],DataMix[,7,]))
dados

attach(dados)
dim(dados)
names(dados)


cor <- cor(dados)
cor

G1 <- as.matrix(cbind(DataMix[,2,],DataMix[,3,]))
G2 <- as.matrix(cbind(DataMix[,6,],DataMix[,7,]))

corca <- cancor(G1,G2)
corca

## Manualmente ##

r11 <- cor[1:2,1:2]
r11

r22 <- cor[3:4,3:4]
r22

r12 <- cor[1:2,3:4]
r12

A <- (solve(r11)) %*% r12 %*% (solve(r22)) %*% (t(r12))
A

B <- (solve(r22)) %*% (t(r12)) %*% (solve(r11)) %*% r12
B

eigen1 <- eigen(A)
eigen2 <- eigen(B)

eigen1
eigen2

sqrt(eigen1$values)      ## correlação canônica ##
sqrt(eigen2$values)      ## correlação canônica ##
```

```
## estimativas das variáveis canônicas ##
## G1: médias, anos

G1m <- (X1 - mean(X1))/(sd(X1))
G1m

G1a <- (X2 - mean(X2))/(sd(X2))
G1a

z1 <- as.matrix(cbind(G1m,G1a))
z1

a1 <- as.matrix(eigen1$vectors[,1])
a1

u1 <- z1 %*% a1
u1

a2 <- as.matrix(eigen1$vectors[,2])
a2

u2 <- z1 %*% a2
u2

cbind(u1,u2)


## G2: especiais, comerciais

G2e <- (X3 - mean(X3))/(sd(X3))
G2e

G2c <- (X4 - mean(X4))/(sd(X4))
G2c

z2 <- as.matrix(cbind(G2e,G2c))
z2

b1 <- as.matrix(eigen2$vectors[,1])
b1

v1 <- z2 %*% b1
v1

b2 <- as.matrix(eigen2$vectors[,2])
b2

v2 <- z2 %*% b2
v2

cbind(v1,v2)

## Correlação entre U e V

cor(u1,v1)    ## correlação canônica ##
cor(u1,v2)    ## correlação canônica ##
```

```
cor(u2,v1)    ## correlação canônica ##
cor(u2,v2)    ## correlação canônica ##


par(mfrow = c(2,2))
plot(u1,v1)
plot(u1,v2)
plot(u2,v1)
plot(u2,v2)


## correlação das variáveis canônicas U com as originais ##

cor(u1,X1); cor(u1,X2)
cor(u1,G1m); cor(u1,G1a) #Igual a anterior

cor(u2,X1); cor(u2,X2)
cor(u2,G1m); cor(u2,G1a) #Igual a anterior

cor(u1,X3); cor(u1,X4)
cor(u1,G2e); cor(u1,G2c) #Igual a anterior

cor(u2,X3); cor(u2,X4)
cor(u2,G2e); cor(u2,G2c) #Igual a anterior

rbind(cor(u1,G1m), cor(u1,G1a), cor(u2,G1m), cor(u2,G1a),
cor(u1,G2e), cor(u1,G2c), cor(u2,G2e), cor(u2,G2c))


## correlação das variáveis canônicas V com as originais ##

cor(v1,X3); cor(v1,X4)
cor(v1,G2e); cor(v1,G2c) #Igual a anterior

cor(v2,X1); cor(v2,X2)
cor(v2,G1m); cor(v2,G1a) #Igual a anterior

cor(v1,X1); cor(v1,X2)
cor(v1,G1m); cor(v1,G1a) #Igual a anterior

cor(v2,X3); cor(v2,X4)
cor(v2,G2e); cor(v2,G2c) #Igual a anterior

rbind(cor(v1,G2e), cor(v1,G2c), cor(v2,G1m), cor(v2,G1a),
cor(v1,G1m), cor(v1,G1a), cor(v2,G2e), cor(v2,G2c))


## medida de qualidade do modelo ##

p <- 2  ## nº variáveis no 1º grupo ##
q <- 2  ## nº variáveis no 2º grupo ##


cor_u1 <- (cor(u1,G1m)^2) + (cor(u1,G1a)^2)
prop_u1 <- 100 * (cor_u1/p)
prop_u1
```

```
cor_v1 <- (cor(v1,G2e)^2) + (cor(v1,G2c)^2)
prop_v1 <- 100 * (cor_v1/p)
prop_v1
```

## ANEXO B – Script software R (Questão 2)

```
### QUESTÃO 2 ###

## Dados ##

dados <- read.table("dados_questao2.txt", header = T)   ## lendo um conjunto de dados em txt ##
dados

attach(dados)

tab <- matrix(NumPess, ncol = 4, nrow = 3, byrow = T)
tab

colnames(tab) <- c("Modelo1", "Modelo2","Modelo3", "Modelo4")
rownames(tab) <- c("F/nt","F/t","M/t")
tab

## teste Qui-Quadrado ##

qui <-chisq.test(tab)
qui


## ANÁLISE DE CORRESPONDÊNCIA ##

n <- sum(tab)
n


P <- tab/n        ## matriz de correspondência ##
P

r1 <- sum(tab[1,])/n
r2 <- sum(tab[2,])/n
r3 <- sum(tab[3,])/n

r <- c(r1,r2,r3)       ## vetor coluna ##
r

c1 <- sum(tab[,1])/n
c2 <- sum(tab[,2])/n
c3 <- sum(tab[,3])/n
c4 <- sum(tab[,4])/n

c <- c(c1,c2,c3,c4)     ## vetor linha ##
c

Dr <- diag(r)
Dr

Dc <- diag(c)
Dc

sDr <- diag((r^(-0.5)))    ## Dr^(-1/2) ##
sDr
```

```
sDc <- diag((c^(-0.5)))    ## Dc^(-1/2) ##
sDc

R <- sDr %*% P %*% sDc
R

## COORDENADAS PRINCIPAIS DAS LINHAS ##

W <- t(R) %*% R
W

eigen(W)        ## inércias: 2º e 3º autovalores ##

an <- diag((r^(-1))) %*% P %*% diag((c^(-0.5)))
an

score1 <- an %*% eigen(W)$vectors
score1          ## 2ª e 3ª colunas são as coordenadas principais das  linhas ##


## COORDENADAS PRINCIPAIS DAS COLUNAS ##

T <- R %*% t(R)
T

eigen(T)        ## inércias: 2º e 3º autovalores ##

an2 <- diag((c^(-1))) %*% t(P) %*% diag((r^(-0.5)))
an2

score2 <- an2 %*% eigen(T)$vectors
score2          ## 2ª e 3ª colunas são as coordenadas principais das colunas ##


## PACOTE FactoMineR ##

require(FactoMineR)
corresp <- CA(tab)      ## roda apenas para nº de categorias por variável acima de 2 ##
summary(corresp)
```

## ANEXO C – Script software R (Questão 3)

```
### Questão 3 ###

## Análise de regressão linear multivariada

## Dados ##

dados <- read.table("dados_questao3.txt", header = T)
dados
attach(dados)
names(dados)

summary(dados)

x <- cbind(dados$P,dados$K,dados$Ca) #variaveis preditoras
x

y1 <- dados$Altura #primeira variavel resposta = Altura
y1
y2 <- dados$Prod #segunda variavel resposta = Produção
y2

## Correlação entre as variáveis respostas ##

cor(y1,y2)
plot(y1,y2)


# Correlação entre as variáveis respostas e independentes #
## Y1
cor(P,y1)
cor(K,y1)
cor(Ca,y1)

par(mfrow = c(1,3))
plot(P,y1,xlab = "Fósforo", ylab = "Altura")
plot(K,y1,xlab = "Potássio", ylab = "Altura")
plot(Ca,y1,xlab = "Cálcio", ylab = "Altura")

## Y2
cor(P,y2)
cor(K,y2)
cor(Ca,y2)

par(mfrow = c(1,3))
plot(P,y2,xlab = "Fósforo", ylab = "Produção")
plot(K,y2,xlab = "Potássio", ylab = "Produção")
plot(Ca,y2,xlab = "Cálcio", ylab = "Produção")


# Correlação entre as variáveis independentes #

cor(P,K)
cor(P,Ca)
cor(K,Ca)
```

```
## Teste de normalidade dos dados - variáveis independentes ##

# TESTE DE SHAPIRO-WILKS UNIVARIADO #

shapiro.test(P)
shapiro.test(K)
shapiro.test(Ca)

# TESTE DE Anderson-Darling #

require(mvsf)
ad.test(P)
ad.test(K)
ad.test(Ca)


## Coeficientes da regressão linear ##

require(systemfit)

eq1 <- y1 ~ x    ## altura em função das variáveis independentes
eq2 <- y2 ~ x    ## produção em função das variáveis independentes

eqSystem <- list(Altura = eq1, Prod = eq2)

fit_ols <- systemfit(eqSystem)

fit_ols            ## estimativas dos coeficientes ##

model <- lm(cbind(y1,y2) ~ x) # lm-linear model, variavel resposta em função das preditoras
model                                        # apresenta os coeficiente associadas as variaveis preditoras

summary(model)
summary(fit_ols)

names(fit_ols)
fit_ols$coefCov       ## matriz de variâncias e covariâncias das estimativas dos coeficientes ##

fitted(fit_ols)       ## valores estimados ##

cbind(y1,y2,fitted(fit_ols)) #tabela com os valores reais e estimados#

e <- as.matrix(residuals(fit_ols))     ## resíduos ##
e
ee <- t(e) %*% e
ee/(length(y1)-1-1)              ## estimativa matriz covariâncias resíduos ##

fit_ols$residCov                 ## estimativa matriz covariâncias resíduos ##


X <- cbind(rep(1,length(x)),x)
X

invXX <- solve(t(X) %*% X)
invXX
```

```
fit_ols$residCov[1,1] * invXX
fit_ols$residCov[2,2] * invXX
fit_ols$coefCov              ## matriz covariâncias estimativas dos coeficientes ##


confint(fit_ols)   ## intervalos de confiança para as estimativas coeficientes ##


## TESTE SE x É SIGNIFICATIVO NO MODELO ##

fit_ols            ## estimativas dos coeficientes com x no modelo##

eq12 <- y1 ~ 1
eq22 <- y2 ~ 1

eq12
eq22

eq2System <- list(Altura = eq12, Prod = eq22)

fit_ols2 <- systemfit(eq2System)

fit_ols2           ## estimativas dos coeficientes ##
mean(y1)
mean(y2)
summary(fit_ols2)

## teste da razao de verossimilhanca comparando os dois modelos ##

lrtest(fit_ols, fit_ols2) # teste de Lambda de Wilks #
# p-valor significativo 5% = variaveis significativas para explicar produtividade
```