Lista 1 - MAE0330

Guilherme NºUSP: 8943160 e Leonardo NºUSP: 9793436

Exercício 1

Em um estudo 24 tanques de água foram aleatoriamente distribuídos a 4 grupos (cada grupo recebeu 6 tanques). Os tanques de cada grupo foram tratados com os reagentes T1, T2, T3 e T4, respectivamente. Os dados a seguir referem-se a medidas de clorofila (Y_1) e oxigênio dissolvido na água (Y_2), mensuradas na superfície dos 24 tanques experimentais. Construa a tabela de MANOVA para estudar o efeito do tratamento para o conjunto das variáveis observadas.

Comente os resultados.

T_1		T_2		T_3		T_4	
Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
6.2	432	12.7	431	7	522	8.3	600
4.8	405	11.3	426	4.4	513	7.1	513
3.0	324	9.3	438	3.8	507	4.7	539
5.6	310	9.5	312	5.0	410	10.0	456
7.4	326	11.7	326	5.5	350	8.5	504
4.8	375	15.3	447	3.2	547	12.4	548

Resolução

Tabela de Manova:

```
##
## Type II MANOVA Tests:
##
## Sum of squares and products for error:
##          tanquey1    tanquey2
## tanquey1 78.78167    97.38333
## tanquey2 97.38333 72195.50000
##
## ----------------------------------------
##
## Term: factor(reagentes)
##
## Sum of squares and products for the hypothesis:
##           tanquey1     tanquey2
## tanquey1 180.65458    -85.94583
## tanquey2 -85.94583 100117.45833
##
## Multivariate Tests: factor(reagentes)
##                  Df test stat approx F num Df den Df     Pr(>F)
## Pillai            3  1.277404 11.78532      6     40 1.4285e-07 ***
## Wilks             3  0.127017 11.43721      6     38 2.8195e-07 ***
## Hotelling-Lawley  3  3.688953 11.06686      6     36 5.6910e-07 ***
## Roy               3  2.311499 15.41000      3     20 1.9871e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nós podemos observar com a tabela de MANOVA que para as duas variáveis reposta (medidas de clorofila e oxigênio dissolvido na água) temos efeitos significantes de tratamento (fixando um nível de significância de 5%) (em todos os testes realizados), para fazer a verificação das suposições do modelo, iremos realizar o teste igualdade de matrizes de covariaĉnias de M-Box e o de normalidade multivariada baseado na assimetria, assim:

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  tanq[1:2]
## Chi-Sq (approx.) = 4.8634, df = 9, p-value = 0.846
```

```
##
##  Multivariate Normality Test Based on Skewness
##
## data:  mnv$res
## U = 1.1894, df = 2, p-value = 0.5517
```

Onde podemos verificar que as suposições de normalidade e homocedasticidade estão satisfeitas, e como encontramos significância estatística nas comparações iremos realizar as comparações múltiplas com a correção de Bonferroni para um α global de 5%, a fim de identificar onde estão as diferenças:

Para o tratamento medidas de clorofila

Diferença entre o Tratamento 1 e 2

```
## [1] -24.51139  11.84472
```

Diferença entre o Tratamento 1 e 3

```
## [1] -17.69472  18.66139
```

Diferença entre o Tratamento 1 e 4

```
## [1] -21.37805  14.97805
```

Diferença entre o Tratamento 2 e 3

```
## [1] -11.36139  24.99472
```

Diferença entre o Tratamento 2 e 4

```
## [1] -15.04472  21.31139
```

Para o tratamento de oxigênio dissolvido na água

Diferença entre o Tratamento 1 e 2

```
## [1] -343.2672  273.9339
```

Diferença entre o Tratamento 1 e 3

```
## [1] -421.4339  195.7672
```

Diferença entre o Tratamento 1 e 4

```
## [1] -473.2672  143.9339
```

Diferença entre o Tratamento 2 e 3

```
## [1] -386.7672  230.4339
```

Diferença entre o Tratamento 2 e 4

```
## [1] -438.6006  178.6006
```

Exercício 2

O plantio de amendoim é bastante importante nos Estados Unidos. Pesquisadores têm muito interesse em desenvolver variedades melhoradas para o plantio, e é bastante comum o delineamento de experimentos para comparar variedades. Os dados apresentados no arquivo T6-17.DAT são referentes a três diferentes variedades do amendoim, plantados em duas diferentes localidades.

As variáveis observadas são:

- X_1: Produção (peso);
- X_2: grãos maduros (peso, em gramas);
- X_3: tamanho da semente (peso, em gramas de 100 sementes).

No arquivo, a primeira coluna refere-se a localização, a segunda é a variedade e as outras são X_1, X_2 e X_3, nesta ordem.

(a) Faça a análise utilizando MANOVA de dois fatores. Teste o efeito de localização, variedade e interação entre variedade e localização.

Resolução

Fazendo uma Manova, obtemos a seguinte tabela:

```
##
## Type II MANOVA Tests:
##
## Sum of squares and products for error:
##         X1      X2      X3
## X1 104.205  49.365  76.480
## X2  49.365 352.105 121.995
## X3  76.480 121.995  94.835
##
## ------------------------------------------
##
## Term: fat1
##
```

```
## Sum of squares and products for the hypothesis:
##             X1        X2          X3
## X1   0.7008333  -10.6575    7.129167
## X2 -10.6575000  162.0675 -108.412500
## X3   7.1291667 -108.4125   72.520833
##
## Multivariate Tests: fat1
##                  Df test stat approx F num Df den Df   Pr(>F)
## Pillai            1  0.893484 11.18432      3      4 0.020502 *
## Wilks             1  0.106516 11.18432      3      4 0.020502 *
## Hotelling-Lawley  1  8.388243 11.18432      3      4 0.020502 *
## Roy               1  8.388243 11.18432      3      4 0.020502 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ----------------------------------------
##
## Term: fat2
##
## Sum of squares and products for the hypothesis:
##          X1        X2       X3
## X1 196.1150  365.1825  42.6275
## X2 365.1825 1089.0150 414.6550
## X3  42.6275  414.6550 284.1017
##
## Multivariate Tests: fat2
##                  Df test stat  approx F num Df den Df    Pr(>F)
## Pillai            2  1.709109  9.792388      6     10 0.0010562 **
## Wilks             2  0.012444 10.619086      6      8 0.0019275 **
## Hotelling-Lawley  2 21.375675 10.687838      6      6 0.0054869 **
## Roy               2 18.187611 30.312685      3      5 0.0012395 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ----------------------------------------
##
## Term: fat1:fat2
##
## Sum of squares and products for the hypothesis:
##          X1       X2        X3
## X1 205.1017 363.6675 107.78583
## X2 363.6675 780.6950 254.22000
## X3 107.7858 254.2200  85.95167
##
## Multivariate Tests: fat1:fat2
##                  Df test stat  approx F num Df den Df   Pr(>F)
## Pillai            2  1.290861  3.033867      6     10 0.058708 .
## Wilks             2  0.074300  3.558197      6      8 0.050794 .
## Hotelling-Lawley  2  7.544290  3.772145      6      6 0.065517 .
## Roy               2  6.824094 11.373490      3      5 0.011340 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

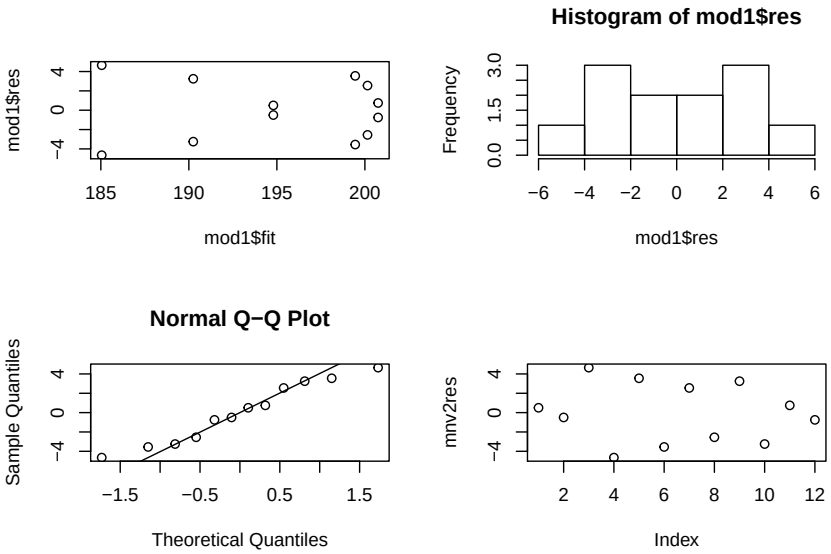
Onde ao fixarmos um nível de significância de 5%, podemos notar os fatores localização e variedade são estatísticamente significante, porém o efeito de interação não é estatísticamente significante.

(b) Faça análise de resíduos do modelo selecionado no item (a).

Resolução

Fazendo uma análise de resíduos onde estamos supondo um modelo de Anova para cada variável, temos:

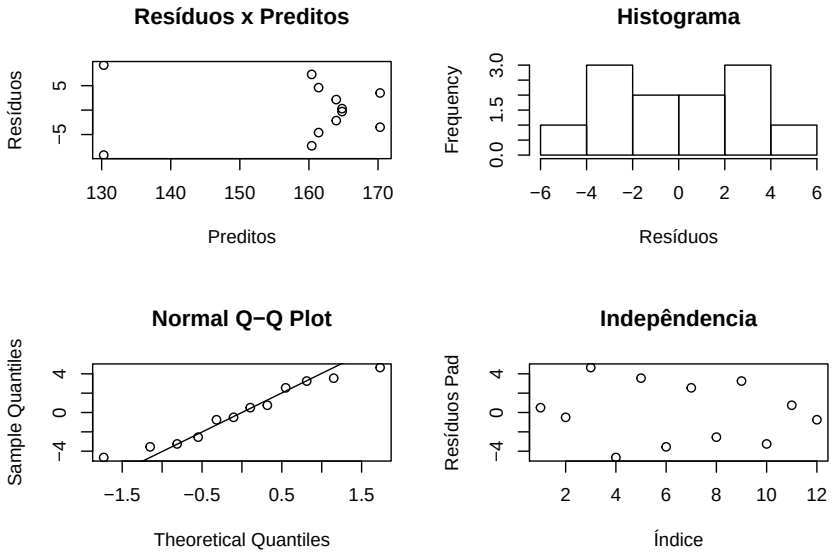
Análise de resíduos para X_1: Produção (peso) como variável resposta




```
##
##  Shapiro-Wilk normality test
##
## data:  mnv2res
## W = 0.95124, p-value = 0.6552

##
##  Bartlett test of homogeneity of variances
##
## data:  X1 by trat
## Bartlett's K-squared = 3.6671, df = 5, p-value = 0.5983
```

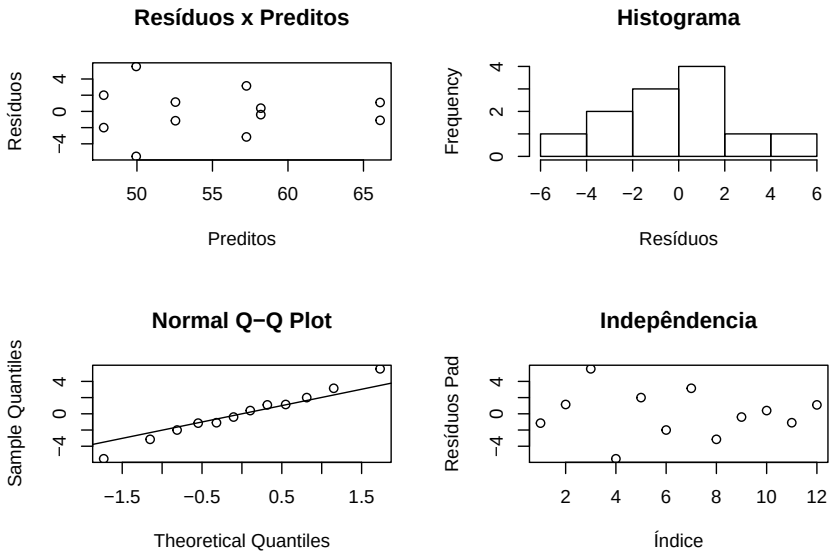
Análise de resíduos para X_1: Produção (peso) como variável resposta




```
##
##  Shapiro-Wilk normality test
##
## data:  mod2res
## W = 0.95124, p-value = 0.6552

##
##  Bartlett test of homogeneity of variances
##
## data:  X2 by trat
## Bartlett's K-squared = 5.1691, df = 5, p-value = 0.3956
```

Análise de resíduos para X_3: tamanho da semente (peso, em gramas de 100 sementes) como variável resposta




```
##
##  Shapiro-Wilk normality test
##
## data:  mod3res
## W = 0.99274, p-value = 1

##
##  Bartlett test of homogeneity of variances
##
## data:  X3 by trat
## Bartlett's K-squared = 4.7931, df = 5, p-value = 0.4416
```

Assim como podemos ver nos gráficos de resíduos e testes temos todas as suposições (normalidade, independência e homocedasticidade) satisfeitas.

(c) Faça a análise dos dados utilizando 3 ANOVAS univariadas de dois fatores. Compare os resultados.

Resolução

Considerando um primeiro modelo de Anova com o X_1: Produção (peso) e os fatores localização e variedade obtermos a seguinte tabela:

Table 1: Anova Produção

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
fat1	1	0.70	0.70	0.04	0.85
fat2	2	196.11	98.06	5.65	0.04
fat1:fat2	2	205.10	102.55	5.90	0.04
Residuals	6	104.20	17.37	NA	NA

Com a tabela de Anova acima podemos concluir que considerando a produção (peso), não temos efeito de localização, porém o efeito de variação dos graõs e a interação entre eles é sifnificante. Assim abaixo segue as comparações múltiplas com a correção de Bonferroni para um α global de 5%

Considerando um primeiro modelo de Anova com o X_2: grãos maduros (peso, em gramas) e os fatores localização e variedade obtermos a seguinte tabela:

Table 2: Anova grãos maduros

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
fat1	1	162.07	162.07	2.76	0.15
fat2	2	1089.01	544.51	9.28	0.01
fat1:fat2	2	780.70	390.35	6.65	0.03
Residuals	6	352.10	58.68	NA	NA

Com a tabela de Anova acima podemos concluir que considerando os grãos maduros (peso), não temos efeito de localização, porém o efeito de variação dos graõs e a interação entre eles é sifnificante. Assim abaixo segue as comparações múltiplas com a correção de Bonferroni para um α global de 5%

Considerando um primeiro modelo de Anova com o X_3: tamanho da semente (peso, em gramas de 100 sementes) e os fatores localização e variedade obtermos a seguinte tabela:

Table 3: Anova tamanho da semente

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
fat1	1	72.52	72.52	4.59	0.08
fat2	2	284.10	142.05	8.99	0.02
fat1:fat2	2	85.95	42.98	2.72	0.14
Residuals	6	94.83	15.81	NA	NA

Com a tabela de Anova acima podemos concluir que considerando tamanho da semente (peso, em gramas de 100 sementes), não temos efeito de localização e a interação entre eles, porém o efeito de variação dos graõs é estatísticamente significante. Assim abaixo segue as comparações múltiplas com a correção de Bonferroni para um α global de 5%

Exercício 3

3. Considere os dados de poluição do ar disponíveis no arquivo T1-5.dat. Esses dados são referentes a 52 medidas obtidas ao meio-dia em uma estação em Los Angeles em diversos dias. As variáveis observadas são:

- 1^a coluna: vento
- 2^a coluna: radiação solar
- 3^a coluna: CO
- 4^a coluna: NO
- 5^a coluna: NO_2
- 6^a coluna: O_3
- 7^a coluna: HC

Considere que as variáveis resposta são $Y_1 = NO_2$ e Y2 = O_3 e considere como preditoras apenas as variáveis Z_1 = vento e Z_2 = radiação solar.

(a) Ajuste um modelo de regressão utilizando apenas a primeira variável resposta Y_1. Obtenha um intervalo de predição para NO_2 correspondente a $z_1 = 10$ e $z_2 = 80$ com confiança igual a 95%.

Resolução

Ajustando o modelo sob condições de normalidade, independência entre observações e homocedasticidade, obtermos a equação $Y_1 = 10.115 - 0.211z_1 + 0.02z_2$, assim temos para $z_1 = 10$ e $z_2 = 80$ o valor de 9.646 com o intervalo de predição na tabela abaixo:

Table 4: Intervalo de Predição

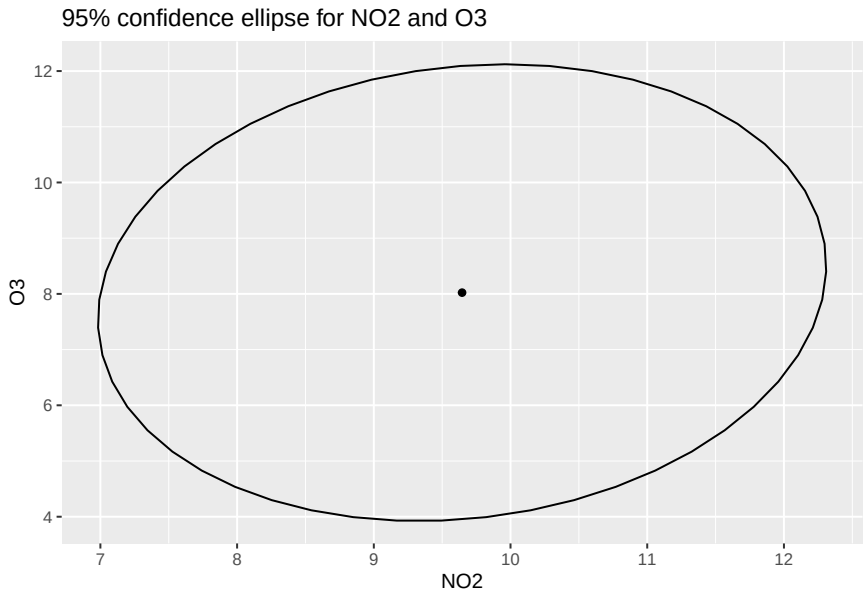
fit	lwr	upr
9.645627	2.427199	16.86406

(b) Ajuste um modelo de regessão multivariado com as respostas Y_1 e Y_2. Leia a seção do livro do Johnson "Predictions from Multivariate Multiple Regressions" (página 399 da sexta edição) e obtenha uma elipse de predição para Y_1 e Y_2 correspondente a $z_1 = 10$ e $z_2 = 80$. Compare com o resultado do item anterior.

Observação: No arquivo Elipse-predicao.r, existe uma função que faz o gráfico da elipse utilizando os pacotes car e ggplot2 e um exemplo com os dados apresentados em aula.

Resolução

Seja $\mathbf{Y} = [Y_1, Y_2]$ e tomando o modelo $\mathbf{Y} = \mathbf{Z}\beta + \epsilon$ onde $Z = [z_1, z_2]$, $\beta = [\beta_{(1)}, \beta_{(2)}]$ e $\epsilon = [\epsilon_1, \epsilon_2]$, sob suposição de normalidade multivariada, chegamos a uma elipse de predição abaixo, onde podemos notar que com o modelo múltiplo multivariado nós temos um intervalo de predição mais preciso olhando apenas para a variável resposta Y_1.



Exercício 4

O arquivo pottery.csv contém informações sobre análise química de cerâmicas Romano-Britânicas encontradas em 3 diferentes regiões. As nove primeiras colunas dos dados contém informações sobre nove óxidos e a última coluna sobre a localização onde a cerâmica foi encontrada (Região 1 correspondem a localização ou kiln 1, Região 2 contém localizações 2 e 3, e Região 3 é formada pelas localizações 4 e 5). A descrição dos dados está a seguir:

Romano-British Pottery Data

Description

Chemical composition of Romano-British pottery.

Format

A data frame with 45 observations on the following 9 chemicals.

Al2O3 aluminium trioxide.

Fe2O3 iron trioxide.

MgO magnesium oxide.

CaO calcium oxide.

Na2O natrium oxide.

K2O calium oxide.

TiO2 titanium oxide.

MnO mangan oxide.

BaO barium oxide.

kiln site at which the pottery was found.

Details

The data gives the chemical composition of specimens of Romano-British pottery, determined by atomic absorption spectrophotometry, for nine oxides.

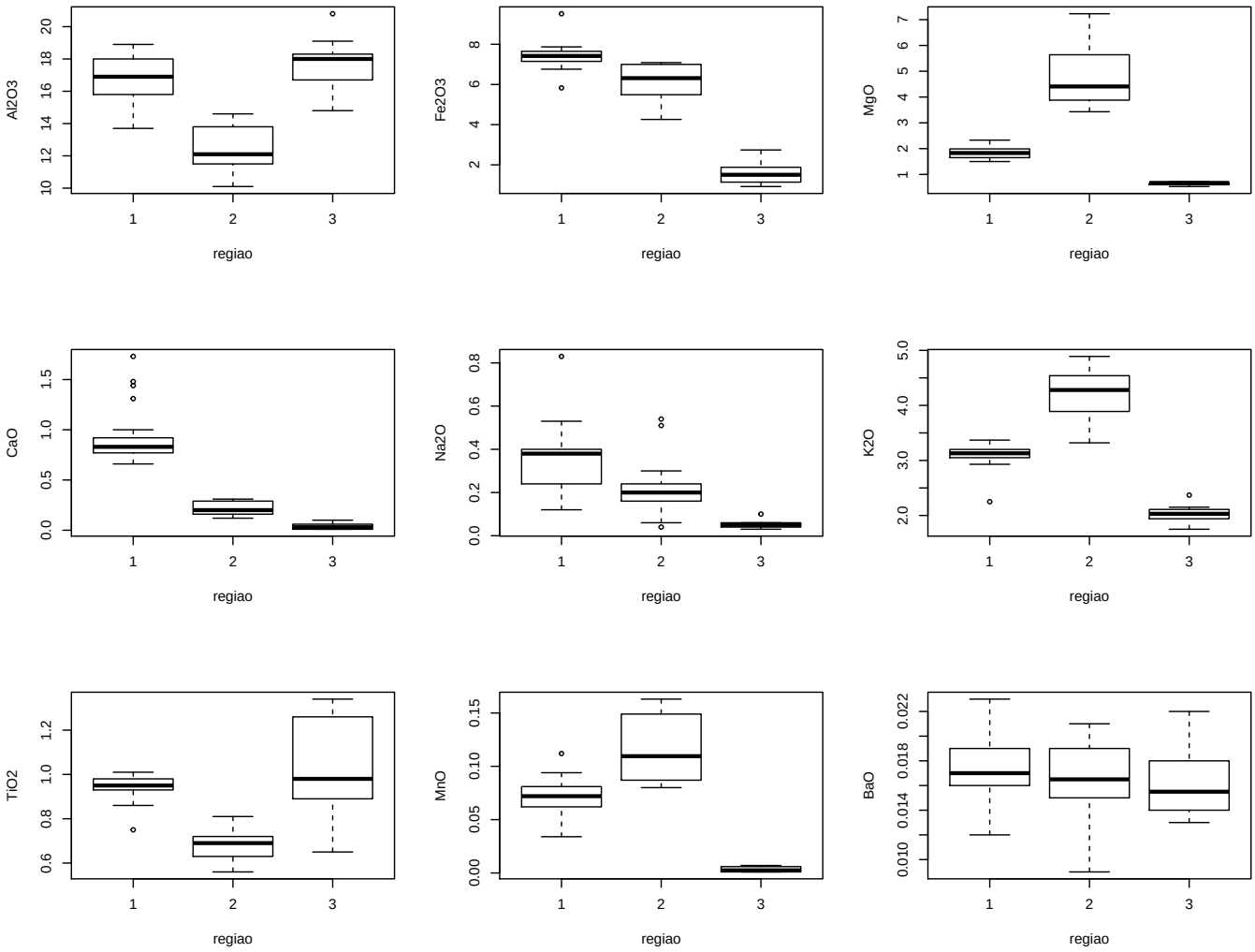
Source

A. Tubb and N. J. Parker and G. Nickless (1980), The analysis of Romano-British pottery by atomic absorption spectrophotometry. Archaeometry, 22, 153-171.

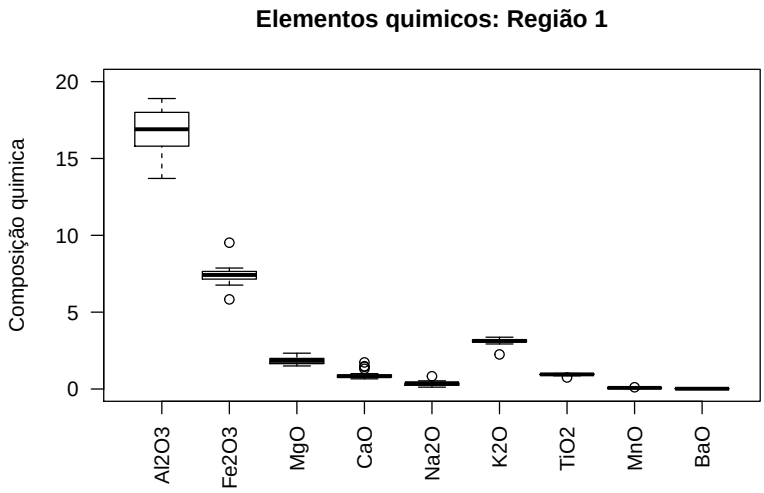
Obtenha representações gráficas multivariadas para visualização dos dados, identificando as regiões onde as cerâmicas foram encontradas. Interprete os resultados.

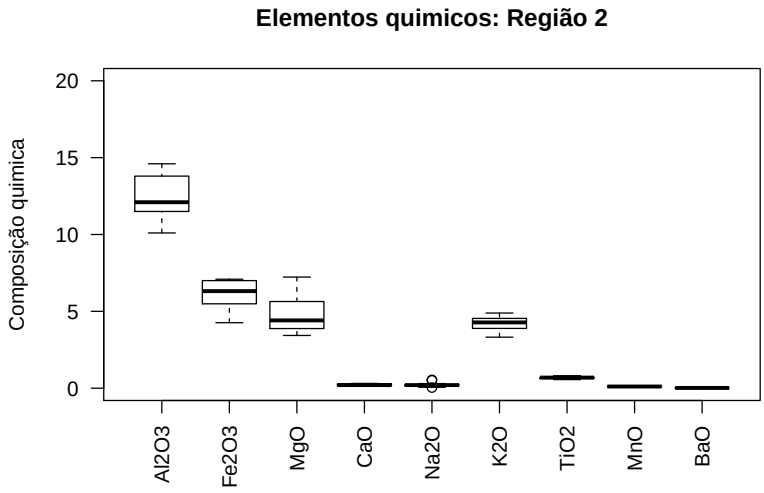
Resolução

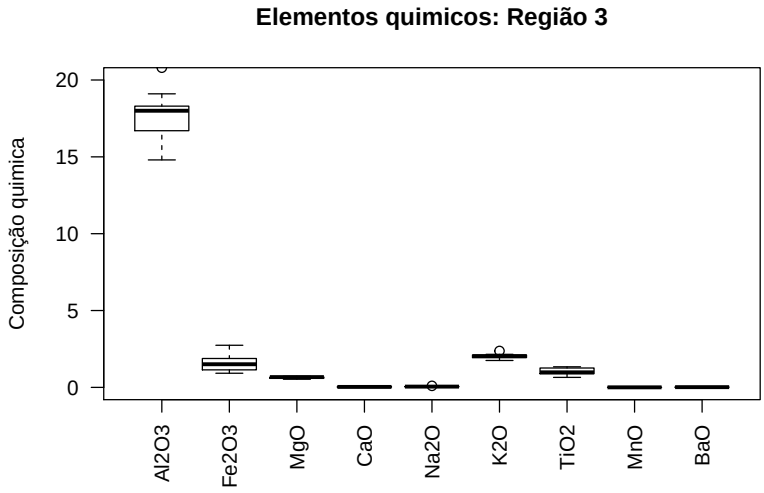
Primeiramente, foi feito os Boxplot's univariados de cada substância por região:



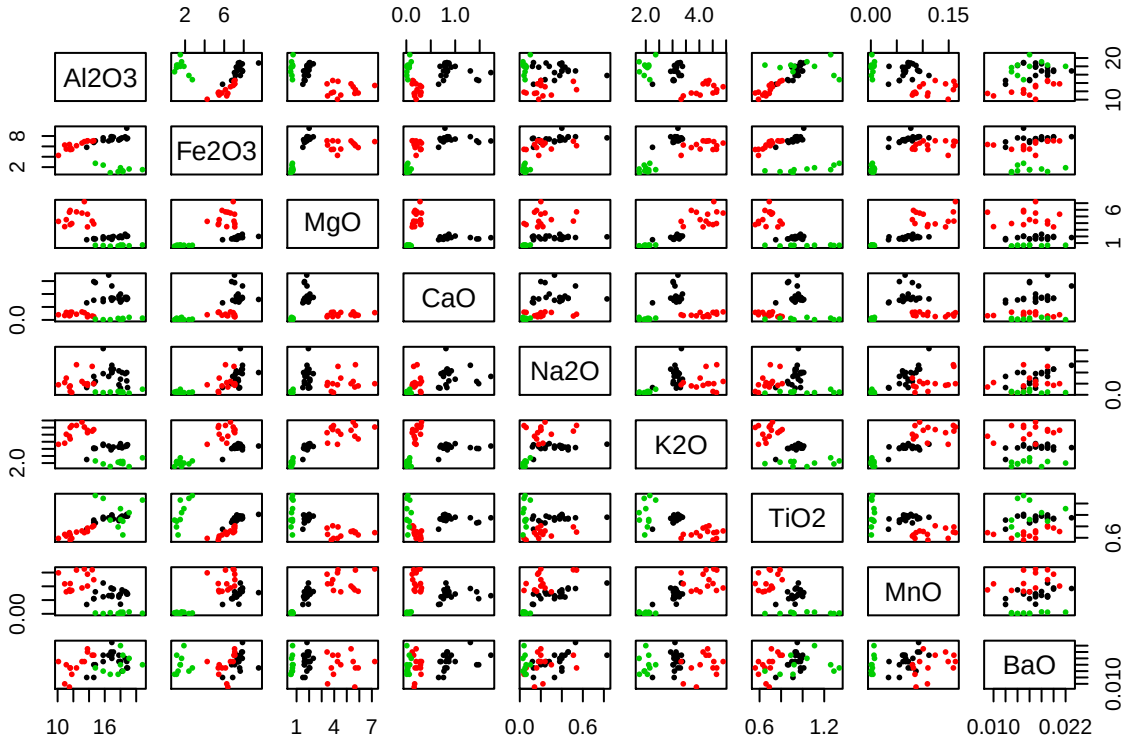
Pode-se ver que o valor mediano é menor na região 2 que nas demais na substância Al_2O_3, entretanto na mesma região há maior concentração de MgO, K_2O e MnO que as demais. Nas substâncias Fe_2O_3, CaO e Na_2O a região 1 é onde há maiores concentrações, seguida pela 2 e por fim a região 3. Para a substância TiO_2 o que chama a atenção é a variabilidade da terceira região ser alta, sua mediana ser próxima a da região 1 e maior que a região 2. Na última substância, BaO, o valor mediano apartenta ser igual para todas as regiões. A seguir foi feito mais Boxplot's mas agora de todas as substâncias em cada região separadamente.


Elementos quimicos: Região 1
Composição quimica
0
5
10
15
20
Al2O3
Fe2O3
MgO
CaO
Na2O
K2O
TiO2
MnO
BaO


Elementos quimicos: Região 2
Composição quimica
0
5
10
15
20
Al2O3
Fe2O3
MgO
CaO
Na2O
K2O
TiO2
MnO
BaO


Elementos quimicos: Região 3
Composição quimica
0
5
10
15
20
Al2O3
Fe2O3
MgO
CaO
Na2O
K2O
TiO2
MnO
BaO

É visível que em todas as regiões há uma grande concentração de Al_2O_3 com uma variabilidade maior que as demais substâncias. Na região 1 podemos verificar também que a segunda maior concentração é de Fe_2O_3 e as demais possuem concentrações abaixo de 5. Na segunda região notamos a média concentração de Fe_2O_3, MgO e K_2O e as demais substâncias com valores bem abaixo destas. Na última região não existe destaque de outras substâncias além da Al_2O_3. Realiza-se então o diagrama de disperção para as substâncias duas a duas.



Os dados em preto representa a primeira região, em vermelho a região 2 e por fim, a última região é representada pela cor verde. Neste gráfico é possível visualizar todas as conclusões já extraidas nos gráficos anteriores mas em apenas um gráfico.