UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Aplicação da Regressão por Redução de Posto para Identificação de Padrões Alimentares

Autor: Vanessa Schierholt da Silva
Orientador: Professora Dra. Suzi Alves Camey
Co-orientador: Professora Msc. Vanessa Leotti Torman

Porto Alegre, 18 de Julho de 2011.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Aplicação da Regressão por Redução de Postos para Identificação de Padrões Alimentares

Autor: Vanessa Schierholt da Silva

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professora Dra. Marilda Borges Neutzling

Porto Alegre, 18 de Julho de 2011.

Dedico este trabalho a Jesus Cristo, que é meu Mestre, meu Senhor, meu Salvador, meu Conselheiro e meu melhor Amigo.

"Porque Deus amou ao mundo de tal maneira que deu seu Filho unigênito, para todo aquele que nele crê não pereça, mas tenha a vida eterna."
João 3:16

Agradecimentos

Mãe, obrigada por todas as orações, por conversar comigo, por me ouvir, por ser minha amiga. Mas, principalmente, porque você foi a primeira pessoa a me conduzir ao nosso amado Senhor Jesus, sem isso a vida não teria o menor sentido, na verdade, eu não teria vida. Você é a mulher mais guerreira e fiel que eu conheço. Quando eu crescer, quero ser como você. Te amo!

Ingrid, minha irmã predileta, preciso muito de você. Obrigada por você ter cedido o seu quarto para eu fazer a monografia. Você me ajudou nessa jornada a perseverar, a ver que havia esperança mesmo quando eu achava que não eu não ia conseguir e que a força não vinha de mim. É um privilégio ter você como irmã e melhor amiga.

Pai, mesmo tu estando longe, esteve presente. A cada ligação, a cada história que você nos contava, cada vez que você se mostrava interessado em saber como nós estávamos isso te fez ser presente em nossas vidas. Obrigada por todo suporte financeiro, por nunca ter deixado faltar nada e dar além do que eu merecia ou precisava, por ter permitido que eu parasse de trabalhar só para estudar. Te amo!

Professora Suzi e professora Vanessa obrigada por terem aceitado me orientar e por terem tido a maior paciência comigo. Admiro a inteligência e o conhecimento estatístico que vocês têm. E também a todos os excelentes professores que tive durante a graduação.

À família Stobbe que me emprestou o computador e me deixou ficar até altas horas, para que eu pudesse terminar este trabalho quando o meu computador havia estragado.

À minha família e todos os meus amados irmãos em Cristo.

Aos meus colegas de curso Greice, Gabriel, Eduardo, Elaine, Letícia e Andrigo que me ajudaram em várias etapas do curso.

Obrigada Senhor Jesus, pelo teu amor, pela tua bondade e pela misericórdia.

Resumo

A regressão por redução de posto (RRR) é uma técnica que vem sendo empregada na epidemiologia nutricional desde 2004. O objetivo dela é encontrar padrões alimentares associados a algum desfecho. Assim, ela é considerada uma técnica que combina informações a priori e a posteriori. A informação a priori é um conhecimento prévio da associação entre as variáveis intermediárias (biomarcadores, nutrientes) e o desfecho (doença), e a posteriori é a combinação entre as variáveis intermediárias e o consumo alimentar. A RRR tenta explicar o máximo possível da variação das variáveis resposta através das variáveis preditoras. Este trabalho fornece uma breve revisão teórica da técnica e rotinas computacionais em SAS. Uma análise ilustrativa também é abordada, utilizando-se dados do estudo "Condições de saúde das mulheres: estudo de base populacional na região do Vale do Rio dos Sinos". Neste estudo o desfecho considerado foi a hipertensão arterial; as variáveis intermediárias consideradas foram os nutrientes sódio, potássio e gordura saturada; as variáveis preditoras foram a frequência de consumo de 70 tipos de alimentos. Uma das suposições do método é que as variáveis intermediárias sejam normais multivariadas. O número máximo de fatores que a técnica encontra é igual ao número de variáveis resposta existentes.

Palavras-chave: Regressão por redução de posto (RRR), padrão alimentar, desfecho.

Abstract

The reduced rank regression (RRR) is a technique that has been used in nutritional epidemiology since 2004. Its goal is to find some food patterns associated with outcome. Thus, it is considered a technique that combines prior and posterior information. The prior information is a prior knowledge of the association between the intermediate variables (biomarkers, nutrients) and the outcome (disease), and subsequently is the combination of the intermediate variables and food consumption. The RRR tries to explain as much as possible of the variation in responses across the predictors. This work provides a brief theoretical review of the technical and computational routines in SAS. An illustrative analysis is also addressed, using data from the study "Health conditions women: population-based study in the Vale do Rio dos Sinos." In this study the outcome was hypertension, the intermediate variables considered were the nutrients sodium, potassium and saturated fat; predictor variables were the frequency of consumption of 70 foods. One of the assumptions of the method is that the intermediate variables are multivariate normal. The maximum number of factors that the technique finds is equal to the number of variables response.

Keywords: reduced rank regression (RRR), dietary patterns, outcome.

Sumário

1 Introdução 8
2 Regressão por Redução de Posto (RRR) 10
2.1 Histórico 10
2.2 Modelo Geral 11
2.2.1 Suposições do Modelo 12
2.3 Estimação 12
3 Aplicação do Modelo RRR 16
3.1 Origem dos Dados 16
3.2 Descrição do Problema 16
3.3 Uso do Pacote Estatístico SAS 17
3.3.1 Normalidade Multivariada 17
3.3.2 Regressão por Redução de Posto 18
3.3.3 Resultados do Teste de Normalidade Multivariada 20
3.3.4 Resultados da Análise por Redução de Posto (RRR) 23
3.3.5 Gráficos da Análise por Redução de Posto (RRR) 31
3.4 Interpretação dos Fatores 47
3.5 Relacionando os Fatores da RRR com o Desfecho 50
4 Conclusão 54
5 Referências 56

1 Introdução

Os métodos de regressão são as ferramentas estatísticas, que talvez sejam as mais amplamente usadas na análise de dados[1]. As regressões linear simples, logística, Poisson, multivariada e entre outras são alguns exemplos de modelos de regressão.

Temos uma regressão multivariada quando duas ou mais variáveis resposta são estudadas simultaneamente, ou seja, quando se deseja estudar dois ou mais eventos ao mesmo tempo. A descrição usual do modelo de regressão multivariado é que um conjunto de m variáveis respostas está relacionado a um conjunto de n variáveis; assume-se, então, que a matriz dos coeficientes de regressão $m \times n$ é uma matriz de posto completo. Logo, quando o modelo possui muitas variáveis resposta e preditoras, temos que interpretar ao mesmo tempo um grande número de coeficientes de regressão, o que pode tornar-se uma tarefa bastante complicada. Então, em muitas situações, é necessário reduzir o número de parâmetros no modelo de regressão multivariada. O modelo de regressão por redução de posto (RRR) traz uma solução para essa questão.

Uma das áreas onde a redução de posto é especialmente importante é na nutrição. Nesta área, uma das necessidades existentes é identificar padrões alimentares associados a algum tipo de doença.

Segundo a Sociedade Brasileira de Cardiologia[2]:

> "Padrão alimentar é definido como o perfil do consumo de alimentos pelo indivíduo ao longo de um determinado período de tempo. É utilizado no estudo da relação entre a ingestão de certos nutrientes e o risco de doenças, pois permite uma compreensão mais clara sobre a alimentação como um todo, em lugar de se considerarem os nutrientes individualmente."

As técnicas estatísticas usualmente utilizadas para encontrar padrões alimentares são análise de componentes principais[3,4,5,6], regressão por mínimo quadrado parcial[5,6], análise fatorial[6], análise de clusters[6] e, recentemente, regressão por redução de posto[5,6,7,8]. Hoffmann (2004) diz que:

> "Uma problema da PCA diz respeito à interpretação dos fatores que abrangem seu papel na etiologia das doenças. Se o padrão alimentar obtido através da PCA for um fator de risco para uma determinada doença, uma justificativa razoável para isso geralmente é difícil de ser encontrada. Embora saibamos qual grupo de alimentos contribuem substancialmente para os fatores, olhando para as cargas fatoriais altas, ainda não é clara a razão desses alimentos serem importantes no desenvolvimento da doença."

O objetivo da RRR na nutrição é identificar padrões alimentares associados a algum desfecho, por exemplo, hipertensão, diabetes, doenças coronárias e etc. Em geral, nessa área, as variáveis preditoras são o consumo de alimentos ou grupos de alimentos e as variáveis resposta, ou ainda variáveis intermediárias, são biomarcadores, consumo de nutrientes ou outras variáveis associadas com o desenvolvimento da doença que estiver sendo estudada (desfecho). Assim, a técnica tenta explicar, através das variáveis preditoras, o máximo possível da variação das variáveis resposta. Desse modo a RRR soluciona a debilidade da PCA apontada por Hoffmann (2004).

Esta monografia está organizada da seguinte forma: no capítulo 2, apresentaremos um breve histórico do modelo, abordaremos o modelo RRR e suas suposições e, por fim, descreveremos o método de estimação utilizado para o modelo.

No capítulo 3, mostraremos a aplicação da RRR na nutrição apresentando as rotinas no SAS. Para isso, utilizamos dados do estudo "Condições de saúde de mulheres adultas residentes na região do Vale do Rio dos Sinos, RS"[9].

2 Regressão por Redução de Posto (RRR)

2.1 Histórico

Em 1951, Anderson foi o primeiro a considerar o problema da regressão por redução de posto para os casos em que o conjunto de variáveis preditoras é fixo[1,10]. Izenman em 1975 introduziu o termo *reduced rank regression* (regressão por redução de posto - RRR)[11,12]. De acordo com Reinsel (1998): "O modelo RRR e suas propriedades estatísticas foram examinados por Robinson (1973, 1974), Tso (1981), Davies e Tso (1982), Zhou (1994) e Geweke (1996)".

Segundo o mesmo autor, os conceitos do modelo RRR foram considerados por vários autores em diferentes contextos e terminologias. Rao (1964) estudou componentes principais e apresentou resultados que podem ser relacionados ao RRR, referindo-se a combinação linear das variáveis preditoras como componentes principais de variáveis instrumentais. Fortier (1966) considerou o RRR como modelagem de predição linear simultânea. Wollenberg (1977) discutiu o mesmo procedimento como uma alternativa à análise de correlação canônica, conhecida como análise de redundância. Reinsel (1998) afirma que os resultados encontrados nesses trabalhos são essencialmente os mesmos e são um caso particular da estimação por redução de posto. Anos mais tarde, em 2004, a técnica foi, pela primeira vez, aplicada na epidemiologia nutricional por Hoffmann[5].

2.2 Modelo Geral

O modelo de regressão multivariado assume a seguinte forma:

$$Y_k = CX_k + \varepsilon_k, \;\; k = 1, \dots, T \tag{2.1}$$

Para obtermos a regressão por redução de posto (RRR) o modelo (2.1) deve ser combinado com a seguinte restrição:

$$posto(C) = r \leq min\,(m, n) \tag{2.2}$$

onde:

$X_k = (x_{1k}, x_{2k}, \dots, x_{nk})'$ é um vetor das n variáveis preditoras do k-ésimo indivíduo/unidade;

$Y_k = (y_{1k}, y_{2k}, \dots, y_{mk})'$ é um vetor das m variáveis resposta do k-ésimo indivíduo/unidade;

$\varepsilon_k = (\varepsilon_{1k}, \varepsilon_{2k}, \dots, \varepsilon_{mk})'$ é o vetor dos m erros aleatórios do k-ésimo indivíduo/unidade;

C é uma matriz de coeficientes de regressão $m \times n$;

T é o número de indivíduo/unidade.

Assim, devido à restrição (2.2) C pode ser fatorada[1, 12] em

$$C = AB$$

sendo que:

A tem dimensão $m \times r$ e B tem dimensão $r \times n$.

Então, podemos reescrever o modelo (2.1) como:

$$Y_k = A(BX_k) + \varepsilon_k, \;\; k = 1, \dots, T, \tag{2.3}$$

onde:

(BX_k) tem dimensão reduzida com r componentes.

Apesar de ser necessária a coleta da mesma quantidade de dados tanto para regressão multivariada quanto para a regressão por redução de posto, existe um ganho em simplicidade e interpretação através da RRR. Isto porque ela só necessita de r combinações lineares das variáveis preditoras para modelar a variação das variáveis resposta enquanto a regressão multivariada precisa de todas as n combinações lineares[1].

2.2.1 Suposições do Modelo

As suposições feitas no modelo RRR (2.1) são as seguintes:

i) Os erros ε_k são independentes e têm distribuição $N(0, \Sigma_{\varepsilon\varepsilon})$, onde $\Sigma_{\varepsilon\varepsilon}$ é a matriz de covariância $m \times m$ positiva definida;

ii) Y_k tem distribuição normal multivariada.

2.3 Estimação

Nessa seção será descrito sucintamente o método de estimação dos parâmetros da RRR. Abordagens mais aprofundadas sobre esse assunto podem ser encontradas em Reinsel[1], Aldrin[10], Izenman[11], Tso[12], Anderson[13], Hansen[14] e Johansen[15].

De acordo com Aldrin[10], iremos, primeiramente, encontrar as estimativas de C do modelo de regressão multivariado (2.1) através da minimização do critério de mínimos quadrados generalizados que é dado por:

$$Q(\hat{C}) = traço\left[\left(Y_k - X_k\hat{C}\right)' W \left(Y_k - X_k\hat{C}\right)\right] \tag{2.4}$$

onde W é uma matriz de pesos previamente escolhida. O método dos mínimos quadrados ordinários (OLS) utiliza W como sendo uma matriz

identidade. Assim, o estimador de C do modelo de regressão multivariado sem restrição no posto de C é dado por:

$$\hat{C}^{OLS} = S_{xx}^{-1} S_{xy} \tag{2.5}$$

onde $S_{xx} = \frac{1}{T} X_k{}' X_k$ e $S_{xy} = \frac{1}{T} X_k{}' Y_k$.

Com r conhecido e satisfazendo a restrição de que $r = posto(C) \leq min\,(m,n)$, para estimarmos C através da máxima verossimilhança, precisamos assumir que os erros são normais. Então, queremos minimizar (2.4) com W dado por:

$$W = \hat{\Sigma}_{\varepsilon\varepsilon}^{-1} = \left(\left(\frac{1}{T} \right) \left(Y_k - X_k \hat{C}^{OLS} \right)' \left(Y_k - X_k \hat{C}^{OLS} \right) \right)^{-1} \tag{2.6}$$

Logo, através da decomposição do valor singular da matriz de posto $min(m,n)$ é que a RRR estima C. Assim,

$$D = S_{xx}^{-1/2} S_{xy} W^{1/2} = S_{xx}^{1/2} \hat{C}^{OLS} W^{1/2} = \sum_{j=1}^{min\,(m,n)} \lambda_j u_j v_j' \tag{2.7}$$

onde:

λ_j são os autovalores de D;

u_j e v_j são os autovetores a direita e a esquerda, respectivamente, de D.

Multiplicando D por $S_{xx}^{-1/2}$ e depois por $W^{1/2}$ obtemos a decomposição de $\hat{C}^{OLS}$ em:

$$\hat{C}^{OLS} = \sum_{j=1}^{min\,(m,n)} \hat{C}_j \tag{2.8}$$

onde:

$\hat{C}_j = \lambda_j S_{xx}^{-1/2} u_j v_j' W^{-1/2} = s_j v_j' W^{-1/2}$;

$s_j = \lambda_j S_{xx}^{-1/2} u_j$.

Conseqüentemente, o modelo RRR estima C da seguinte forma:

$$\hat{C}_r^{RRR} = \sum_{j=1}^{r} \hat{C}_j \tag{2.9}$$

lembrando que $r \leq \min(m, n)$.

Assim, $\hat{C}_1^{RRR}$ explica o máximo possível da variância de Y, $\hat{C}_2^{RRR}$ explica o máximo da variação restante e assim, sucessivamente, até $\hat{C}_r^{RRR}$.

Anderson (1999) mostra que para r conhecido o estimador de máxima verossimilhança de C é assintoticamente não viciado e normal, mesmos se os erros não forem Gaussianos.

Assim podemos ver que quando $r = min(m, n)$ temos que $\hat{C}^{OLS} = \hat{C}_r^{RRR}$. Além disso, a variabilidade de $\hat{C}_r^{RRR}$ é menor que a variabilidade de $\hat{C}^{OLS}$, mas também o $\hat{C}_r^{RRR}$ é, em geral, um estimador viesado. No entanto, se a redução da variabilidade é bastante alta em comparação ao viés, então os mínimos quadrados da RRR podem dar estimações ou predições mais precisas que os mínimos quadrados ordinários[10].

Dessa forma, o modelo (2.1) estimado é dado por:

$$Y_k = X_k \hat{C}_r^{RRR} + E_k = \sum_{j=1}^{r} X_k s_j v_j' W^{-1/2} + E_k \tag{2.10}$$

onde E_k é o vetor de erros ajustados. Por isso, multiplicando ambos os lados de (2.10) por $t_j = W^{1/2} v_j$ temos que:

$$Y_k t_j = X_k s_j + E_k t_j, \qquad j = 1, \ldots, r \tag{2.11}$$

em que:

$Y_k t_j$ são os escores de Y;

$X_k s_j$ são os de escores X;

t_j são as cargas de Y;

s_j são as cargas de X.

Isso significa que, para cada componente j em (2.11), uma combinação linear de Y_k é explicada por uma combinação linear de X_k mais algum erro.

Para determinar o número de fatores do modelo RRR, ou seja, determinar r, Reinsel[1] utiliza o teste de razão de máxima verossimilhança. Com o mesmo objetivo, o software SAS tem implementado o teste de Van der Voet[16] que será mostrado no capítulo de aplicação.

3 Aplicação do Modelo RRR

3.1 Origem dos Dados

Os dados que serão utilizados nessa seção foram cedidos pela coordenadora do projeto "Condições de saúde das mulheres: estudo de base populacional na região do Vale do Rio dos Sinos"[9], Maria Teresa Anselmo Olinto. Ao todo foram estudadas 1026 mulheres adultas de 20 a 60 anos residentes na zona urbana da cidade de São Leopoldo, RS, Brasil, que responderam um Questionário de Frequência Alimentar (QFA).

3.2 Descrição do Problema

O objetivo desta aplicação é identificar padrões alimentares em mulheres adultas residentes em São Leopoldo RS associados com hipertensão.

Assim, para aplicarmos a RRR precisamos ter um conhecimento a priori da associação das variáveis resposta e o desfecho[5]. Sabe-se, então, que os nutrientes sódio e gordura saturada são fatores de risco para hipertensão e o nutriente potássio é fator protetor[17]. E os nutrientes também estão associados com os alimentos.

Então consideramos a hipertensão arterial como o desfecho, os nutrientes sódio, potássio e gordura saturada como variáveis resposta e a frequência de consumo de 70 tipos de alimentos como as variáveis preditoras.

3.3 Uso do Pacote Estatístico SAS

A seguir apresentaremos as rotinas do SAS para testar a normalidade multivariada e para fazer a RRR.

3.3.1 Normalidade Multivariada

O teste da normalidade multivariada é feito através de uma macro que deve ser baixada no site[18] do SAS. Os termos em negrito referem-se, particularmente, ao caso estudado.

```
1. %inc "F:\UFRGS\TCC\analises\multnorm.sas";
2. %multnorm(
     a. data=_proj_.banco_dados0,
     b. var= SODIO POTASSIO SATURADA,
     c. plot=mult)
```

Apresentaremos uma breve explicação dos comandos[18] a seguir:

1. Indica o arquivo que contém a macro MULTNORM.

2. Roda a macro MULTNORM

   a. Indica o conjunto de dados a ser utilizado;

   b. Indica quais variáveis que serão testadas quanto a normalidade multivariada;

   c. Solicita o gráfico do Quantil Qui-Quadrado *versus* o quadrado das distâncias Mahalanobis das observações.

3.3.2 Regressão por Redução de Posto

A rotina PROC PLS[19] do *software* SAS 9.2 foi utilizada para fazermos a análise RRR.

```
1. ods html;
2. ods graphics on;
3. proc pls
    a. data=_proj_.banco_dados0
    b. method=rrr
    c. varss
    d. details
    e. plots= ALL
    f. plots=(corrload(trace=off))
    g. cv=one
    h. cvtest;

4. model SODIO POTASSIO SATURADA = AVEFARM2 PAOCENM2
   PAOFANM2 PAOCASM2 LEITEINTM2 LEITEDESM2 LEITEFERM2
   IOGURTEM2 QUEIJOM2 KASM2 ABACAM2 ABACAXIM2
   BANANAM2 MAMAOM2 MACAM2 AMEIXAM2 CAQUVAM2
   BERGAM2 LARANJAM2 LIMARM2 MANGAM2 MELAOM2
   MORANGM2 ARROZINTM2 RROZBM2 MASM2 MASINM2 FEIJM2
   FRANGOM2 PEIXEM2 GADOM2 PORCOM2 FIGADOM2 OVOSM2
   PRESUNM2 LINGUIM2 ABOBORAM2 AGRIAOM2 ALHOM2
   BATATAM2 AIPIMM2 BERINJM2 BROCOLISM2 COUVEM2
   OUTVEGM2 SOJAM2 BANHAM2 CREMLEITM2 MAIOCASM2
   MAIOINDM2 MANTM2 MARGM2 NATAM2 FRITM2 SOBRM2
   SORVM2 CHOCOM2 BISDOCM2 BISSALM2 CUCAM2 AVELAM2
   MCM2 IPOCAM2 MELM2 ACUCM2 ACUMASCM2 SUCNATM2
   SUCINDM2 VINTINM2 CHOPM2;

5. output
    a. out=pattern
    b. xscore=scorex
    c. yscore=scorey
    d. stdy=ypad1-ypad3;
6. run;
7. ods graphics off;
8. ods html close;
```

Agora descreveremos cada comando.

1. Inicia a criação da saída em HTML.

2. Solicita o ODS GRAPHICS, para que sejam feitos os gráficos do procedimento PLS.

3. Determina que o procedimento a ser utilizado seja o PLS.

    a. Indica o conjunto de dados a ser utilizado;

    b. Indica que o método a ser utilizado é RRR;

    c. Solicita a exibição da variação explicada por cada variável resposta e preditora;

    d. Solicita a exibição dos detalhes do modelo ajustado para cada fator;

    e. Constrói todos os gráficos existentes no procedimento PLS;

    f. Retira as linhas do gráfico de correlação das cargas (*Correlation Loading Plot*);

    g. Faz a validação cruzada *one-at-a-time* que é a técnica mais comumente utilizada para comparar número de fatores a serem extraídos. Outras técnicas podem ser encontradas no site do SAS[16];

    h. Especifica que o teste a ser utilizado é o teste de Van der Voet. Ele testa modelos com diferentes números de fatores extraídos contra o modelo que minimiza a soma dos resíduos ao quadrado preditos (PRESS)[16].

4. Indica o modelo a ser ajustado através da expressão <variáveis resposta> = <variáveis preditoras>.

5. Especifica um conjunto de dados para receber as quantidades que podem ser computadas para cada observação:

a. Nomeia o banco que conterá os dados originais e as variáveis criadas na sequência;

b. Cria as variáveis que recebem os valores dos escores das variáveis preditoras para cada fator, no caso três;

c. Cria as variáveis que recebem os valores dos escores das variáveis resposta para cada fator extraído, no caso três;

d. Cria as variáveis que recebem os valores das variáveis resposta padronizadas para cada indivíduo.

6. Indica que todos os comandos acima devem ser executados.

7. Finaliza a criação dos gráficos.

8. Finaliza a criação das saídas dos gráficos em HTML.

3.3.3 Resultados do Teste de Normalidade Multivariada

Os resultados produzidos pelos comandos descritos para o teste de normalidade multivariada são apresentados a seguir.

MULTNORM macro: Univariate and Multivariate Normality Tests

The MODEL Procedure

	Normality Test			
	Equation	**Test Statistic**	**Value**	**Prob**
1.	**SODIO**	Shapiro-Wilk W	0.96	<.0001
2.	**POTASSIO**	Shapiro-Wilk W	0.98	0.0589
3.	**SATURADA**	Shapiro-Wilk W	0.98	<.0001
4.	**System**	Mardia Skewness	85.88	<.0001
5.		Mardia Kurtosis	0.61	0.5429
6.		Henze-Zirkler T	9.26	<.0001

1. Refere-se ao teste de normalidade de Shapiro-Wilk apenas para a variável **SODIO**. Assim, rejeita-se a hipótese de normalidade para **SODIO** (Shapiro-Wilk W= 0,96; p<0,0001).

2. Refere-se ao teste de normalidade de Shapiro-Wilk para a variável **POTASSIO**. Assim, rejeita-se a hipótese de normalidade para **POTASSIO** (Shapiro-Wilk W= 0,98; p=0,0589).

3. Refere-se ao teste de normalidade de Shapiro-Wilk para a variável **SATURADA**. Assim, rejeita-se a hipótese de normalidade para **SATURADA** (Shapiro-Wilk W= 0,98; p<0,0001).

4. Refere-se ao teste de Mardia para o coeficiente de assimetria da normal multivariada. Assim, rejeita-se a hipótese de assimetria normal multivariada (Mardia *Skewness*= 85,88; p<0,0001).

5. Refere-se ao teste de Mardia para o coeficiente de curtose da normal multivariada. Assim, não se rejeita a hipótese de curtose normal multivariada (Mardia *Kurtosis*= 0,61; p= 0,5429).

6. Refere-se ao teste de Henze-Zirkler T para testar a normalidade multivariada. Assim, rejeita-se a hipótese de normalidade multivariada (Henze-Zirkler T= 9,26; p<0,0001).

A Figura 1 apresenta o quantil-quantil qui-quadrado do quadrado das distâncias de Mahalanobis das observações em relação ao vetor das médias. Esse gráfico confirma que as variáveis resposta não são normais multivariadas, pois se os dados tivessem distribuição normal multivariada, os pontos cairiam sobre a reta[19].

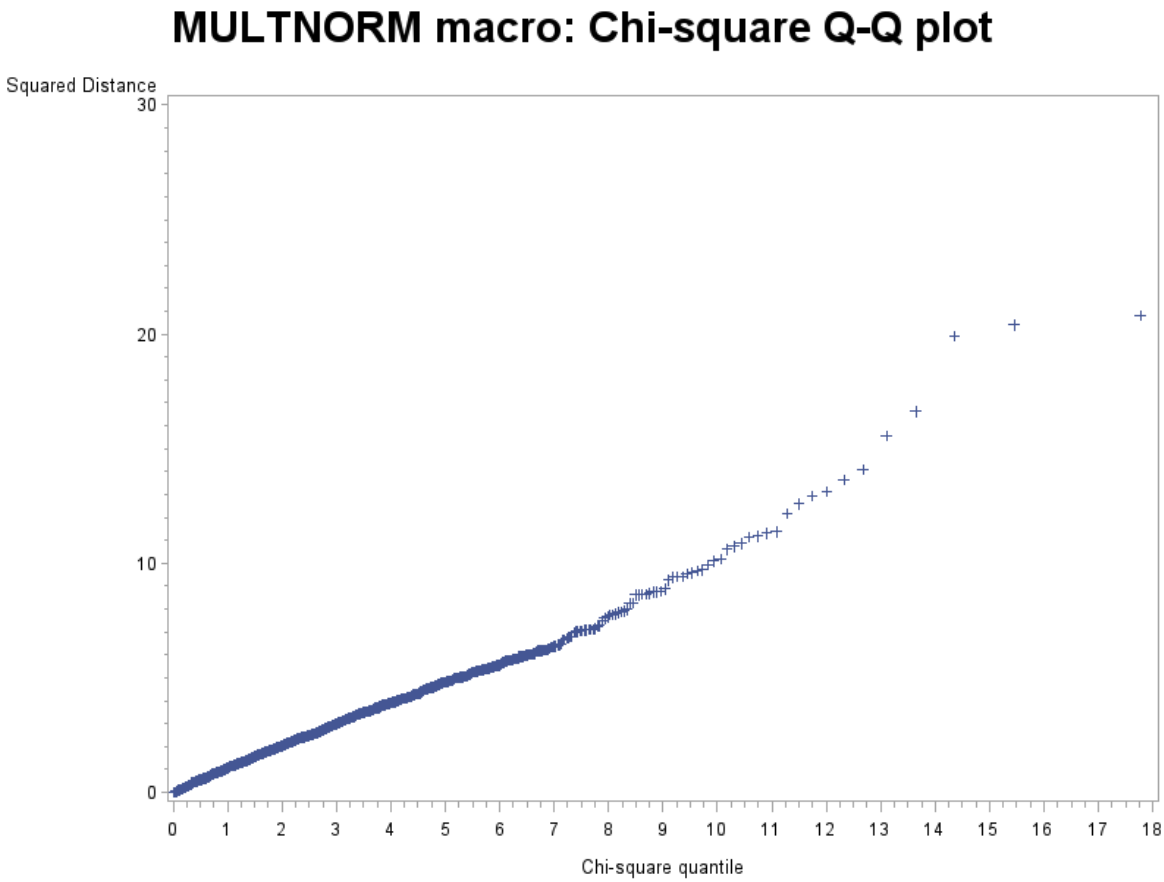


Figura 1: Quantil-quantil qui-quadrado *versus* o quadrado da distância de mahalanobis.

Apesar das variáveis resposta não apresentarem distribuição normal multivariada, percebemos pela Figura 1 que não há um grande desvio da normalidade. Dessa forma, procederemos a análise sem fazer nenhuma transformação nas variáveis.

3.3.4 Resultados da Análise por Redução de Posto (RRR)

Os resultados produzidos pelos comandos descritos na seção 3.3.2 são apresentados a seguir.

RRR

Tabela 1:Resumo do procedimento PLS.

	The PLS Procedure	
0.	**Data Set**	_PROJ_.BANCO_DADOS0
2.	**Factor Extraction Method**	Reduced Rank Regression
3.	**Number of Response Variables**	3
4.	**Number of Predictor Parameters**	70
5.	**Missing Value Handling**	Exclude
6.	**Maximum Number of Factors**	3
7.	**Validation Method**	Leave-one-out Cross Validation
8.	**Validation Testing Criterion**	Prob T**2 > 0.1
9.	**Number of Random Permutations**	1000
10.	**Random Permutation Seed**	903883001

1. Mostra o nome do banco de dados em que foi feita a análise.

2. Mostra que o método utilizado para fazer a análise foi a Regressão por Redução de Posto (RRR).

3. Mostra o número de variáveis respostas utilizadas na análise: 3 (**SODIO**, **POTASSIO** e **SATURADA**).

4. Mostra o número de variáveis preditoras utilizadas na análise: 70 frequências diárias de consumo dos alimentos.

5. Mostra que os dados faltantes (*missings*) foram excluídos.

6. Indica o número de fatores que podem ser extraídos na análise: 3.

7. Indica o método de validação utilizado: Validação Cruzada *Leave-one-out*.

8. Indica que o teste utilizado como critério de validação foi a estatística T^2 de Hotelling com p>0,1 como ponto de corte para uma diferença não significativa.

9. Indica o número de permutações aleatórias utilizado: 1000.

10. Indica a semente para permutação aleatória utilizada: 903883001.

Tabela 2: Número de observações lidas e usadas.

11.	**Number of Observations Read**	1026
12.	**Number of Observations Used**	1023

11. O número de observações do banco de dados: 1026.

12. O número de observações utilizado na análise: 1023, uma vez que os dados faltantes foram excluídos da análise.

Tabela 3: Validação cruzada para o número de fatores extraídos.

Cross Validation for the Number of Extracted Factors			
Number of Extracted Factors	**Root Mean PRESS**	**T**2**	**Prob > T**2**
0	1,000978	475.7195	<.0001
1	0.557117	512.1546	<.0001
2	0.337668	329.8757	<.0001
3	1.03E-14	0	1.0000

13.	**Minimum root mean PRESS**	1.03E-14
14.	**Minimizing number of factors**	3
15.	**Smallest number of factors with p > 0,1**	3

13. O *root mean PRESS* mínimo: 1,03E-14.

14. A minimização do número de fatores: 3.

15. O menor número de fatores com p>0,1: 3.

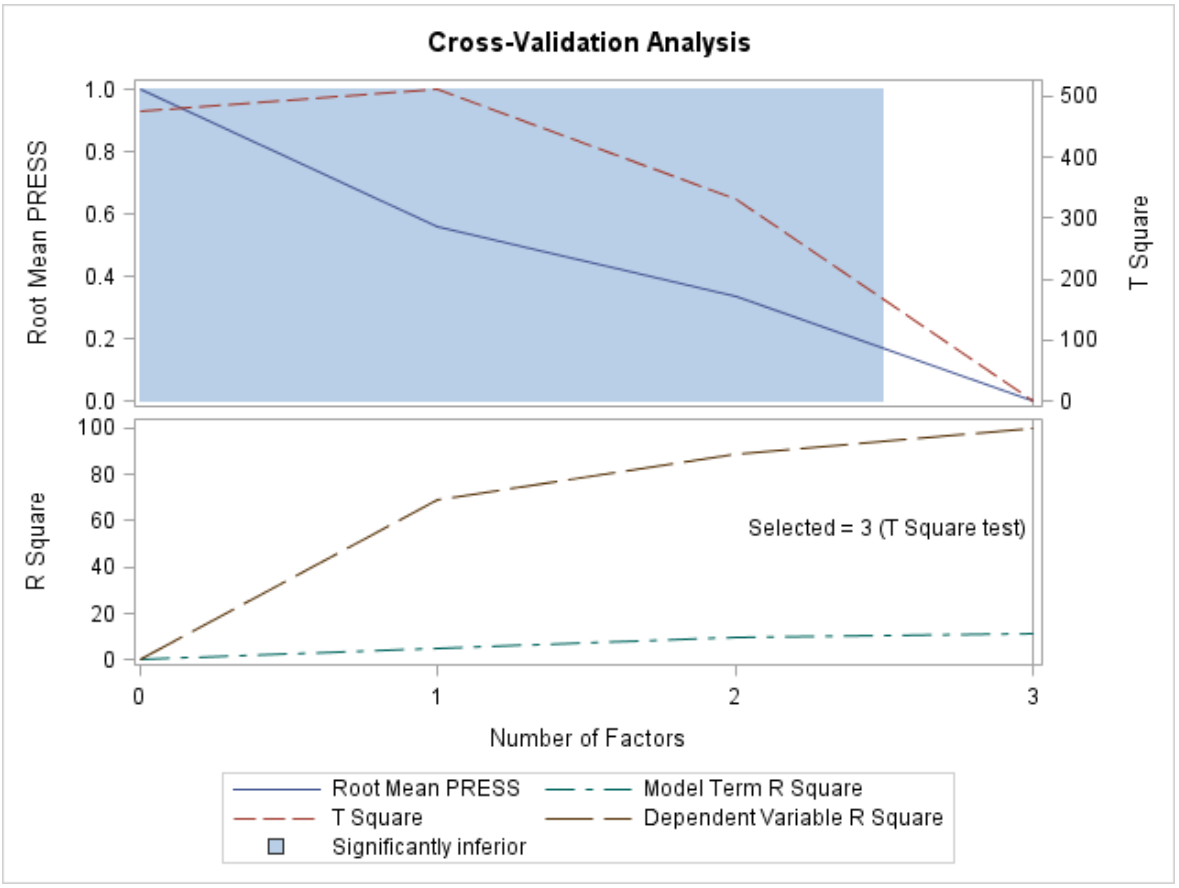


Figura 2: Gráfico de validação cruzada.

O gráfico na parte superior da Figura 2 é apenas a ilustração das informações contidas na Tabela 3. A parte inferior da figura representa o gráfico da proporção da variação explicada (ou R^2) pelo número de fatores. Note que a escala do eixo do *R Square* está em percentual. A linha azul (*Model Term R Square*) representa a proporção da variação das variáveis preditoras explicada pelos fatores. A linha marrom (*Dependent Variable R Square*) representa a proporção das variáveis resposta explicada pelos fatores.

Tabela 4: Percentual de variação explicada pelas variáveis preditoras e pelas variáveis resposta para fatores da RRR.

16. Percent Variation Accounted for by Reduced Rank Regression Factors									
Number of Extracted	16.1. Model Effects				16.2. Dependent Variables				
	a.	b.	c.	d.	a.	b.	c.	d.	e.
Factors	aveia[1]	...	Current	Total	SODIO	POTASSIO	SATURADA	Current	Total
1	1.8139	...	5.2885	5.2885	76.5201	57.8701	72.8369	69.0757	69.0757
2	8.5469	...	4.0816	9.3701	81.4232	99.6234	85.0176	19.6123	88.6880
3	8.5564		1.6218	10.9919	100.0000	100.0000	100.0000	11.3120	100.0000

16. Percentual de variação explicada pelos fatores da RRR;

16.1. Variáveis preditoras:

a. O primeiro fator explica 1,81% da variação da variável **aveia**. Os fatores 1 e 2 conjuntamente explicam 8,55% da variação da variável **aveia**. Os três fatores explicam 8,56% da variação da variável **aveia**, conjuntamente.

b. A tabela foi suprimida, mas na saída do SAS é mostrado o percentual de variação de cada variável preditora, que entrou no modelo;

c. O primeiro fator explica 5,29% da variação das variáveis preditoras o segundo fator explica 4,08% e o terceiro fatore explica 1,62% da variação das variáveis preditoras;

d. Os três fatores explicam conjuntamente 10,99% da variação de todas as variáveis preditoras.

16.2. Variáveis resposta:

a. O primeiro fator explica 76,52%, os fatores 1 e 2 explicam conjuntamente 81,42% e os três fatores explicam conjuntamente 100% da variação da variável **SODIO**;

[1] A saída no software SAS traz o nome da variável, neste trabalho trocamos os nomes das variáveis com o objetivo de facilitar a compreensão.

b. O primeiro fator explica 57,87%, os fatores 1 e 2 explicam conjuntamente 99,62% e os três fatores explicam conjuntamente 100% da variação da variável da variação da variável **POTASSIO**;

c. O primeiro fator explica 72,84%, os fatores 1 e 2 explicam conjuntamente 85,02% e os três fatores explicam conjuntamente 100% da variação da variável da variação da variável **SATURADA**;

d. O primeiro fator explica 69,08%, o segundo fator explica 19,61% e o terceiro fatores explica 11,31% da variação de todas variáveis resposta;

e. Os fatores 1 e 2 explicam conjuntamente 88,69% da variação das variáveis resposta e os três fatores explicam 100% da variação das variáveis resposta conjuntamente.

Aqui cabe uma observação de que os consumos de **SODIO**, **POTASSIO** e **SATURADA** foram calculados a partir da frequência de consumo dos 70 itens alimentares. Por essa razão o modelo atinge 100% de explicação. Em geral, o consumo dos nutrientes deve ser medido de forma independente do consumo dos alimentos.

Tabela 5: Cargas das variáveis preditoras

20. Model Effect Loadings			
Number of Extracted Factors	**aveia**	**...**	**chop**
1	0.069998	...	0.043374
2	0.153512	...	-0.033164
3	0.009148		-0.003358

20. A carga da variável **aveia** no fator 1 é 0,07; no fator 2 é 0,15 e no fator 3 é 0,01.

Tabela 6: Peso das variáveis preditoras e das variáveis resposta.

Weights						
Number of Extracted	21. Model Effect			22. Dependent Variable		
	a.		b.	a.	b.	
Factors	aveia	...	chop	SODIO	POTASSIO	SATURADA
1	0.049953	...	0.004494	0.607665	0.528450	0.592860
2	0.070321	...	0.030137	-0.288674	0.842403	-0.454999
3	0.013203	...	0.005322	-0.739872	0.105344	0.664449

21. Peso das variáveis preditoras:

    a. O peso da variável **aveia** é igual a 0,05 no fator 1; 0,07 no fator 2 e 0,01 no fator 3;

    b. O peso da variável **chop** é igual a 0,004 no fator 1; 0,03 no fator 2 e 0,005 no fator 3;

22. Peso das variáveis resposta, esses valores serão utilizado para fazer o cálculo dos escores de y (ver seção 3.4):

    a. O peso da variável **SODIO** é igual a 0,61 no fator 1; no fator 2 igual a -0,29 e no fator 3 igual a -0,74;

    b. O peso da variável **POTASSIO** é igual a 0,53 no fator 1; 0,84 no fator 2 e 0,11 no fator 3.

    c. O peso da variável **SATURADA** é igual a 0,59 no fator 1; -0,45 no fator 2 e 0,66 no fator 3.

Tabela 7: Coeficientes de regressão para um fator extraído.

Coded Regression Coefficients for 1 Extracted Factor			
	23. SODIO	**24. POTASSIO**	**25. SATURADA**
aveia	0.0303548244	0.0263977649	0.0296152657
...	...	...	...
chop	0.0027309074	0.0023749059	0.0026643721

Considerando a extração apenas do fator 1:

23. O coeficiente de regressão para a equação de **SODIO** e da variável preditora **aveia** é 0,03;

24. O coeficiente de regressão para a equação de **POTASSIO** e da variável preditora **aveia** é 0,003;

25. O coeficiente de regressão para a equação de **SATURADA** e da variável preditora **aveia** é 0,03;

Tabela 8: Coeficientes de regressão para dois fatores extraídos.

Coded Regression Coefficients for 2 Extracted Factors			
	26. SODIO	**27. POTASSIO**	**28. SATURADA**
aveia	0.0100550609	0.0856361458	-0.0023805675
...	...	...	
chop	0.0011945814	0.0068581830	0.0002428645

Considerando a extração apenas dos fatores 1 e 2:

26. O coeficiente de regressão para a equação de **SODIO** e da variável preditora **aveia** é 0,01;

27. O coeficiente de regressão para a equação da **POTASSIO** e da variável preditora **aveia** é 0,09;

28. O coeficiente de regressão para a equação da **SATURADA** e da variável preditora **aveia** é -0,002;

Tabela 9: Coeficientes de regressão para três fatores extraídos.

Coded Regression Coefficients for 3 Extracted Factors			
	SODIO	**POTASSIO**	**SATURADA**
aveia	0.0002862533	0.0870270377	0.0063924081
...	...	...	
chop	0.0014650138	0.0068196786	0.0000000000

Considerando a extração dos três fatores:

29. O coeficiente de regressão para a equação de **SODIO** e da variável preditora **aveia** é 0,0003;

30. O coeficiente de regressão para a equação da **POTASSIO** e da variável preditora **aveia** é 0,09;

31. O coeficiente de regressão para a equação da **SATURADA** e da variável preditora **aveia** é -0,006;

3.3.5 Gráficos da Análise por Redução de Posto (RRR)

A seguir apresentaremos os gráficos gerados pela RRR e suas respectivas interpretações.

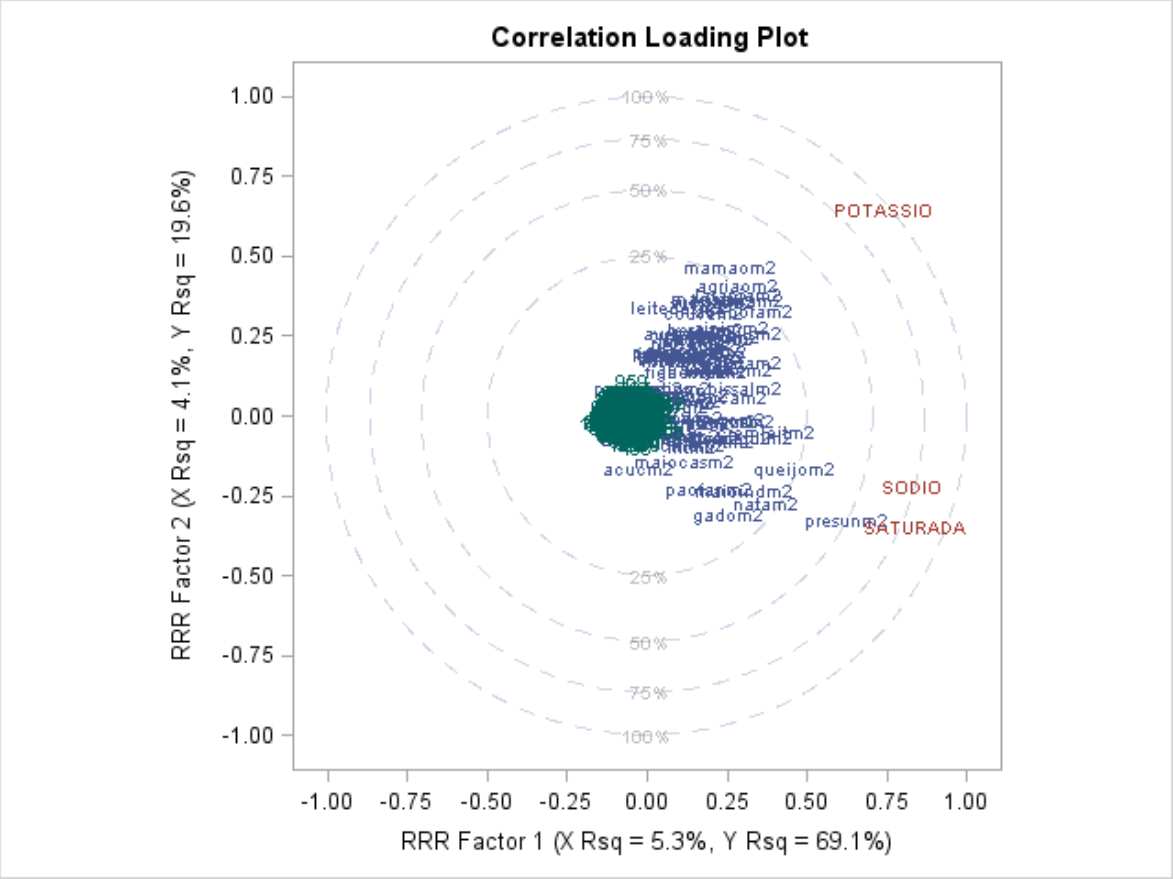


Figura 3: Gráfico de correlação das cargas.

A Figura 3 permite extrair muitas informações[20]:

i) Encontrar padrões e agrupamentos dos indivíduos (cada indivíduo tem um número e ele é colocado no gráfico em cor verde). No nosso exemplo podemos ver claramente um único agrupamento dos indivíduos (parte verde).

ii) Ver qualidade da explicação da variável resposta. Assim, quanto maior for a explicação de cada variável resposta (cor vermelha), mais próxima do círculo 100% ela estará.

iii) As cargas mostram o quanto da variação de cada variável é explicada pelos primeiros dois fatores, conjuntamente pela distância da variável até a origem (onde a variável se encontra no círculo) e individualmente pelas projeções dessa variável sobre o eixo horizontal e vertical. Assim, as variáveis **SODIO** e **SATURADA** são altamente relacionadas ao fator 1 e pouco relacionadas com o fator 2, já a variável **POTASSIO** é altamente relacionada com os dois fatores.

iv) É possível utilizar o gráfico para analisar a relação das variáveis umas com as outras. Por exemplo, a variável **presunm2** (presunto) está altamente relacionada às variáveis resposta **SATURADA** e **SODIO**.

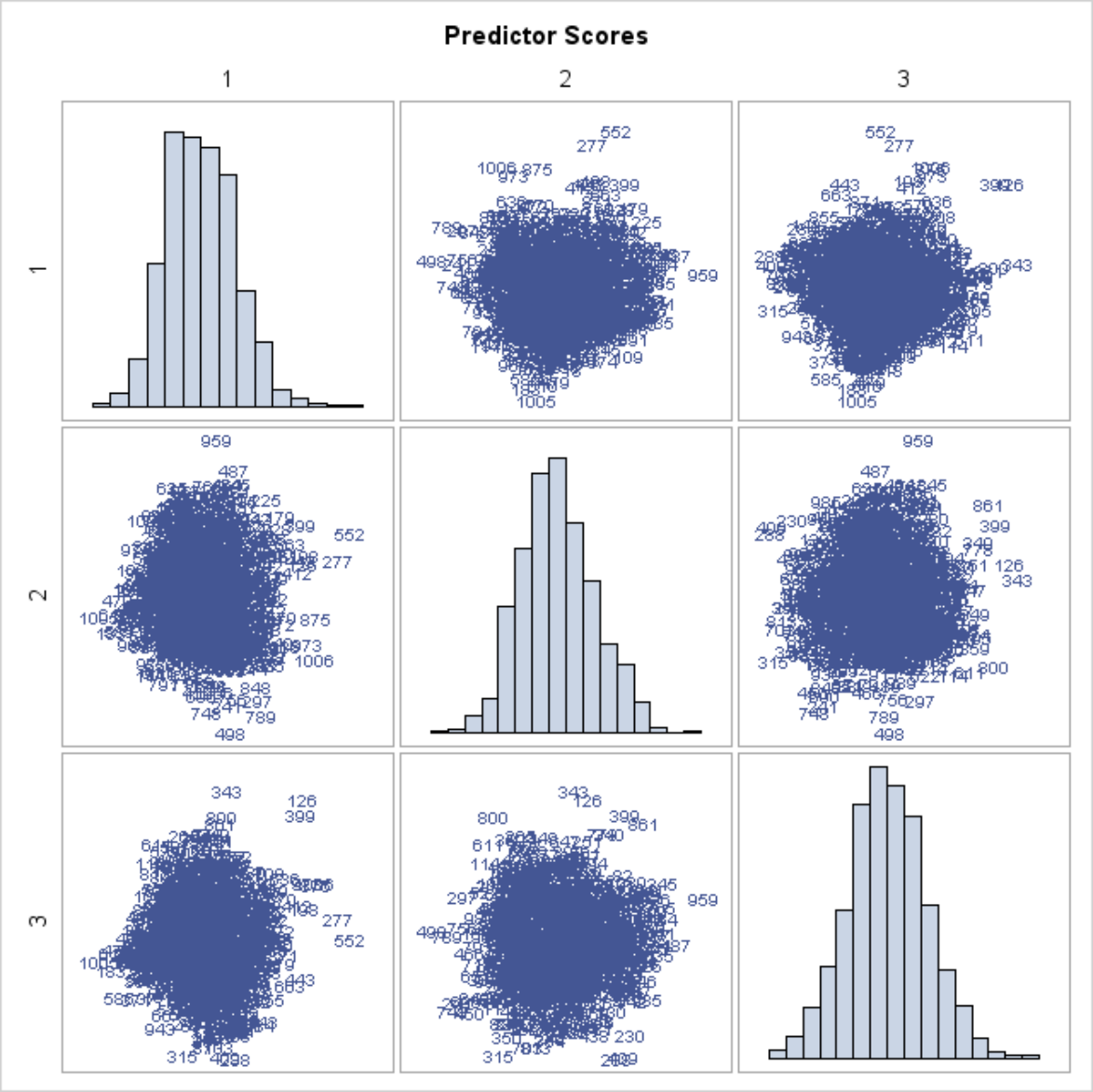


Figura 4: Matriz de gráficos dos escores das variáveis preditoras de cada fator.

A Figura 4 mostra a matriz de gráficos dos escores das variáveis preditoras de cada fator para cada indivíduo. Na diagonal principal tem-se os histogramas dos escores de cada fator. Nas demais posições tem-se gráficos de dispersão dos escores das variáveis preditoras para dois fatores, onde cada ponto é identificado pelo número do indivíduo. Por exemplo, na linha 1 coluna 1, temos os escores das variáveis preditoras do fator 1 *versus* os do fator 2, não encontramos nenhum padrão neles (nuvem), mostrando a independência entre os fatores.

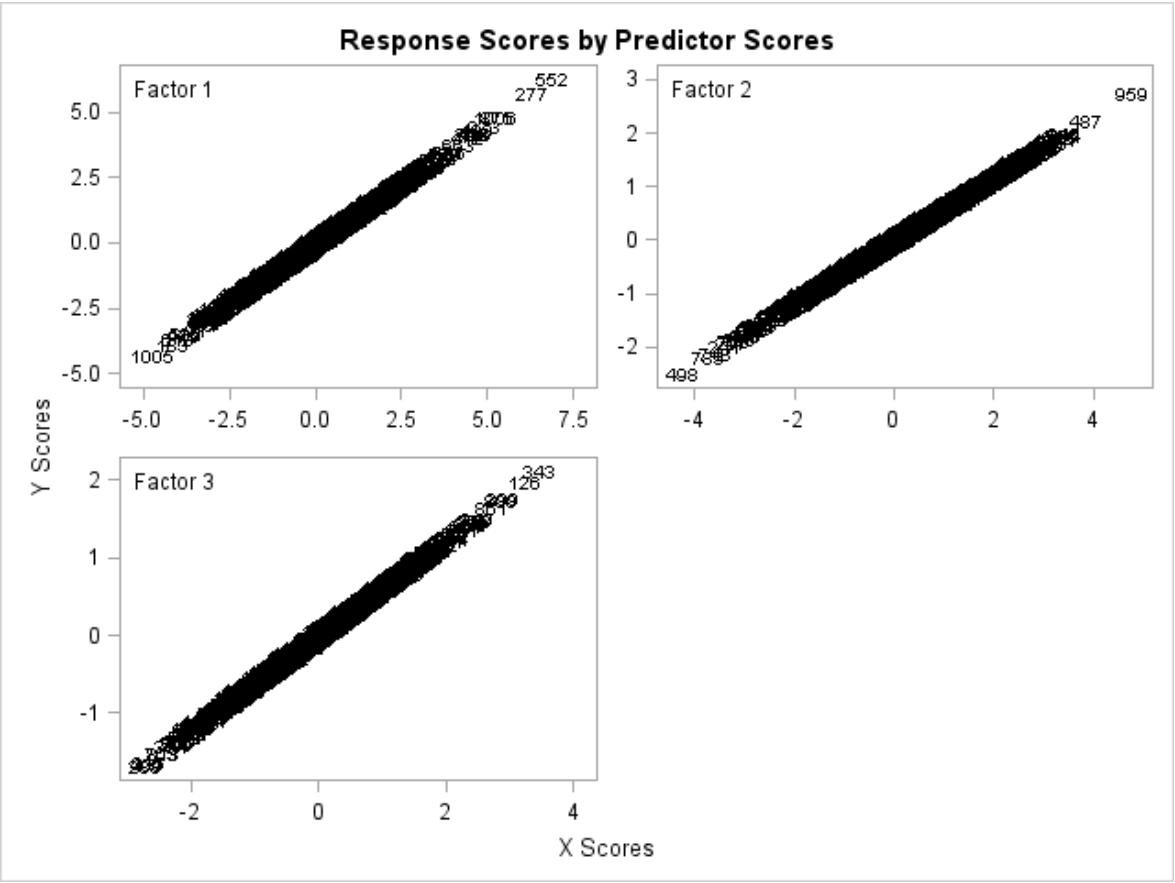


Figura 5: Gráficos dos escores das respostas *versus* os escores das preditoras.

A Figura 5 mostra os escores das variáveis resposta *versus* os escores das variáveis preditoras de cada fator. Nesse exemplo, existe alta correlação entre os escores de X e Y para os fatores 1, 2 e 3. Isso se deve ao fato, já mencionado, das variáveis resposta terem sido calculadas a partir das variáveis preditoras.

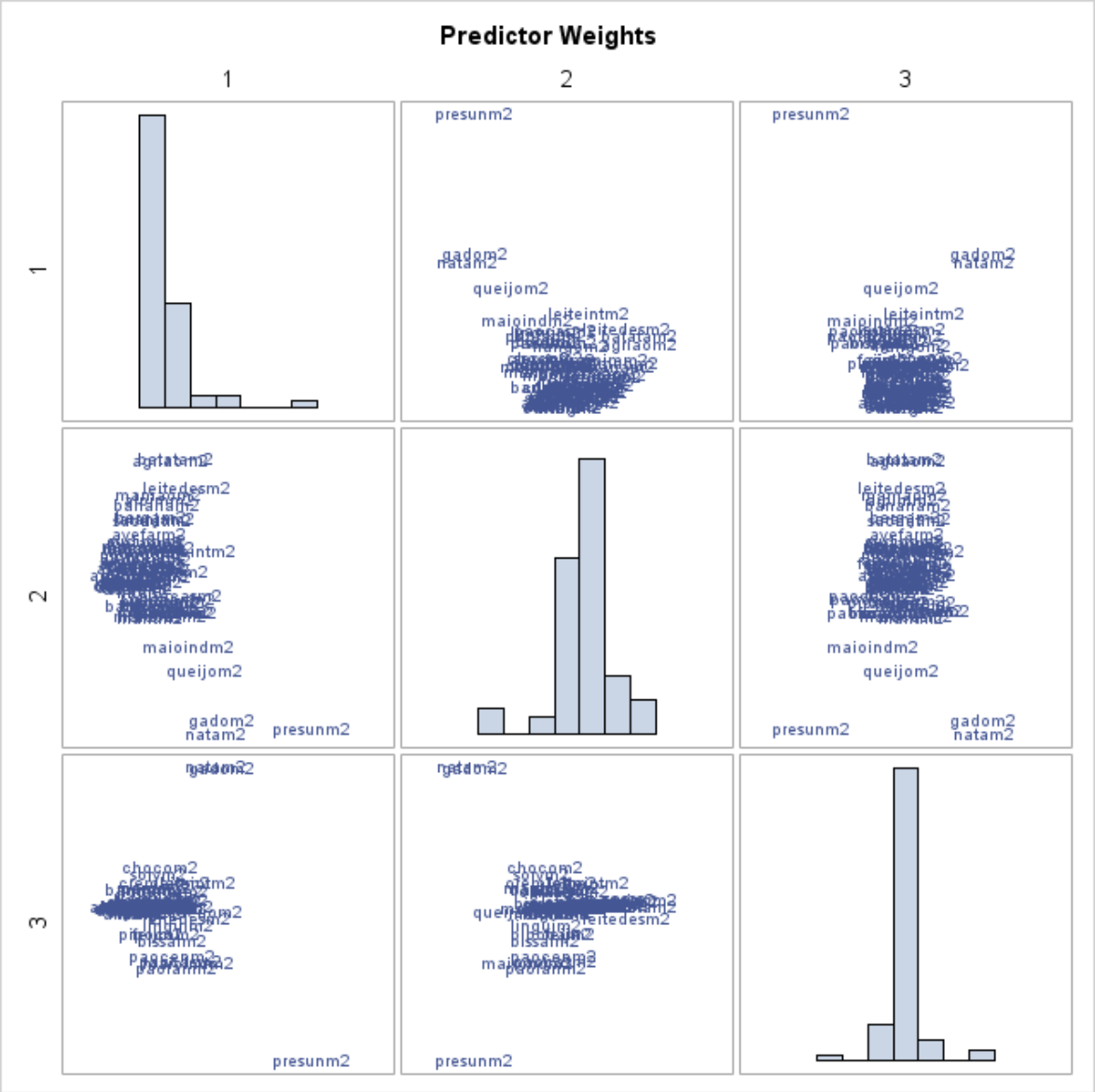


Figura 6: Matriz de gráficos dos pesos das variáveis preditoras.

A Figura 6 mostra a matriz de gráficos dos pesos das variáveis preditoras de cada fator. Na diagonal principal tem-se os histogramas dos pesos de cada fator. Nas demais posições tem-se gráficos de dispersão dos pesos das variáveis preditoras. Na linha 1 coluna 1, temos os pesos das variáveis preditoras do fator 1 *versus* os do fator 2. Em particular, podemos notar que a variável **presunm2** tem um alto peso no fator 1 e baixo peso no fator 2.

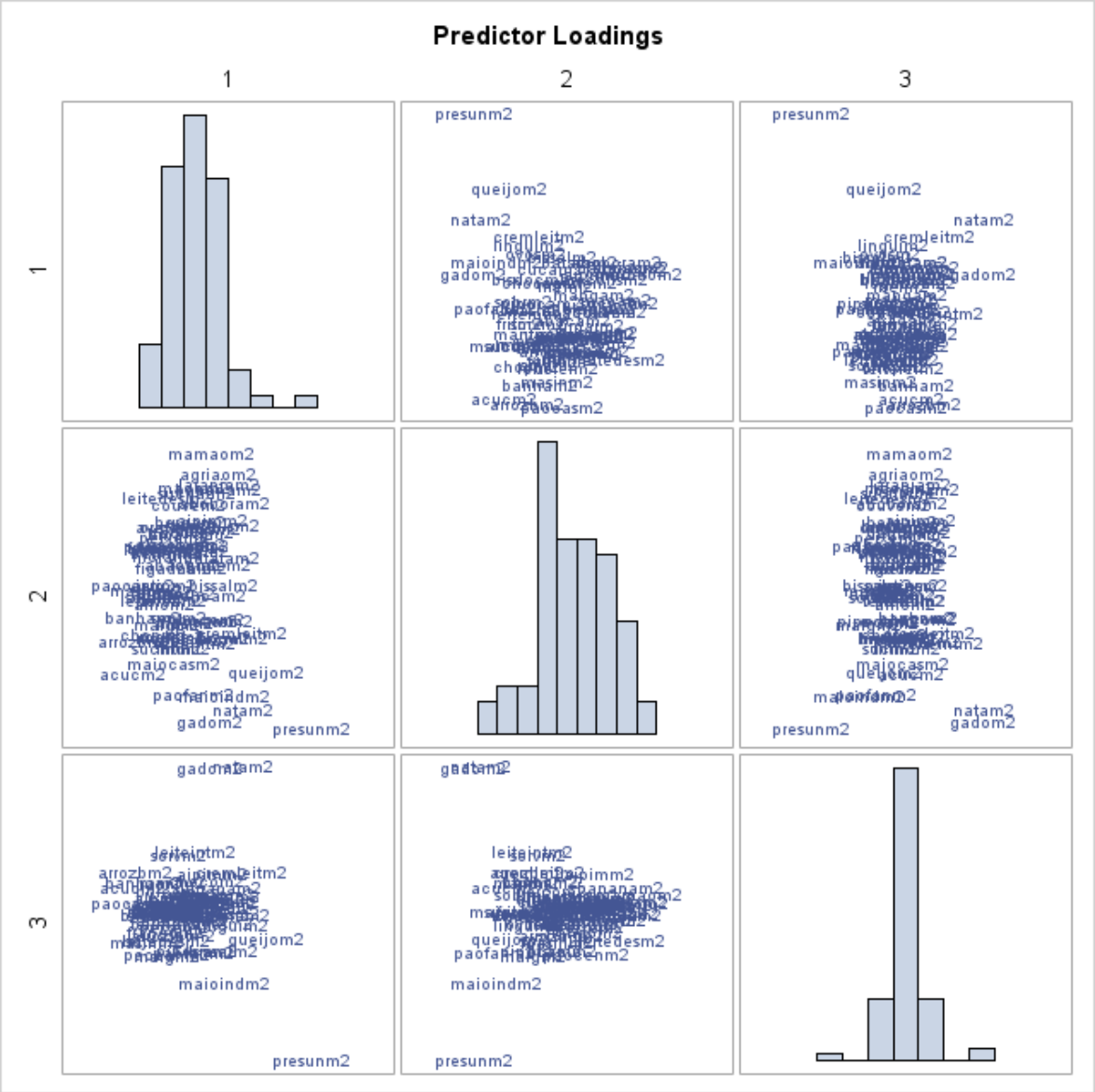


Figura 7: Matriz de gráficos das cargas das variáveis preditoras.

A Figura 7 mostra a matriz de gráficos das cargas das variáveis preditoras de cada fator. Na diagonal principal tem-se os histogramas das cargas de cada fator. Nas demais posições tem-se gráficos de dispersão das cargas das variáveis preditoras. A interpretação é análoga a Figura 6.

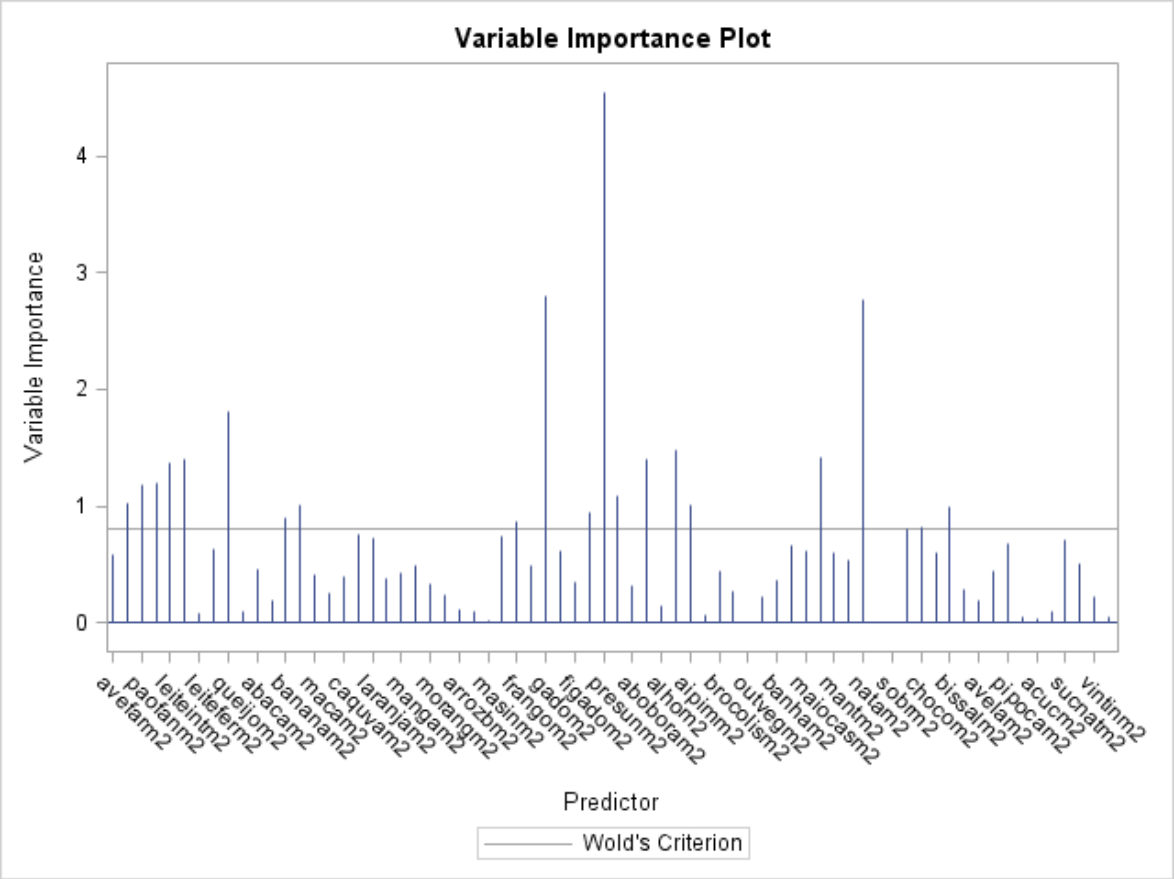


Figura 8: Gráfico de importância das variáveis (VIP).

A Figura 8 representa a contribuição de cada variável preditora no modelo ajustado RRR. Assim, se uma variável preditora tem um coeficiente em módulo muito pequeno e a importância dela é baixa (Critério de Wold < 0,8 – linha cinza horizontal no gráfico) então ela é uma forte candidata a ser retirada do modelo[20]. Esse gráfico deve ser analisado conjuntamente com a Figura 9.

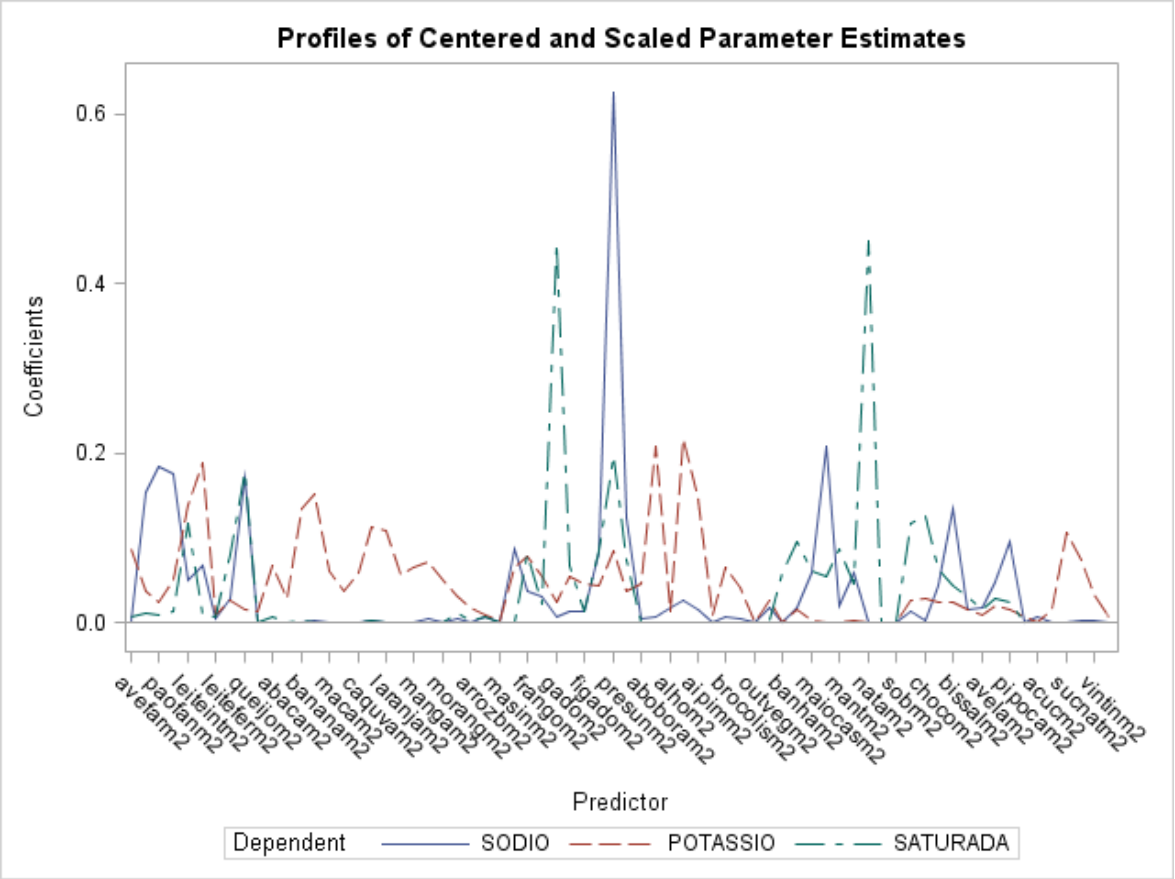


Figura 9: Perfil das estimativas dos parâmetros centrados e escalonados.

A Figura 9 apresenta os coeficientes absolutos das variáveis preditoras para cada variável resposta.

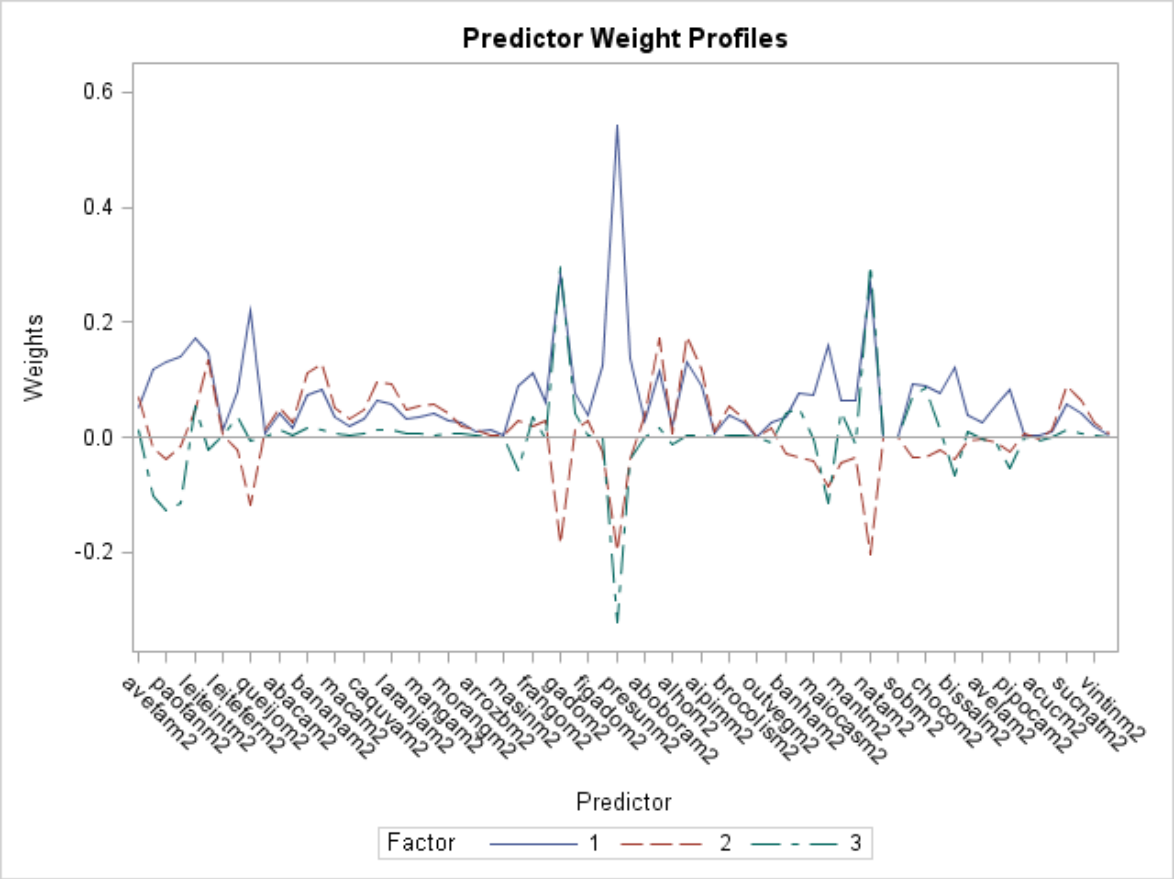


Figura 10: Perfil dos pesos das variáveis preditoras.

A Figura 10 é o gráfico dos pesos das variáveis preditoras para cada fator. Podemos observar quais são as variáveis preditoras que têm as maiores e menores pesos em cada fator. Este gráfico refere-se a Tabela 6.

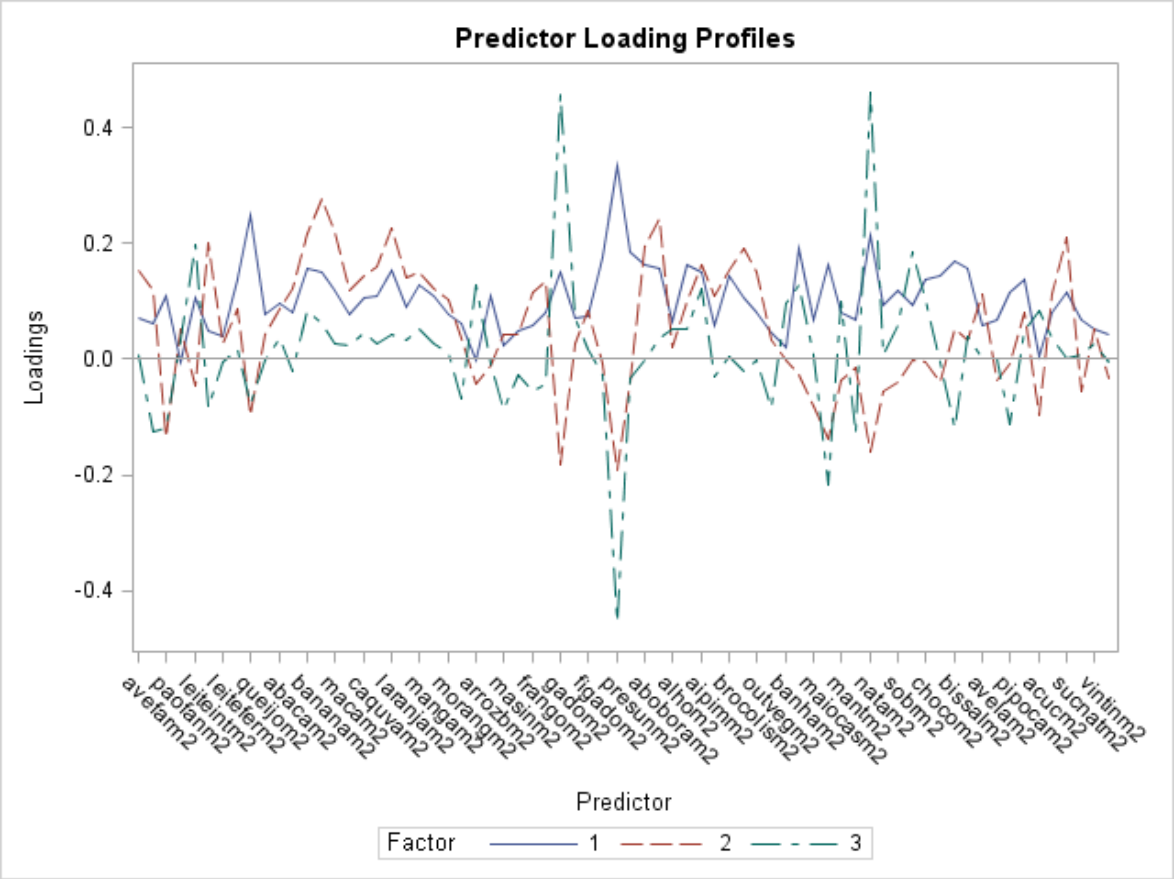


Figura 11: Perfil das cargas das variáveis preditoras.

A Figura 11 é o gráfico das cargas das variáveis preditoras para cada fator. Podemos observar quais são as variáveis preditoras que têm as maiores e menores cargas em cada fator21. Este gráfico refere-se a Tabela 5.

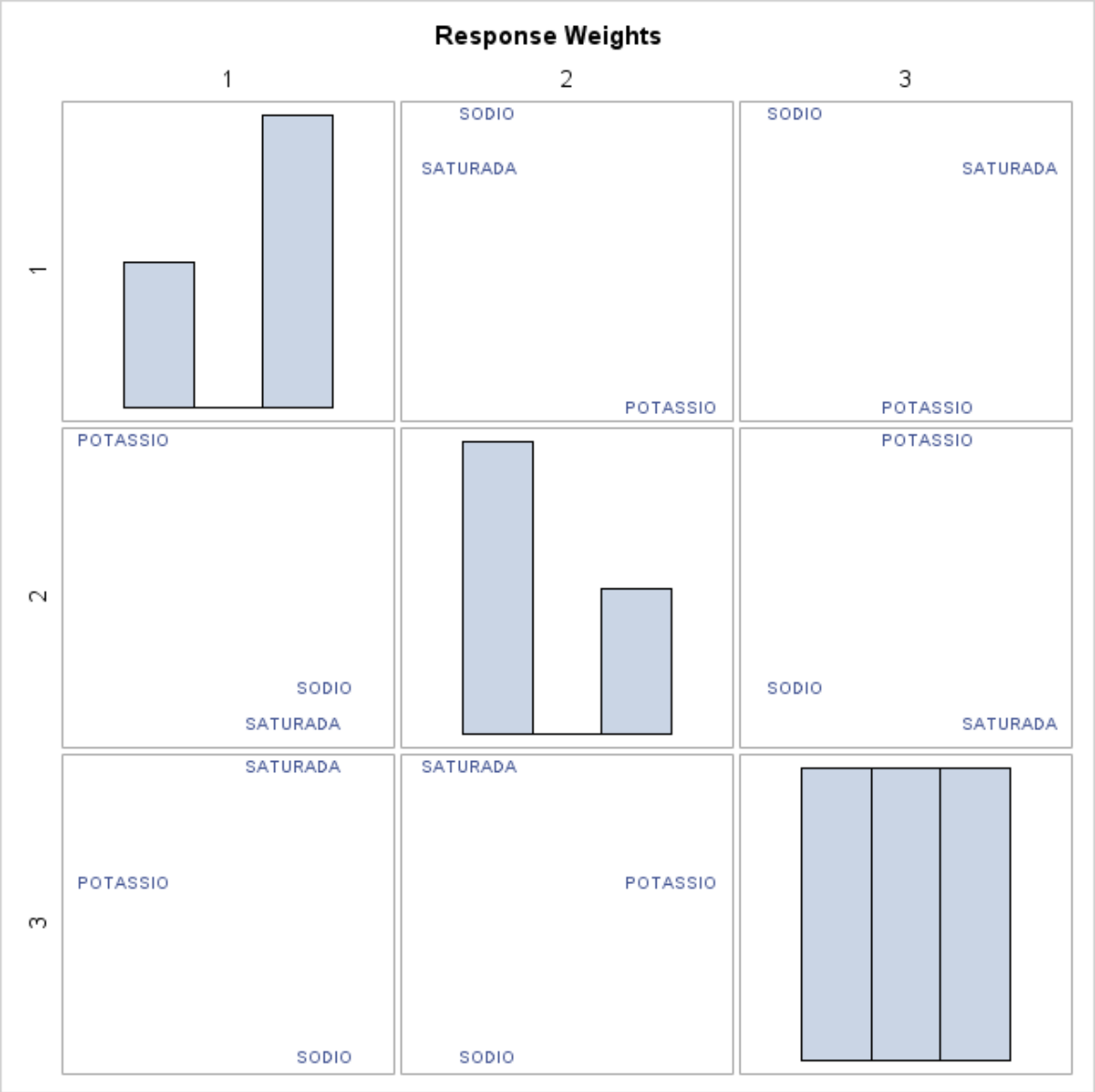


Figura 12: Matriz de gráficos dos pesos das variáveis resposta.

A Figura 12 mostra a matriz de gráficos dos pesos das variáveis resposta. Na diagonal principal tem-se os histogramas dos pesos de cada fator. Nas demais posições tem-se gráficos de dispersão dos pesos das variáveis resposta. Na linha 1 coluna 1, temos os pesos das variáveis resposta do fator 1 *versus* os do fator 2. As variáveis **SODIO** e **SATURADA**, por exemplo, têm um alto peso no fator 1 e baixo peso no fator 2, já com a variável **POTASSIO**, ocorre o inverso.

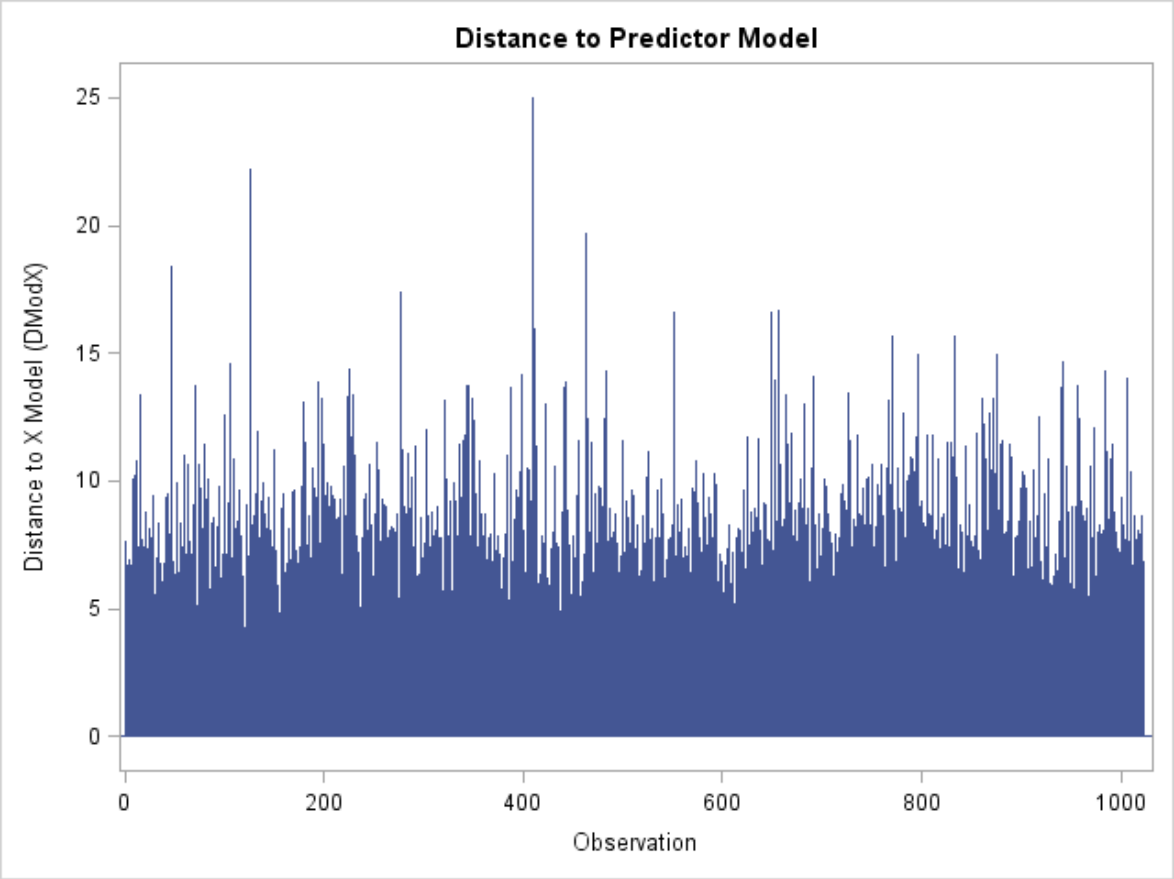


Figura 13: Distância de cada observação até o modelo para as preditoras.

O gráfico da Figura 13 mostra a distância de cada observação até o modelo para as preditoras[22]. Assim, o modelo não é adequado para as observações que tem distâncias muito maiores em relação às outras.

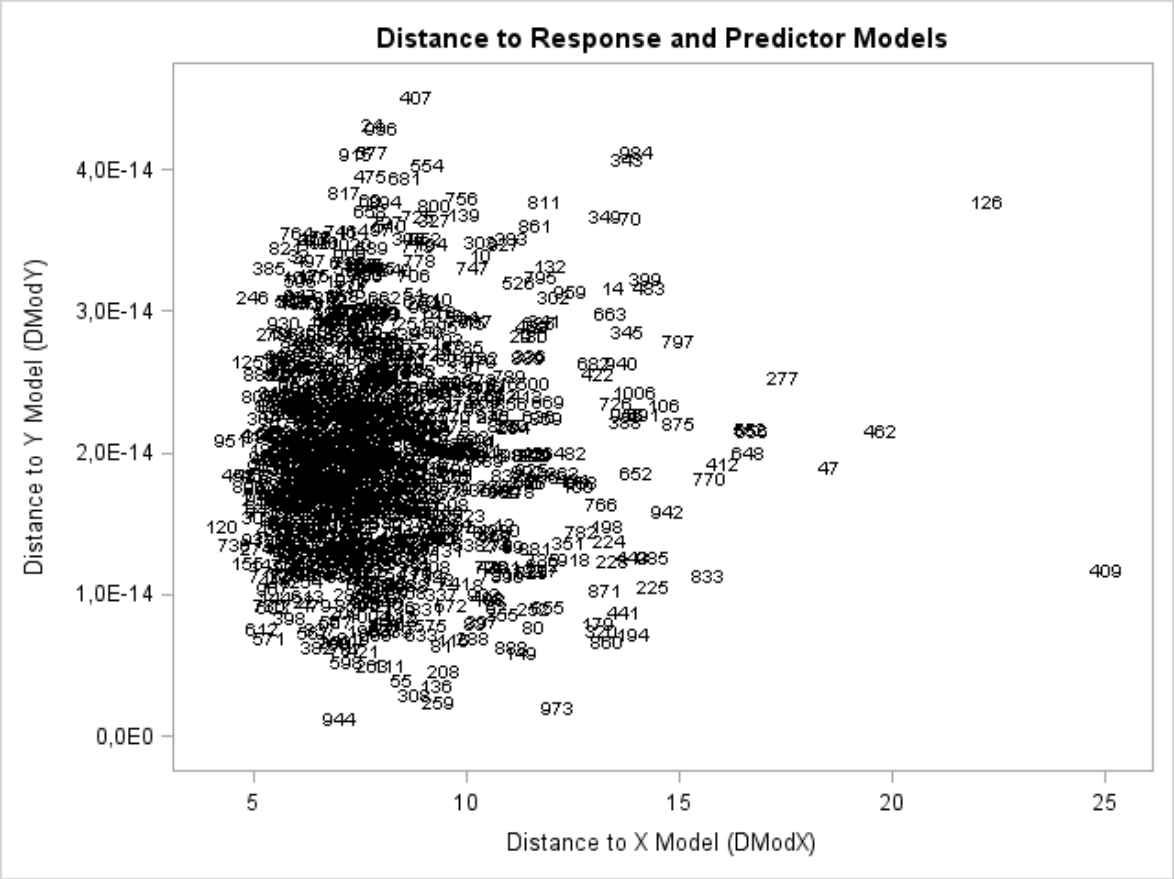


Figura 14: Distância das observações até o modelo para as resposta *versus* a distância das observações até o modelo para as preditoras.

Na Figura 14 podemos ver qual é a distância das observações até o modelo para as resposta e para as preditoras. Por exemplo, a observação 126 tem uma distância muito grande em relação ao modelo para as preditoras e para as resposta, ou seja, ambos os modelos não são adequados para essa observação.

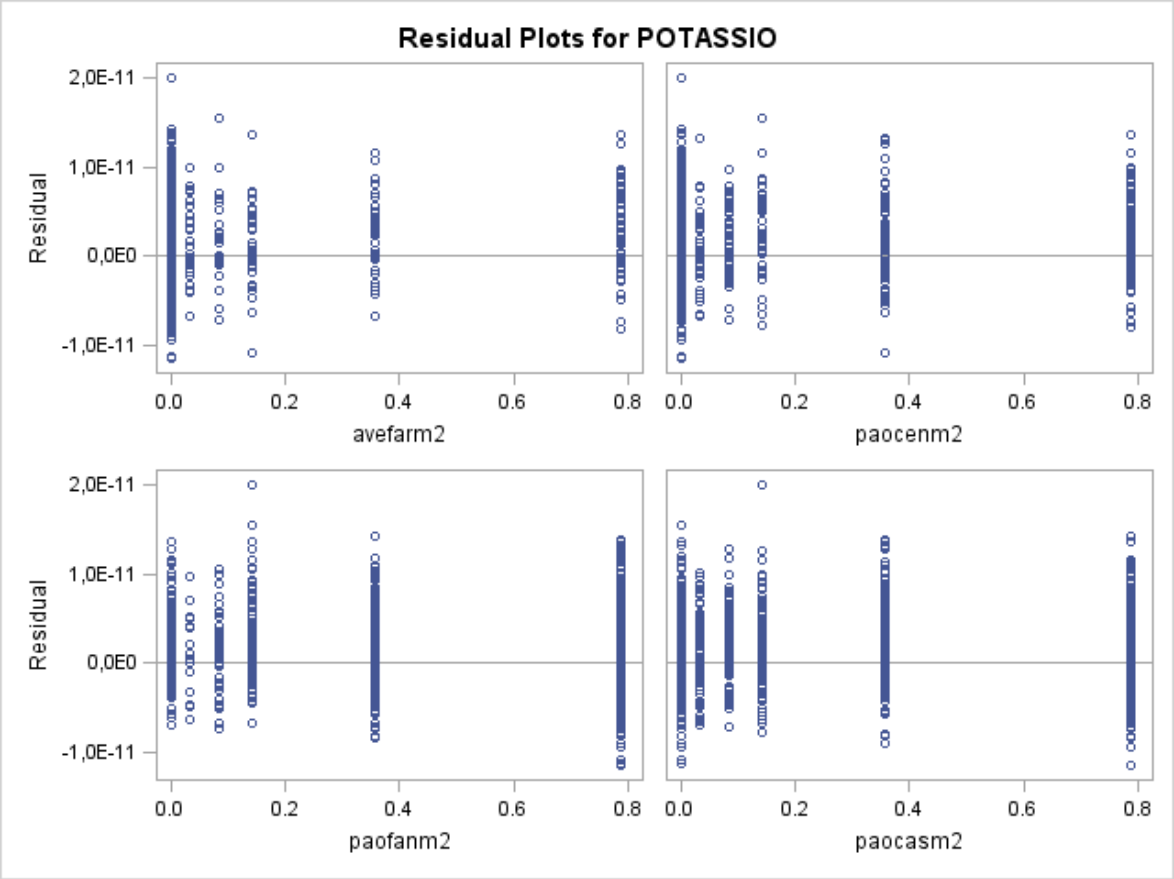


Figura 15: Gráfico dos resíduos *versus* variáveis preditoras para **POTASSIO**.

O programa gera gráficos para todas as variáveis preditoras e serão omitidos nesse trabalho. A Figura 15 é o gráfico dos resíduos *versus* as variáveis preditoras. Nos quatro gráficos os resíduos se distribuem aleatoriamente em torno da média zero, pode-se dizer que o modelo é adequado.

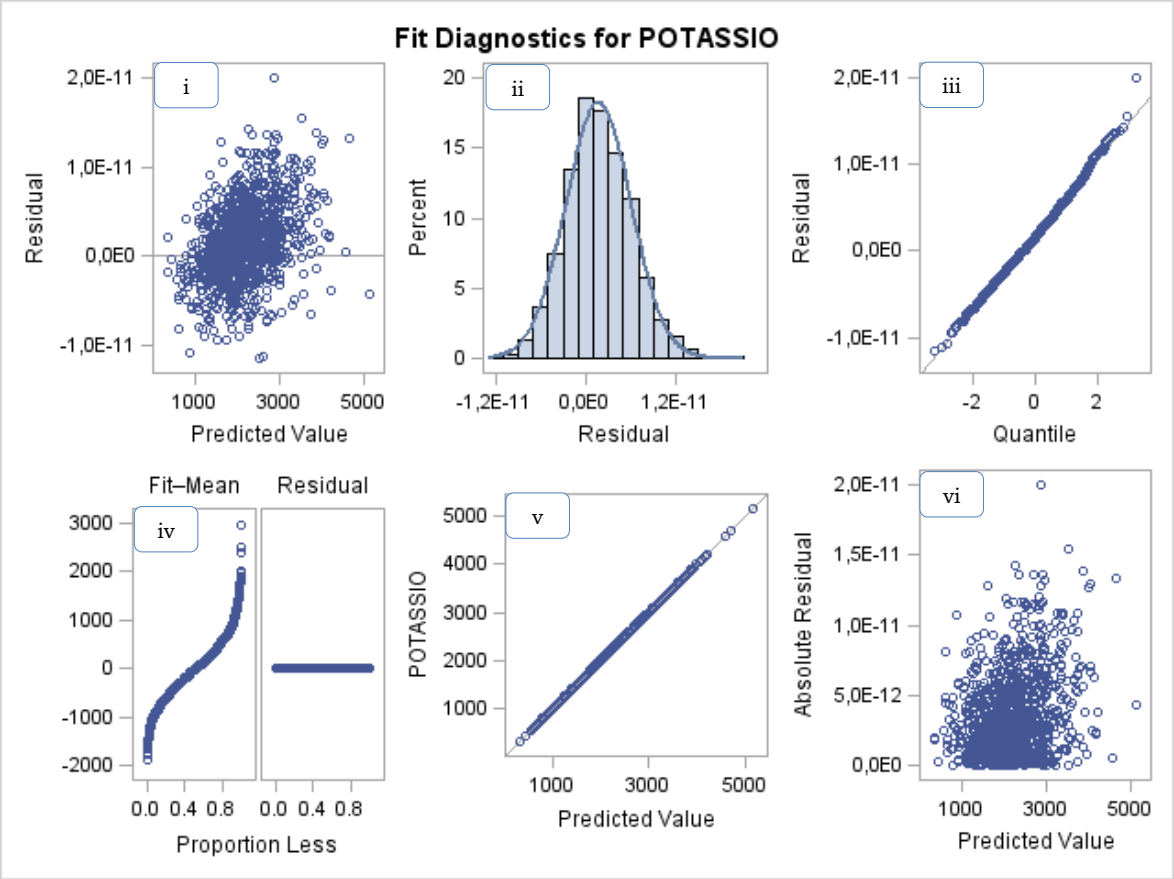


Figura 16: Gráficos de diagnóstico para a variável **POTASSIO**.

A Figura 16 traz os gráficos de diagnóstico de ajuste do modelo para a variável **POTASSIO**. O programa gera os gráficos para as variáveis **SODIO** e **SATURADA**, mas eles serão omitidos.

i) É o gráfico dos valores preditos *versus* os resíduos. Ele apresenta um comportamento errático, ou seja, os resíduos não têm nenhum padrão definido. Também podemos observar a presença de valores discrepantes (*outliers*);

ii) É o histograma dos resíduos. Através dele, podemos verificar se os resíduos têm aproximadamente distribuição normal. O gráfico mostra um bom ajuste.

iii) É o gráfico de probabilidade normal, que também é útil para verificar a suposição de normalidade dos resíduos. No caso, está bem ajustado (valores em cima da linha diagonal traçada).

iv) Esse gráfico é chamado de *Residual-Fit* ou *RF Plot*. Nesse gráfico podemos avaliar se o modelo é inadequado ou não. Se a dispersão dos resíduos for maior que a dispersão do ajuste centrado, o modelo é considerado inadequado[23].

v) Gráfico da variável resposta **POTASSIO** *versus* valores preditos. Confirma os resultados dos gráficos anteriores, que o modelo ajustado faz boas predições (Valores em torno da linha diagonal).

vi) Gráfico dos resíduos absolutos *versus* valores preditos. Tem uma interpretação semelhante ao do gráfico (i), porém os resíduos negativos são projetados na parte positiva, assim podemos observar um comportamento errático dos resíduos e a presença de *outliers*.

A saída completa do SAS pode ser encontrada em: www.mat.ufrgs.br\~camey\RRR_nutrientes.

3.4 Interpretação dos Fatores

Para interpretar os fatores construímos a Tabela 10 que contém os alimentos com valores absolutos das cargas acima de 0,2.

Tabela 10: Variáveis preditoras com maiores cargas em valor absoluto para o fator 1.

	Fator 1	**Fator 2**	**Fator 3**
Presunto	0,3356	*	-0,4561
Queijo	0,2482	*	*
Nata	0,2133	*	0,4615
Mamão	*	0,2772	*
Agrião	*	0,2430	*
Laranja	*	0,2251	*
Maçã	*	0,2192	*
Banana	*	0,2163	*
Suco Natural	*	0,2119	*
Leite Desnatado	*	0,2013	*
Gado	*	*	0,4589
Maionese Industrial	*	*	-0,2169

*Valores em módulo menores que 0,2.

Na Tabela 10 são apresentadas as variáveis preditoras com as cargas em valor absoluto maiores que 0,20 para o fator 1 para o fator 2 e para o fator 3. Desta forma podemos ver quais são os alimentos que mais influenciam em cada fator. Assim, podemos ver que os alimentos que mais influenciam no fator 1 são o presunto, o queijo e a nata. No fator 2 são o mamão, o agrião, a laranja, a maçã, a banana, o suco natural e o leite desnatado. No fator 3 são a nata, o gado, o presunto e a maionese industrial, sendo os dois últimos no sentido inverso.

É importante sabermos como os escores das variáveis resposta são calculados, pois isso facilita a interpretação dos fatores. A seguir exemplificaremos o cálculo para o sujeito 1. Primeiro temos que padronizar o consumo de cada nutriente, ou seja, calcular o escore Z de cada um. Para o sujeito 1 temos que os escores Z de cada nutriente são:

$$Z_{SODIO} = \frac{1851{,}71 - 1388{,}22}{528{,}44} = 0{,}877$$

$$Z_{POTASSIO} = \frac{3088{,}08 - 2192{,}61}{700{,}05} = 1{,}279$$

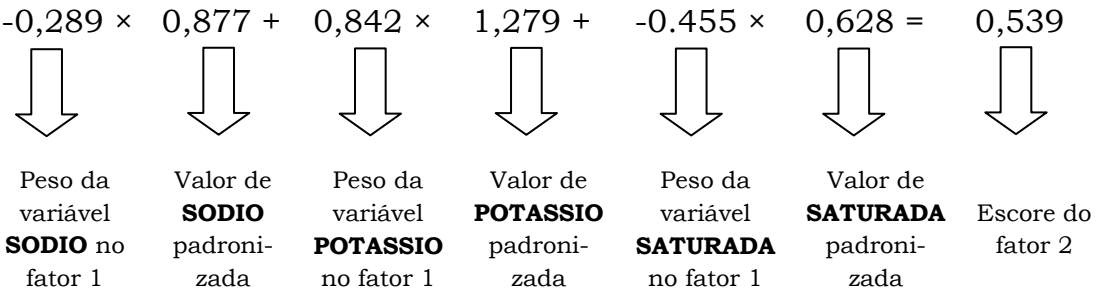
$$Z_{SATURADA} = \frac{35{,}37 - 28{,}63}{10{,}73} = 0{,}628$$

Esse indivíduo, portanto, ingere acima da média sódio, potássio e gordura saturada.

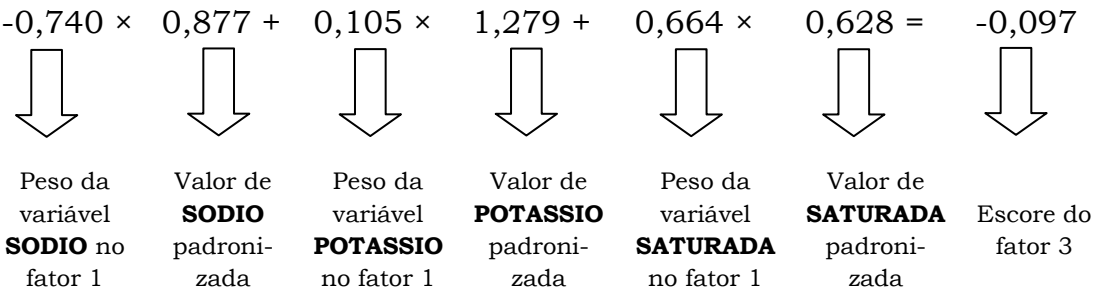
Assim, para o sujeito 1, o escore do fator 1 é:

0,608 ×	0,877 +	0,528 ×	1,279 +	0,593 ×	0,628 =	1,581
Peso da variável **SODIO** no fator 1	Valor de **SODIO** padroni-zada	Peso da variável **POTASSIO** no fator 1	Valor de **POTASSIO** padroni-zada	Peso da variável **SATURADA** no fator 1	Valor de **SATURADA** padroni-zada	Escore do fator 1

E o escore do fator 2 é:



E o escore do fator 3 é:



Portanto, o escore do fator 1 será alto para indivíduos que consomem sódio, potássio e gordura saturada acima da média. O escore do fator 2 será alto para indivíduos que consomem potássio acima da média e gordura saturada abaixo da média. E o escore do fator 3 será alto para indivíduos que consomem gordura saturada acima da média e sódio abaixo da média.

Na análise, os valores dos escores e os valores padronizados das variáveis resposta podem ser encontrados no novo banco de dados **pattern** criado através da linha de comando (5.a), nas colunas de nomes scorey1, scorey2 e scorey3, ypad1, ypad2 e ypad3, respectivamente. Os pesos das variáveis resposta de cada fator encontram-se na Tabela 6.

3.5 Relacionando os Fatores da RRR com o Desfecho

Apenas para ilustrar a utilização dos escores fatoriais os comparamos entre os sujeitos com ou sem hipertensão autorreferida.

Foi utilizado o software SPSS 18 para fazermos a relação entre os fatores obtidos na RRR e a hipertensão arterial autorreferida (**temhas**).

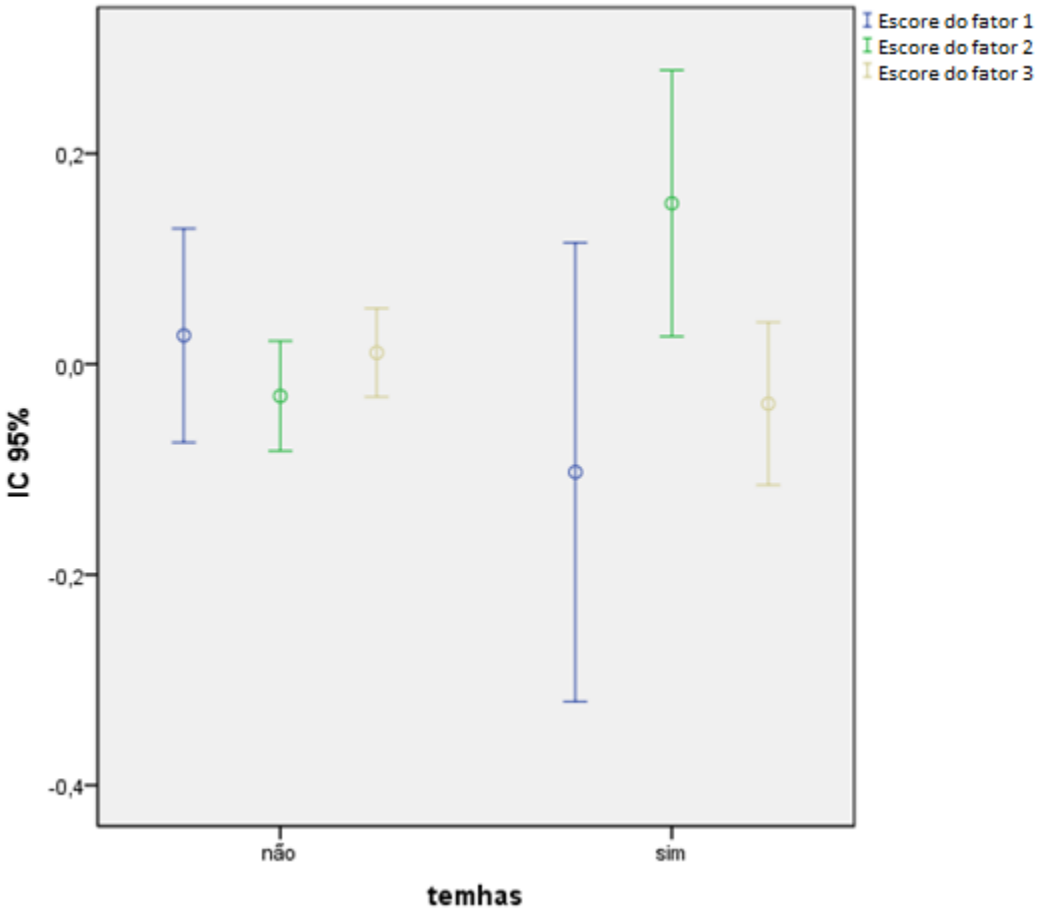


Figura 17: Gráfico de erros.

Não há evidências amostrais de que as médias do escore do fator 1 para indivíduos que se autorreferem hipertensos ou não são diferentes (p=0,276). E também, não há evidências amostrais que as médias do escore do fator 3 para indivíduos que se autorreferem hipertensos ou não são diferentes (p=0,312). Por fim, há evidências amostrais de que as médias do escore do fator 2 para indivíduos que se autorreferem hipertensos ou não, são diferentes (p=0,009)

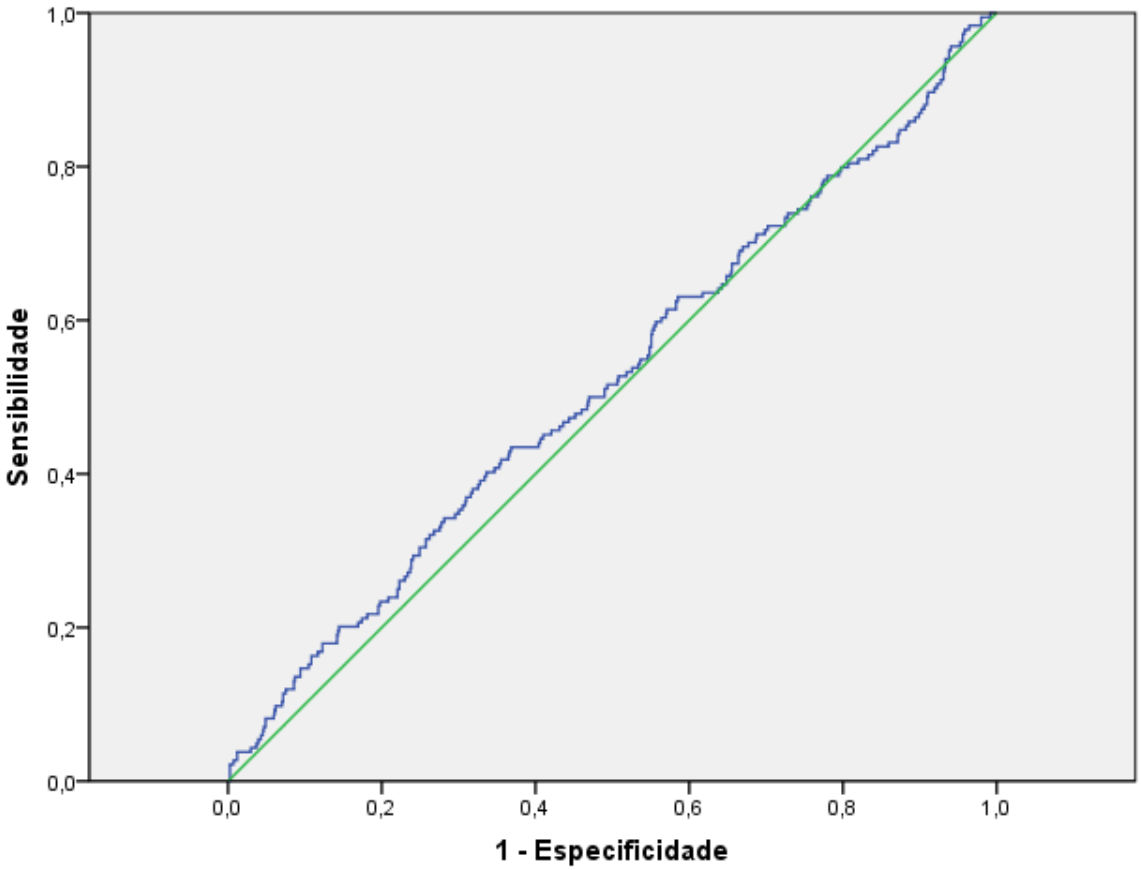


Figura 18: Curva característica de operação (ROC) do fator 1.

A curva ROC foi construída para verificar quão bem os fatores extraídos da RRR predizem a hipertensão. Para o fator 1 a área sob a curva é igual a 0,521 (IC95%: 0,473; 0,569), isso significa que este fator não deve discriminar bem os indivíduos hipertensos dos não hipertensos.

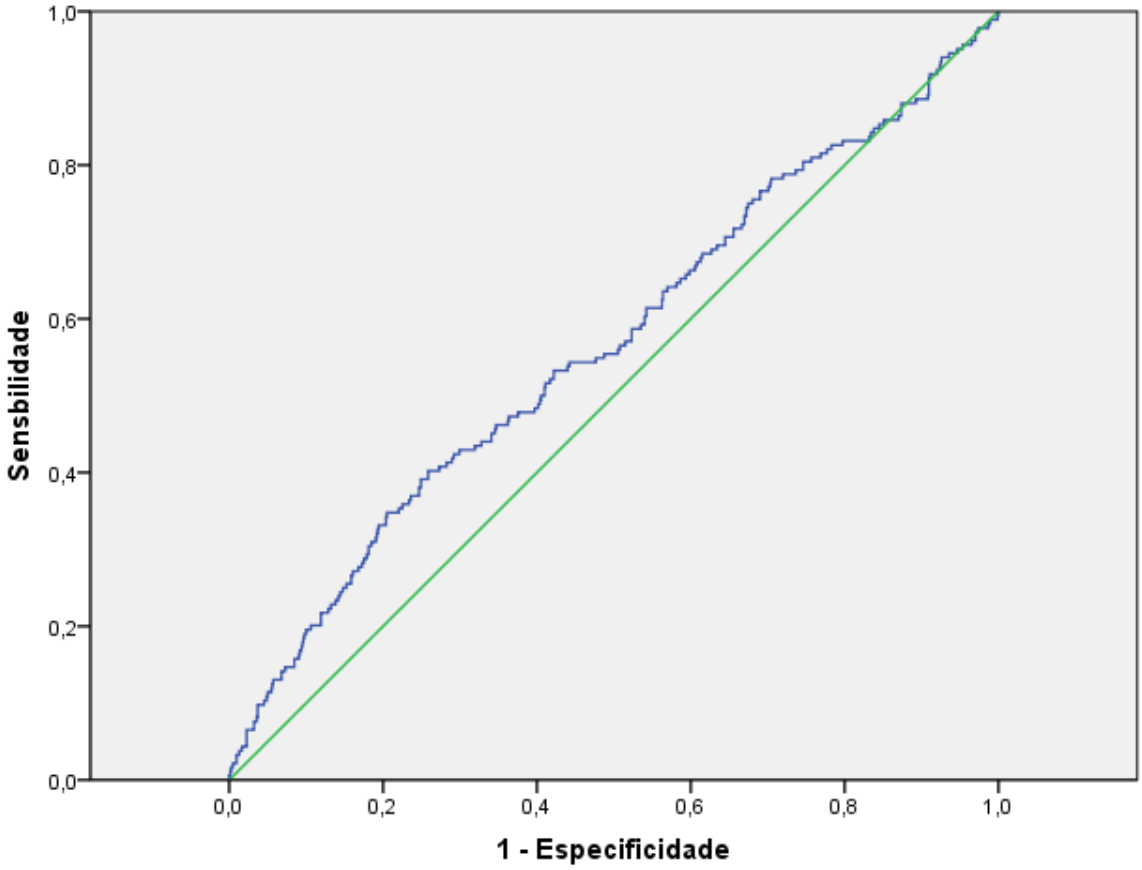


Figura 19: Curva característica de operação (ROC) do fator 2.

Para o fator 2, a área sob a curva ROC é igual a 0,566 (IC95%: 0,517; 0,614), o que significa que o fator 2 não é um bom critério de diagnóstico para hipertensão.

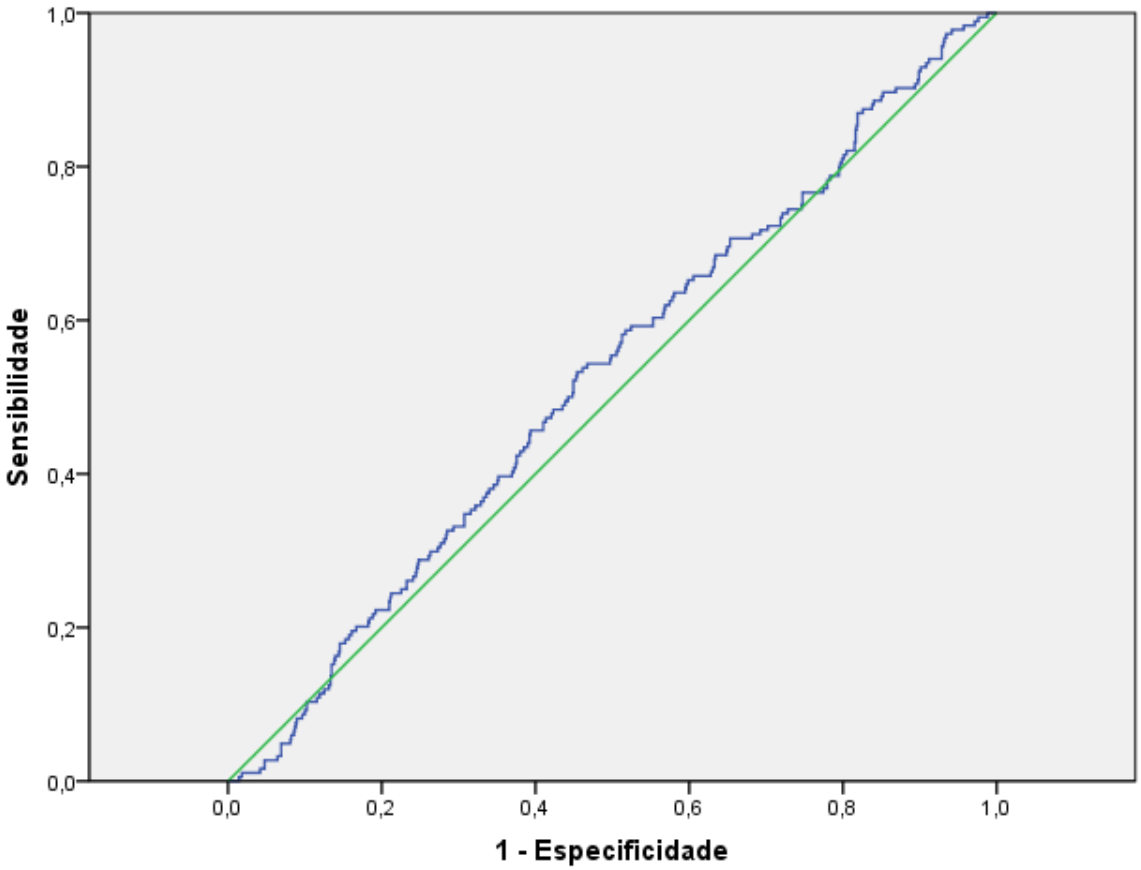


Figura 20: Curva característica de operação (ROC) do fator 3.

Para o fator 3, a área sob a curva ROC é igual a 0,527 (IC95%: 0,482; 0,572), o que significa que o fator 3 não é um bom critério de diagnóstico para hipertensão.

4 Conclusão

Neste trabalho, a Regressão por Redução de Postos (RRR) foi estudada e exemplificada no contexto de epidemiologia nutricional. Foi visto que o método RRR combina informação a priori e a posteriori, ou seja, é a priori, pois utiliza a informação existente da associação entre o desfecho e as variáveis intermediárias e é a posteriori, pois utiliza a informação dos dados do estudo de consumo alimentar.

No entanto, se não houver a informação a priori da associação das variáveis intermediárias com o desfecho, devemos preferir a técnica exploratória de análise de componentes principais (PCA) à RRR, pois sem esse conhecimento a priori não podemos justificar o uso dessas variáveis resposta[5].

Como visto nesse trabalho, uma das suposições do modelo é que as variáveis resposta tenham distribuição normal multivariada, por isso é importante que seja feito o teste de normalidade multivariada para as variáveis resposta. E também precisamos ser cuidadosos em relação aos dados faltantes das variáveis resposta e das variáveis preditoras, uma vez que a RRR trabalha com matrizes de completas, ou seja, se houver algum dado faltante nas variáveis resposta e/ou nas variáveis preditoras de um indivíduo, esse sujeito não entrará na análise; logo, perderemos informações. Também é fundamental que seja feita uma boa análise de resíduos para investigarmos a adequabilidade do modelo ajustado.

É importante ressaltar que na RRR, o número máximo de fatores que a técnica encontra é igual ao número de variáveis resposta existentes[5]. No entanto, podemos ter a redução do número de fatores, para isso precisamos fazer um teste para saber qual é o número de fatores ideal para ser utilizado na análise[16].

Uma desvantagem da RRR apontada por Hoffmann[5] é que os coeficientes dos escores dos fatores são estimados através dos dados disponíveis e eles não podem ser reproduzidos para dados de outra população que estiver sendo estudada.

Nesse trabalho foi utilizado o software SAS, mas a RRR pode ser feita nos softwares R e S-PLUS <http://lib.stat.cmu.edu/S/rrr.s>.

5 Referências

1. REINSEL, G. **Multivariate Reduced-Rank Regression: Theory and Applications**. New York: Springer, 1998. ISBN 9780387986012.

2. **SBC / DHA - Departamento de Hipertensão Arterial**. Disponível em: <http://departamentos.cardiol.br/dha/vdiretriz/07-tratamento.pdf>. Acesso em: 18 jun. 2011.

3. MCCANN, S. E.; MCCANN, W. E.; HONG, C.-C.; et al. Dietary Patterns Related to Glycemic Index and Load and Risk of Premenopausal and Postmenopausal Breast Cancer in the Western New York Exposure and Breast Cancer Study. **The American Journal of Clinical Nutrition**, 2007. v. 86, n. 2, p. 465 -471.

4. NETTLETON, J. A.; STEFFEN, L. M.; SCHULZE, M. B.; et al. Associations Between Markers of Subclinical Atherosclerosis and Dietary Patterns Derived by Principal Components Analysis and Reduced Rank Regression in the Multi-Ethnic Study of Atherosclerosis (MESA). **The American Journal of Clinical Nutrition**, 1 Jun 2007. v. 85, n. 6, p. 1615 -1625.

5. HOFFMANN, K. Application of a New Statistical Method to Derive Dietary Patterns in Nutritional Epidemiology. **American Journal of Epidemiology**, 2004. v. 159, n. 10, p. 935-944.

6. DIBELLO, J. R.; KRAFT, P.; MCGARVEY, S. T.; et al. Comparison of 3 Methods for Identifying Dietary Patterns Associated With Risk of Disease. **American Journal of Epidemiology**, 2008. v. 168, n. 12, p. 1433-1443.

7. CUNHA, D. B.; ALMEIDA, R. M. V. R. DE; PEREIRA, R. A. A Comparison of Three Statistical Methods Applied in the Identification of Eating Patterns. **Cadernos de Saúde Pública**, 2010. v. 26, n. 11, p. 2138-2148.

8. MCNAUGHTON, S. A.; MISHRA, G. D.; BRUNNER, E. J. Dietary Patterns, Insulin Resistance, and Incidence of Type 2 Diabetes in the Whitehall II Study. **Diabetes Care**, 2008. v. 31, n. 7, p. 1343-1348.

9. ALVES, A. L. S. OLINTO, M. T. A. COSTA, J. S. D. DA; BAIRROS, F. S. DE; BALBINOTTI, M. A. A. Padrões alimentares de mulheres adultas

residentes em área urbana no sul do Brasil. **Revista de Saúde Pública**, v. 40, n. 5, out 2006.

10. ALDRIN, M. **Encyclopedia of Environmetrics**. Chichester, New York: Wiley, 2002. ISBN 9780471899976.

11. IZENMAN, A. J. Reduced-Rank Regression for the Multivariate Linear Model. **Journal of Multivariate Analysis,** 1875. v. 5, n. 2, p. 248-264.

12. TSO, M. K.-S. Reduced-Rank Regression and Canonical Analysis. **Journal of the Royal Statistical Society. Series B (Methodological)**, 1981. v. 43, p. 183-189.

13. ANDERSON, T. W. Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions. **The Annals of Mathematical Statistics**, 1951. v. 22, n. 3, p. 327-351.

14. HANSEN, P. R. On the Estimation of Reduced Rank Regressions. **Brown University, Department of Economics**, 2002b.

15. JOHANSEN, S. Reduced rank regression. In: DURLAUF, S. N.; BLUME, L. E. (Eds.). **The New Palgrave Dictionary of Economics**. Basingstoke: Nature Publishing Group, 2008. 2nd ed., p. 50-53. ISBN 978-0-333-78676-5.

16. **The PLS Procedure: Cross Validation**. Disponível em: <http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_pls_sect015.htm>. Acesso em: 14 jun. 2011.

17. **SBC / DHA - Departamento de Hipertensão Arterial** <http://www.scielo.br/pdf/abem/v43n4/11752.pdf>

18. **24983 - Macro to test multivariate normality**. Disponível em: <http://support.sas.com/kb/24/983.html>. Acesso em: 11 jun. 2011.

19. **The PLS Procedure: Regression Methods**. Disponível em: <http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_pls_sect014.htm>. Acesso em: 11 jun. 2011.

20. **The PLS Procedure: Examining Model Details**. Disponível em: <http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_pls_sect022.htm>. Acesso em: 13 jun. 2011.

21. **The PLS Procedure: Choosing a PLS Model by Test Set Validation**. Disponível em: <http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_pls_sect023.htm>. Acesso em: 13 jul. 2011.

22. **SAS Papers**. Disponível em: <http://support.sas.com/rnd/app/papers/plsex.pdf>. Acesso em: 13 jul. 2011.

23. **The REG Procedure: ODS Graphics**. Disponível em: <http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_reg_sect052.htm>. Acesso em: 18 jun. 2011.

24. ALDRIN, M. **Função RRR S-PLUS**. Disponível em: <http://lib.stat.cmu.edu/S/rrr.s>. Acesso em: 18 jun. 2011.

25. SCHULZ, M.; NÖTHLINGS, U.; HOFFMANN, KURT; BERGMANN, M. M.; BOEING, H. Identification of a Food Pattern Characterized by High-Fiber and Low-Fat Food Choices Associated with Low Prospective Weight Change in the EPIC-Potsdam Cohort. **The Journal of Nutrition**, PMID: 15867301, Mai 2005. v. 135, n. 5, p. 1183-1189.

26. WEIKERT, C. HOFFMANN, K. DIERKES, J. et al. A Homocysteine Metabolism-Related Dietary Pattern and the Risk of Coronary Heart Disease in Two Independent German Study Populations. **The Journal of Nutrition**, 2005. v. 135, n. 8, p. 1981 -1988.

27. MARGOTTO, P. R. **Curva ROC SPSS**. Disponível em: <http://www.paulomargotto.com.br/documentos/Curva_ROC_SPSS.pdf>. Acesso em: 16 jun. 2011.

28. BRAAK, C. J. F. T.; LOOMAN, C. W. N. Biplots in Reduced-Rank Regression. **Biometrical Journal**, 1994. v. 36, n. 8, p. 983-1003.

29. BRILLINGER, D. R. The Canonical Analysis of Stationary Time Series. **Multivariate Analysis**. New York: Academic Press, 1969. 2nd ed., p. 331-350.

30. DAVIES, P. T.; TSO, M. K.-S. Procedures for Reduced-Rank Regression. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, 1982. v. 31, n. 3, p. 244-255.

31. HANSEN, P. R. Generalized Reduced Rank Regression. **SSRN Electronic Journal**, 2002a.

32. HOFFMANN, KURT; ZYRIAX, B.-C.; BOEING, H.; WINDLER, E. A Dietary Pattern Derived to Explain Biomarker Variation is Strongly Associated with the Risk of Coronary Artery Disease. **The American Journal of Clinical Nutrition**, 2004. v. 80, n. 3, p. 633 -640.

33. SCHULZE, M. B; HOFFMANN, K. Methodological Approaches to Study Dietary Patterns in Relation to Risk of Coronary Heart Disease and Stroke. **British Journal of Nutrition**, 2006. v. 95, n. 5, p. 860-869.

34. SCHULZE, M. B; HOFFMANN, K; MANSON, J. E; et al. Dietary Pattern, Inflammation, and Incidence of Type 2 Diabetes in Women. **The American Journal of Clinical Nutrition**, 2005. v. 82, n. 3, p. 675 -684.

35. **S-PLUS StatLib: rrr**. Disponível em: <http://lib.stat.cmu.edu/S/rrr.s>. Acesso em: 9 jul. 2011.

36. Anderson, T. W. Asymptotic distribution of de reduced rank estimator under general conditions. **Annals of Mathematical Statistics**, 1999. v. 27, p. 1141-1154.