Multicolinearidade em Modelos de Regressão

Simone A. Miloca [1] **, Paulo D. Conejo** [2]

[1]Colegiado do Curso de Matemática - Centro de Ciências Exatas e Tecnológicas da
Universidade Estadual do Oeste do Paraná
Caixa Postal 711 - 85819-110 - Cascavel - PR - Brasil
smiloca@unioeste.br, pconejo@unioeste.br

Resumo. *As técnicas da análise multivariada de dados têm sido regularmente aplicada em problemas de diversas áreas. A escolha de uma determinada técnica normalmente é determinada segundo os objetivos da investigação a ser realizada. Uma teoria abordada na análise multivariada de dados é a construção e validação de modelos de regressão linear. Tais modelos surgem em problemas em que o interesse de estudo está em saber qual o comportamento das variáveis em questão e qual relação existente entre elas. Na construção de tais modelos, alguns pressupostos devem ser verificados e um deles é a dependência entre os regressores (variáveis independentes). Se tais dependências forem fortes pode existir multicolinearidade, provocando efeitos nas estimativas dos coeficientes de regressão e na aplicabilidade geral do modelo estimado. Este trabalho traz uma discussão inicial sobre o problema de multicolinearidade, apresentando através de um exemplo uma das formas de se detectar e solucionar (técnica de Análise Fatorial) o problema.*

Palavras Chaves. *multicolinearidade, análise fatorial, dependência linear*

1. Problema de Multicolinearidade em Modelos de Regressão

A escolha de um determinado método multivariado é determinada segundo os objetivos da investigação a ser realizada, [9], [10], [12], [14], [13]. Uma teoria abordada na análise multivariada de dados é a construção e validação de modelos de regressão linear. Tais modelos surgem em problemas em que o interesse de estudo está em saber qual o comportamento das variáveis em questão e qual relação existente entre elas. Deseja-se construir um modelo matemático que melhor represente tal relacionamento.

Quando se tem pares de observações de duas variáveis, é possível avaliar o relacionamento entre elas fazendo-se um gráfico de dispersão e assim, indicar como seria o modelo matemático. Em muitos problemas, os modelos a serem construídos são lineares.

A teoria de Regressão Linear é importante principalmente quando se tem duas ou mais variáveis envolvidas no problema (tanto na variável resposta quanto nas covariáveis).

Um modelo de regressão linear, tem por equação

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{B}_{p\times 1} + \epsilon_{n\times 1}$$

onde,

- **Y** é o vetor das n observações (variável dependente).
- **X** é a matriz das variáveis independentes.
- ϵ é a matriz dos erros aleatórios e representa a influência de outros fatores não considerados no modelo, bem como os erros de medição da variável resposta Y. Em geral é suposto que $\epsilon \approx N(0, \sigma^2)$, ou seja, os erros experimentais são independentes e normalmente distribuídos.
- **B** é a matriz dos coeficientes desconhecidos do modelo que devem ser estimados.

O que se espera em tais modelos, é encontrar dependências entre a variável resposta Y_i e os regressores X_j. Em tais problemas, deve-se fazer uma avaliação das suposições exigidas para aplicação do modelo. Segundo HAIR [14] existem diversos fatores que podem influenciar na busca do melhor modelo de regressão. Neste sentido, algumas etapas devem ser seguidas, a primeira delas é a especificação dos objetivos da análise de regressão, que inclui a seleção das variáveis dependentes e independentes. A segunda etapa inclui determinação do tamanho da amostra. A seguir, as suposições inerentes à análise de regressão (normalidade, linearidade, homocesdasticidade e independência dos termos de erro) devem ser testadas para as variáveis individuais e se todas forem atendidas, o modelo deverá ser estimado. Após obtenção dos resultados, faz-se análises diagnósticas no sentido de verificar se o modelo geral atende às suposições de regressão e que nenhuma observação tenha influência excessiva sobre os dados. A próxima etapa é interpretar a variável estatística de regressão e examinar o papel de cada variável independente na previsão da medida dependente. Por fim, os resultados são validados para garantir generalização para a população. Nas referências [9] e [14] pode-se encontrar informações sobre cada uma dessas etapas.

O enfoque deste trabalho está centrado em um aspecto a ser considerado na primeira destas etapas. Trata-se de discutir a seleção de variáveis. Uma variável independente adicional pode melhorar a previsão da variável dependente, essa melhoria está relacionada não somente com a correlação existente com a variável dependente, mas também com a correlação desta variável com as demais variáveis independentes existentes no modelo. Assim, deve-se investigar se existe dependências entre os regressores X_j. Em situações onde essas dependências forem fortes, dizemos que existe **multicolinearidade**. A multicolinearidade refere-se à correlação entre três ou mais variáveis independentes. O que precisa ser feito é procurar variáveis independentes que tenham baixa multicolinearidade com as outras variáveis independentes, mas também apresentem correlações elevadas com a variável dependente. Segundo HAIR (2005), além dos efeitos na explicação, a multicolinearidade pode ter sérios efeitos nas estimativas dos coeficientes de regressão e na aplicabilidade geral do modelo estimado.

Algumas indicações da presença de multicolinearidade são:

1. valores altos do coeficiente de correlação;
2. grandes alterações nas estimativas dos coeficientes de regressão, quando uma variável independente for adicionada ou retirada do modelo, ou quando uma observação for alterada ou eliminada;
3. a rejeição da hipótese $H_o : \beta_1 = \beta_2 = ... = \beta_k = 0$, mas nenhuma rejeição das hipóteses $H_o : \beta_i = 0$, $i = 1, 2, ..., k$, sobre os coeficientes individuais de regresssão;

4. obtenção de estimativas para os coeficientes de regressão com sinais algébricos contrários àqueles que seriam esperados a partir de conhecimentos teóricos disponíveis ou de experiências anteriores sobre o fenômeno estudado e
5. obtenção de intervalos de confiança com elevadas amplitudes para os coeficientes de regressão, associados a variáveis independentes importantes.

A presença de multicolinearidade pode ser detectada de várias maneiras. Duas medidas mais comumente utilizadas são o valor de tolerância ou seu inverso, chamada fatores de inflação da variância (VIF) definido pela equação $F(\widehat{\beta}_j) = \frac{1}{1-R^2{}_j}$. É uma medida do grau em que cada variável independente é explicada pelas demais variáveis independentes. Quanto maior for o fator de inflação da variância, mais severa será a multicolinearidade. Sugerem-se [9] e [14] que se qualquer fator de inflação da variância exceder 10, então a multicolinearidade causará efeitos nos coeficientes de regressão. Outros autores sugerem que os fatores de inflação da variância não devem exceder 4 ou 5, isso dependerá do conhecimento teórico do pesquisador sobre o assunto estudado.

Várias medidas têm sido propostas para resolver o problema de multicolinearidade.

Hair [14] destaca as seguintes:

- excluir uma ou mais variáveis independentes altamente correlacionadas e identificar outras variáveis independentes para ajudar na previsão. Esse procedimento deve ser feito com cautela pois, neste caso, há o descarte de informações, contida nas variáveis removidas;
- usar o modelo com variáveis independentes altamente correlacionadas apenas para previsão, ou seja, não interpretar os coeficientes de regressão;
- usar as correlações simples entre cada variável independente e a dependente para compreender a relação entre variáveis independentes e dependente e
- usar um método mais sofisticado de análise como a regresão Bayesiana (ou um caso especial - regressão ridge) ou a regressão sobre componentes principais para obter um modelo que reflita mais claramente os efeitos simples das variáveis independentes.

Segundo Aranha [13], o problema de multicolinearidade pode ser contornado utilizando-se a Análise de Componentes Principais (ACP), que transforma os X_j em componentes ortogonais (não correlacionados) que são utilizados como variáveis explicativas da regressão. O problema que pode ocorrer com esta prática está na dificuldade de interpretar o significado dos componentes e por consequência, dificuldade de interpretação dos coeficientes de regressão. Comenta ainda que os escores fatoriais, obtidos através de outra técnica denominada Análise Fatorial, podem ser utilizados como variáveis de interesse em modelos de regressão, sendo esta prática realizada com cautela, devido principalmente a interpetação dos chamados fatores.

Outro autor, Alpert *apud* Hair [14], escreve um artigo abordando questões referentes ao modelo de regressão e sugere, como forma de redução do número de variáveis independentes, a técnica de Análise Fatorial.

Neste contexto, a proposta deste trabalho é apresentar a técnica de Análise Fatorial bem como um exemplo ilustrativo, utilizando-se dados de um modelo de regressão cujas variáveis independentes apresentam multicolinearidade.

2. Análise Fatorial

A Análise Fatorial (AF) é uma técnica da análise multivariada que tem por objetivo explicar a correlação entre um conjunto grande de variáveis aleatórias em termos de um conjunto de poucas variáveis aleatórias não observáveis chamadas fatores.

As variáveis aleatórias observadas são agrupadas de acordo com suas correlações. Dentro de um grupo as variáveis aleatórias são altamente correlacionadas entre si e de um grupo para outro as correlações são baixas.

A idéia central é condensar a informação contida em diversas variáveis originais em um conjunto menor (fatores) com pequena perda de informação, com a vantagem de não haver correlação entre os fatores e eles estarem nas direções de maior variabilidade.

O modelo fatorial ortogonal é

$$X_{p\times 1} = \mu_{p\times 1} + L_{p\times m}F_{m\times 1} + \varepsilon_{p\times 1}$$

onde,

- $X = (x_1, x_2, ..., x_p)^t$ é um vetor aleatório,
- μ_i é a média da i-ésima variável,
- ε_i é o i-ésimo erro ou fator específico,
- F_j é o j-ésimo fator comum, e
- l_{ij} é o peso ou carregamento na i-ésima variável x_i do j-ésimo fator F_j

sendo,

- F e ε independentes;
- $E(F) = 0$ e $Cov(F) = I_{m\times m}$ e
- $E(\varepsilon) = 0$, $Cov(\varepsilon) = \psi$, onde ψ é uma matriz diagonal com variância específica ψ_i na diagonal principal.

Postula-se que X é LD de algumas variáveis aleatórias não observáveis $F_1, ..., F_m$ e p fontes de variação $\varepsilon_i, ..., \varepsilon_p$

Proposição 1 *A estrutura para o modelo fatorial ortogonal satisfaz:*

1. $Cov(X) = LL' + \psi$, *ou*

$$Var(X_i) = l_{i1}^2 + ... + l_{im}^2 + \psi_i$$

$$Cov(X_i, X_k) = l_{i1}l_{k1} + ... + l_{im}l_{km}$$

2. $Cov(X, F) = L$, *ou*

$$Cov(X_i, F_j) = l_ij$$

Observações:

- ${h_i}^2 = l_{i1}^2 + ... + l_{im}^2$ é denominada comunalidade específica e é a porção da variância da variável x_i que é distribuída pelos *m* fatores comuns.
- ψ_i é denominada variância específica e é a porção da variância $Var(X_i)$ devida ao fator específico. Logo, pode-se escrever

$$Var(X_i) = {h_i}^2 + \psi_i = {\sigma_i}^2 = \sigma_{ii}.$$

- A exigência $m \leq p$ implica que a estrutura dos dados (fatores comuns) não é mais complicada do que aquela para os dados observados. Se fosse, não haveria o que se ganhar com a AF.
- l_{ij} é a covariância da i-ésima variável X_i com o j-ésimo fator comum F_j. Se a matriz de correlação for usada, então l_{ij} será o coeficiente de correlação entre a i-ésima variável e o j-ésimo fator comum.

2.1. Métodos de Estimação

O modelo fatorial ortogonal procura representar adequadamente os dados com um pequeno número de fatores não observáveis. A matriz de covariância S é um estimador da matriz populacional Σ desconhecida. Se em determinado problema, os elementos fora da diagonal de S são pequenos, ou equivalentemente, os elementos fora da diagonal da matriz de correlação R são essencialmente nulos, então as variáveis não são correlacionadas e portanto a análise fatorial não será útil ao problema. Agora, se Σ desvia significativamente de uma matriz diagonal, então o modelo fatorial pode ser utilizado, e o problema inicial é fazer uma estimação dos carregamentos fatoriais l_{ij} e as variâncias específicas ψ_i. Os métodos mais populares são: o método das componentes principais e o método de máxima verossimilhança. A seguir descreve-se o método das componentes principais.

Seja S a matriz de covariância amostral e $(\lambda_1, e_1), (\lambda_2, e_2), \cdots, (\lambda_p, e_p)$ os pares de autovalores e autovetores de S, sendo $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Seja $m \leq p$ o número de fatores comuns. A matriz de pesos (ou cargas) estimadas, dos fatores l_{ij} é dada por

$$L = CD_{\lambda}^{\frac{1}{2}}$$

onde

$$C_{p\times p} = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{22} & \cdots & e_{2n} \\ e_{31} & e_{32} & \cdots & e_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ e_{p1} & e_{p2} & \cdots & e_{pp} \end{bmatrix}$$

e

$$D_{\lambda}^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_p} \end{bmatrix}$$

A obtenção de tais matrizes deve-se ao teorema da decomposição espectral, que permite escrever a matriz de covariância Σ (que normalmente é estimada e denotada anteriormente por S) na forma

$$\Sigma = P\Lambda P'$$

onde P é uma matriz de autovetores e Λ é uma matriz diagonal de autovalores, ou seja,

$$\Sigma = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \cdots + \lambda_p e_p e_p'$$

que na forma matricial fatorada fica

$$\Sigma = [e_1\sqrt{\lambda_1} \vdots \cdots \vdots e_p\sqrt{\lambda_p}] \begin{bmatrix} \sqrt{\lambda_1}e_1' \\ \vdots \\ \sqrt{\lambda_p}e_p' \end{bmatrix}$$

Escreve-se

$$\Sigma = LL'$$

O que ocorre no modelo que queremos é explicar a estrutura de covariância em termos de poucos fatores, ou seja, tomando $m \leq p$. Desta forma exclui-se a contribuição de $\lambda_{m+1}e_{m+1}e'_{m+1} + \cdots + \lambda_p e_p e'_p$ e toma-se a aproximação

$$\Sigma = [e_1\sqrt{\lambda_1}\vdots\cdots\vdots e_m\sqrt{\lambda_m}] \begin{bmatrix} \sqrt{\lambda_1}e'_1 \\ \vdots \\ \sqrt{\lambda_m}e'_m \end{bmatrix} = L_{p\times m}L'_{m\times p}$$

Observações:

- as variâncias específicas ψ são dadas pelos elementos da diagonal principal da matriz $\Sigma - LL'$, e
- na aplicação deste modelo costuma-se subtrair de cada observação $X_1, ..., x_n$ a média amostral para que as observações fiquem centradas no mesmo ponto. Também é usual padronizar as variáveis

$$Z_j = \begin{bmatrix} \frac{x_{ij}-\overline{x_1}}{\sqrt{s_{11}}} \\ \vdots \\ \frac{x_{pj}-\overline{x_p}}{\sqrt{s_{pp}}} \end{bmatrix}$$

Tal solução pode ser rotacionada para a obtenção de uma estrutura mais simples, na qual a interpretação dos fatores seja bem mais visível e simplificada. Isto pode ser feito por diversos métodos sendo os mais utilizados, o método Varimax ou Varimax Normal, [9].

Quanto a escolha do número de fatores comuns, dentre os critérios utilizados, pode-se citar:

1. o número de fatores *m* é igual ao número de autovalores de R (matriz de correlação), maiores do que 1 (um) (critério de Kaiser), e
2. o número de fatores *m* é escolhido de acordo com o percentual da variância explicada sendo,
   - $\frac{\lambda_j}{tr(S)}$ quando a análise é feita a partir de S, e
   - $\frac{\lambda_j}{p}$ quando a análise é feita a partir da matriz de correlação

3. Exemplo

Para exemplificar a técnica de Análise Fatorial, supõe-se que se deseja construir um modelo de regressão para tentar explicar o relacionamento entre uma variável resposta *y*=preço e um conjunto de variáveis explicativas dadas por índices (variáveis quantitativas) de qualidade industrial do trigo ([4], [3]), obtidos através de testes específicos realizados em laboratórios. Os dados para análise foram extraídos de [7], perfazendo um total de 7 variáveis explicativas (*w, pl, fn, gs, gu, aa* e *est*), cada uma contendo 18 informações, que podem ser vistas como 18 lotes.

Uma idéia inicial é optar por um modelo do tipo

$$y = \beta_0 + \beta_1 w + \beta_2 pl + \beta_3 fn + \beta_4 gs + \beta_5 gu + \beta_6 aa + \beta_7 est.$$

Nosso objetivo aqui será voltar nossa atenção ao fato de haver uma dependência entre as variáveis explicativas.

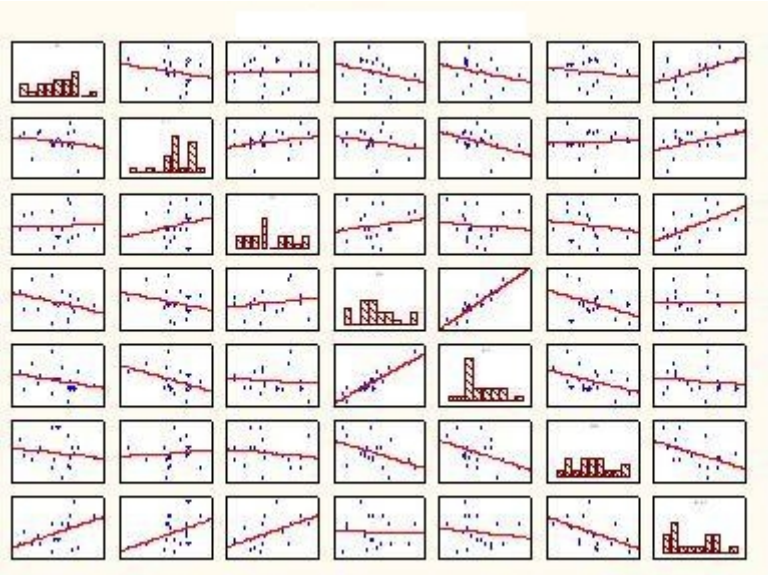
As correlações estimadas a partir dos dados originais é apresentado na tabela 1.

Tabela 1: correlações a partir dos dados originais.

	w	pl	fn	gs	gu	aa	est
w	1,00	-0,20	0,05	-0,32	-0,27	-0,16	0,46
pl	-0,20	1,00	0,24	-0,26	-0,41	0,08	0,41
fn	0,05	0,24	1,00	0,17	-0,09	-0,16	0,55
gs	-0,32	-0,26	0,17	1,00	0,91	-0,45	-0,02
gu	-0,27	-0,41	-0,09	0,91	1,00	-0,43	-0,14
aa	-0,16	0,08	-0,16	-0,45	-0,43	1,00	-0,50
est	0,46	0,41	0,55	-0,02	-0,14	-0,50	1,00

O gráfico de dispersão para as variáveis do problema é dado pela figura 1.

Figura 1: gráfico de dispersão.



Observe pela figura 1 e tabela 1, que algumas variáveis como *gs* e *gu* apresentam alta correlação.

A seguir apresenta-se o *scree plot* (figura 2). É um gráfico que auxilia na decisão do número de fatores. O critério associado ao uso do *scree plot* para determinar a quantidade de fatores é considerar o número de autovalores à esquerda do "ponto de cotovelo", isto é, o ponto onde ocorre uma forte mudança da inclinação da linha que une as representações dos autovalores. Tal critério costuma ser coerente com o critério de Kayser, mas isso não é uma regra. Observa-se que a solução com três fatores para sete variáveis atende ao critério do "cotovelo".

Utilizando o critério de Kayser, temos três autovalores maiores do que 1, explicando em torno de 83% da variância total dos dados (tabela 2).

Figura 2: gráfico scree plot.

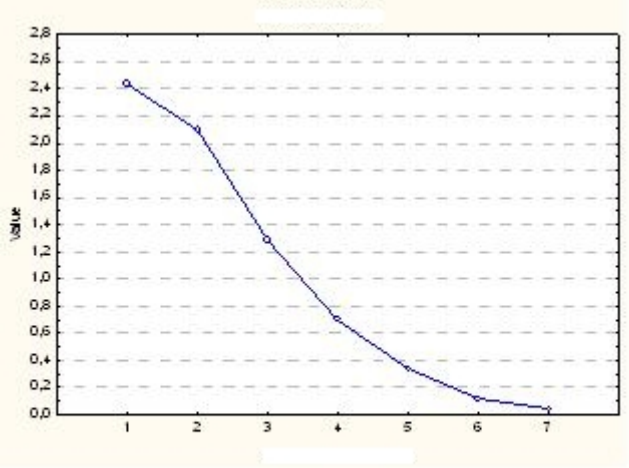


Tabela 2: autovalores da matriz de correlação.

Autovalores	% variância explicada	% variância cumulativa
2,4343	34,7757	34,7757
2,0907	29,8677	64,6434
1,2837	18,3393	82,9827

Na tabela 3 apresenta-se a matriz de carregamentos(pesos) e comunalidades.

Tabela 3: matriz de pesos (ou carregamentos) e comunalidades estimados pelo método das componentes principais não rotacionado.

Variáveis	cargas fatoriais estimadas			Comunalidades			R^2 múltiplo
	Fator 1	Fator 2	Fator 3	Fator 1	Fator 2	Fator 3	
w	0,343	-0,396	**0,797**	0,117	0,274	0,909	0,626
pl	0,496	-0,318	**-0,671**	0,246	0,348	0,799	0,644
fn	0,077	**-0,689**	-0,339	0,006	0,481	0,597	0,620
gs	**-0,924**	-0,208	-0,189	0,854	0,897	0,933	0,911
gu	**-0,965**	-0,041	-0,006	0,931	0,933	0,933	0,915
aa	0,480	**0,664**	-0,211	0,231	0,672	0,717	0,559
est	0,219	**-0,932**	0,034	0,048	0,918	0,919	0,806
Exp.Var	2,434	2,091	1,284				
Prp.Tot	0,347	0,299	0,183				

Observa-se a associação das variáveis *gs* e *gu* como o primeiro fator F1, *est, aa* e *fn*, como segundo fator F2 e *w, p*l como F3. Tenta-se procurar uma solução mais interpretável, rotacionando o eixo que descreve o espaço solução. A idéia é a seguinte: os resultados do modelo fatorial pode ser representado em um espaço ortogonal onde cada eixo é representado por um fator e cada variável é representada por um ponto cujas coordenadas são descritas por uma *n-upla*. No nosso problema a variável *w* será representada

em R^3 pelo ponto $(0,342; -0,396; 0,797)$, e assim por diante. O que se fez foi procurar um outro sistema de eixos ortogonais (atráves de uma rotação do sistema original) que também descrevesse as configurações das variáveis envolvidas, preservando as distâncias relativas entre os pontos, de modo que se possa perceber a associação das variáveis a um dos fatores mais claramente. Isto é feito via rotação varimax (tabela 4).

Tabela 4: dados obtidos pela rotação varimax.

Variáveis	cargas fatoriais estimadas			Comunalidades			R^2 múltiplo	VIF
	Fator 1	Fator 2	Fator 3	Fator 1	Fator 2	Fator 3		
w	-0,234	0,145	**-0,913**	0,055	0,075	0,909	0,626	2,67
pl	-0,417	**0,636**	0,470	0,174	0,578	0,799	0,644	2,81
fn	0,080	**0,765**	0,069	0,006	0,592	0,596	0,620	2,63
gs	**0,945**	0,063	0,189	0,893	0,897	0,933	0,911	11,24
gu	**0,950**	-0,163	0,079	0,900	0,927	0,933	0,915	11,76
aa	**-0,622**	-0,426	0,384	0,387	0,569	0,717	0,559	2,27
est	0,0006	**0,882**	-0,377	0,000	0,777	0,912	0,806	5,15
Exp.Var	2,416	2,001	1,391					
Prp.Tot	0,345	0,285	0,198					

Os fatores encontrados são identificados pelas suas cargas em suas respectivas variáveis, ficando o fator F1 formado pelas variáveis (*gs, gu, aa*), o fator F2, por (*pl, fn* e *est*) e o fator F3 por (*w*). A diferença agora é que a variável *pl* passou a fazer parte do fator F2, e a variável *aa*, do fator F1. Também pode-se verificar tal identificação através do gráfico da matriz de cargas fatoriais.

Observa-se que o R^2 múltiplo, permite obter o VIF, definido na seção 1, indicando presença de multicolinearidade.

O cálculo dos escores fatoriais podem ser feitos com base nas cargas fatoriais estimadas, e são apresentados na tabela 5.

Tabela 5: escores fatoriais

F1	-0,42	1,77	0,18	-1,58	-0,56	1,23	2,19	-0,38	-0,57	-0,29	-1,59	0,61	-0,09	-0,69	0,43	0,21	0,07	-0,52
F2	0,11	-0,46	-0,87	-1,04	-0,92	0,35	0,76	0,59	-0,26	0,38	1,05	-1,84	1,08	1,30	-0,38	1,91	-0,68	-1,07
F3	0,95	0,61	-0,26	0,58	0,48	-0,74	0,15	1,39	-1,09	-1,74	0,16	-1,57	-0,64	-0,68	1,92	0,03	1,02	-0,56

4. Considerações finais

O modelo fatorial ortogonal pode ser utilizado para redução de um conjunto de dados. Os escores fatoriais podem ser utilizados para substituir as variáveis independentes com alta correlação em modelos de regressão, devendo esta substituição ser feita com cautela, devido as interpretações dos coeficientes de regressão.

Referências

[1] BRUNETTA, D.; DOTTO, S. R.; FRANCO, F. de A; BASSOI, M. C. *Cultivares de trigo no Paraná: rendimento, características agronômicas e qualidade industrial*. Londrina: EMBRAPA - CNPSo, 1997. 48p. (EMBRAPA-CNPSo. Circular Técnica, 18).

[2] MORETTIN, P. A.; BUSSAB, W. O. *Estatística básica*. 5. ed. São Paulo: Saraiva, 2004. 526p.

[3] INSTITUTO AGRONÔMICO DO PARANÁ. Londrina, PR. *Informações técnicas para a cultura do trigo no Paraná* - Londrina 2000 152p. ilus. (IAPAR.Circular, 109)

[4] GUARIENTI, E. M. *Efeito de variáveis meteorológicas na qualidade industrial de trigo.* Tese de doutorado. Unicamp - 2001.

[5] KOEHLER, H. S. *Estatística Experimental*. UFPR - Curitiba. 1998. P 36-43.

[6] MILOCA, Simone A., *Aplicação da Teoria de Correlação Canônica e método PROMETHEE num problema de qualidade industrial do trigo*, Curitiba, 2002, UFPR.

[7] MILOCA, Simone A., *Relação entre variáveis meteorológicas e a qualidade industrial do trigo*, Revista Ciência Rural. Santa Maria, v.37, n.1, p.31-37, jan-fev, 2007.

[8] CHAVES NET0, Anselmo. *Probabilidade e estatística aplicada*. Curitiba. UFPR, 1998. (notas de aula)

[9] JOHNSON, R.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. New Jersey: Prentice Hall International, Inc. 1988. 642p.

[10] JOHNSON, R.; WICHERN, D. W. *Business Statistics*. John Wiley and Sons, Inc. 1997.

[11] MOOD,A.; GRAYBILL, F. *Introduction to the theory of statistics.* Mc-Graw-Hill, Inc. 1986.

[12] MARDIA, Kantilal Varichand. *Multivariate Analysis*. (Probability and mathematical statistics). New York. 1980.

[13] Aranha, Francisco; Zambaldi, Felipe. *Análise Fatorial em Administração*. São Paulo: CENGAGE Learning, 2008.

[14] Hair, Jr., J. H.; Anderson, R. E.; Tatham, R. L.; Black, W. C. trad. Adonai Schlup Sant'Ana e Anselmo Chaves Neto. *Análise Multivariada de Dados*. 5 ed. Porto Alegre: Bookman. 2005.