Synergismus scyentifica UTFPR, Pato Branco, 04 (2) . 2009
13 º Encontro Regional de Matemática Aplicada e Computacional - XIII ERMAC

ANÁLISE FATORIAL E A MULTICOLINEARIDADE EM MODELOS DE REGRESSÃO

Simone Aparecida Miloca[1], Paulo Domingos Conejo[2]

1-Professora do curso de Licenciatura em Matemática da UNIOESTE; 2-Professor do curso de Licenciatura em Matemática da UNIOESTE

Resumo - A proposta deste trabalho é apresentar um exemplo de como identificar e tratar o problema de multicolinearidade na construção de um modelo de regressão linear múltipla . Em tais modelos, alguns pressupostos devem ser verificados sendo um deles a não existência de dependências entre os regressores, denominado problema de multicolinearidade, abordado aqui através da técnica multivariada de Análise Fatorial, segundo o método das componentes principais via rotação ortogonal do tipo Varimax. O modelo de regressão foi construído com base nos escores fatoriais, atendendo as suposições teóricas para sua existência e utilização.

Palavras-Chave: regressão, multicolinearidade, análise fatorial.

FACTORIAL ANALYSIS AND THE MULTICOLLINEARITY IN LINEAR REGRESSION MODELS

Abstract- The proposition of the present work is to present an example of how to identify and deal with the problem of multicollinearity in the construction of a multiple linear regression model. In models like that, a bunch of presumptions must be verified such as the non-existence of dependency between the regressors, which is called multicollinearity problem, covered in our essay throughout the Factorial Analysis multivariate tecnique and according to the main component method through orthogonal rotation from the Varimax type. The regression model was built up based on the factorial scores, meeting the theoretical assumptions for its existence and usage.

KeyWords: regression, multicollinearity, factorial analysis**.**

1. INTRODUÇÃO

Problemas envolvendo estudos de relacionamento entre variáveis surgem com frequência em diversas área do conhecimento. O embasamento teórico para a construção de tais modelos podem ser encontrados com maiores detalhes em Hair (2005), Johnson (1988) e Mood Graybill (1986). O objetivo deste trabalho é a construção de um modelo de regressão para tentar explicar o relacionamento entre uma variável resposta, que se refere a preço de automóveis e um conjunto de variáveis explicativas assim nominadas: aceleração em segundos de 0-60 milhas por hora (acel), aceleração em segundos para percorrer ¼ de milha (acel14), velocidade máxima em milhas por hora (velmax), distância de frenagem em pés a 70 milhas por hora (disfren), consumo em milhas por galão, (consum), aderência (ader) durante a curva, medida em g (gravidade). A intenção é exemplificar uma das formas de se identificar a multicolinearidade, muitas vezes presentes em tais modelos, e também ilustrar a aplicabilidade e importância de técnicas da Análise Multivariada, como a Análise Fatorial, em problemas desta natureza. O mesmo exemplo foi utilizado por Moreira (2008), segundo outra abordagem.

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. Pode-se detectar a presença de multicolinearidade 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 VIF= $(1/(1-R^2_j))$, onde R^2_j é o coeficiente de determinação múltipla. 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. Em Moreira (2008), sugere que se qualquer fator de inflação da variância exceder 10, então a multicolinearidade será um problema. Outros autores, como Hair (2005), sugerem que os fatores de inflação da variância não devem exceder 4 ou 5, isso dependerá do conhecimento teórico do pesquisador sobre o problema estudado.
Várias medidas têm sido propostas para resolver o problema de multicolinearidade, como:

I. excluir uma ou mais variáveis independentes altamente correlacionadas e identificar outras variáveis independentes para ajudar na previsão, tal procedimento deve ser feito com cautela pois, neste caso, há o descarte de informações, contida nas variáveis removidas;
II. usar o modelo com variáveis independentes altamente correlacionadas apenas para previsão, ou seja, não interpretar os coeficientes de regressão;
III. usar as correlações simples entre cada variável independente e a dependente para compreender a relação entre variáveis independentes e dependente;
IV. 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;
V. Aranha (2008), sugere que os escores fatoriais, obtidos através da técnica denominada Análise Fatorial, podem ser utilizados como variáveis de interesse em modelos de regressão.

O problema de multicolinearidade neste trabalho será tratado segundo o enfoque (v), dentre os citados acima.

A Análise Fatorial é uma técnica estatística cujo objetivo é condensar a informação contida nas variáveis originais em um conjunto menor de variáveis hipotéticas e latentes (não mensuráveis diretamente), denominados fatores comuns, com pequena perda de informação. Existem dois tipos de modelos, o fatorial ortogonal que é quando os fatores não estão correlacionados e o fatorial obliquo, quando há correlação entre os fatores (Aranha e Zambaldi,2008).

2. MATERIAL E MÉTODOS

Os dados foram extraídos de Moreira (2008), a análise parte de uma base de dados de 113 informações a respeito de automóveis de marcas e modelos diversos.

A matriz de correlações das variáveis independentes apresentou valores maiores que 0,30 (HAIR 2005), assim há indícios de que existe inter-relações entre as variáveis, nesta situação, a Análise Fatorial pode ser utilizada.

Aplicou-se o modelo fatorial ortogonal aos dados utilizando-se o método das componentes principais via Rotação Ortogonal do tipo Varimax, sendo o número de fatores a extrair obtidos segundo o critério de Kayser e o scree-plot. Após, foi ajustado um modelo de regressão linear múltipla.

3. RESULTADOS E DISCUSSÃO

As correlações estimadas a partir dos dados originais é dada na figura 1 bem como o gráfico de dispersão.

Correlações entre as variáveis

Variável	Preço	Acel	Acel14	Velmax	Distfren	Consum	Ader
Preço	1,00	-0,35	-0,39	0,54	-0,34	-0,55	0,09
Acel	-0,35	1,00	0,99	-0,81	0,53	0,01	-0,70
Acel14	-0,39	0,99	1,00	-0,83	0,54	0,06	-0,70
Velmax	0,54	-0,81	-0,83	1,00	-0,63	-0,12	0,60
Distfren	-0,34	0,53	0,54	-0,63	1,00	0,10	-0,63
Consum	-0,55	0,01	0,06	-0,12	0,10	1,00	0,24
Ader	0,09	-0,70	-0,70	0,60	-0,63	0,24	1,00

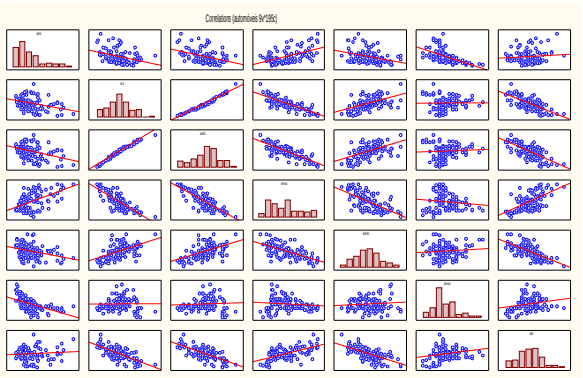

Figura 1. Matriz de correlações e gráfico de dispersão

Após o cálculo da matriz de correlação e identificação de correlações significativas ($p<0,05$) entre as variáveis, partiu-se para o modelo de regressão linear múltipla, de primeira ordem linear nos parâmetros na forma:
$Y = b_0 + b_1 acer + b_2 acer14 + b_3\ velmax + b_4\ distfren + b_5\ consum + b_6\ ader$

Os resultados são apresentados na tabela 1.

Tabela 1. Modelo de Regressão Linear para os dados dos automóveis

Variável dependente		Coeficientes	Teste t	p-valor	ANOVA	R múltiplo	R^2	R^2 ajustado
Y	Constante	99946,4	1,08627	0,2798	F(6,106)=21,42	0,74	0,55	0,52
	Acel	3795,3	0,87446	0,3838				
	Acel14	-5132,3	-0,80782	0,4210	p valor: 0,0000			
	Velmax	519,6	4,37173	0,0000				
	Distfren	-57,9	-0,49522	0,6215				
	Consum	-1895,7	-5,19424	0,0000				
	Ader	-42650,0	-1,44498	0,1514				
Durbin-Watson	1,50368							

Observa-se que existe relação de regressão, a estatística F na ANOVA foi significativa, mas nem todas as variáveis são relevantes conforme mostra o p-valor do teste t para os coeficientes b_j, j=1,...,6. Tem-se, na tabela 3, o VIF, apresentando valores acima de 10 para as variáveis acel e acel14.

Para contornar o problema da multicolinearidade aqui, utilizou-se a técnica da Análise Fatorial (Johnson,1988), apresentada a seguir.

Um modelo com dois fatores mostrou-se adequado para representar a estrutura de covariância inicial explicando 84% da variabilidade dos dados originais. Na tabela 2 apresentam-se os autovalores, as porcentagens das variâncias explicadas e acumuladas, após rotação ortogonal do tipo varimax. A figura 2 mostra o *scree plot* utilizado para determinar a quantidade de fatores. Toma-se como número de fatores comuns o número de autovalores à esquerda do "ponto de cotovelo", ou seja, o ponto onde ocorre uma forte mudança da inclinação da linha que une as representações dos autovalores (Aranha,2008).

Tabela 2. Autovalores da matriz de correlação e variância explicada pelos fatores

Autovalores	Variância (%)	Variância Acumulada (%)
3,812608	63,54347	63,54347
1,115877	18,59796	82,14142

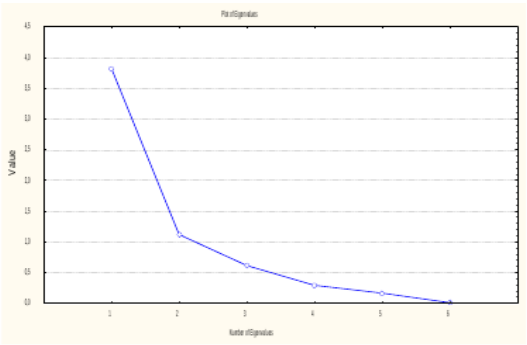

Figura 2. Scree-plot

Com tais resultados, buscou-se identificar quais variáveis mais influenciam em cada fator, ou seja, aquelas que possuem maiores cargas fatoriais. Os resultados estão na tabela 3.

Pode-se observar que o primeiro fator possui pesos mais altos nas variáveis acel, acel14, velmax, disfren e ader, já o segundo fator possui carga fatorial mais alta somente na variável consum.

A soma dos quadrados das cargas fatoriais para cada variável resulta num valor estimado da comunalidade (h^2), que é a parte da variância total explicada pelo fator, avaliando assim o quanto a variância em uma dada variável é explicada pela solução fatorial. Como apresentado na tabela 3, todas as variáveis tem importância moderada na estrutura de covariância, acima de 0,60.

Pelo coeficiente de determinação R^2, pode-se obter o fator de inflação da variância (VIF). Os valores obtidos, conforme apresentados na tabela 3, indicam presença de multicolinearidade severa. (valores maiores do que 10).

Tabela 3. Matriz de pesos (ou carregamentos) e comunalidades estimados pelo método das componentes principais via rotação varimax.

Variáveis	Cargas fatoriais estimadas		Comunalidades ($h^{2)}$	R^2 múltiplo	VIF
	Fator 1	Fator 2			
Acel	**0,937**	-0.007	0,878	0,98	25,25
acel14	**0.944**	0,046	0,894	0,98	25,25
velmax	**-0,895**	-0,152	0,824	0,74	2,21
distfren	**0,749**	0,082	0,568	0,55	1,43
consum	0,027	**0,990**	0,961	0,34	1,13
ader	**-0,823**	0,351	0,800	0,69	1,91
Expl. Var	3,8123	1,1161			
Prop. Totl	0,6354	0,1860			

Os escores fatoriais foram calculados com base nas cargas fatoriais e substituídos pelas variáveis independentes. Os resultados do modelo de regressão apresentam-se na tabela 4.

Tabela 4. Modelo de Regressão Linear

Variável dependente		Coeficientes	Teste t	p-valor	ANOVA	R múltiplo	R^2	R^2 ajustado
y	Constante	32157,46	30,75	0,000000	F(2,110)=54,99	0,71	0,49	0,49
	Fator 1	-6282,16	-5,98	0,000000				
	Fator 2	-9049,04	-8,61	0,000000	p valor: 0,00			

A análise individual de cada fator através do teste t e p-valor mostra que os fatores 1 e 2 são significativos ao nível de 0,05 de significância para a variável y. Para avaliar a significância geral do modelo, realizou-se a ANOVA. O modelo é significativo (rejeita-se a hipótese de não haver regressão) com nível de significância de 0,05 (p-valor = 0,00) e conclui-se que pelo menos uma variável explicativa esta relacionada com a variável y.

Avaliando-se as medidas de ajuste, observa-se que os valores do coeficiente de correlação múltiplo (R múltiplo), o coeficiente de determinação (R^2) e o coeficiente de determinação ajustado (R^2 múltiplo), não são altos, indicando pouca correlação da variável dependente com as independentes, desta forma o modelo não tem alta explicação na

variabilidade do valor de y. O coeficiente de determinação ajustado indica a proporção da variação de y que é explicada através do conjunto de variáveis explicativas selecionadas, ou seja, 49%. Os outros 51% são explicados por outras variáveis não incluídas no modelo.

A equação de regressão linear múltipla para os dados que descreve o relacionamento entre y e os dois fatores é dado por

y = 32157,46 – 6282,16 *escore do fator 1 – 9049,04*escore do fator 2

Ou ainda, a equação estimada para obtenção do preço y é dada por

y = 32909,20 – 5277,59 *(0,937acel + 0,944acel14 - 0,895velmax + 0,749distfren - 0,823ader) – 9938,28*(0,990consum)

Análise dos resíduos

As suposições do modelo com a equação de regressão ajustada podem ser verificadas através de uma análise gráfica, utilizando-se os resíduos, a fim de procurar evidências sobre violações das suposições de homocedasticidade e normalidade dos resíduos. A figura 3 mostra o conjunto de pontos distribuídos aleatoriamente em torno de uma reta horizontal que passa pela origem, sem qualquer padrão. Esta disposição dos pontos indica que a suposição de variância constante é atendida.

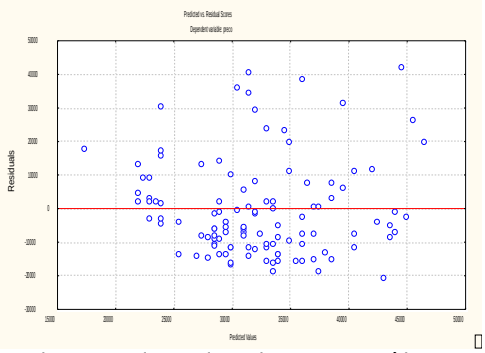

Figura 3. Valores ajustados versus resíduos.

A figura 4 mostra um conjunto de pontos distribuídos de modo aleatório ao redor de uma linha reta, o que indica que a suposição de normalidade dos erros é atendida.

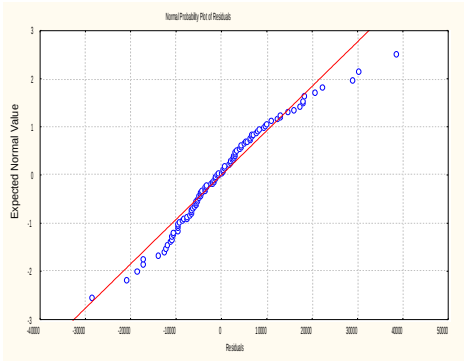

Figura 4. Valor normal esperado versus resíduos

4. CONCLUSÕES

Neste trabalho utilizou-se de técnicas de Análise Multivariada para construir um modelo de regressão linear múltipla que pudesse explicar uma variável em função de outras. A transformação e redução do número de variáveis foram obtidas via Análise Fatorial, sem perda significativa de informações, e obtendo-se variáveis não correlacionadas, por meio de dois fatores. Com os escores obtidos da Análise Fatorial construiu-se um modelo de regressão linear múltipla para a variável y. O modelo ajustado não se mostrou eficiente para previsão, haja vista que o coeficiente de determinação ajustado indicou que somente 49% da variação de y foi explicada através do conjunto de variáveis explicativas selecionadas, isto indica que outros fatores não mencionados influenciam no preço dos automóveis. Destaca-se ainda que outros modelos, como os não lineares, podem ser investigados.

REFERÊNCIAS

ARANHA, Francisco; Zambaldi, Felipe. **Análise Fatorial em Administração**. São Paulo: CENGAGE Learning, 2008.

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.

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

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

MOREIRA, L.F. **Multicolineariadade em Análise de Regressão**. XII ERMAC, 2008.