Robustez na análise de dados financeiros: análise fatorial associada à regressão em painel

Robust analysis of financial data: factor analysis associated with panel regression

Flávia Vital Januzzi[1]
Mariana de Freitas Coelho[2]
Carlos Alberto Gonçalves[3]
Leandro Martins Vieira[4]

Resumo

Uma das limitações encontradas nos estudos acadêmicos em Finanças baseados em dados secundários é a falta de dados completos para análise. Este artigo tem como objetivo discutir o uso de duas técnicas que, em conjunto, podem auxiliar na mitigação do problema de falta de dados para os pesquisadores. A análise fatorial tem como premissa a redução de fatores e pode contribuir ao priorizar os fatores de uma pesquisa determinada. A regressão em painel é utilizada quando existem muitas unidades de análise com um número limitado de informações, e a estimação deve ser feita para dois ou mais períodos de tempo. Dentre os modelos de regressão em painel, o pesquisador deve escolher entre: (1) modelo empilhado; (2) modelo de efeitos fixos; (3) modelo de efeitos aleatórios, ou (4) modelo de efeitos mistos. Portanto, há técnicas alternativas que viabilizam estudos de análise financeira, dando robustez às pesquisas que possuem dados heterogêneos e dados incompletos,

[1] Mestrado em Administração, Universidade Federal de Minas Gerais. – UFMG. Professora assistente da Faculdade de Administração e Contabilidade da Universidade Federal de Juiz de Fora. E-mail: flavia_januzzi@yahoo.com.br

[2] Mestrado em Administração, Universidade Federal de Minas Gerais, UFMG. Doutorado em andamento em Administração, Universidade Federal de Minas Gerais, UFMG. E-mail: marifcoelho@gmail.com

[3] Doutor em Administração pela Universidade de São Paulo. Professor associado na Universidade Federal de Minas Gerais e na Universidade FUMEC, nas Faculdades de Ciências Econômicas e Ciências Empresariais.

[4] Mestrado em Administração Universidade Federal de Minas Gerais, UFMG. E-mail: leandrovieira@globo.com

desde que os trabalhos sejam embasados no rigor metodológico de cada uma das técnicas.

Palavras-chave: Análise Fatorial. Regressão em Painel. Análise Financeira.

Abstract

A limitation found in finance studies based on secondary data is the lack of full data for analysis. This article aims discussing the use of two combined techniques to help mitigating the problem of lack of data for researchers. Factor analysis aims the reduction of factors and contributes to prioritize them ina research. The regression panel is used when there are many analytical units with a limited amount of information and the estimation must be done for two or more time periods. Among the panel regression models, the researcher must choose between: (1) pooled model; (2) fixed effects model; (3) random effects model, or (4) mixed effects model. Therefore,this paper presents alternative techniques, which enable stronger financialstudies analysis, giving robustness to the researches with heterogeneous data and incomplete data, when grounded in the methodological accuracy of each technique.

Keywords: *Factor Analysis. Panel Regression. Financial Analysis.*

1 Introdução

Uma das fontes de dados em pesquisas em Finanças é proveniente de dados secundários, ou seja, os dados preexistentes não são coletados pelo pesquisador. Dentre as possíveis fontes de dados utilizados estão a BMF&FBovespa, IPEADATA, Banco Central do Brasil, Dow Jones Sustentability Index, entre vários outros. Isso pode proporcionar um *gap* em relação aos dados necessários para a solução de um problema de pesquisa. Outro possível problema a ser enfrentado é unir dados de fontes diferentes para utilização em pesquisa acadêmica.

Assim, um dos desafios enfrentados pelos pesquisadores é o problema de falta de dados, seja por inexistência (a empresa, por exemplo, ainda não tinha aberto capital em determinada data), seja por impossibilidade de comparabilidade da informação (em virtude das inúmeras mudanças da legislação contábil das sociedades anônimas sofridas ao longo do tempo), seja pela periodicidade da divulgação das

informações (muitos dados financeiros possuem apenas periodicidade anual), entre outros aspectos.

Dado o contexto, diversos estudos – seja em Finanças Corporativas, Mercado de Capitais ou até mesmo na mais recente, Finanças Comportamentais – passam a ter sua robustez e capacidade de inferência comprometida em virtude das limitações estatísticas decorrentes da modelagem aplicada a pequenas amostras.

Para amenizar tal efeito e garantir maior robustez estatística nas análises, a medida de regressão em painel pode ser utilizada, pois considera tanto as informações ao longo do tempo (*time series*) quanto seccionais (*cross section*). Heij et al. (2004) destacam que, ao fundir dados seccionais com séries temporais, a regressão em painel permite ao observador ter acesso a mais informações, maior grau de liberdade (derivado do maior número de observações) e, principalmente, uma maior eficiência, isto é, maior probabilidade de se obter um estimador não enviesado com variâncias menores, para todos os possíveis valores do parâmetro. Autores como Costa, Bandeira-de-Mello e Marcon (2013); Fortunato, Funchal e Mota (2012); Giacomoni e Sheng (2013); e Suliani et al. (2012) utilizaram a regressão em painel em seus estudos.

Stock e Watson (2004) destacam três principais vantagens no uso da regressão em painel, a saber: i) controle dos efeitos fixos não observáveis (específicos a uma empresa, país, indivíduos, entre outros) que podem enviesar as estimativas dos coeficientes; ii) uma amostra maior possibilita obter estimativas mais precisas dos coeficientes; iii) com uma amostra maior é possível trabalhar com variáveis defasadas sem comprometer a robustez do modelo.Além disso, a regressão em painel também pode contribuir para pesquisas longitudinais, uma vez que permite a análise temporal com maior consistência.

Não obstante, outro problema que também surge na área financeira decorre do fato do pesquisador possuir uma extensa gama de possíveis variáveis explicativas que deverão ter sua significância testada uma a uma. Isto é, por mais que se tenha um número de observações escasso para uma determinada unidade de análise (seja ela empresa,

ou um país, por exemplo) que representa sua variável dependente, como mencionado anteriormente, é possível encontrar dificuldades em propor um modelo parcimonioso, baseado em teoria. Com isso, é possível reduzir o número de variáveis, sem incorrer em perda de informação?

Fávero et al. (2009) descrevem a análise fatorial, ou análise de fator comum, como uma técnica multivariada de interdependência que tem como propósito principal sintetizar as relações observadas entre um conjunto de variáveis inter-relacionadas, buscando identificar fatores comuns. Seu objetivo principal é representar um conjunto de variáveis originais observadas por meio de um número menor de fatores intrínsecos.

Em função do exposto, esse trabalho tem como objetivo proporcionar ao pesquisador as inúmeras áreas temáticas da teoria financeira a refletir sobre o uso conjunto dessas duas técnicas metodológicas como subsídio para suas pesquisas no contexto em que a realidade de poucas observações temporais (seja das variáveis dependente ou independentes), se faz presente conjuntamente com um cenário de um número demasiado de potenciais variáveis explicativas a serem testadas.

Além dessa introdução, o respectivo ensaio é composto de um referencial teórico, nos quais ambas as técnicas são discutidas individualmente, seguido de uma conclusão, na qual o uso conjunto dessas ferramentas é pormenorizado.

2 Referencial teórico

Nesse tópico as duas principais ferramentas de análise de dados foram discutidas em maior profundidade, a saber: a análise fatorial e a regressão em painel.

2.1 A metodologia de análise fatorial

Como uma técnica presente dentro da análise multivariada, Hair et al. (1998) definem que a análise fatorial está relacionada ao estudo da

estrutura dos dados quando esta possui um grande número de variáveis, auxiliando na definição de fatores que representem dimensões relativas comuns. Em outras palavras, consiste em um método estatístico que utiliza uma matriz de correlação para identificar maiores níveis de correlação e agregar variáveis em um construto chamado fator. Dessa forma, o fator passa a representar um aglomerado de variáveis que possuem correlação entre si.

Por conseguinte, as duas principais funções da análise fatorial acabam por ser definidas pela sumarização e redução dos dados. A sumarização dos dados ocorre quando um número menor de conceitos representa o número de variáveis que é maior. A redução dos dados ocorre quando se calcula ponderações sobre as variáveis originais e as agrupa em seus fatores (HAIR et al., 1998).

Fávero et al.(2009) discorrem que o grande trunfo da análise fatorial é permitir a simplificação ou a redução de um grande número de dados, por intermédio da determinação das dimensões latentes, também denominadas fatores. E como resultado, possibilitar ao pesquisador a criação de indicadores inicialmente não observáveis compostos de agrupamento de variáveis.

Já Johnson e Wichern (2007) descrevem que, na análise fatorial (AF), o principal critério empregado para o agrupamento das variáveis são as correlações. Dessa forma, as variáveis que compõem determinado fator devem ser altamente correlacionadas entre si e fracamente correlacionadas com as variáveis que sintetizam outro fator qualquer. Logo, a AF pode ser entendida como uma técnica de agrupamento de variáveis com base na correlação, que supõem normalidade e linearidade entre os dados (desvios na normalidade e na linearidade podem reduzir as correlações observadas entre as variáveis e prejudicar a solução).

Para se iniciar o processo de análise fatorial, deve-se primeiramente conhecer a amostra de dados para se verificar basicamente se: i) os dados deverão ser agrupados em termos de respondentes (Q-type) ou por variáveis (R-type); ii) a forma como a amostra se apresenta; e iii) o tamanho necessário da amostra.

Com relação ao primeiro questionamento, existem dois tipos de abordagens para se calcular a matriz de correlação. A primeira, baseada na organização por respondentes é amplamente utilizada na área de pesquisa em *marketing,* pois é capaz de identificar indivíduos similares, contanto que eles possuam uma variável de identificação individual. Já a abordagem por variáveis possibilita identificar as correlações entre as variáveis e agrupá-las em fatores, ao contrário da abordagem por respondentes, que permite apenas agregar em perfis semelhantes de respondentes.

Sob o segundo aspecto, Hair et al. (1998) citam que geralmente assume-se que as variáveis adotadas para o uso de análise fatorial são do tipo métricas, isto é, numerais capazes de medir determinado aspecto. Em alguns casos variáveis, *dummy*, embora consideradas não métricas, podem ser utilizadas.

Finalmente, a última observação, que diz respeito ao tamanho da amostra, tanto em termos absolutos quanto em termos relativos, é sempre uma função do número de variáveis em análise. Geralmente, não se usa análise fatorial para menos de 50 observações. Hair et al. (1998) continuam ponderando que, como regra geral, deve-se ter no mínimo cinco vezes mais observações do que o número de variáveis analisadas, sendo que 10 observações por número de variáveis é altamente recomendável.

Fávero et al. (2009) sintetizaram o modelo AF por meio da seguinte equação:

$$X_i = a_{i1} F_1 + a_{i2} F_2 + \cdots + a_{im} F_m + \varepsilon_i \quad (i=1,\ldots,p) \qquad (1)$$

Em que X_i representa as variáveis padronizadas, a_i as cargas fatoriais, F_m os fatores comuns e ε_i os fatores específicos.

No entanto, conforme Fávero et al. (2009), as seguintes premissas são assumidas durante a estruturação da equação 1:

a) os fatores comuns (F_m) são independentes e igualmente distribuídos com média 0 e variância 1;

b) os fatores específicos (ε_i) são independentes e igualmente distribuídos, com média 0 e variância ωi;

c) F_m e ε_i são independentes.

Para estimar a equação 1, é necessário que sejam seguidas as seguintes etapas:

a) análise da matriz de correlação e adequação da utilização da AF;
b) extração dos fatores iniciais e determinação do número de fatores;
c) rotação dos fatores;
d) interpretação dos dados.

Na primeira etapa, busca-se examinar a matriz de correlação e averiguar se existem valores significativos para justificar a aplicação da AF. Caso a correlação entre todas as variáveis sejam baixas, é mais provável que a análise fatorial não seja o método mais adequado (FÁVERO et al., 2009).

Durante a etapa de definição do método de extração de fatores, Johnson e Wichern (2007) destacam que dois métodos principais poderão ser utilizados: a análise de componentes principais (ACP) e a análise de fatores comuns (AFC). Enquanto na ACP considera-se a variância total dos dados, ao contrário da AFC, cujos fatores são estimados com base na variância comum. Dessa forma, a ACP procura uma combinação linear das variáveis observadas, de maneira a maximizar a variância total explicada, ou seja, combinam-se as variáveis para formar um fator que explique a maior quantidade de variância da amostra.

No que tange à verificação da quantidade de fatores que serão extraídos da amostra, o método mais utilizado é o Critério de Raiz Latente, em que cada variável contribui com 1 para o total do autovalor, dessa forma, apenas fatores que possuam raízes latentes ou autovalores maiores que 1 são considerados significantes e os que forem menores são descartados (HAIR Jr. et al., 1998).

Para se interpretar os fatores é necessário compreender a rotação fatorial, o critério para a significância das cargas fatoriais e a interpretação da matriz fatorial. A carga fatorial é a maneira pela qual se interpreta o papel de cada variável para se explicar cada fator, em termos estatísticos, ou seja, é a correlação de cada variável com o fator. As cargas indicam o peso entre a variável e o fator, logo, quanto maior a carga, mais representativa a variável é. A matriz fatorial contém esses pesos e é inicialmente não rotacionada. Logo, a rotação fatorial se refere ao processo de se girar os eixos fatoriais de modo a se encontrar uma solução melhor para o fator. Quando se estima a matriz fatorial inicial, segundo Hair Jr. et al. (1998), o pesquisador simplesmente busca a melhor combinação linear entre as variáveis para se explicar a variância.

Salienta-se que os fatores são estimados sempre do mais explicativo para o menos explicativo, daí há a necessidade de muitas vezes se realizar uma rotação ortogonal fatorial. O objetivo da última é garantir que a variância seja redistribuída dos anteriores para os posteriores de modo a alcançar um modelo teoricamente mais significativo.

Reis (2001) destaca que os fatores produzidos na fase de extração nem sempre são interpretados com facilidade. A aplicação de um método de rotação objetiva viabilizar a transformação dos coeficientes dos componentes principais retidos em uma estrutura simplificada.

O modo mais simples de rotacionar os fatores é o ortogonal (REIS, 2001). O termo “ortogonal” é um conceito matemático que denota independência, ou não correlação (correlação igual à zero), entre os eixos fatoriais que para tal, devem ser mantidos a 90 graus. Conforme exposto, o objetivo de todos os métodos de rotação é simplificar as linhas e colunas da matriz fatorial para facilitar a interpretação, isto é, tornar o número de fatores o mais próximo possível de zero fazendo com que as cargas fatoriais aumentem e agreguem maior parte do conteúdo das variáveis.

O método ortogonal VARIMAX maximiza a soma das variâncias necessárias para os pesos da matriz fatorial. O método é mais facilmente interpretado, pois trata de um intervalo de -1 a +1 para se interpretar as

correlações na matriz fatorial. Correlações próximas a 0 possuem pouca associação entre a variável e o fator, ao passo que valores próximos a -1 (ou +1) possuem relação forte negativa (positiva) entre a variável e o fator. Segundo Hair et al. (1998), o método VARIMAX se mostra mais eficiente que outros métodos de rotação ortogonal como o EQUIMAX e o QUARTIMAX.

Para se determinar a significância das cargas fatoriais, Hair et al. (1998) sugerem que se utilize, em primeira análise: cargas fatoriais maiores que ± 0,30 podem ser consideradas com o nível mínimo; ± 0,40 já são considerados mais importantes, e ± 0,50 são boas candidatas para representar a matriz fatorial. Visto que a carga fatorial é a correlação da variável com o fator, o quadrado da carga é a quantidade da variância total da variável explicada pelo fator. Uma carga fatorial de 0,30 reflete aproximadamente 9% da variância ($0{,}30^2 = 0{,}09$). Uma carga fatorial pode até mesmo exceder 0,80, e explicar 50% da variância, todavia este resultado é pouco provável de acontecer na prática.

Essa abordagem é prática, não estatística e só aplicável quando o tamanho da amostra é maior que 100 unidades observacionais (HAIR et al., 1998). Notar-se-á que essa interpelação aproxima-se da supracitada no momento em que o tamanho da amostra gradativamente aumenta. Considerando-se um poder do teste da inferência estatística de 80%, ao nível de significância de 5% ($\alpha = 0{,}05$) a tabela a seguir contém os tamanhos amostrais necessários para que cada carga fatorial possa ser considerada significante.

Tabela 1: Diretriz para a identificação de pesos fatoriais significantes

Carga fatorial	Tamanho amostral necessário para significância*
0,3	350
0,35	250
0,4	200
0,45	150
0,5	120
0,55	100
0,6	85
0,65	70
0,7	60
0,75	50

* significância baseada em 0,05 de nível de significância (α), nível de poder inferencial de 80% e erros padrão assumidos como sendo duas vezes os coeficientes de correlação tradicionais

Fonte: Hair et al. (1998, p. 112)

Ao se estabelecer a carga fatorial, deve-se identificar quais delas são significantes para serem aplicadas ao estudo. O próprio *software* de análise estatística se encarrega de agrupá-las, utilizando-se de procedimentos matemáticos para tal, entretanto cabe uma percepção humana do significado dos fatores.

Uma discussão importante se relaciona à quais variáveis devem ser consideradas para o diagnóstico. É comum que a análise fatorial seja rodada por subconstructo, ou seja, no caso de um modelo com os construtos que já possuem subdimensões propostas pela literatura, alguns pesquisadores optam por efetuar o exame das cargas fatoriais e agrupamentos por subdimensão, conforme literatura anterior. No entanto, se o objetivo da AF é reduzir e agrupar variáveis, sugere-se que o pesquisador tenha em mente, construtos formados por duas ou mais variáveis, deve-se ponderar a seleção de todas as variáveis para a avaliação de cargas e dimensões.

Realizada a análise fatorial e definidos os fatores, o próximo passo a ser adotado pelo pesquisador consiste na estimação do modelo econométrico de dados em painel, que foi detalhado na subseção 2.2.

2.2 O modelo econométrico de dados em painel

Os dados são formatados em painel quando existem “n” entidades (que podem ser representadas por empresas, fundos de investimento, países, por exemplo) para dois ou mais períodos de tempo (STOCK; WATSON, 2004). Supondo duas variáveis financeiras X e Y, suas observações em função da entidade “i” e do tempo “t” poderiam ser assim representadas:

$$(X\_it, Y\_it), i=1,\cdots, net=1,\ldots,T \quad (2)$$

Greene (1997) argumenta que a principal vantagem de dados em painel sobre dados de seção cruzada é que o painel proporciona ao pesquisador maior flexibilidade no que tange à modelagem de diferenças de comportamento entre as entidades. Dessa forma, a estrutura básica de dados em painel é um modelo de regressão representado por:

$$y_{it} = x'_{it}\,\beta + z'_i\,\alpha + \varepsilon_{it} \quad (3)$$

Em que β representa os K regressores em sem incluir o termo constante, mede o efeito individual ou a heterogeneidade, contém um termo constante e um conjunto de indivíduos ou grupo específico de variáveis que são constantes ao longo do tempo e é o termo de erro estocástico.

Os dados são ditos balanceados quando há uma quantidade de períodos iguais de observação para todos os elementos contidos na amostra. O uso dos dados no formato de painel possibilita ao pesquisador estruturar um modelo por meio da utilização de diversas variáveis preditoras sob duas perspectivas: entre as entidades e ao longo do tempo (FÁVERO et al., 2014).

Os modelos de regressão em painel podem ser classificados conforme os seguintes grupos, a saber:

1. modelo empilhado ou *pooled*;
2. modelo de efeitos fixos;
3. modelo de efeitos aleatórios;
4. modelos de efeitos mistos.

Cada item será descrito com mais detalhes nas subseções seguintes.

2.2.1 O modelo empilhado ou *pooled*

O modelo mais simples, passível de ser obtido por meio de regressão em painel é apresentado por Heij et al., (2004) como modelo *pooled*. Este é expresso por uma regressão simples, que supõe constância no comportamento do coeficiente angular (α_i) e do coeficiente linear (β_{it}) para todas as entidades analisadas. Consequentemente, o modelo pode ser expresso como:

$$y_{it} = \alpha + x'_t \ \beta + \varepsilon_t \quad (4)$$

Nakamura et al.(2007) destacam que em situações nas quais os parâmetros não apresentam nenhuma variação durante o período, é possível reunir os dados (*pooling*) e aplicar o Mínimo Quadrados Ordinários (*Ordinary Least Squares* – OLS) à amostra de dados em painel, mantendo-se as hipóteses clássicas do modelo de regressão linear. Tal modelo é conhecido como *pooled OLS*.

Pela característica dos dados de variabilidade entre setores e ao longo do tempo, não se espera que o painel empilhado seja a melhor especificação para a pesquisa, porém isto será confirmado por meio de testes a serem realizados na amostra.

2.2.2 O modelo de efeitos fixos

A fim de considerar a diferença entre as entidades observadas, os termos constantes designados por α poderiam variar entre elas, apesar de se manterem inalterados ao longo do tempo. Nesse modelo, o termo de erro dado por εi é considerado não correlacionado (nem ao longo do

tempo e nem entre as entidades) e homocedástico (HEIJ et al., 2004). Em termos matemáticos o modelo designado sobre essas premissas é denominado efeito fixo, sendo assim representado:

$$y_{it} = \alpha_i + x'_{it}\,\beta + \varepsilon_{it} \quad (5)$$

Em que:

y_{it} : vetor da variável dependente expresso para cada entidade "i" ao longo do tempo t;

α_i : termo constante designado para cada variável "i";

$x'_{it}\ \beta$: vetor de variáveis explanatórias expresso para cada entidade "i" ao longo do tempo t que não inclui o termo constante multiplicado pelos seus respectivos betas;

ε_{it} : termo de erro expresso para cada entidade "i" ao longo do tempo t.

O modelo é de fácil estimação e apresenta como vantagem a possibilidade de tratar as diferenças individuais de forma sistemática e testável.

O modelo de efeitos-fixos, conforme Holland (2005), também conhecido por abordagem variável *dummy* de mínimos quadrados (*Least Square Dummy Variable – LSDV*), consiste em uma generalização de um modelo do tipo "constante-intercepto-inclinação", pois insere uma variável *dummy* para capturar os efeitos das variáveis omitidas, que permanecem constantes no tempo. Nesta especificação, os efeitos individuais podem ser livremente correlacionados com os demais regressores.

À medida que é feita a estimação de um modelo de regressão múltipla com variáveis binárias para cada uma das N unidades de análise (também denominadas entidades),passa-se a ter um intercepto da regressão diferente para cada uma destas unidades, a fim de captar a heterogeneidades existentes entre elas. O modelo é de fácil estimação, e apresenta como vantagem a possibilidade de tratar as diferenças individuais de forma sistemática e testável.

2.2.3 O modelo de efeitos aleatórios

No modelo de efeito aleatório, o termo constante (α_i) não representa um parâmetro fixo, mas sim um parâmetro aleatório não observável. Ao contrário do modelo de efeitos fixos que consideram que as diferenças entre os indivíduos são captadas pela parte constante, os modelos aleatórios consideram que tais diferenças são captadas no termo de erro. O modelo de efeitos aleatórios pode ser expresso, conforme Heij et al. (2004):

$$y_{it} = \alpha_i + x'_{it}\,\gamma + w_{it} \qquad w_{it} = \varepsilon_{it} + \eta_i \tag{6}$$

y_{it} : vetor da variável dependente expresso para cada entidade "i" ao longo do tempo t;
α_i : termo randômico: $\alpha_i = \alpha + \eta_i$;
$x'_{it}\,\beta$: vetor de variáveis explanatórias expresso para cada entidade "i" ao longo do tempo t que não inclui o termo constante multiplicado pelos seus respectivos betas;
ε_{it} : termo de erro expresso para cada entidade "i" ao longo do tempo t;
η_i : termo aleatório normalmente distribuído, com média zero e variância constante.

Nesta última especificação, pressupõe-se que o comportamento específico dos indivíduos e períodos de tempo é desconhecido, não podendo ser observado, nem medido. Assim, o modelo representa estes efeitos individuais ou temporais específicos sob a forma de uma variável aleatória normal (GREENE, 1997).

A especificação do modelo de efeitos aleatórios trata os efeitos específicos individuais como se fossem variáveis aleatórias (HOLLAND, 2005). Neste modelo, a premissa básica é de que não existe correlação entre os efeitos individuais e as demais variáveis aleatórias. O principal método de estimação seria o de mínimos quadrados generalizados (GLS).

2.1.4 O modelo de efeitos mistos

O modelo de painel efeitos mistos tem como premissa básica o fato de que algum subconjunto dos parâmetros da regressão variar aleatoriamente entre os indivíduos, sendo, então, parte da heterogeneidade natural entre os indivíduos. Uma característica marcante desta especificação é a de que a resposta média é modelada como uma combinação de características da população, β, que são assumidas como sendo comuns a todos os indivíduos, e efeitos específicos que são particulares a cada indivíduo. Assim, o modelo incorpora na sua especificação tanto os efeitos fixos quanto aleatórios, ao distinguir explicitamente em sua especificação as fontes de variação inter e intra-indivíduos, em uma especificação geral dada por Fitzmaurice (2004):

$$y_{it} = x'_{it}\,\beta + x'_{it}\,b_i + \eta_i + \omega_{it} \qquad (7)$$

Quando o vetor de efeitos aleatórios $\boldsymbol{b}_i$ tem média igual a zero, capta a variabilidade entre indivíduos para o vetor β. Assim, no modelo, o vetor β (efeitos fixos) é o mesmo para todos os indivíduos ao passo que o vetor $\boldsymbol{b}_i$, quando combinado com o efeito fixo correspondente, capta os efeitos específicos ao indivíduo. Desse modo, esses efeitos aleatórios, quando combinados com os efeitos fixos, descrevem a heterogeneidade da resposta média para qualquer indivíduo na amostra.

2.1.5 Validação dos modelos de regressão em painel

Para avaliar o melhor modelo a ser empregado, testes de especificação e validação deverão ser realizados. No que tange à especificação, o primeiro teste aplicado é o de Breusch-Pagan, que tem como objetivo avaliar se o modelo de efeitos aleatórios se sobrepõe ao modelo mais simples (*pooled*). Trata-se de um teste do multiplicador de Lagrange que segue uma distribuição. Como hipótese nula, preconiza-se que o modelo *pooled* é preferível ao modelo aleatório. Ao rejeitar-se a hipótese nula, o modelo aleatório será mais adequado (WOOLDRIGE, 2010).

É possível, por meio do teste de Chow (teste F), avaliar qual será a melhor especificação, dentre duas alternativas: modelo *pooled* versus modelos de efeito fixo. Sob a hipótese nula tem-se que o *pooled* é preferível ao de efeitos fixos. Posteriormente, a fim de avaliar a preponderância do modelo de efeitos fixos sobre o modelo de efeitos aleatórios, poderá ser empregado o teste de Hausman, que avalia a presença de correlação entre o termo de erro e as variáveis explicativas (variáveis independentes). A hipótese nula preconiza que o modelo de efeitos aleatórios é mais consistente e eficiente, sendo, portanto, aplicável à base de dados averiguada. Na hipótese alternativa, presume-se que os estimadores com efeitos aleatórios são não consistentes, e, por conseguinte, o modelo de efeitos fixos seria mais apropriado (WOOLDRIGE, 2010).

A validação da especificação do modelo obtido deve abranger os seguintes testes, com base no termo de erro: normalidade, homocedasticidade e autocorrelação dos resíduos. A fim de se avaliar o pressuposto de normalidade dos resíduos, pode-se construir um histograma para os erros padrões e compará-lo com a distribuição normal padrão, empregando testes de aderência. Para o pressuposto de homocedasticidade do termo de erro, pode ser testado via aplicação do teste de Wald (específico para painel), que supõe como hipótese nula que o termo é homocedástico. Por fim, quanto à análise das autocorrelações dos resíduos pode-se empregar o teste de Wooldridge (teste baseado no multiplicador de Lagrange), cuja hipótese nula aponta para a presença de autocorrelação. Todos os testes estão disponíveis e podem ser operacionalizados com base no *software* Stata® versão 13.0 (WOOLDRIGE, 2010).

A regressão em painel pode contribuir para a análise de dados incompletos, desde que os testes recomendados sejam efetuados. Entretanto, cabe ao pesquisador tomar decisões sobre qual modelo será escolhido, além de definir quais variáveis serão utilizadas (com base na análise fatorial) e corroborar ou contradizer a literatura.

Deste modo, mesmo com o auxílio das técnicas apresentadas, dos testes e de parâmetros estatísticos, as escolhas fundamentadas em literatura e estudos anteriores é de extrema importância para a solidez das análises financeiras. Contudo, essas orientações não devem deixar de considerar o contexto de inovação e implicações gerenciais das pesquisas em Finanças, uma vez que o papel de contribuições teóricas e práticas não deve ficar em segundo plano em pesquisas que implementarem as técnicas sugeridas neste artigo.

Considerações finais

Devido à necessidade de encontrar saídas para situações indesejadas nos estudos de Finanças, como a presença de dados heterogêneos e/ou incompletos, é necessário o estudo de novas abordagens e métodos. Propõe-se, então, a utilização conjunta da análise fatorial e da regressão em painel para maior robustez na análise dos dados.

Fávero et al. (2009) salientam que a análise fatorial representa uma relevante técnica presente no contexto da análise multivariada que tem como objetivo principal sintetizar as relações observadas entre um conjunto de variáveis inter-relacionadas, buscando identificar fatores comuns. Dessa forma, a técnica se propõe a sintetizar um conjunto de variáveis originais em um número menor de fatores.Tal ferramenta assume extrema relevância no contexto de estudos em Finanças, visto que muitas vezes é levantado pelo pesquisador um número extenso de variáveis candidatas a variáveis explicativas, que quando inseridas no modelo financeiro podem gerar problemas de multicolinariedade, quando as variáveis independentes possuem relações lineares exatas ou aproximadamente exatas.

Apesar do uso de Análise Fatorial Exploratória já ser amplamente utilizado por pesquisadores na área de Finanças, são necessárias maiores reflexões e rigor ao seguir os passos propostos pelos autores citados neste trabalho. Ainda, defende-se que a avaliação da AFE por

subconstrutos pode não ser apropriada, já que o objetivo da técnica é justamente reduzir e reagrupar variáveis.

Após a realização da análise fatorial e identificação dos fatores mais relacionados com a variável dependente, prossegue-se para a estimação do modelo. Para tal, a metodologia de regressão em painel pode ser aplicada. Todavia, a técnica de Regressão em Painel ainda se mostra menos utilizada em pesquisas na área de Finanças e merece maior atenção dos pesquisadores, sobretudo, aqueles que trabalham com dados secundários e pesquisas longitudinais onde há dados faltantes.

A metodologia de dados em painel consegue evitar alguns problemas existentes nas estimações de *cross section*, apresentando vantagens significativas para a estimação dos parâmetros. Um dos benefícios encontrado diz respeito à relevância da heterogeneidade individual. Assim, por meio desta metodologia, considera-se a existência de características diferenciadoras das empresas, que podem ser ou não constantes ao longo do tempo (DAHER, 2004).

Daher (2004) ainda destaca que, não obstante, as estimações mediante dados em painel provêm um maior número de informações, maior variabilidade dos dados, aumento dos graus de liberdade, diminuição dos efeitos de colinearidade das variáveis independentes e maior eficácia na estimação. Dessa forma, a técnica pode ser aplicada para o contexto em que existem muitas unidades de análise (empresas, fundos de investimento, investidores) com um limitado número de informações.

A utilização de dados em painel para construir e testar modelos comportamentais complexos, possibilitar a identificação e medir efeitos que não são detectáveis em estudos que empregam a metodologia exclusiva de *cross section* ou de série temporal (DAHER, 2004).No entanto, Marques (2000) destaca que o emprego de dados em painel também apresenta algumas deficiências. No caso da utilização de dados em painel não balanceado, por exemplo, aumenta-se o risco de se ter

amostras incompletas ou com problemas significativos na coleta de dados.

Outras duas desvantagens que são destacadas recorrentemente pelo uso de modelos com dados em painel são o enviesamento resultante da heterogeneidade entre os indivíduos e da seleção dos indivíduos que constituem a amostra, que não garantirá a constituição de uma amostra aleatória, comprometendo a estimação dos parâmetros dos modelos (MARQUES, 2000). Desse modo, novos estudos devem ser feitos para atestar a aplicabilidade dos métodos em questão e para burlar os desafios encontrados durante as pesquisas em Finanças.

Referências

COSTA, M.; BANDEIRA-DE-MELLO, R.; MARCON, R.. Influência da conexão política na diversificação dos grupos empresariais brasileiros. **Rev. Adm. Empres.,** São Paulo, v. 53, n. 4, p.376-387, jul.-ago. 2013.

DAHER, C. E. **Testes empíricos de teorias alternativas sobre a determinação da estrutura de capital das empresas brasileiras**. 2004. 107f. Dissertação (Mestrado em Ciências Contábeis) - Universidade de Brasília, Brasília, 2004.

FITZMAURICE G.; LAIRD N.; WARE, J.. **Applied Longitudinal Analysis**. Hoboken, New Jersey: John Wiley & Sons, 2004.

FÁVERO, L. P.; BELFIORE, P.; SILVA, P.; CHAN, B. **Análise de dados**: modelagem multivariada para tomada de decisões. Rio de Janeiro: Campos Elsevier, 2009.

FÁVERO, L. P. L.; ALMEIDA, J. E. F.; TAKAMATSU, R. T. Propensity for growth of stock prices in emerging markets: a logit panel approach**. Business and Economics Journal**, India, v. 5, n. 12, p. 1-9, may, 2014.

FORTUNATO, G.; FUNCHAL, B.; MOTTA, A. P. Impacto dos investimentos no desempenho das empresas brasileiras. **RAM-Rev. Adm. Mackenzie**, São Paulo, v. 13, n. 4, p. 75-98, ago. 2012.

GIACOMONI, B. H.; SHENG, H. H. O impacto da liquidez nos retornos esperados das debêntures brasileiras. **Rev. Adm.**, São Paulo , v. 48, n. 1, p. 80-97, jan./fev./mar. 2013.

GREENE, W. H. **Econometric Analysis**. 5. ed. New York University: Prentice Hall, 1997.

HAIR, J. F. JR., et al. **Multivariate Data Analysis**, 5. ed., New Jersey: Prentice Hall, 1998.

HEIJ, C., et al. **Econometric methods with applications in business and economics.** New York: Oxford, 2004.

HOLLAND, M.; XAVIER, C. L. Dinâmica e competitividade setorial das exportações brasileiras: uma análise de painel para o período recente. **Economia e Sociedade**, Campinas, v. 14, n. 24, p.85-108, jan./jun.2005.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. New Jersey: Prentice Hall, 2007. 773p.

MARQUES, L. D. **Modelos dinâmicos com dados em painel**: revisão de literatura. Faculdade de Economia do Porto, Portugal, 2000. Disponível em: <http://www.fep.up.pt/investigacao/workingpapers/wp100.pdf?origin=publication_detai>. Acesso em: 9 dez. 2014.

NAKAMURA, W. T., et al. Determinantes de estrutura de capital no mercado brasileiro – análise de regressão com painel de dados no período 1999-2003**. Revista de Contabilidade e Finanças da USP**, São Paulo, v.18, n. 44, p. 72-85, maio/ago., 2007

PYNDICK, R. S.; RUBINFELD, D. L. **Econometria:** modelos e previsões. Rio de Janeiro: Elsevier, 2004.

REIS, E.. **Estatística multivariada aplicada**. 2. ed. Lisboa: Edições Silabo, 2001.

ROVER, S., et al. Explicações para a divulgação voluntária ambiental no Brasil utilizando a análise de regressão em painel. **Rev. Adm.**, São Paulo, v. 47, n. 2, jun. 2012.

STOCK, J. H.; WATSON, M. W. **Econometria.** 1. ed. São Paulo: Pearson, 2004.

WOOLDRIGE, J. M. **Econometric analysis of cross section and panel data**. Cambridge: MIT Press, 2010.

Artigo recebido em: 13/03/2015
Aprovado em: 17/06/2015