

SACHIKO ARAKI LIRA

**EFEITOS DO ERRO AMOSTRAL NAS ESTIMATIVAS DOS PARÂMETROS
DO MODELO FATORIAL ORTOGONAL**

**Tese apresentada como requisito parcial à
obtenção do grau de “Doutora em Ciências”
no Programa de Pós-Graduação em Métodos
Numéricos em Engenharia, dos Setores de
Tecnologia e de Ciências Exatas da
Universidade Federal do Paraná.**

Orientador: Prof. Dr. Anselmo Chaves Neto

CURITIBA

2008

Lira, Sachiko Araki

Efeitos do erro amostral nas estimativas dos parâmetros do modelo fatorial ortogonal / Sachiko Araki Lira – Curitiba, 2008.

193 p.

Orientador: Anselmo Chaves Neto.

Tese (Doutorado) - Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Setores de Tecnologia e de Ciências Exatas Universidade Federal do Paraná.

1. Análise fatorial. 2. Modelo fatorial ortogonal. 3. Simulação Monte Carlo. 4. Precisão relativa. 5. Erro total relativo. I. Título.

CDU 519.237.7

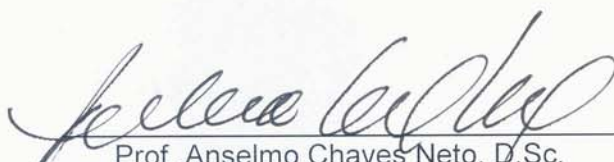
TERMO DE APROVAÇÃO

Sachiko Araki Lira

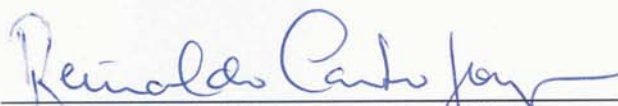
“Efeitos do Erro Amostral nas Estimativas dos Parâmetros do Modelo Fatorial Ortogonal”

Tese aprovada como requisito parcial à obtenção do grau de Doutora em Ciências no Curso de Pós-Graduação em Métodos Numéricos em Engenharia – Área de Concentração em Programação Matemática, Setores de Tecnologia e de Ciências Exatas da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientador:



Prof. Anselmo Chaves Neto, D.Sc.
Departamento de Estatística da UFPR



Prof. Reinaldo Castro Souza, Ph.D.
Departamento de Engenharia Elétrica da PUC/RIO



Prof. Jair Mendes Marques, D.Sc.
PPGMNE/UFPR



Prof. Osmir José Lavoranti, D.Sc.
EMBRAPA - Curitiba/PR



Prof. Inácio Andruski Guimarães, D.Sc.
Departamento de Matemática da UTFPR

Curitiba, 22 de fevereiro de 2008

AGRADECIMENTOS

Ao orientador Prof. Anselmo Chaves Neto, pela confiança depositada em mim, pela motivação, pelo apoio, acompanhamento e orientação na realização deste trabalho.

Aos professores e colegas do Programa de Pós-Graduação em Métodos Numéricos em Engenharia.

À Maristela Bandil, do PPGMNE, sempre prestativa no atendimento da Secretaria da Coordenação do curso.

Ao Instituto Paranaense de Desenvolvimento Econômico e Social (IPARDES), pela liberação, durante o horário das aulas, para o cumprimento dos créditos.

Aos amigos e colegas da Diretoria de Estatística do Ipardes, que compartilharam comigo os bons e maus momentos, durante vários anos em que trabalhamos juntos.

À Eliane M. D. Mandu, pelo apoio nos momentos difíceis pelos quais passamos, no decorrer do desenvolvimento do projeto deste trabalho.

À Estelita S. de Matias e Ana Rita B. Nogueira, que me auxiliaram na revisão e editoração final do texto.

Ao meu esposo Herbert, pela compreensão, pela paciência, pelo carinho e apoio irrestrito, em todos os momentos. Sem o seu apoio, não teria conseguido atingir este objetivo.

Aos meus filhos, Herbert Júnior e Bernard, pelo carinho, pela motivação e principalmente pela paciência e compreensão da importância de minha dedicação a este trabalho.

RESUMO

O presente estudo tem como objetivo avaliar os efeitos do erro amostral nas estimativas dos parâmetros do modelo fatorial ortogonal por componentes principais. A precisão das estimativas foi avaliada pelo coeficiente de variação. As populações normais multivariadas foram geradas pelo Método de Simulação Monte Carlo. Para cada tamanho de amostra dimensionado, para estimar o vetor de médias populacional, adotando-se nível de confiança de 95% e margens de erros relativos fixados em 5%, 10% e 15%, foram retiradas 1.000 amostras aleatórias, com reposição. Outra medida avaliada foi a raiz quadrada do erro quadrático médio relativa (erro total relativo) das estimativas. O estudo considerou todos os fatores (autovalores maiores do que 1, definido pelo Critério de Kaiser). Optou-se por utilizar o maior coeficiente de variação e a maior raiz quadrada do erro quadrático médio relativa das estimativas, pois, para cada modelo fatorial estimado, têm-se diferentes números de componentes (fatores e variáveis). Desta forma, está-se avaliando a menor precisão e o maior erro total relativos das estimativas. Ajustaram-se os modelos de regressão linear múltipla para analisar a relação existente entre coeficiente de variação e raiz quadrada do erro quadrático médio relativa, com as variáveis explicativas: estimativas dos autovalores, autovetores, cargas fatoriais e comunalidades, tamanhos de amostra, números de variáveis e de fatores e estimativa da explicação dos fatores. Todas as variáveis explicativas são determinantes na precisão das estimativas. Em situações cujas estimativas são pequenas, tanto o coeficiente de variação quanto a raiz quadrada do erro quadrático médio relativa são grandes. Constatou-se a existência de viés nas estimativas, sendo consideravelmente maior nos autovetores e cargas fatoriais, principalmente quando o número de variáveis é grande. A medida indicada para avaliar a qualidade das estimativas do modelo fatorial ortogonal é erro total relativo, ou a raiz quadrada do erro quadrático médio relativa.

Palavras-chave: Análise Fatorial; Modelo Fatorial Ortogonal; Simulação Monte Carlo; Precisão relativa; Erro total relativo.

ABSTRACT

The present study aims at assessing sampling error effects on the estimates of Orthogonal Factor Model parameters based on the Principal Components Method. Estimate precision was assessed through the coefficient of variation. We also produced multivariate normal populations through the Monte Carlo Simulation Method. In order to estimate the mean population vector, it was used a 95% confidence level and 5%, 10% and 15% margin of relative error for each sample dimensioned size. The study selected 1.000 samples with replacement randomly. This work also assessed the relative root mean square error (relative total error) of the estimates and took into consideration every factor (eigenvalue higher than 1), as defined by the Kaiser Criterion. We chose to use the highest coefficient of variation and the relative root mean square error (relative total error) of the estimates, since each factor model estimated has a different number of components (factor and variables), thus we assessed the estimate least precision. Multiple Linear Regression models were adjusted so that the study could analyze the relation between the coefficient of variation and the relative root mean square error (relative total error), with the following explanatory variables: eigenvalue estimates, eigenvectors, factor loads and communalities, sample sizes, variable and factor number, and factor explanation estimates. All the explanatory variables are essential for the precision of the estimates. In situations where estimates are low, both the coefficient of variation and the relative root mean square error (relative total error) are relatively high. In the estimates there was evidence of bias, which was considerably higher in the eigenvectors and factor loads, mainly when number of variables is large. Relative total error, or relative root mean square error (relative total error), is the best measurement to assess the estimates of Orthogonal Factor Model parameters.

Key-words: Factor Analysis; Orthogonal Factor Model; Monte Carlo Simulation; Relative Precision; Relative Total Error.

LISTA DE QUADROS

1 - TAMANHOS DE AMOSTRAS PARA DIFERENTES ERROS RELATIVOS E NÍVEL DE CONFIANÇA DE 95%, SEGUNDO VARIÁVEIS	89
2 - AUTOVALORES, AUTOVETORES, CARGAS FATORIAIS E COMUNALIDADES DO MODELO FATORIAL ORTOGONAL	92
3 - ESTIMATIVAS DOS AUTOVALORES, VIÉS, VARIÂNCIA E ERRO QUADRÁTICO MÉDIO, SEGUNDO TAMANHOS DE AMOSTRAS	93
4 - ESTIMATIVAS DOS AUTOVETORES, VIÉS, VARIÂNCIA E ERRO QUADRÁTICO MÉDIO, SEGUNDO TAMANHOS DE AMOSTRAS	94
5 - ESTIMATIVAS DAS CARGAS FATORIAIS, VIÉS, VARIÂNCIA E ERRO QUADRÁTICO MÉDIO, SEGUNDO TAMANHOS DE AMOSTRAS	94
6 - ESTIMATIVAS DAS COMUNALIDADES, VIÉS, VARIÂNCIA E ERRO QUADRÁTICO MÉDIO, SEGUNDO TAMANHOS DE AMOSTRAS	95
7 - COEFICIENTE DE VARIAÇÃO E RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DOS AUTOVALORES, SEGUNDO TAMANHOS DE AMOSTRAS.....	96
8 - COEFICIENTE DE VARIAÇÃO E RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DOS AUTOVETORES, SEGUNDO VARIÁVEIS	97
9 - COEFICIENTE DE VARIAÇÃO E RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DAS CARGAS FATORIAIS, SEGUNDO VARIÁVEIS.....	97
10 - COEFICIENTES DE VARIAÇÃO E RAÍZES QUADRADAS DO ERRO QUADRÁTICO MÉDIO RELATIVAS DAS ESTIMATIVAS DAS COMUNALIDADES, SEGUNDO VARIÁVEIS.....	98
11 - DESCRIÇÃO DAS VARIÁVEIS	99
12 - VALORES MÍNIMO, PERCENTIL 25, MEDIANA, PERCENTIL 75 E MÁXIMO, SEGUNDO VARIÁVEIS.....	100
13 - ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO, ERRO PADRÃO, ESTATÍSTICAS t E F , VALOR- p E COEFICIENTE DE DETERMINAÇÃO DO MODELO AJUSTADO PARA O MAIOR COEFICIENTE DE VARIAÇÃO DOS AUTOVALORES ESTIMADOS.....	102
14 - ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO, ERRO PADRÃO, ESTATÍSTICAS t E F , VALOR- p E COEFICIENTE DE DETERMINAÇÃO DO MODELO AJUSTADO PARA A MAIOR RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DOS AUTOVALORES ESTIMADOS	103
15 - ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO, ERRO PADRÃO, ESTATÍSTICAS t E F , VALOR- p E COEFICIENTE DE DETERMINAÇÃO DO MODELO	

AJUSTADO PARA O MAIOR COEFICIENTE DE VARIAÇÃO DOS AUTOVETORES ESTIMADOS.....	105
16 - ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO, ERRO PADRÃO, ESTATÍSTICAS t E F, VALOR- p E COEFICIENTE DE DETERMINAÇÃO DO MODELO AJUSTADO PARA A MAIOR RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DOS AUTOVETORES ESTIMADOS.....	106
17 - ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO, ERRO PADRÃO, ESTATÍSTICAS t E F, VALOR- p E COEFICIENTE DE DETERMINAÇÃO DO MODELO AJUSTADO PARA O MAIOR COEFICIENTE DE VARIAÇÃO DAS CARGAS FATORIAIS ESTIMADAS.....	108
18 - ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO, ERRO PADRÃO, ESTATÍSTICAS t E F, VALOR- p E COEFICIENTE DE DETERMINAÇÃO DO MODELO AJUSTADO PARA A MAIOR RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS CARGAS FATORIAIS ESTIMADAS.....	109
19 - ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO, ERRO PADRÃO, ESTATÍSTICAS t E F, VALOR- p E COEFICIENTE DE DETERMINAÇÃO DO MODELO AJUSTADO PARA O MAIOR COEFICIENTE DE VARIAÇÃO DAS COMUNALIDADES ESTIMADAS.....	111
20 - ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO, ERRO PADRÃO, ESTATÍSTICAS t E F, VALOR- p E COEFICIENTE DE DETERMINAÇÃO DO MODELO AJUSTADO PARA A MAIOR RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS COMUNALIDADES ESTIMADAS.....	112
21 - MODELOS AJUSTADOS PARA O MAIOR COEFICIENTE DE VARIAÇÃO E A MAIOR RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA E OS COEFICIENTES DE DETERMINAÇÃO, SEGUNDO OS ESTIMADORES DO MODELO FATORIAL ORTOGONAL.....	114

SUMÁRIO

1 INTRODUÇÃO	11
1.1 JUSTIFICATIVA	13
1.2 OBJETIVOS	14
1.2.1 Objetivo Geral.....	14
1.2.2 Objetivos Específicos	14
2 REVISÃO DE LITERATURA	15
2.1 PRECISÃO E ACURÁCIA DAS ESTIMATIVAS	15
2.2 AUTOVALORES E AUTOVETORES DA MATRIZ QUADRADA	18
2.3 TEOREMA DA DECOMPOSIÇÃO ESPECTRAL.....	19
2.4 JACOBIANO DA MATRIZ DE TRANSFORMAÇÃO.....	22
2.4.1 Propriedades do Jacobiano.....	22
2.5 DISTRIBUIÇÃO NORMAL MULTIVARIADA	24
2.5.1 Função Densidade de Probabilidade.....	24
2.5.2 Propriedades da Distribuição Normal Multivariada.....	27
2.5.3 Estimadores de Máxima Verossimilhança da Distribuição Normal Multivariada	31
2.5.4 Estimadores não Viesados da Distribuição Normal Multivariada	34
2.5.5 Distribuição Amostral de \bar{X} e S	36
2.5.6 Avaliação da Suposição de Normalidade (Gaussianidade).....	37
2.5.7 Inferência sobre Vetor de Médias.....	39
2.5.8 Região de Confiança com Largura Fixa	42
2.6 MÉTODO DE MONTE CARLO.....	44
2.7 ANÁLISE DE REGRESSÃO LINEAR MÚLTIPLA	45
2.7.1 Estimação pelo Método dos Mínimos Quadrados.....	47
2.7.2 Inferência sobre os Parâmetros de Regressão	53
2.7.3 Teste para o Relacionamento Modelável por Regressão.....	54
2.8 TESTES PARA AVALIAR AS SUPOSIÇÕES SOBRE A COMPONENTE ERRO	55
2.8.1 Teste de Multicolinearidade.....	55
2.8.2 Teste de Homogeneidade de Variância	55
2.8.3 Teste de Gaussianidade de Kolmogorov-Smirnov com Correção de Lilliefors	56

2.9 IDENTIFICAÇÃO DOS <i>OUTLIERS</i> E PONTOS INFLUENTES	57
2.9.1 Resíduos Studentizados Externamente	57
2.9.2 Pontos de Alavanca ou de Alto <i>Leverage</i>	58
2.9.3 Medida de Influência	58
3 ANÁLISE FATORIAL	60
3.1 INTRODUÇÃO	60
3.2 MODELO FATORIAL ORTOGONAL.....	62
3.3 MÉTODO DA MÁXIMA VEROSSIMILHANÇA	66
3.3.1 Teste para o Número de Fatores Comuns em Grandes Amostras	70
3.4 MÉTODO DAS COMPONENTES PRINCIPAIS	72
3.4.1 Análise Fatorial para População.....	73
3.4.2 Análise Fatorial para Amostra	75
3.5 NÚMERO DE FATORES.....	75
3.5.1 Número de Fatores Definido com Base no Grau de Explicação dos Autovalores Estimados.....	76
3.5.2 Número de Fatores Definido com Base no Critério de Kaiser.....	76
3.6 ROTAÇÃO DOS FATORES	76
3.7 ESCORES FATORIAIS	78
3.7.1 Método dos Mínimos Quadrados Ponderados	78
3.7.2 Método da Regressão	79
3.8 SIGNIFICÂNCIA ESTATÍSTICA DA MATRIZ DE CORRELAÇÃO	81
3.8.1 Teste de Esfericidade de Bartlett	81
3.8.2 Medida de Adequabilidade da Amostra de Kaiser-Meyer-Olkin.....	81
4 MATERIAL E MÉTODO	83
4.1 MÉTODO DE DETERMINAÇÃO DA POPULAÇÃO.....	83
4.1.1 População Normal Multivariada com 5 Variáveis	87
4.2 MÉTODO DE OBTENÇÃO DAS AMOSTRAS	89
4.3 MÉTODO DE AVALIAÇÃO DAS ESTIMATIVAS DOS PARÂMETROS DO MODELO FATORIAL ORTOGONAL.....	90
5 RESULTADOS E DISCUSSÃO	92
5.1 APLICAÇÃO DA ANÁLISE FATORIAL EM DADOS POPULACIONAIS.....	92
5.2 APLICAÇÃO DA ANÁLISE FATORIAL EM DADOS AMOSTRAIS	93

5.2.1	Coeficiente de Variação e Raiz Quadrada do Erro Quadrático Médio Relativa das Estimativas dos Autovalores.....	95
5.2.2	Coeficiente de Variação e Raiz Quadrada do Erro Quadrático Médio Relativa das Estimativas dos Autovetores.....	96
5.2.3	Coeficiente de Variação e Raiz Quadrada do Erro Quadrático Médio Relativa das Estimativas das Cargas Fatoriais	97
5.2.4	Coeficiente de Variação e Raiz Quadrada do Erro Quadrático Médio Relativa, das Estimativas das Comunalidades.....	98
5.3	ANÁLISE DO COEFICIENTE DE VARIAÇÃO E RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DOS AUTOVALORES.....	101
5.3.1	Coeficiente de Variação das Estimativas dos Autovalores.....	101
5.3.2	Raiz Quadrada do Erro Quadrático Médio Relativa das Estimativas dos Autovalores.....	103
5.4	ANÁLISE DO COEFICIENTE DE VARIAÇÃO E RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DOS AUTOVETORES	104
5.4.1	Coeficiente de Variação das Estimativas dos Autovetores	104
5.4.2	Raiz Quadrada do Erro Quadrático Médio Relativa das Estimativas dos Autovetores	106
5.5	ANÁLISE DO COEFICIENTE DE VARIAÇÃO E RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DAS CARGAS FATORIAIS.....	107
5.5.1	Coeficiente de Variação das Cargas Fatoriais Estimadas.....	107
5.5.2	Raiz Quadrada do Erro Quadrático Médio Relativa das Cargas Fatoriais Estimadas.....	109
5.6	ANÁLISE DO COEFICIENTE DE VARIAÇÃO E RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DAS COMUNALIDADES....	110
5.6.1	Coeficiente de Variação das Estimativas das Comunalidades.....	110
5.6.2	Raiz Quadrada do Erro Quadrático Médio Relativa das Estimativas das Comunalidades.....	112
5.7	MODELOS AJUSTADOS PARA O MAIOR COEFICIENTE DE VARIAÇÃO E MAIOR RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DOS AUTOVALORES, AUTOVETORES, CARGAS FATORIAIS E COMUNALIDADES	113

CONCLUSÕES E RECOMENDAÇÕES	117
REFERÊNCIAS	119
BIBLIOGRAFIAS CONSULTADAS	122
APÊNDICE 1 - PARÂMETROS PARA SIMULAÇÃO MONTE CARLO	123
APÊNDICE 2 - TESTE DE ESFERICIDADE DE BARTLETT E ESTATÍSTICA DE ADEQUABILIDADE DA AMOSTRA (MSA)	152
APÊNDICE 3 - <i>SCRIPTS</i> DO SISTEMA R	154
APÊNDICE 4 - MATRIZES DE CORRELAÇÃO DAS POPULAÇÕES 27 E 51 E DAS RESPECTIVAS AMOSTRAS	167
APÊNDICE 5 - MAIORES COEFICIENTES DE VARIAÇÃO E MAIORES RAÍZES QUADRADAS DO ERRO QUADRÁTICO MÉDIO RELATIVAS E AS RESPECTIVAS ESTIMATIVAS DOS AUTOVALORES, AUTOVETORES, CARGAS FATORIAIS E COMUNALIDADES	170
APÊNDICE 6 - MÉDIA E DESVIO PADRÃO DAS VARIÁVEIS DOS MODELOS AJUSTADOS	180
APÊNDICE 7 - AVALIAÇÃO DAS SUPOSIÇÕES DO MODELO DE REGRESSÃO LINEAR MÚLTIPLA E IDENTIFICAÇÃO DE <i>OUTLIERS</i> E PONTOS INFLUENTES	181
APÊNDICE 8 - AUTOVALOR, FORMA QUADRÁTICA E PROPRIEDADE DOS DETERMINANTES	192

1 INTRODUÇÃO

A Análise Fatorial é, atualmente, aplicada nas diversas áreas do conhecimento. É particularmente útil na área das Engenharias, com aplicações muito importantes na Engenharia de Produção (MÜLLER e CHAVES NETO, 2007; ZANELLA et al., 2007), na Engenharia Agrícola e Ambiental (KURTZ et al., 2001; FURTADO et al., 2003; BRITO et al., 2006 e GIRÃO et al., 2007), entre outras.

Isto se deve à sua grande utilidade, que permite descrever a estrutura de covariância dos relacionamentos existentes entre muitas variáveis, por meio de um número menor de fatores. Os fatores são combinações lineares das variáveis originais, podendo ser correlacionados (fatores oblíquos) ou não (fatores ortogonais), de maneira a conservar o máximo das informações originais.

Devido à complexidade não só da Análise Fatorial, mas das demais técnicas da Análise Multivariada, a teoria da Estatística Inferencial tem sido pouco desenvolvida neste campo.

O presente trabalho tem por objetivo avaliar a precisão das estimativas dos parâmetros do modelo fatorial ortogonal. Utilizou-se o Método das Componentes Principais para estimar os carregamentos (pesos) dos fatores e definiu-se o número de fatores pelo Critério de Kaiser.

A estimação da matriz das cargas fatoriais, L , pelo Método das Componentes Principais, necessita das estimativas da matriz diagonal dos autovalores Λ e da matriz ortogonal dos autovetores P . Outro elemento a ser estimado, que é importante na decisão da escolha das variáveis que permanecerão no modelo, é a soma dos quadrados das cargas fatoriais estimadas, $\sum_{i=1}^m \hat{\ell}_i^2$, para cada variável. Este valor é conhecido como comunalidade.

Foram considerados, neste trabalho, todos os fatores (autovalores maiores que 1, pelo Critério de Kaiser) definidos para o modelo, possibilitando, desta forma, conhecer a precisão real das estimativas. O que ocorre, normalmente, é considerar

os fatores mais importantes, ou seja, que explicam a maior proporção da variância total. No entanto, os demais são também componentes do modelo e, portanto, sujeitos aos erros amostrais.

Definiu-se o número de variáveis p entre 5 e 20, e então foram geradas as populações normais multivariadas, pelo Método de Monte Carlo, utilizando-se dos parâmetros previamente definidos. A partir destas populações, retiraram-se 1.000 amostras aleatórias simples, com reposição, de diferentes tamanhos. Adotou-se os tamanhos de amostras para estimar o vetor de médias populacional, com nível de confiança de 95% e margens de erros relativos de 5%, 10% e 15%. Os tamanhos das amostras n variam entre 24 e 984, assim a razão entre o tamanho da amostra e o número de variáveis n/p está compreendida entre 3,7 e 49,5.

O presente estudo traz contribuições importantes para os pesquisadores e profissionais que utilizam a Análise Fatorial no desenvolvimento de suas pesquisas, no tocante à questão do erro amostral nas estimativas dos autovalores, autovetores, cargas fatoriais e comunalidades.

O trabalho está estruturado em 5 capítulos. Além desta introdução, onde constam a justificativa e os objetivos deste estudo, tem-se mais quatro capítulos.

No capítulo 2 apresenta-se uma revisão de literatura, onde se comentam trabalhos sobre as estimativas das cargas fatoriais do modelo fatorial ortogonal pelo Método das Componentes Principais, entre outros. Também é feita uma revisão de conceitos e definições fundamentais ao desenvolvimento da metodologia deste estudo.

No capítulo 3 apresenta-se a metodologia da Análise Fatorial, bem como o teste para avaliar a significância estatística da matriz de correlação e medida de adequabilidade da amostra, para aplicação do método.

O capítulo 4 traz o método de determinação das populações (universos) e obtenção das amostras, além da Análise de Regressão Linear Múltipla, utilizada para o ajuste dos modelos matemáticos.

No capítulo 5 são apresentados os modelos matemáticos que relacionam o coeficiente de variação e raiz quadrada do erro quadrático médio relativa das estimativas dos parâmetros do modelo fatorial ortogonal, com as variáveis explicativas: estimativas dos parâmetros, tamanho da amostra, número de variáveis, número de fatores e estimativa da explicação dos fatores. Finalmente, apresentam-se a conclusão e sugestões para pesquisas futuras.

1.1 JUSTIFICATIVA

O emprego da Análise Fatorial na área das Engenharias está se tornando imperativo na resolução de alguns problemas. É possível citar o caso da identificação de instrumentos inoperantes na barragem da Usina Hidrelétrica de Itaipu (VILLWOCK et al., 2007). Ainda, as metodologias de classificação (ranqueamento) de áreas especialmente protegidas, na Engenharia Ambiental (FURTADO et al., 2003), e de fornecedores, na Engenharia de Produção (MÜLLER e CHAVES NETO, 2007). Então, em análises que envolvem dados amostrais é fundamental conhecer a precisão das estimativas associada ao tamanho da amostra. E, apesar das preocupações com essa questão da Análise Fatorial, ainda são poucos os trabalhos desenvolvidos sobre o tema. Diferentes autores discutem a importância tanto do tamanho da amostra, quanto do número de fatores, no modelo, mas não existe consenso quanto aos números ideais. Assim, estudos sobre esse tema são imprescindíveis.

Segundo FABRIGAR et al. (1999), apesar das atuais facilidades computacionais, para se utilizar a Análise Fatorial é necessário que o pesquisador tome decisões importantes com relação a algumas questões metodológicas. Entre elas, está a definição do tamanho da amostra e do número de fatores a serem incluídos no modelo.

Diante dessas considerações e da necessidade de estudos que possam trazer contribuições para a solução dessas questões, pretende-se, neste trabalho, avaliar os efeitos do tamanho da amostra, do número de variáveis, do número de

fatores e da estimativa da explicação dos fatores na precisão e no erro total dos autovalores, autovetores, cargas fatoriais e comunalidades, estimados pelo Modelo Fatorial Ortogonal, utilizando o Método das Componentes Principais.

A medida de precisão utilizada foi o coeficiente de variação e, do erro total relativo, a raiz quadrada do erro quadrático médio relativa.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Este trabalho tem como objetivo geral avaliar a precisão das estimativas dos parâmetros do modelo fatorial ortogonal, que são: autovalores, autovetores, cargas fatoriais e comunalidades.

1.2.2 Objetivos Específicos

Os objetivos específicos são:

- a) ajustar um modelo para estimar a precisão das estimativas, medida pelo coeficiente de variação, dos parâmetros do Modelo Fatorial Ortogonal, em função das variáveis explicativas: estimativas dos parâmetros, tamanhos de amostra, estimativa da explicação dos fatores, número de variáveis e número de fatores;
- b) ajustar um modelo para estimar o erro total relativo das estimativas, medido pela raiz quadrada do erro quadrático médio relativa dos parâmetros do Modelo Fatorial Ortogonal, em função das variáveis explicativas: estimativas dos parâmetros, tamanhos de amostra, estimativa da explicação dos fatores, número de variáveis e número de fatores;
- c) propor a melhor medida para avaliar a qualidade das estimativas dos autovalores, autovetores, cargas fatoriais e comunalidades.

2 REVISÃO DE LITERATURA

2.1 PRECISÃO E ACURÁCIA DAS ESTIMATIVAS

Os resultados de levantamentos por amostragem estão sempre sujeitos a um certo grau de incerteza, pois apenas uma parte da população é avaliada, devendo-se considerar, também, erros de medida (COCHRAN, 1977). Essa incerteza pode ser reduzida à medida que se aumenta o tamanho da amostra e utilizam-se melhores instrumentos de medida.

A qualidade da estimativa pode ser avaliada através do erro quadrático médio, apresentado na definição 2.1.

Definição 2.1:

Seja o parâmetro θ e o seu estimador $\hat{\theta}$. Então, uma medida do desempenho de $\hat{\theta}$ é dada por:

$$\text{EQM}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 \quad (2.1)$$

O erro quadrático médio pode ser expresso em uma forma que evidencie as duas componentes da variabilidade dos dados, ou seja, a variância do estimador (para a precisão) e o vício do estimador (para a acurácia).

Resultado 2.1:

O erro quadrático médio, apresentado na definição 2.1, pode ser expresso como sendo: $\text{EQM}(\hat{\theta}) = V(\hat{\theta}) + b^2(\hat{\theta})$.

Prova:

Tem-se da definição 2.1 que o erro quadrático médio é dado por:

$$\text{EQM}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

Assim, subtraindo e adicionando $E(\hat{\theta})$, na expressão anterior, tem-se:

$$EQM(\hat{\theta}) = E\left[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta\right]^2 \quad (2.2)$$

$$EQM(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2 + (E(\hat{\theta}) - \theta)^2 \quad (2.3)$$

$$EQM(\hat{\theta}) = V(\hat{\theta}) + b^2(\hat{\theta}) \quad (2.4)$$

em que:

$V(\hat{\theta})$ é a variância da distribuição amostral do estimador $\hat{\theta}$ (precisão);

$b(\hat{\theta})$ é o viés do estimador $\hat{\theta}$ (acurácia).

A raiz quadrada da variância da distribuição amostral do estimador é chamada de erro padrão $EP(\hat{\theta})$ e indica a precisão das estimativas. O erro padrão mede o erro de natureza aleatória, inerente ao processo de amostragem, ou seja, o erro amostral. Quanto menor o erro padrão, maior será a precisão das estimativas obtidas.

Em algumas situações, é útil considerar medidas relativas de variação, ao invés das absolutas, principalmente quando as unidades de medida dificultam as comparações (KISH, 1965). Uma medida relativa comum é o coeficiente de variação $CV(\hat{\theta})$. Assim, a precisão relativa pode ser avaliada por:

$$CV(\hat{\theta}) = \frac{\sqrt{V(\hat{\theta})}}{E(\hat{\theta})} = \frac{EP(\hat{\theta})}{E(\hat{\theta})} \quad (2.5)$$

De acordo com KISH (1965) e SILVA (1998), a raiz quadrada do erro quadrático médio é denominada de erro total (ET). Sua expressão é:

$$ET = \sqrt{V(\hat{\theta}) + b^2(\hat{\theta})} \quad (2.6)$$

O erro total (ET) é uma medida que inclui o viés (vício) e a variância. Portanto, quando o viés não for desprezível, o erro total é melhor do que a variância como medida para avaliar as estimativas.

A determinação da precisão desejada pode ser feita através da quantidade de erro que se dispõe a aceitar nas estimativas amostrais. Esta quantidade é definida de acordo com a utilização que se pretende fazer da estimativa.

Na análise fatorial, poucos são os estudos envolvendo a Inferência Estatística. A aplicação de conceitos e técnicas da Estatística inferencial tem sido pequena (CLIFF e HAMBURGER, 1967). Segundo COSTA (2006), não existem testes adequados para a comprovação da significância estatística, na Análise Fatorial, devido à dificuldade de especificação dos parâmetros teóricos dos modelos de distribuição por amostragem, das estatísticas envolvidas.

Os estatísticos, engenheiros e outros profissionais têm lutado durante décadas com a questão do tamanho da amostra na Análise Fatorial e Análise de Componentes Principais. Alguns se preocupam com o tamanho da amostra (n) e outros com a razão entre o número de observações (tamanho da amostra) e o de variáveis (p) (OSBORNE e COSTELLO, 2004). Ainda, de acordo com os autores, amostras grandes são melhores do que as pequenas, pois as primeiras tendem a minimizar a probabilidade de erros, maximizar a acurácia das estimativas e aumentar as possibilidades de generalização dos resultados.

Segundo HAIR et al. (1998), na Análise Fatorial a amostra não deve ter menos do que 50 observações e preferencialmente deve ser maior do que 100. Como regra geral, o número de observações deve ser, no mínimo, 5 vezes o número de variáveis em análise, e o mais aceitável é a razão de dez para uma, ou seja, o número de observações deve ser 10 vezes o número de variáveis. Alguns propõem um mínimo de 20 observações para cada variável. Nota-se que os critérios variam muito.

De acordo com FABRIGAR et al. (1999), quando cada fator é representado por 3 ou 4 variáveis e as comunalidades são altas, podem ser obtidas boas estimativas mesmo com amostras pequenas, como as de tamanho igual a 100. Entretanto, em condições mais moderadas, pode ser necessário o uso de amostras

de pelo menos 200 observações. Quando se tem muitas variáveis e comunalidades moderadas, é possível que mesmo amostras de tamanho grande, como, por exemplo, entre 400 e 800, não sejam suficientes.

A utilização da Análise Fatorial requer a definição de outros critérios, além da questão do tamanho da amostra e número de variáveis. Um deles é quanto ao número de fatores a serem considerados, que é uma decisão importante. Se adotado o método das Componentes Principais, faz-se necessário estimar os pares de autovalores-autovetores, cujas distribuições para grandes amostras estão apresentadas em JOHNSON e WICHERN (1988). A partir das estimativas dos pares de autovalores-autovetores obtêm-se as cargas fatoriais, que são as correlações de cada variável com o fator.

CLIFF e HAMBURGER (1967) apresentam algumas evidências dos erros amostrais, nas estimativas das cargas fatoriais, sem e com a rotação dos fatores, utilizando o Método de Simulação Monte Carlo. No estudo desenvolvido por um dos autores, observou-se viés grande nas estimativas das cargas fatoriais. Ocorreram tendências de várias subestimações, principalmente para as cargas fatoriais maiores.

Recentemente, COSTA (2006) utilizou os procedimentos *jackknife* e *bootstrap* para estabelecer um critério para significância das cargas fatoriais. Em seu estudo, obteve o viés, a variância e o erro quadrático médio das cargas fatoriais, para os primeiros dois fatores. A partir destes resultados, construiu os intervalos de confiança e testes de hipóteses para as estimativas obtidas.

2.2 AUTOVALORES E AUTOVETORES DA MATRIZ QUADRADA

Sejam $A_{p \times p}$ uma matriz quadrada e $I_{p \times p}$ a matriz identidade. Então os escalares $\lambda_1, \lambda_2, \dots, \lambda_p$ satisfazendo à equação polinomial

$$q(\lambda) = |A - \lambda I| = 0 \quad (2.7)$$

são chamados de autovalores (ou raízes características) da matriz A .

A equação $q(\lambda) = |A - \lambda I| = 0$ (como uma função de λ) é chamada de equação característica. E, para cada autovalor λ , existe um autovetor (vetor característico) correspondente $\underline{x} \neq \underline{0}$ que satisfaz $A\underline{x} = \lambda\underline{x}$. Em geral obtém-se o autovetor padronizado, dada a indeterminação do sistema de equações no cálculo dos componentes do autovetor, ou seja, com comprimento unitário. Assim, se $A\underline{x} = \lambda\underline{x}$, faz-se $\underline{e} = \frac{\underline{x}}{\sqrt{\underline{x}'\underline{x}}}$, tendo o autovetor correspondente de λ .

Tem-se que o coeficiente de λ^p em $q(\lambda)$ é igual a $(-1)^p$, logo é possível escrever $q(\lambda)$ em termos de suas raízes, na forma:

$$q(\lambda) = \prod_{i=1}^p (\lambda_i - \lambda) \quad (2.8)$$

e igualando as expressões (2.7) e (2.8) e com $\lambda = 0$, tem-se:

$$|A| = \prod_{i=1}^p \lambda_i \quad (2.9)$$

De forma que o determinante de A é igual ao produto dos autovalores de A . De maneira semelhante, a soma dos autovalores da matriz A é igual ao traço de A , representado por:

$$\sum_{i=1}^p a_{ii} = \text{tr } A = \sum_{i=1}^p \lambda_i \quad (2.10)$$

2.3 TEOREMA DA DECOMPOSIÇÃO ESPECTRAL

O teorema da decomposição espectral ou da decomposição de Jordan tem grande importância nas técnicas de Análise Multivariada. Sendo assim, é apresentado a seguir.

Resultado 2.2:

Qualquer matriz simétrica A ($p \times p$) pode ser escrita como

$$A = \Gamma \Lambda \Gamma' = \sum_{i=1}^p \lambda_i \underline{\gamma}_i \underline{\gamma}_i' \quad (2.11)$$

onde Λ é a matriz diagonal dos autovalores (λ_i) de A e Γ é uma matriz ortogonal cujas colunas são os autovetores padronizados ($\underline{\gamma}_i$).

Prova:

Suponha que seja possível encontrar vetores ortonormais $\gamma_1, \gamma_2, \dots, \gamma_p$ tal que $A\underline{\gamma}_i = \lambda_i \underline{\gamma}_i$, para algum valor λ_i . Então

$$\underline{\gamma}_i' A \underline{\gamma}_j = \lambda_j \underline{\gamma}_i' \underline{\gamma}_j = \begin{cases} \lambda_i, & i = j \\ 0, & i \neq j \end{cases} \quad (2.12)$$

$$\text{ou na forma matricial } \Gamma' A \Gamma = \Lambda \quad (2.13)$$

Pré e pós multiplicando a expressão (2.13) por Γ e Γ' , respectivamente, tem-se:

$$\Gamma \Gamma' A \Gamma \Gamma' = \Gamma \Lambda \Gamma' \quad (2.14)$$

$$A = \Gamma \Lambda \Gamma' \quad (2.15)$$

Tem-se, neste caso, que A e Λ têm os mesmos autovalores, conforme mostra a expressão A.8.1, do apêndice 8. Então, os elementos de Λ são exatamente os autovalores de A com as mesmas multiplicidades.

É preciso achar bases ortonormais dos autovetores. Note que, se $\lambda_i \neq \lambda_j$ são autovalores distintos, com autovetores \underline{x} e \underline{y} , respectivamente, então $\lambda_i \underline{x}'\underline{y} = \underline{x}'A\underline{y} = \underline{y}'A\underline{x} = \lambda_j \underline{y}'\underline{x}$, de modo que $\underline{y}'\underline{x} = 0$. Portanto, para a matriz simétrica, autovetores correspondendo a autovalores distintos são ortogonais entre si.

Supondo que existem k autovalores distintos de A com autoespaços correspondentes H_1, H_2, \dots, H_k de dimensões r_1, r_2, \dots, r_k .

Seja

$$r = \sum_{j=1}^k r_j \quad (2.16)$$

Já que distintos autoespaços são ortogonais, existe um conjunto ortonormal de vetores e_1, e_2, \dots, e_r , tal que os vetores denominados

$$\sum_{i=1}^{j-1} r_i + 1, \dots, \sum_{i=1}^j r_i \quad (2.17)$$

formam uma base para H_j . Tem-se que r_j é menor ou igual à multiplicidade do autovalor correspondente, conforme apresentado em MARDIA, KENT e BIBBY (1982, p.467): seja λ_1 um autovalor particular de $A_{p \times p}$, com autoespaço H de dimensão r . Se k representa a multiplicidade de λ_1 em H , então $1 \leq r \leq k$.

Portanto, reordenando os autovalores λ_i , se necessário, pode-se supor que:

$$A \underline{e}_i = \lambda_i \underline{e}_i, \quad i = 1, 2, \dots, r \quad (2.18)$$

e $r \leq p$.

Se $r = p$, tem-se que $\underline{\gamma}_i = \underline{e}_i$, o que prova o teorema.

É necessário mostrar que se $r < p$, cai-se numa contradição, o que não pode ocorrer.

Sem perda de generalidade, pode-se supor que todos os autovalores de A são estritamente positivos (se não, pode-se substituir A por $A + \alpha I$, para um α adequado, pois ambos têm os mesmos autovetores).

$$\text{Seja } B = A - \sum_{i=1}^r \lambda_i \underline{e}_i \underline{e}'_i \quad (2.19)$$

Então tem-se que: $\text{tr} B = \text{tr} A - \sum_{i=1}^r \lambda_i \underline{e}_i \underline{e}'_i = \sum_{i=r+1}^p \lambda_i > 0$, desde que $r < p$. Então B tem pelo menos um autovalor diferente de zero, chamado θ . Seja $\underline{x} \neq \underline{0}$ autovetor correspondente. Então para $1 \leq j \leq r$,

$$\theta \underline{e}'_j \underline{x} = \underline{e}'_j B \underline{x} = \left\{ \lambda_j \underline{e}'_j - \sum_{i=1}^r \lambda_i (\underline{e}'_i \underline{e}_i) \underline{e}'_j \right\} \underline{x} = 0 \quad (2.20)$$

de modo que \underline{x} é ortogonal a \underline{e}_j , $j = 1, 2, \dots, r$. Então

$$\theta \underline{x} = B \underline{x} = \left(A - \sum_{i=1}^r \lambda_i \underline{e}_i \underline{e}'_i \right) \underline{x} = A \underline{x} - \sum_{i=1}^r \lambda_i (\underline{e}'_i \underline{x}) \underline{e}_i = A \underline{x} \quad (2.21)$$

de modo que \underline{x} é também autovetor de A . Assim, $\theta = \lambda_i$ para algum i e \underline{x} é uma combinação linear para algum dos \underline{e}_i , que contradiz a ortogonalidade entre \underline{x} e \underline{e}_i .

Assim, fica demonstrado o teorema da decomposição espectral, e, como consequência, a prova do resultado 2.2.

2.4 JACOBIANO DA MATRIZ DE TRANSFORMAÇÃO

Definição 2.2:

Suponha que X e Y sejam matrizes que têm o mesmo número de elementos distintos r . Então se $Y = f(X)$, o jacobiano da transformação é definido como sendo:

$$J(Y \rightarrow X) = \|A\|, \text{ onde } A = \left(\frac{\partial y_i}{\partial x_j} \right), \quad i, j = 1, 2, \dots, r \quad (2.22)$$

Tem-se que $\|A\|$ é o valor absoluto de $|A|$ e (x_1, x_2, \dots, x_r) e (y_1, y_2, \dots, y_r) são os distintos valores de X e Y , respectivamente.

Listam-se, a seguir, algumas propriedades importantes do Jacobiano da transformação, apresentadas em PRESS (1982). Os Jacobianos são utilizados com frequência na Análise Multivariada, para obter densidades de funções de vetores e matrizes aleatórios.

2.4.1 Propriedades do Jacobiano

1. Se $\underline{y}_{p \times 1}$, $\underline{x}_{p \times 1}$, $A_{p \times p}$ e $\underline{y} = A \underline{x}$, então tem-se que

$$J(\underline{y} \rightarrow \underline{x}) = |A| \text{ é o jacobiano da transformação linear do vetor.} \quad (2.23)$$

2. Se $Y_{p \times q}$, $A_{p \times p}$, $X_{p \times q}$ e $Y = AX$, então tem-se que

$$J(Y \rightarrow X) = |A|^q \quad \text{onde } q \text{ colunas de } Y \text{ são transformações de } q \quad (2.24)$$

colunas de X .

3. Se $Y_{p \times q}$, $X_{p \times q}$, $B_{q \times q}$ e $Y = XB$, então tem-se que

$$J(Y \rightarrow X) = |B|^p \quad \text{é análoga à propriedade 2, exceto que as trans-} \quad (2.25)$$

formações são aplicadas nas p linhas.

4. Se $Y_{p \times q}$, $A_{p \times p}$, $B_{q \times q}$, $X_{p \times q}$ e $Y = AXB$, então tem-se que

$$J(Y \rightarrow X) = |A|^q |B|^p \quad (2.26)$$

5. Se $Y_{p \times p}$, $X_{p \times p}$, $X = X'$ e $Y = AXA'$, então tem-se que

$$J(Y \rightarrow X) = |A|^{p+1}, \quad |A| \neq 0 \quad (2.27)$$

6. Se $Y_{p \times q}$, $X_{p \times q}$, “ a ” é um escalar e $Y = aX$, então tem-se que

$$J(Y \rightarrow X) = a^{pq} \quad (2.28)$$

Esta transformação corresponde à mudança de escala da unidade de todos os elementos de Y .

7. Se $Y_{p \times p}$, $X_{p \times p}$, $Y = Y'$, $X = X'$, “ a ” é um escalar e $Y = aX$, então tem-se que

$$J(Y \rightarrow X) = a^{p(p+1)/2} \quad (2.29)$$

8. Se $|A| \neq 0$, $\partial A^{-1} = -A^{-1}(\partial A)A^{-1}$.

Se $X = A^{-1}$, então tem-se que

$$J(A \rightarrow X) = |X|^{-(p+1)}, \text{ onde } X_{p \times p} \text{ sendo } X = X'. \quad (2.30)$$

9. Se $f(X) \cong \text{tr}(AX'\Sigma^{-1}X)$, onde $A_{q \times q}$ sendo $A = A'$, $X_{p \times q}$, $\Sigma_{p \times p} > 0$, então tem-se

$$\text{que } \frac{\partial f(x)}{\partial X} = 2\Sigma^{-1}XA \quad (2.31)$$

2.5 DISTRIBUIÇÃO NORMAL MULTIVARIADA

2.5.1 Função Densidade de Probabilidade

A função densidade de probabilidade da distribuição normal multivariada é uma generalização da normal univariada para $p \geq 2$ dimensões (JOHNSON e WICHERN, 1988).

Relembrando, a função densidade de probabilidade (f.d.p.) da distribuição normal univariada é:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad \mu \in \mathbb{R}, \sigma > 0, x \in \mathbb{R} \quad (2.32)$$

O expoente da f.d.p. da distribuição normal univariada pode ser desenvolvido em: $\left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu)$. Esta expressão mede a distância quadrática de x em relação a μ , em unidades do desvio padrão. E, esta distância pode ser generalizada para o caso multivariado, onde \underline{x} é um vetor de dimensão p . Então, tem-se:

$$(\underline{x} - \underline{\mu})' \Sigma^{-1} (\underline{x} - \underline{\mu}) \quad (2.33)$$

O vetor $\underline{\mu}$ na expressão (2.33), de dimensão p , representa o valor esperado do vetor aleatório \underline{X} , e a matriz Σ , de ordem $p \times p$, simétrica e definida positiva, é a matriz de covariância desse vetor. A expressão (2.33) é conhecida como distância de Mahalanobis (D^2).

Ao substituir a expressão (2.33), na função densidade de probabilidade dada em (2.32), a constante de normalização $\sigma\sqrt{2\pi}$ deve ser trocada, de forma que o volume sob a superfície da densidade multivariada seja igual à unidade, para qualquer p . Segundo ANDERSON (1958, p.12), esta constante é $(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}$. Deste modo, a f.d.p. da distribuição normal multivariada é dada por:

$$f(\underline{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})'\Sigma^{-1}(\underline{x}-\underline{\mu})}, \quad (2.34)$$

sendo $\underline{\mu} \in \mathbb{R}^p$, Σ é definida positiva e $\underline{x} \in \mathbb{R}^p$.

Assim, se o vetor aleatório p -dimensional \underline{X} tem distribuição normal multivariada, a sua função densidade de probabilidade é representada por $N_p(\underline{\mu}, \Sigma)$.

O vetor médio e a matriz de covariância do vetor aleatório p -dimensional \underline{X} são apresentados adiante:

$$E(\underline{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \underline{\mu} \quad (2.35)$$

$$e \Sigma = \text{COV}(\underline{X}) = E(\underline{X}-\underline{\mu})(\underline{X}-\underline{\mu})' = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix} \quad (2.36)$$

É comum separar as informações contidas nas variâncias σ_i^2 daquelas contidas nas medidas de associação, em particular o coeficiente de correlação populacional ρ_{ik} .

O coeficiente de correlação populacional, ρ_{ik} é definido como segue:

$$\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_i^2} \sqrt{\sigma_k^2}} = \frac{\text{COV}(X_i, X_k)}{\sigma_i \sigma_k} \quad (2.37)$$

em que:

ρ_{ik} é o coeficiente de correlação entre as variáveis X_i e X_k ;

σ_{ik} é a covariância entre as variáveis X_i e X_k ;

σ_i^2 é a variância da variável X_i ;

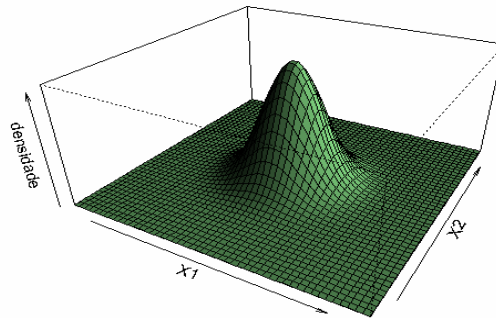
σ_k^2 é a variância da variável X_k .

Assim, a matriz de correlação populacional ρ é obtida e apresentada a seguir:

$$\rho = \begin{bmatrix} \frac{\sigma_1^2}{\sqrt{\sigma_1^2} \sqrt{\sigma_1^2}} & \frac{\sigma_{12}}{\sqrt{\sigma_1^2} \sqrt{\sigma_2^2}} & \dots & \frac{\sigma_{1p}}{\sqrt{\sigma_1^2} \sqrt{\sigma_p^2}} \\ \frac{\sigma_{21}}{\sqrt{\sigma_1^2} \sqrt{\sigma_2^2}} & \frac{\sigma_2^2}{\sqrt{\sigma_2^2} \sqrt{\sigma_2^2}} & \dots & \frac{\sigma_{2p}}{\sqrt{\sigma_2^2} \sqrt{\sigma_p^2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{p1}}{\sqrt{\sigma_1^2} \sqrt{\sigma_p^2}} & \frac{\sigma_{p2}}{\sqrt{\sigma_2^2} \sqrt{\sigma_p^2}} & \dots & \frac{\sigma_p^2}{\sqrt{\sigma_p^2} \sqrt{\sigma_p^2}} \end{bmatrix} = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \dots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & 1 \end{bmatrix} \quad (2.38)$$

Apresenta-se a seguir, como ilustração, a figura da função densidade de probabilidade da distribuição normal bivariada, com variâncias iguais, $\sigma_1^2 = \sigma_2^2$, e correlação nula, $\rho = 0$. Para a obtenção da figura utilizou-se o sistema R, cujo *script* encontra-se no apêndice 3.

FIGURA 1 - FUNÇÃO DENSIDADE DE PROBABILIDADE DA DISTRIBUIÇÃO NORMAL BIVARIADA



FONTE: A autora

NOTA: $\sigma_1^2 = \sigma_2^2$ e $\rho = 0$

2.5.2 Propriedades da Distribuição Normal Multivariada

São apresentadas a seguir, como resultados, propriedades bastante úteis da distribuição normal multivariada.

Resultado 2.3:

Seja \underline{X} um vetor aleatório, com distribuição normal p -variada, ou seja, $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$. Tem-se que:

$$\underline{Y} = C\underline{X} \quad (2.39)$$

onde C é uma matriz não singular. Então, \underline{Y} tem distribuição $N_p(C\underline{\mu}, C\Sigma C')$.

Prova:

A função densidade de probabilidade do vetor aleatório \underline{X} é dada por:

$$f(\underline{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})' \Sigma^{-1}(\underline{x}-\underline{\mu})} \quad (2.40)$$

em que $\underline{\mu} \in \mathbb{R}^p$, Σ é definida positiva e $\underline{x} \in \mathbb{R}^p$.

E, a densidade de \underline{Y} é obtida a partir da densidade de \underline{X} , fazendo a seguinte substituição:

$$\underline{x} = \mathbf{C}^{-1}\underline{y} \quad (2.41)$$

e multiplicando pelo jacobiano de transformação da expressão (2.41). Assim, tem-se que:

$$J = \left\| \frac{\partial \underline{x}}{\partial \underline{y}} \right\| = \|\mathbf{C}^{-1}\|, \text{ que é o Jacobiano de transformação, como apresentado}$$

na seção 2.4.

$$\text{Fazendo } \|\mathbf{C}^{-1}\| = \frac{1}{\|\mathbf{C}\|} = \sqrt{\frac{1}{|\mathbf{C}|^2}} = \sqrt{\frac{|\Sigma|}{|\mathbf{C}||\Sigma||\mathbf{C}'|}} = \frac{|\Sigma|^{\frac{1}{2}}}{|\mathbf{C}\Sigma\mathbf{C}'|^{\frac{1}{2}}} \quad (2.42)$$

A forma quadrática¹ do expoente de $f(\underline{x})$ da expressão (2.40) é:

$$Q = (\underline{x} - \underline{\mu})' \Sigma^{-1} (\underline{x} - \underline{\mu}) \quad (2.43)$$

Substituindo (2.41) em (2.43) e desenvolvendo, tem-se:

$$Q = (\mathbf{C}^{-1}\underline{y} - \underline{\mu})' \Sigma^{-1} (\mathbf{C}^{-1}\underline{y} - \underline{\mu}) \quad (2.44)$$

$$Q = (\mathbf{C}^{-1}\underline{y} - \mathbf{C}^{-1}\mathbf{C}\underline{\mu})' \Sigma^{-1} (\mathbf{C}^{-1}\underline{y} - \mathbf{C}^{-1}\mathbf{C}\underline{\mu}) \quad (2.45)$$

$$Q = [\mathbf{C}^{-1}(\underline{y} - \mathbf{C}\underline{\mu})]' \Sigma^{-1} [\mathbf{C}^{-1}(\underline{y} - \mathbf{C}\underline{\mu})] \quad (2.46)$$

$$Q = (\underline{y} - \mathbf{C}\underline{\mu})' (\mathbf{C}^{-1})' \Sigma^{-1} \mathbf{C}^{-1} (\underline{y} - \mathbf{C}\underline{\mu}) \quad (2.47)$$

$$Q = (\underline{y} - \mathbf{C}\underline{\mu})' (\mathbf{C}\Sigma\mathbf{C}')^{-1} (\underline{y} - \mathbf{C}\underline{\mu}) \quad (2.48)$$

¹ Maiores detalhes sobre forma quadrática poderão ser obtidos no apêndice 8 deste trabalho.

Portanto, a densidade de \underline{Y} será:

$$f(\underline{y}) = f(\underline{x}) \left\| C^{-1} \right\| = \left[\frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{C}^{-1}\underline{y}-\underline{\mu})' \Sigma^{-1}(\underline{C}^{-1}\underline{y}-\underline{\mu})} \right] \frac{|\Sigma|^{\frac{1}{2}}}{|\underline{C}\Sigma\underline{C}'|^{\frac{1}{2}}} \quad (2.49)$$

$$f(\underline{y}) = f(\underline{x}) \left\| C^{-1} \right\| = \frac{1}{(2\pi)^{\frac{p}{2}} |\underline{C}\Sigma\underline{C}'|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{y}-\underline{C}\underline{\mu})' (\underline{C}\Sigma^{-1}\underline{C})^{-1}(\underline{y}-\underline{C}\underline{\mu})} \quad (2.50)$$

$$f(\underline{y}) = f(\underline{x}) \left\| C^{-1} \right\| = f(\underline{C}\underline{x}) \quad (2.51)$$

$$\text{Logo, } \underline{Y} \sim N_p(\underline{C}\underline{\mu}, \underline{C}\Sigma\underline{C}') \quad (2.52)$$

Resultado 2.4:

Seja \underline{X} , com distribuição $N_p(\underline{\mu}, \Sigma)$ com $|\Sigma| > 0$. Então, tem-se que a distância de Mahalanobis $(\underline{x}-\underline{\mu})' \Sigma^{-1}(\underline{x}-\underline{\mu})$ é distribuída conforme distribuição qui-quadrado com p graus de liberdade, ou seja, $(\underline{x}-\underline{\mu})' \Sigma^{-1}(\underline{x}-\underline{\mu}) \sim \chi_p^2$.

Prova:

Tem-se que $Z_1^2 + Z_2^2 + \dots + Z_p^2 = \sum_{i=1}^p Z_i^2 \sim \chi_p^2$, onde Z_1, Z_2, \dots, Z_p são variáveis independentes $N(0,1)$ e pelo teorema da decomposição espectral (ver seção 2.3) tem-se que $\Sigma = \sum_{i=1}^p \lambda_i \underline{e}_i \underline{e}_i'$, e a sua inversa é dada por $\Sigma^{-1} = \sum_{i=1}^p \frac{1}{\lambda_i} \underline{e}_i \underline{e}_i'$.

Pré-multiplicando por $(\underline{x}-\underline{\mu})'$ ambos os membros da igualdade da segunda expressão e pós-multiplicando por $(\underline{x}-\underline{\mu})$, tem-se:

$$(\underline{x}-\underline{\mu})' \Sigma^{-1}(\underline{x}-\underline{\mu}) = \sum_{i=1}^p \frac{1}{\lambda_i} (\underline{x}-\underline{\mu})' \underline{e}_i \underline{e}_i' (\underline{x}-\underline{\mu}) = \sum_{i=1}^p \frac{1}{\lambda_i} [\underline{e}_i' (\underline{x}-\underline{\mu})]^2 \quad (2.53)$$

$$(\underline{x}-\underline{\mu})' \Sigma^{-1}(\underline{x}-\underline{\mu}) = \sum_{i=1}^p \left[\frac{1}{\sqrt{\lambda_i}} \underline{e}_i' (\underline{x}-\underline{\mu}) \right]^2 \quad (2.54)$$

Fazendo $\underline{Z} = A(\underline{x} - \underline{\mu})$ com $A = \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} \underline{e}'_1 \\ \frac{1}{\sqrt{\lambda_2}} \underline{e}'_2 \\ \vdots \\ \frac{1}{\sqrt{\lambda_i}} \underline{e}'_i \end{bmatrix}$, a expressão acima será

escrita na forma:

$$(\underline{x} - \underline{\mu})' \Sigma^{-1} (\underline{x} - \underline{\mu}) = \sum_{i=1}^p Z_i^2 \quad (2.55)$$

De modo que $(\underline{x} - \underline{\mu})$ tem distribuição $N_p(\underline{0}, \Sigma)$. Então, pelo resultado 2.3, tem-se que $\underline{Z} = A(\underline{x} - \underline{\mu})$ é distribuído como $N_p(\underline{0}, A\Sigma A')$. É necessário mostrar que $A\Sigma A' = I$, pois assim tem-se que $\underline{Z} \sim N_p(\underline{0}, I)$.

Então, fazendo o produto das matrizes, $A\Sigma A'$, resultará em:

$$A \Sigma A' = \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} \underline{e}'_1 \\ \frac{1}{\sqrt{\lambda_2}} \underline{e}'_2 \\ \vdots \\ \frac{1}{\sqrt{\lambda_p}} \underline{e}'_p \end{bmatrix} \left[\sum_{i=1}^p \lambda_i \underline{e}_i \underline{e}'_i \right] \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} \underline{e}_1 & \frac{1}{\sqrt{\lambda_2}} \underline{e}_2 & \cdots & \frac{1}{\sqrt{\lambda_p}} \underline{e}_p \end{bmatrix} \quad (2.56)$$

$$A \Sigma A' = \begin{bmatrix} \sqrt{\lambda_1} \underline{e}'_1 \\ \sqrt{\lambda_2} \underline{e}'_2 \\ \vdots \\ \sqrt{\lambda_p} \underline{e}'_p \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} \underline{e}_1 & \frac{1}{\sqrt{\lambda_2}} \underline{e}_2 & \cdots & \frac{1}{\sqrt{\lambda_p}} \underline{e}_p \end{bmatrix} = I \quad (2.57)$$

Portanto, $\underline{Z} \sim N_p(\underline{0}, I)$. Logo, $(\underline{x} - \underline{\mu})' \Sigma^{-1} (\underline{x} - \underline{\mu})$ tem distribuição χ_p^2 . (2.58)

2.5.3 Estimadores de Máxima Verossimilhança da Distribuição Normal Multivariada

O resultado a seguir apresenta os estimadores de máxima verossimilhança da distribuição normal multivariada.

Resultado 2.5:

Os estimadores de máxima verossimilhança do vetor médio $\underline{\mu}$ e da matriz de covariância Σ são, respectivamente:

$$\hat{\underline{\mu}} = \bar{\underline{X}} \quad (\text{vetor de médias}) \quad (2.59)$$

$$\hat{\Sigma} = \frac{1}{n} \mathbf{V} = \frac{1}{n} \sum_{i=1}^n (\underline{x}_i - \bar{\underline{X}})(\underline{x}_i - \bar{\underline{X}})' \quad (\text{matriz de covariância}) \quad (2.60)$$

Prova:

Sejam os vetores $(p \times 1)$, $\underline{X}_1, \underline{X}_2, \dots, \underline{X}_n$, que formam uma amostra aleatória da população normal multivariada com vetor médio $\underline{\mu}$ e matriz de covariância Σ , ou seja, $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$. Como $\underline{X}_1, \underline{X}_2, \dots, \underline{X}_n$ são mutuamente independentes, a função densidade conjunta da amostra é o produto das densidades marginais normais. Então, tem-se:

$$f(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{x}_1 - \underline{\mu})' \Sigma^{-1}(\underline{x}_1 - \underline{\mu})} \times \dots \times \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{x}_n - \underline{\mu})' \Sigma^{-1}(\underline{x}_n - \underline{\mu})} \quad (2.61)$$

e, a expressão acima pode ser escrita da seguinte forma:

$$f(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n) = \prod_{i=1}^n \left[\frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{x}_i - \underline{\mu})' \Sigma^{-1}(\underline{x}_i - \underline{\mu})} \right] \quad (2.62)$$

e, finalmente, tem-se que:

$$f(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n) = \frac{1}{(2\pi)^{\frac{n \times p}{2}} |\Sigma|^{\frac{n}{2}}} e^{-\frac{1}{2} \sum_{i=1}^n (\underline{x}_i - \underline{\mu})' \Sigma^{-1} (\underline{x}_i - \underline{\mu})} \quad (2.63)$$

a expressão acima é função de $\underline{\mu}$ e Σ , e para o conjunto fixo de observações $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n$ é chamada de verossimilhança, e é apresentada abaixo.

$$L(\underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{n \times p}{2}} |\Sigma|^{\frac{n}{2}}} e^{-\frac{1}{2} \sum_{i=1}^n (\underline{x}_i - \underline{\mu})' \Sigma^{-1} (\underline{x}_i - \underline{\mu})} \quad (2.64)$$

Escrevendo o expoente na forma de traço, e adicionando e subtraindo \bar{X} , tem-se:

$$L(\underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{n \times p}{2}} |\Sigma|^{\frac{n}{2}}} e^{-\frac{1}{2} \text{tr} \left[\Sigma^{-1} \left(\sum_{i=1}^n (\underline{x}_i - \bar{X})(\underline{x}_i - \bar{X})' + n(\bar{X} - \underline{\mu})(\bar{X} - \underline{\mu})' \right) \right]} \quad (2.65)$$

reescrevendo a expressão acima, tem-se:

$$L(\underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{n \times p}{2}} |\Sigma|^{-1 \times \frac{n}{2}}} e^{-\frac{1}{2} \text{tr} \left[\Sigma^{-1} \left(\sum_{i=1}^n (\underline{x}_i - \bar{X})(\underline{x}_i - \bar{X})' + n(\bar{X} - \underline{\mu})(\bar{X} - \underline{\mu})' \right) \right]} \quad (2.66)$$

e aplicando a propriedade do determinante, tem-se que $|\Sigma|^{-1} = |\Sigma^{-1}|$, portanto a expressão acima será:

$$L(\underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{n \times p}{2}} |\Sigma^{-1}|^{\frac{n}{2}}} e^{-\frac{1}{2} \text{tr} \left[\Sigma^{-1} \left(\sum_{i=1}^n (\underline{x}_i - \bar{X})(\underline{x}_i - \bar{X})' + n(\bar{X} - \underline{\mu})(\bar{X} - \underline{\mu})' \right) \right]} \quad (2.67)$$

E, definindo $\Lambda \equiv \Sigma^{-1}$, tem-se que:

$$L(\underline{\mu}, \Lambda) = \frac{1}{(2\pi)^{\frac{n \times p}{2}} |\Lambda|^{\frac{n}{2}}} e^{-\frac{1}{2} \text{tr} \left[\Lambda \left(\sum_{i=1}^n (\underline{x}_i - \bar{X})(\underline{x}_i - \bar{X})' + n(\bar{X} - \underline{\mu})(\bar{X} - \underline{\mu})' \right) \right]} \quad (2.68)$$

Definindo $nV = \sum_{i=1}^n (\underline{x}_i - \bar{X})(\underline{x}_i - \bar{X})'$, e substituindo na expressão (2.68), tem-se:

$$L(\underline{\mu}, \Lambda) = \frac{1}{(2\pi)^{\frac{n \times p}{2}}} |\Lambda|^{\frac{n}{2}} e^{-\frac{1}{2} \text{tr}[\Lambda(nV + n(\bar{X} - \underline{\mu})(\bar{X} - \underline{\mu})']]} \quad (2.69)$$

Agora, escrevendo a expressão (2.69) na forma logarítmica, tem-se:

$$\ln L(\underline{\mu}, \Lambda) = -\frac{np}{2} \ln(2\pi) + \frac{n}{2} \ln |\Lambda| - \frac{n}{2} \text{tr} \Lambda V - \frac{n}{2} \text{tr} \Lambda (\bar{X} - \underline{\mu})(\bar{X} - \underline{\mu})' \quad (2.70)$$

Derivando a expressão (2.70) em relação a $\underline{\mu}$, tem-se:

$$\frac{\partial \ln L(\underline{\mu}, \Lambda)}{\partial \underline{\mu}} = -\frac{n}{2} \frac{\partial [(\bar{X} - \underline{\mu}) \Lambda (\bar{X} - \underline{\mu})']}{\partial \underline{\mu}} \quad (2.71)$$

$$\frac{\partial \ln L(\underline{\mu}, \Lambda)}{\partial \underline{\mu}} = -\frac{n}{2} 2\Lambda(\bar{X} - \underline{\mu}) \quad (2.72)$$

Igualando a zero tem-se:

$$-\frac{n}{2} 2\Lambda(\bar{X} - \underline{\mu}) = 0 \quad (2.73)$$

Como foi definido que $\Lambda \cong \Sigma^{-1}$, sendo Λ definida positiva, a expressão (2.73) será igual a zero somente se $(\bar{X} - \underline{\mu}) = \underline{0}$. Logo:

$$\hat{\underline{\mu}} = \bar{X} \quad (2.74)$$

Derivando a expressão (2.70) em relação a Λ , tem-se:

$$\frac{\partial \ln L(\underline{\mu}, \Lambda)}{\partial \Lambda} = \frac{\partial \left[\frac{n}{2} \ln |\Lambda| - \frac{n}{2} \text{tr} \Lambda V - \frac{n}{2} \text{tr} \Lambda (\bar{X} - \underline{\mu})(\bar{X} - \underline{\mu})' \right]}{\partial \Lambda} \quad (2.75)$$

$$\frac{\partial \ln L(\underline{\mu}, \Lambda)}{\partial \Lambda} = \frac{n}{2} [2\Lambda^{-1} - \text{diag} \Lambda^{-1}] - \frac{n}{2} [2V - \text{diag} V] - \frac{n}{2} [2(\bar{X} - \underline{\mu})(\bar{X} - \underline{\mu})' - \text{diag}(\bar{X} - \underline{\mu})(\bar{X} - \underline{\mu})'] \quad (2.76)$$

Definiu-se que $\Lambda \equiv \Sigma^{-1}$, então, $\Lambda^{-1} \equiv (\Sigma^{-1})^{-1} = \Sigma$, assim:

$$\frac{\partial \ln L(\underline{\mu}, \Lambda)}{\partial \Lambda} = n\Sigma - \frac{n}{2} \text{diag } \Sigma - nV + \frac{n}{2} \text{diag } V - \frac{n}{2} [2(\bar{X} - \underline{\mu})(\bar{X} - \underline{\mu})' - \text{diag}(\bar{X} - \underline{\mu})(\bar{X} - \underline{\mu})'] \quad (2.77)$$

Como $\hat{\underline{\mu}} = \bar{X}$, $\text{diag } \Sigma = \text{diag } (V)$, logo, tem-se:

$$\frac{\partial \ln L(\underline{\mu}, \Lambda)}{\partial \Lambda} = n\Sigma - nV \quad (2.78)$$

Igualando a zero, resultará em:

$$\hat{\Sigma} = V = \frac{1}{n} \sum_{i=1}^n (\underline{x}_i - \bar{X})(\underline{x}_i - \bar{X})' \quad (2.79)$$

2.5.4 Estimadores não Viesados da Distribuição Normal Multivariada

Os estimadores não viesados da distribuição normal multivariada são apresentados no resultado 2.6.

Resultado 2.6:

Sejam os vetores de dimensão p , $\underline{X}_1, \underline{X}_2, \dots, \underline{X}_n$, que formam uma amostra aleatória da população normal multivariada com vetor médio $\underline{\mu}$ e matriz de covariância Σ , ou seja, $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$. Então \bar{X} e $S = \frac{1}{n-1} \sum_{i=1}^n (\underline{x}_i - \bar{X})(\underline{x}_i - \bar{X})'$ são os estimadores não viesados de $\underline{\mu}$ e Σ .

Prova:

$$\text{Seja : } \bar{X} = \frac{1}{n} [\underline{X}_1 + \underline{X}_2 + \dots + \underline{X}_n] \quad (2.80)$$

Então, tem-se que:

$$E(\bar{X}) = E\left\{\frac{1}{n}[\underline{X}_1 + \underline{X}_2 + \dots + \underline{X}_n]\right\} \quad (2.81)$$

$$E(\bar{X}) = \frac{1}{n}E[\underline{X}_1 + \underline{X}_2 + \dots + \underline{X}_n] \quad (2.82)$$

$$E(\bar{X}) = \frac{1}{n}[\underline{\mu} + \underline{\mu} + \dots + \underline{\mu}] \quad (2.83)$$

$$E(\bar{X}) = \underline{\mu} \quad (2.84)$$

E, o estimador não viesado de Σ é S, conforme apresentado a seguir:

$$\text{Tem-se que: } S = \frac{1}{n-1} \sum_{i=1}^n (\underline{x}_i - \bar{X})(\underline{x}_i - \bar{X})' \quad (2.85)$$

Então:

$$E(S) = E\left[\frac{1}{n-1} \sum_{i=1}^n (\underline{x}_i - \bar{X})(\underline{x}_i - \bar{X})'\right] \quad (2.86)$$

$$E(S) = \frac{1}{n-1} E\left[\sum_{i=1}^n \underline{x}_i \underline{x}_i' - n \bar{X} \bar{X}'\right] \quad (2.87)$$

$$E(S) = \frac{1}{n-1} \left[\sum_{i=1}^n E(\underline{x}_i \underline{x}_i') - n E(\bar{X} \bar{X}') \right] \quad (2.88)$$

$$E(S) = \frac{1}{n-1} \left[n(\underline{\mu} \underline{\mu}' + \Sigma) - n \left(\underline{\mu} \underline{\mu}' + \frac{1}{n} \Sigma \right) \right] \quad (2.89)$$

$$E(S) = \frac{1}{n-1} [n \underline{\mu} \underline{\mu}' + n \Sigma - n \underline{\mu} \underline{\mu}' - \Sigma] \quad (2.90)$$

$$E(S) = \frac{(n-1)}{n-1} \Sigma \quad (2.91)$$

$$E(S) = \Sigma \quad (2.92)$$

2.5.5 Distribuição Amostral de \bar{X} e S

Seja $X' = [X_1, X_2, \dots, X_n]$ uma amostra aleatória da normal univariada, ou seja, $X \sim N(\mu, \sigma^2)$. Então, no caso univariado ($p = 1$), \bar{X} é normalmente distribuída com média μ e variância $\frac{\sigma^2}{n}$, ou seja, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. Tem-se ainda que

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Resultado 2.7:

Para o caso multivariado ($p \geq 2$), \bar{X} tem distribuição normal com média $\underline{\mu}$ e matriz de covariância $\frac{1}{n} \Sigma$.

Prova:

Da seção 2.5.4, tem-se que $E(\bar{X}) = \underline{\mu}$

$$E, \text{ tem-se que: } \text{COV}(\bar{X}) = E(\bar{X} - \underline{\mu})(\bar{X} - \underline{\mu})' \quad (2.93)$$

Portanto:

$$\text{COV}(\bar{X}) = E\left[\frac{1}{n} \sum_{i=1}^n (\underline{x}_i - \underline{\mu}) \frac{1}{n} \sum_{i=1}^n (\underline{x}_i - \underline{\mu})'\right] \quad (2.94)$$

$$\text{COV}(\bar{X}) = \frac{1}{n^2} E\left[\sum_{i=1}^n (\underline{x}_i - \underline{\mu}) (\underline{x}_i - \underline{\mu})'\right] \quad (2.95)$$

$$\text{COV}(\bar{X}) = \frac{1}{n} \Sigma \quad (2.96)$$

Definição 2.3:

Segue-se a definição da distribuição de Wishart apresentada por Mardia, Kent e Bibby (MARDIA, KENT e BIBBY, 1982, p.66):

Se $M_{p \times p}$ pode ser escrita como $M = XX'$, onde $X_{m \times p}$ é a matriz de dados da distribuição $N_p(\underline{0}, \Sigma)$, então M tem distribuição de Wishart com matriz de escala Σ e graus de liberdade m , e representa-se $M \sim W(\Sigma, m)$. Quando $\Sigma = I_p$, a distribuição é dita ser na forma padrão.

A distribuição amostral da matriz de covariância amostral (S) é chamada de Distribuição de Wishart. De acordo com JOHNSON e WICHERN (1988), $(n-1)S$ é distribuída como matriz aleatória de Wishart com $(n-1)$ graus de liberdade.

Outra distribuição importante é a de $n(\bar{X} - \underline{\mu})' S^{-1}(\bar{X} - \underline{\mu})$, para a construção do intervalo de confiança para o vetor médio $\underline{\mu}$. De acordo com JOHNSON e WICHERN (1988), a distribuição χ^2 é aproximadamente a distribuição amostral de $n(\bar{X} - \underline{\mu})\Sigma^{-1}(\bar{X} - \underline{\mu})'$, quando \bar{X} é aproximadamente normalmente distribuído. E, ainda segundo os autores, quando o tamanho da amostra n é grande e é muito maior que o número de variáveis p , a substituição de Σ^{-1} por S^{-1} não afeta seriamente a aproximação da distribuição χ^2 . Então, tem-se:

$$n(\bar{X} - \underline{\mu})' S^{-1}(\bar{X} - \underline{\mu}) \text{ é aproximadamente } \chi_p^2. \quad (2.97)$$

2.5.6 Avaliação da Suposição de Normalidade (Gaussianidade)

Tendo em vista que muitas técnicas multivariadas dependem da suposição de Gaussianidade, é prudente checar essa premissa. Em situações em que o tamanho da amostra é grande, e as técnicas dependem unicamente do comportamento de \bar{X} , ou das distâncias envolvendo \bar{X} , da forma $n(\bar{X} - \underline{\mu})' S^{-1}(\bar{X} - \underline{\mu})$, a suposição de normalidade é menos crucial (JOHNSON e WICHERN, 1988, p.142 e 146). Mas, até certo ponto, a qualidade das inferências feitas por esses métodos depende de quanto se aproximam da distribuição normal multivariada.

Os gráficos são sempre úteis para qualquer análise estatística de dados. O gráfico *Q-Q plots* pode ser utilizado para avaliar a suposição de normalidade. Este gráfico pode ser construído para distribuições marginais das observações amostrais de cada variável. Trata-se de um gráfico do quantil amostral *versus* quantil esperado da distribuição normal. Quando os pontos estão bastante próximos da reta, a suposição de normalidade pode ser aceita, pois $Z_i = \frac{X_i - \mu}{\sigma}$ e $X_i = \sigma Z_i + \mu$ (equação da reta).

Uma outra possibilidade é aplicar testes baseados nas medidas de assimetria e curtose multivariada de Mardia (MARDIA, 1970). Para qualquer distribuição normal multivariada essas medidas são obtidas respectivamente pelas expressões a seguir:

$$\beta_{1,p} = E \left\{ (\underline{y} - \underline{\mu})' \Sigma^{-1} (\underline{x} - \underline{\mu}) \right\}^3 \quad (2.98)$$

sendo \underline{x} independente de \underline{y} , mas com a mesma distribuição e,

$$\beta_{2,p} = E \left\{ (\underline{y} - \underline{\mu})' \Sigma^{-1} (\underline{y} - \underline{\mu}) \right\}^2 \quad (2.99)$$

Assim, quando a distribuição é normal multivariada tem-se que $\beta_{1,p} = 0$ e $\beta_{2,p} = p(p+2)$. As estimativas de $\beta_{1,p}$ e $\beta_{2,p}$, para amostras de tamanho n , podem ser obtidas através de:

$$\hat{\beta}_{1,p} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n g_{ij}^3 \quad (2.100)$$

$$\hat{\beta}_{2,p} = \frac{1}{n} \sum_{i=1}^n g_{ii}^2 = \frac{1}{n} \sum_{i=1}^n d_i^4 \quad (2.101)$$

$$\text{onde } g_{ij} = (\underline{y}_i - \bar{\underline{y}})' S^{-1} (\underline{y}_j - \bar{\underline{y}}) \quad (2.102)$$

$$\text{e } d_i = \sqrt{g_{ii}} \quad (2.103)$$

MARDIA (1970) demonstrou que para grandes amostras tem-se:

$$\kappa_1 = \frac{n\hat{\beta}_{1,p}}{6} \sim \chi^2_v \quad (2.104)$$

$$\text{onde } v = \frac{p(p+1)(p+2)}{6} \quad (2.105)$$

$$\kappa_2 = \frac{\hat{\beta}_{2,p} - p(p+2)}{\left(\frac{8p(p+2)}{n}\right)^{\frac{1}{2}}} \sim N(0,1) \quad (2.106)$$

Estas estatísticas podem ser utilizadas para testar a hipótese nula de multinormalidade. Rejeita-se a hipótese nula para valores pequenos de κ_1 e κ_2 .

2.5.7 Inferência sobre Vetor de Médias

Quando o objetivo é fazer inferências sobre a média populacional, é possível utilizar dois métodos: testes de significância e intervalos de confiança. Será abordada aqui a obtenção do intervalo de confiança para a média populacional.

2.5.7.1 Intervalo de confiança

Sejam os vetores $\underline{X}_1, \underline{X}_2, \dots, \underline{X}_n$, que compõem uma amostra independente e normalmente distribuída com média $\underline{\mu}$ e matriz de covariância Σ . Então as estimativas de $\underline{\mu}$ e Σ são obtidas do vetor \underline{X} a partir dos estimadores não viesados (viciados):

$$\bar{\underline{X}} = \frac{1}{n} \sum_{i=1}^n \underline{x}_i \quad (2.107)$$

$$S = \frac{1}{n-1} \sum_{i=1}^n (\underline{x}_i - \bar{\underline{X}})(\underline{x}_i - \bar{\underline{X}})' \quad (2.108)$$

No caso univariado, a região de confiança é um intervalo na reta. Já, no caso multivariado é uma região elipsoidal $R(\underline{X})$. A região $R(\underline{X})$ é chamada de $100(1-\alpha)\%$ de confiança se, antes da amostra ser selecionada, $P\left[R(\underline{X}) \text{ cobrir o verdadeiro } \underline{\theta}\right] = 1-\alpha$, onde $\underline{\theta}$ representa o vetor de parâmetros desconhecidos. A região de confiança para $\underline{\mu}$ com nível de $(1-\alpha)100\%$, segundo JOHNSON e WICHERN (1988), é dada por:

$$P\left[n(\bar{\underline{X}}-\underline{\mu})'S^{-1}(\bar{\underline{X}}-\underline{\mu}) \leq \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\right] = 1-\alpha \quad (2.109)$$

onde $n(\bar{\underline{X}}-\underline{\mu})'S^{-1}(\bar{\underline{X}}-\underline{\mu}) = T^2$, e é chamado de T^2 de Hotelling, sendo uma generalização da distância quadrática $t^2 = n(\bar{X}-\mu)(S^2)^{-1}(\bar{X}-\mu)$ do caso univariado.

Tem-se que:

$$T^2 \sim \frac{(n-1)p}{n-p}F_{p,n-p} \quad (2.110)$$

Resultado 2.8:

Sejam $\underline{X}_1, \underline{X}_2, \dots, \underline{X}_n$ amostras aleatórias da população normal $N_p(\underline{\mu}, \Sigma)$, com Σ definida positiva. Assim, os intervalos simultâneos de confiança de $1-\alpha$, para a cobertura dos valores paramétricos, para todo $\underline{\ell}$, são dados pela expressão a seguir:

$$P\left[\underline{\ell}'\bar{\underline{X}} - \sqrt{\frac{(n-1)p}{n(n-p)}F_{p,n-p}(\alpha)}\underline{\ell}'S \underline{\ell} \leq \underline{\ell}'\underline{\mu} \leq \underline{\ell}'\bar{\underline{X}} + \sqrt{\frac{(n-1)p}{n(n-p)}F_{p,n-p}(\alpha)}\underline{\ell}'S \underline{\ell}\right] = 1-\alpha$$

Prova:

Fazendo $T^2 = n(\bar{\underline{X}}-\underline{\mu})'S^{-1}(\bar{\underline{X}}-\underline{\mu}) \leq c^2$, é possível obter-se a seguinte expressão:

$$n \frac{(\underline{\ell}' \bar{X} - \underline{\ell}' \underline{\mu})^2}{\underline{\ell}' \underline{S} \underline{\ell}} \leq c^2 \quad (2.111)$$

$$c \geq \pm \sqrt{n \frac{(\underline{\ell}' \bar{X} - \underline{\ell}' \underline{\mu})^2}{\underline{\ell}' \underline{S} \underline{\ell}}} \quad (2.112)$$

Resolvendo a expressão anterior, obtém-se:

$$\underline{\ell}' \underline{\mu} \geq \underline{\ell}' \bar{X} - c \sqrt{\frac{\underline{\ell}' \underline{S} \underline{\ell}}{n}} \quad (2.113)$$

$$\text{e } \underline{\ell}' \underline{\mu} \leq \underline{\ell}' \bar{X} + c \sqrt{\frac{\underline{\ell}' \underline{S} \underline{\ell}}{n}}, \text{ logo} \quad (2.114)$$

$$\underline{\ell}' \bar{X} - c \sqrt{\frac{\underline{\ell}' \underline{S} \underline{\ell}}{n}} \leq \underline{\ell}' \underline{\mu} \leq \underline{\ell}' \bar{X} + c \sqrt{\frac{\underline{\ell}' \underline{S} \underline{\ell}}{n}} \quad (2.115)$$

Escolhendo-se $c^2 = \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$, tem-se de (2.109) que

$$P\left[T^2 \leq c^2\right] = 1 - \alpha, \text{ logo}$$

$$P\left[\underline{\ell}' \bar{X} - c \sqrt{\frac{\underline{\ell}' \underline{S} \underline{\ell}}{n}} \leq \underline{\ell}' \underline{\mu} \leq \underline{\ell}' \bar{X} + c \sqrt{\frac{\underline{\ell}' \underline{S} \underline{\ell}}{n}}\right] = 1 - \alpha \quad (2.116)$$

Finalmente, os intervalos simultâneos de confiança podem ser obtidos através da expressão:

$$P\left[\underline{\ell}' \bar{X} - \sqrt{\frac{(n-1)p}{n(n-p)} F_{p,n-p}(\alpha) \underline{\ell}' \underline{S} \underline{\ell}} \leq \underline{\ell}' \underline{\mu} \leq \underline{\ell}' \bar{X} + \sqrt{\frac{(n-1)p}{n(n-p)} F_{p,n-p}(\alpha) \underline{\ell}' \underline{S} \underline{\ell}}\right] = 1 - \alpha \quad (2.117)$$

que conterà $\underline{\ell}' \underline{\mu}$, com probabilidade $1 - \alpha$, para todo $\underline{\ell}$, ou seja:

quando $\underline{\ell}' = [1 \ 0 \ \dots \ 0]$, $\underline{\ell}' \underline{\mu} = \underline{\mu}_1$

$\underline{\ell}' = [0 \ 1 \ \dots \ 0]$, $\underline{\ell}' \underline{\mu} = \underline{\mu}_2$

\vdots

$\underline{\ell}' = [0 \ 0 \ \dots \ 1]$, $\underline{\ell}' \underline{\mu} = \underline{\mu}_p$

2.5.7.2 Inferência sobre a média populacional a partir de grandes amostras

Quando a amostra é grande, testes de hipóteses e regiões de confiança podem ser construídos sem a suposição da normalidade da população. Todas as inferências sobre $\underline{\mu}$ a partir de grandes amostras são baseadas na distribuição χ^2 (JOHNSON e WICHERN, 1988, p.190).

Conforme apresentado na seção 2.5.5, $n(\bar{X} - \underline{\mu})' \Sigma^{-1}(\bar{X} - \underline{\mu})$ tem distribuição aproximadamente χ_p^2 . Para n grande, muito maior que p , a distribuição de $n(\bar{X} - \underline{\mu})' S^{-1}(\bar{X} - \underline{\mu})$ é aproximadamente χ_p^2 . Assim, fazendo-se $c^2 = \chi^2$ na expressão (2.116), o intervalo de confiança será:

$$P \left[\underline{\ell}' \bar{X} - \sqrt{\chi_p^2(\alpha)} \sqrt{\frac{\underline{\ell}' S \underline{\ell}}{n}} \leq \underline{\ell}' \underline{\mu} \leq \underline{\ell}' \bar{X} + \sqrt{\chi_p^2(\alpha)} \sqrt{\frac{\underline{\ell}' S \underline{\ell}}{n}} \right] = 1 - \alpha \quad (2.118)$$

para todo $\underline{\ell}$ (JOHNSON e WICHERN, 1988, p.191).

2.5.8 Região de Confiança com Largura Fixa

Desejando-se fixar a largura da região de confiança, faz-se necessário variar o tamanho da amostra. No caso univariado, se X_1, X_2, \dots, X_n são independentes, com distribuição $N(\mu, \sigma^2)$, a média amostral é \bar{X} . Se σ^2 é conhecida, o intervalo de confiança para μ é obtido através da expressão $\left(\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$, para nível de confiança de $100(1-\alpha)\%$. Fixando-se a largura do intervalo tal que seja $< 2d$, para algum valor fixo de d tem-se:

$$Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < d \quad (2.119)$$

Isolando n , tem-se:

$$n > \frac{\sigma^2 Z_{\alpha/2}^2}{d^2} \quad (2.120)$$

Agora, considerando-se as observações multivariadas $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n \sim N_p(\underline{\mu}, \Sigma)$, com Σ conhecida. É possível, segundo SRIVASTAVA e CARTER (1983, p.45), obter k intervalos de confiança para os parâmetros da forma $\underline{\ell}'_i \underline{\mu}$ para algum vetor $\underline{\ell}'_i$, $i = 1, 2, \dots, k$. Se \bar{X} é a média amostral, então os k intervalos de confiança simultâneos para nível de $(1 - \alpha)100\%$, para $\underline{\ell}'_i$, $i = 1, 2, \dots, k$ são dados por:

$$\underline{\ell}'_i \bar{X} \pm \sqrt{\frac{\chi_{p,\alpha}^2 \underline{\ell}'_i \Sigma \underline{\ell}_i}{n}}, \text{ onde } P(\chi_p^2 > \chi_{p,\alpha}^2) = \alpha \quad (2.121)$$

$$\text{Quando } Z_{\alpha/2k} < \chi_{p,\alpha}, \text{ utiliza-se: } \underline{\ell}'_i \bar{X} \pm Z_{\alpha/2k} \sqrt{\frac{\underline{\ell}'_i \Sigma \underline{\ell}_i}{n}} \quad (2.122)$$

Na maioria dos casos, a não ser que k seja grande, tem-se que $Z_{\alpha/2k} < \chi_{p,\alpha}$. Neste caso, para o comprimento do j-ésimo intervalo não ser maior que $2d$, escolhe-se n tal que:

$$n > Z_{\alpha/2k}^2 \left(\frac{\underline{\ell}'_i \Sigma \underline{\ell}_i}{d_i^2} \right), \quad i = 1, 2, \dots, k \quad (2.123)$$

Se utilizar $\chi_{p,\alpha}^2$ ao invés de $Z_{\alpha/2k}^2$, então escolhe-se n tal que:

$$n > \chi_{p,\alpha}^2 \left(\frac{\underline{\ell}'_i \Sigma \underline{\ell}_i}{d_i^2} \right), \quad i = 1, 2, \dots, k \quad (2.124)$$

Se Σ é desconhecida, é possível utilizar sua estimativa S (SRIVASTAVA e CARTER, 1983). Assim, as expressões (2.121) e (2.122) poderão ser expressas como segue:

$$\underline{\ell}'_i \bar{X} \pm \sqrt{\frac{\chi_{p,\alpha}^2 \underline{\ell}'_i S \underline{\ell}_i}{n}} \quad (2.125)$$

$$\text{E, quando } Z_{\alpha/2k} < \chi_{p,\alpha}, \quad \underline{\ell}'_i \bar{X} \pm Z_{\alpha/2k} \sqrt{\frac{\underline{\ell}'_i S \underline{\ell}_i}{n}} \quad (2.126)$$

Para obter o comprimento do j -ésimo intervalo que não seja maior que $2d$, escolhe-se n tal que:

$$n > Z_{\alpha/2k}^2 \left(\frac{\ell'_i S \ell_i}{d_i^2} \right), \quad i = 1, 2, \dots, k \quad (2.127)$$

$$\text{e } n > \chi_{p,\alpha}^2 \left(\frac{\ell'_i S \ell_i}{d_i^2} \right), \quad i = 1, 2, \dots, k \quad (2.128)$$

quando utiliza-se $Z_{\alpha/2k}^2$ e $\chi_{p,\alpha}^2$, respectivamente.

2.6 MÉTODO DE MONTE CARLO

O Método de Monte Carlo tem sido bastante utilizado para obter aproximações numéricas de funções complexas. Envolve a geração de observações de alguma distribuição de probabilidades e a utilização da amostra obtida, para aproximação da função de interesse (EHLERS, 2003).

As aplicações mais comuns do Método de Monte Carlo, em computação numérica, são para avaliar integrais. O propósito do método é escrever a integral que se deseja calcular na forma de esperança matemática, ou seja, do valor esperado. De acordo com EHLERS (2003), a expressão em que a integral é a esperança matemática de uma função $g(X)$, onde X tem função densidade de probabilidade $f(x)$, é dada por:

$$M = \int_a^b g(x)f(x)dx = E[g(X)] \quad (2.129)$$

Assim, é possível obter uma aproximação de M através de:

$$\hat{M} = \frac{1}{n} \sum_{i=1}^n g(x_i) \quad (2.130)$$

Dado que \hat{M} é uma aproximação para a quantidade que se deseja obter, faz-se necessário estudar o erro, $(\hat{M}-M)$. Uma vez que os valores de $g(X)$ são gerados independentemente e pela Lei Forte dos Grandes Números, segue que \hat{M} converge quase certamente para M , ou seja:

$$\frac{1}{n} \sum_{i=1}^n g(x_i) \xrightarrow{n \rightarrow \infty} E[g(X)] \text{ quase certamente} \quad (2.131)$$

Ainda, de acordo com EHLERS (2003), definindo $\sigma^2 = V[g(X)]$ e assumindo a existência da variância, o erro padrão de Monte Carlo é uma estimativa consistente de σ , dada pela expressão:

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n [g(x_i) - \bar{g}]^2} \quad (2.132)$$

A aproximação pode ser tão precisa quanto se deseja, bastando para isso aumentar o valor de n , que é o número de amostras (repetições). Hoje em dia, com o cálculo computacional barato, isto é muito fácil.

Quando se tem o vetor aleatório de dimensão p , $\underline{X} = (X_1, X_2, \dots, X_p)$, com função densidade de probabilidade $f(\underline{x})$, os valores gerados serão também vetores, e o estimador de Monte Carlo é dado por:

$$\underline{\hat{M}} = \frac{1}{n} \sum_{i=1}^n g(\underline{x}_i) \quad (2.133)$$

2.7 ANÁLISE DE REGRESSÃO LINEAR MÚLTIPLA

A Análise de Regressão Linear Múltipla é usada na modelagem do relacionamento entre uma variável resposta (dependente) e k variáveis explicativas (independentes). O modelo linear geral é dado por:

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon} \quad (2.134)$$

em que:

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \text{ é o vetor de respostas de dimensão } n; \quad (2.135)$$

$$\underline{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix} \text{ é a matriz do modelo de ordem } n \times (k+1); \quad (2.136)$$

$$\underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \text{ é o vetor de parâmetros de dimensão } (k+1); \quad (2.137)$$

$$\underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \text{ é o vetor de erros de dimensão } n. \quad (2.138)$$

As suposições sobre a componente erro ($\underline{\varepsilon}$) do modelo de regressão são as seguintes:

$$(i) E(\underline{\varepsilon}) = \underline{0} \quad (2.139)$$

$$(ii) COV(\underline{\varepsilon}) = \sigma^2 I_n \quad (2.140)$$

$$(iii) COV(\varepsilon_i, \varepsilon_j) = 0, \quad i, j = 1, 2, \dots, n, \quad i \neq j \quad (2.141)$$

Ainda, quando há interesse em fazer inferências estatísticas, ou seja, testar hipóteses sobre os parâmetros e construir intervalos de confiança, é necessário atender à suposição de Gaussianidade para os erros ε_i , $i = 1, 2, \dots, n$. Assim, tem-se:

$$(iv) \underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I_n) \quad (2.142)$$

2.7.1 Estimação pelo Método dos Mínimos Quadrados

Um dos objetivos da análise de regressão é desenvolver um modelo que possibilite prever a variável resposta sendo conhecidos os valores das variáveis explicativas (independentes). Desta forma, faz-se necessário ajustar um modelo para a variável observada Y_i e os correspondentes valores conhecidos de X_{ij} , ou seja, é preciso determinar os valores dos coeficientes de regressão $\underline{\beta}$ e a variância do erro σ^2 .

Resultado 2.9:

Seja o modelo $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$. O estimador de mínimos quadrados do vetor de parâmetros $\underline{\beta}$ é dado por:

$$\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{Y} \quad (2.143)$$

Prova:

Dado o modelo $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$, tem-se então que:

$$\underline{\varepsilon} = \underline{Y} - X\underline{\beta} \text{ é o vetor coluna dos erros} \quad (2.144)$$

O método dos mínimos quadrados seleciona o valor $\underline{\beta}$ que minimiza a soma dos quadrados dos erros (ou resíduos), ou seja:

$$\text{SQR} = \sum_{i=1}^n \varepsilon_i^2 = \underline{\varepsilon}' \underline{\varepsilon} \quad (2.145)$$

Substituindo a expressão (2.144) na (2.145), e desenvolvendo-a, tem-se:

$$\text{SQR} = \underline{\varepsilon}' \underline{\varepsilon} = \underline{Y}' \underline{Y} - 2\underline{\beta}' X' \underline{Y} + \underline{\beta}' X' X \underline{\beta} \quad (2.146)$$

Derivando a expressão (2.146) em relação a $\underline{\beta}$, e igualando seu resultado a zero, tem-se:

$$\frac{\partial \text{SQR}}{\partial \underline{\beta}} = -2\underline{X}'\underline{Y} + 2\underline{X}'\underline{X}\underline{\beta} \quad (2.147)$$

$$\text{Logo, } \hat{\underline{\beta}} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{Y} \quad (2.148)$$

Agora, derivando a expressão (2.147), novamente em relação a $\underline{\beta}$, comprova-se que é ponto de mínimo, pois:

$$\frac{\partial^2 \text{SQR}}{\partial \underline{\beta}^2} = 2\underline{X}'\underline{X} > 0 \quad (2.149)$$

2.7.1.1 Propriedades do estimador $\hat{\underline{\beta}}$ e do vetor de erros estimado $\hat{\underline{\varepsilon}}$

As propriedades do estimador $\hat{\underline{\beta}}$ e do vetor de erros $\hat{\underline{\varepsilon}}$ estão demonstradas nos resultados apresentados a seguir.

Resultado 2.10:

O estimador de mínimos quadrados $\hat{\underline{\beta}}$ é não viesado, ou seja, $E(\hat{\underline{\beta}}) = \underline{\beta}$ e sua variância é dada por $V(\hat{\underline{\beta}}) = \sigma^2 (\underline{X}'\underline{X})^{-1}$.

Prova:

Da expressão (2.134) do modelo de regressão linear geral, tem-se que $\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon}$. Substituindo-a na (2.148), tem-se:

$$\hat{\underline{\beta}} = (\underline{X}'\underline{X})^{-1}\underline{X}'(\underline{X}\underline{\beta} + \underline{\varepsilon}) = \underline{\beta} + (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{\varepsilon} \quad (2.150)$$

Aplicando a esperança matemática em ambos os membros da igualdade e desenvolvendo, resultará em:

$$E(\hat{\underline{\beta}}) = \underline{\beta} + (\underline{X}'\underline{X})^{-1}\underline{X}'E(\underline{\varepsilon}) \quad (2.151)$$

Mas, tem-se da suposição (i) que $E(\underline{\varepsilon}) = \underline{0}$, portanto:

$$E(\underline{\hat{\beta}}) = \underline{\beta} \quad (2.152)$$

Da expressão (2.150), tem-se que:

$$(\underline{\hat{\beta}} - \underline{\beta}) = (X'X)^{-1}X'\underline{\varepsilon} \quad (2.153)$$

Mas, a variância do estimador é dada por:

$$V(\underline{\hat{\beta}}) = E\left[(\underline{\hat{\beta}} - \underline{\beta})(\underline{\hat{\beta}} - \underline{\beta})'\right] \quad (2.154)$$

Então, substituindo a expressão (2.153) na (2.154), tem-se:

$$V(\underline{\hat{\beta}}) = E\left[\left((X'X)^{-1}X'\underline{\varepsilon}\right)\left((X'X)^{-1}X'\underline{\varepsilon}\right)'\right] \quad (2.155)$$

Finalmente, desenvolvendo a expressão (2.155), o resultado será:

$$V(\underline{\hat{\beta}}) = \sigma^2(X'X)^{-1} \quad (2.156)$$

Resultado 2.11:

O vetor de erros (ou resíduos), $\underline{\hat{\varepsilon}} = \underline{Y} - X\underline{\hat{\beta}}$, tem as seguintes propriedades:

$$\text{i) } E(\underline{\hat{\varepsilon}}) = \underline{0} \quad (2.157)$$

$$\text{ii) } \text{COV}(\underline{\hat{\varepsilon}}) = \sigma^2(I - X(X'X)^{-1}X') \quad (2.158)$$

$$\text{iii) } E(\underline{\hat{\varepsilon}}'\underline{\hat{\varepsilon}}) = (n - k - 1)\sigma^2 \quad (2.159)$$

$$\text{iv) } E(S^2) = \sigma^2 \quad (2.160)$$

Prova:

i) Substituindo a expressão (2.148) em $\hat{\underline{\varepsilon}} = \underline{Y} - X\hat{\underline{\beta}}$, tem-se:

$$\hat{\underline{\varepsilon}} = \underline{Y} - X(X'X)^{-1}X'\underline{Y} = (I - X(X'X)^{-1}X')\underline{Y} \quad (2.161)$$

Substituindo a expressão (2.134) na (2.161) e desenvolvendo, resultará em:

$$\hat{\underline{\varepsilon}} = (I - X(X'X)^{-1}X')\underline{\varepsilon} \quad (2.162)$$

Aplicando a esperança matemática em ambos os membros da igualdade, tem-se:

$$E(\hat{\underline{\varepsilon}}) = (I - X(X'X)^{-1}X') E(\underline{\varepsilon}) \quad (2.163)$$

Da suposição (i) tem-se que $E(\underline{\varepsilon}) = \underline{0}$, portanto:

$$E(\hat{\underline{\varepsilon}}) = \underline{0}$$

ii) Aplicando a covariância em ambos os membros da igualdade da expressão (2.162), tem-se:

$$\text{COV}(\hat{\underline{\varepsilon}}) = (I - X(X'X)^{-1}X') \text{COV}(\underline{\varepsilon})(I - X(X'X)^{-1}X')' \quad (2.164)$$

$$\text{COV}(\hat{\underline{\varepsilon}}) = \sigma^2(I - X(X'X)^{-1}X') \quad (2.165)$$

iii) Da expressão (2.162) tem-se que $\hat{\underline{\varepsilon}} = (I - X(X'X)^{-1}X')\underline{\varepsilon}$, assim:

$$\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} = [(I - X(X'X)^{-1}X')\underline{\varepsilon}]'[(I - X(X'X)^{-1}X')\underline{\varepsilon}] \quad (2.166)$$

$$\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} = \underline{\varepsilon}'[I - X(X'X)^{-1}X']\underline{\varepsilon} \quad (2.167)$$

Escrevendo a expressão (2.167) na forma de traço, tem-se:

$$\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} = \text{tr}[(I - X(X'X)^{-1}X')\underline{\varepsilon}\underline{\varepsilon}'] \quad (2.168)$$

Aplicando a esperança matemática em ambos os membros da igualdade:

$$E(\underline{\hat{\varepsilon}}' \underline{\hat{\varepsilon}}) = \text{tr}[(I - X(X'X)^{-1}X') E(\underline{\varepsilon}\underline{\varepsilon}')] \quad (2.169)$$

Tem-se que $E(\underline{\varepsilon}\underline{\varepsilon}') = V(\underline{\varepsilon}) = \sigma^2$, portanto, a expressão (2.169) será:

$$E(\underline{\hat{\varepsilon}}' \underline{\hat{\varepsilon}}) = \sigma^2 \text{tr}[I - X(X'X)^{-1}X'] \quad (2.170)$$

Desenvolvendo a expressão (2.170), resultará em:

$$E(\underline{\hat{\varepsilon}}' \underline{\hat{\varepsilon}}) = \sigma^2 n - \sigma^2 (k + 1) = \sigma^2 n - \sigma^2 k - \sigma^2 \quad (2.171)$$

$$E(\underline{\hat{\varepsilon}}' \underline{\hat{\varepsilon}}) = \sigma^2 (n - k - 1) \quad (2.172)$$

iv) Definindo $S^2 = \frac{\underline{\hat{\varepsilon}}' \underline{\hat{\varepsilon}}}{n - (k + 1)}$ (2.173)

Aplicando a esperança matemática em ambos os membros da igualdade da expressão (2.173), resultará em:

$$E(S^2) = E\left(\frac{\underline{\hat{\varepsilon}}' \underline{\hat{\varepsilon}}}{n - (k + 1)}\right) \quad (2.174)$$

$$E(S^2) = \frac{E(\underline{\hat{\varepsilon}}' \underline{\hat{\varepsilon}})}{n - k - 1} \quad (2.175)$$

Mas, tem-se da expressão (2.172) que:

$$E(\underline{\hat{\varepsilon}}' \underline{\hat{\varepsilon}}) = \sigma^2 (n - k - 1) \quad (2.176)$$

Substituindo a expressão (2.176) na (2.175), tem-se o resultado final:

$$E(S^2) = \frac{\sigma^2 (n - k - 1)}{n - k - 1} = \sigma^2 \quad (2.177)$$

2.7.1.2 Decomposição da soma de quadrados total

A soma de quadrados da resposta $\underline{Y}' \underline{Y} = \sum_{i=1}^n y_i^2$ satisfaz a expressão

matricial:

$$\underline{Y}' \underline{Y} = (\underline{\hat{Y}} + \underline{Y} - \underline{\hat{Y}})' (\underline{\hat{Y}} + \underline{Y} - \underline{\hat{Y}}) = (\underline{\hat{Y}} + \underline{\hat{\varepsilon}})' (\underline{\hat{Y}} + \underline{\hat{\varepsilon}}) \quad (2.178)$$

$$\underline{Y}' \underline{Y} = \underline{\hat{Y}}' \underline{\hat{Y}} + \underline{\hat{\varepsilon}}' \underline{\hat{\varepsilon}} \quad (2.179)$$

Finalmente, tem-se:

$$\underline{\hat{Y}}' \underline{\hat{Y}} = \underline{Y}' \underline{Y} - \underline{\hat{\varepsilon}}' \underline{\hat{\varepsilon}} \quad (2.180)$$

A expressão anterior pode ser escrita como a apresentada a seguir (JOHNSON e WICHERN, 1988):

$$\underline{Y}' \underline{Y} - n\bar{Y}^2 = \underline{\hat{Y}}' \underline{\hat{Y}} - n(\bar{\hat{Y}})^2 + \underline{\hat{\varepsilon}}' \underline{\hat{\varepsilon}} \quad (2.181)$$

ou

$$\sum_{i=1}^n (y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{Y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.182)$$

Logo, tem-se que a soma de quadrados total é a soma de quadrados devido à regressão, mais a soma de quadrados dos erros (resíduos).

A qualidade do modelo ajustado pode ser medida pelo coeficiente de determinação (R^2), como segue:

$$R^2 = 1 - \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\sum_{i=1}^n (y_i - \bar{Y})^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{Y})^2}{\sum_{i=1}^n (y_i - \bar{Y})^2}, \text{ sendo } 0 \leq R^2 \leq 1 \quad (2.183)$$

O coeficiente de determinação indica a proporção da variação total em y_i 's explicada ou atribuída a variáveis explicativas, X_1, X_2, \dots, X_k .

2.7.2 Inferência sobre os Parâmetros de Regressão

Testes de significância e intervalos de confiança para os $\hat{\beta}_i$ podem ser construídos. O procedimento segue a suposição (iv), expressão (2.142), ou seja, $\underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I_n)$.

Então, dada a suposição de normalidade para os erros, a função de verossimilhança para a amostra é dada por:

$$L(\underline{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{-\underline{\varepsilon}'\underline{\varepsilon}}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{(\underline{Y}-X\underline{\beta})'(\underline{Y}-X\underline{\beta})}{2\sigma^2}\right] \quad (2.184)$$

E, para um valor fixo de σ^2 , minimizar a verossimilhança com relação a $\underline{\beta}$, equivale a escolher um valor que minimiza a soma de quadrados $(\underline{Y}-X\underline{\beta})'(\underline{Y}-X\underline{\beta})$. O estimador de máxima verossimilhança de $\underline{\beta}$ é o estimador de mínimos quadrados $\hat{\underline{\beta}}$ obtido em (2.148). Além disso, tem-se que:

$$\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{Y} \sim N_{k+1}(\underline{\beta}, \sigma^2(X'X)^{-1}). \quad (2.185)$$

E, $\hat{\underline{\beta}}$ é distribuído independentemente dos resíduos $\hat{\underline{\varepsilon}} = \underline{Y} - X\hat{\underline{\beta}}$. Maiores detalhes podem ser encontrados em JOHNSON e WICHERN (1988, p. 284). O intervalo de confiança de $100(1-\alpha)\%$ para β_i é dado por:

$$P\left[\hat{\beta}_i - \sqrt{\hat{V}(\hat{\beta}_i)}\sqrt{(k+1)F_{k+1, n-k-1}(\alpha)} \leq \beta_i \leq \hat{\beta}_i + \sqrt{\hat{V}(\hat{\beta}_i)}\sqrt{(k+1)F_{k+1, n-k-1}(\alpha)}\right] = 1-\alpha \quad (2.186)$$

sendo $i = 0, 1, \dots, k$.

onde $\hat{V}(\hat{\beta}_i)$ é o elemento da diagonal de $S^2(X'X)^{-1}$ correspondente a $\hat{\beta}_i$ e $F_{k+1, n-k-1}(\alpha)$ é o (100α) -ésimo percentil superior da distribuição F com $(k+1)$ e $(n-k-1)$ graus de liberdade.

Segundo JOHNSON e WICHERN (1988), alguns autores substituem $(k+1)F_{k+1, n-k-1}(\alpha)$ por $t_{n-k-1}(\alpha/2)$ e utilizam o intervalo:

$$P\left[\hat{\beta}_i - t_{n-k-1}(\alpha/2)\sqrt{\hat{V}(\hat{\beta}_i)} \leq \beta_i \leq \hat{\beta}_i + t_{n-k-1}(\alpha/2)\sqrt{\hat{V}(\hat{\beta}_i)}\right] = 1-\alpha, \quad i = 0, 1, \dots, k \quad (2.187)$$

2.7.3 Teste para o Relacionamento Modelável por Regressão

A significância do ajuste do modelo é feita através da Análise da Variância. Esta técnica consiste em decompor a variância total da variável resposta (dependente) em duas componentes: a primeira devida ao modelo de regressão e a segunda devida aos resíduos (erros). Esta decomposição foi apresentada na seção 2.7.1.2, sendo denominadas respectivamente de: Soma de Quadrados Total (SQT), Soma de Quadrados da Regressão (SQRegr) e Soma de Quadrados dos Resíduos (SQR). Para cada uma dessas somas de quadrados existe associado um número de graus de liberdade da distribuição Qui-quadrado (χ^2), uma vez que se admite uma estrutura probabilística para os desvios do ajuste ε_i .

A estatística F é obtida pela razão entre a Soma de Quadrados da Regressão (SQRegr) e a Soma de Quadrados dos Resíduos (SQR), divididos pelos seus respectivos graus de liberdade. Assim, tem-se que:

$$F = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{Y})^2 / (p - 1)}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - p)} \quad (2.188)$$

em que:

$\sum_{i=1}^n (\hat{y}_i - \bar{Y})^2$ é a Soma de Quadrados da Regressão;

$\sum_{i=1}^n (y_i - \hat{y}_i)^2$ é a Soma de Quadrados dos Resíduos;

n é o número de observações;

p é o número de parâmetros estimados.

Um teste com a estatística F acima pode ser feito para testar se existe relacionamento modelável por regressão entre a variável resposta Y e o conjunto de variáveis explicativas X_1, X_2, \dots, X_k (NETER et al., 1996). As hipóteses a serem testadas são:

$$H_0 : \beta_0 = \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \text{algum } \beta_k \text{ difere de } 0$$

2.8 TESTES PARA AVALIAR AS SUPOSIÇÕES SOBRE A COMPONENTE ERRO

2.8.1 Teste de Multicolinearidade

De acordo com MONTEGOMERY e PECK (1982), as raízes características ou autovalores de $X'X$, ou seja, $\lambda_1, \lambda_2, \dots, \lambda_p$, podem ser usados para medir a extensão da multicolinearidade nos dados. Se existem uma ou mais dependências lineares nos dados, então um ou mais autovalores serão pequenos. Alguns analistas preferem examinar a condição de $X'X$, definida como:

$$k = \frac{\lambda_{\text{máx}}}{\lambda_{\text{min}}} \quad (2.189)$$

Geralmente, quando $k < 100$, não existe problema sério de multicolinearidade. Se $100 \leq k \leq 1000$, implica a existência de multicolinearidade moderada para forte, e, se $k > 1000$, indica multicolinearidade severa.

2.8.2 Teste de Homogeneidade de Variância

O teste de Goldfeld-Quandt (GUJARATI, 2000) é utilizado para verificar a homogeneidade da variância, σ_i^2 . Os autores sugerem os seguintes passos:

- (i) deve-se ordenar as observações de acordo com qualquer uma das variáveis explicativas X_i , em ordem crescente;
- (ii) omitir as observações centrais c , valor este especificado a princípio, e dividir as $(n - c)$ observações restantes em dois grupos, cada um com $\frac{(n - c)}{2}$ observações.
- (iii) ajustar as distintas regressões por mínimos quadrados ordinários (MQO) às primeiras $\frac{(n - c)}{2}$ observações e às últimas $\frac{(n - c)}{2}$ observações e obter as respectivas somas de quadrados dos resíduos, SQR_1 e SQR_2 , sendo, respectivamente, a soma dos quadrados dos resíduos da regressão que corresponde aos menores valores de X_i (o

grupo com pequena variância) e a que corresponde aos maiores valores de X_i (o grupo com grande variância). Os graus de liberdade de cada uma das somas dos quadrados são dados por $\frac{(n-c)}{2} - k$, onde k é o número de parâmetros a serem estimados, incluindo o intercepto;

(iv) calcular a razão:

$$\lambda = \frac{SQR_2/|g|}{SQR_1/|g|}, \text{ que segue distribuição F com } \frac{(n-c)}{2} - k \text{ graus de}$$

liberdade no numerador e no denominador.

De acordo com GUJARATI (2000), os valores de $c = 4$ se $n = 30$ e $c = 10$ se $n = 60$ mostram-se satisfatórios na prática.

2.8.3 Teste de Gaussianidade de Kolmogorov-Smirnov com Correção de Lilliefors

A variável do teste é a maior diferença observada entre a função de distribuição acumulada do modelo e a empírica da amostra.

A função de distribuição acumulada do modelo testado, ou função de repartição, fornece as probabilidades acumuladas em cada ponto, ou seja, $F(x) = P(X \leq x)$. A função de distribuição acumulada da amostra corresponderá ao gráfico das freqüências relativas acumuladas (ogiva). Designa-se essa segunda função por $G(x)$. O teste consta simplesmente da verificação do valor:

$$d = \max | F(x) - G(x) | \quad (2.190)$$

e da comparação com um valor crítico tabelado em função do nível de significância (α) e do tamanho da amostra (n).

O teste de Kolmogorov-Smirnov testa a normalidade a partir da média e desvio padrão fornecidos. Lilliefors fez uma adaptação, usando a média e o desvio-padrão calculados a partir do próprio conjunto de dados.

2.9 IDENTIFICAÇÃO DOS *OUTLIERS* E PONTOS INFLUENTES

A verificação do modelo ajustado pode ser feita pela identificação dos pontos que se encontram significativamente afastados dos demais, normalmente chamados de *outliers*, e dos pontos influentes.

De acordo com HAIR et al. (1998), essas observações não são necessariamente ruins, no sentido de que devem ser eliminadas. É preciso identificá-las e avaliar seus impactos, antes de qualquer procedimento. O pesquisador é incentivado a excluir as observações verdadeiramente excepcionais, mas não deve fazê-lo se, embora distintas, elas representam a população. Deve lembrar-se de que o objetivo do ajuste é assegurar o modelo mais representativo para dados amostrais; assim, refletirá melhor a população amostrada.

2.9.1 Resíduos *Studentizados* Externamente

Uma das formas para detectar os *outlier* é através dos resíduos *studentizados* externamente (NETER et al., 1996), obtidos pela expressão abaixo:

$$t_i = e_i \left[\frac{n - p - 1}{\text{SQR}(1 - h_{ii}) - e_i^2} \right]^{1/2}, \quad i = 1, 2, \dots, n \quad (2.191)$$

em que:

e_i é o i -ésimo resíduo;

n é o número de observações;

p é o número de parâmetros estimados;

SQR é a soma de quadrados dos resíduos;

h_{ii} é o i -ésimo elemento da diagonal da matriz H (*hat matrix*).

A matriz H é dada por:

$$H = X(X'X)^{-1}X' \quad (2.192)$$

e o elemento h_{ii} pertence à diagonal dessa matriz, que é:

$$h_{ii} = x_i'(X'X)^{-1}x_i \quad (2.193)$$

Identifica-se uma observação y como *outlier* quando o resíduo studentizado, em valor absoluto, é grande. É possível aplicar o teste de Bonferroni (NETER et al., 1996) para avaliar se este ponto é de fato um *outlier*. O valor crítico de Bonferroni é obtido através da expressão: $t(1-\alpha/2n, n-p-1)$.

Após a obtenção do resíduo *studentizado* externamente (t_i), para as n observações, deve-se calcular o valor crítico de Bonferroni para o maior valor absoluto, comprovando se é ou não *outlier*.

2.9.2 Pontos de Alavanca ou de Alto Leverage

Os pontos de alavanca ou de alto *leverage* são, também, identificados como *outliers*, neste caso, da matriz das variáveis explicativas. O *leverage* é o elemento da diagonal da matriz H, conforme apresentado na expressão (2.192). Os *leverages* maiores que $2p/n$ são considerados outliers (NETER et al., 1996).

2.9.3 Medida de Influência

Os pontos influentes, por sua vez, são aqueles que, quando excluídos, causam mudanças no modelo ajustado. Uma medida usual de influência (NETER et al., 1996) é dada por:

$$DFFITS_i = e_i \left[\frac{n-p-1}{\text{SQR}(1-h_{ii}) - e_i^2} \right]^{1/2} \left(\frac{h_{ii}}{1-h_{ii}} \right)^{1/2} = t_i \left(\frac{h_{ii}}{1-h_{ii}} \right)^{1/2} \quad (2.194)$$

em que:

e_i é o i -ésimo resíduo;

n é o número de observações;

p é o número de parâmetros estimados;

SQR é a soma de quadrados dos resíduos;

h_{ii} é o i -ésimo elemento da diagonal da matriz H (*hat matrix*);

t_i é o i -ésimo resíduo studentizado externamente.

NETER et al. (1996) sugerem que seja considerado um ponto influente aquele cujo valor de DFFITS seja superior a 1, quando o tamanho da amostra é de pequeno para médio, e superior a $2\sqrt{p/n}$, para amostra grande.

3 ANÁLISE FATORIAL

3.1 INTRODUÇÃO

O desenvolvimento inicial da Análise Fatorial surgiu na área da psicologia, com os estudos da habilidade mental. A análise era feita a partir das matrizes de correlações de um conjunto de testes cognitivos.

O estudo dessa técnica teve início com Charles Spearman, em 1904, quando publicou o artigo *“General intelligence objectively determined and measured”*. Neste, ele apresentou a teoria do fator de inteligência geral, o fator “g” (SPEARMAN, 1904).

Estudando os resultados de testes de inteligência, Spearman observou que, geralmente, havia alta correlação entre os escores de inteligência e notas escolares. Por meio dessas observações, ele avançou na sua famosa teoria de inteligência de dois fatores, mais conhecida como teoria do fator “g” (SPEARMAN, 1904).

Um fator é uma variável latente, na qual distribui-se um certo atributo quantitativo dos indivíduos. De acordo com a teoria de Spearman, o fator “g” é a quantidade que expressa a inteligência geral, descrevendo um determinado teste em maior ou menor grau. Já o fator s , específico de cada teste, é, por outro lado, o fator estabelecido pelo assunto ou tipo do teste. Supõe-se que todos os diferentes fatores s não são relacionados, e cada um deles, por sua vez, está relacionado com o teste específico (SPEARMAN, 1904).

De acordo com THURSTONE (1934), a teoria de Spearman, conhecida como método ou teoria de dois fatores, envolve um fator geral, comum a todos os testes ou variáveis, e outro fator que é específico para cada teste ou variável. Ainda, segundo ele, é menos ambíguo referir-se a esse método como sendo de apenas um fator, pois trata-se somente de um fator comum ou geral.

Thurstone, em seu artigo “*Vector of mind*”, descreveu a teoria do fator geral. Para Thurstone, a multidimensionalidade da mente precisava ser reconhecida, antes de se fazer progressos quanto ao isolamento das habilidades e à sua descrição separadamente. Entretanto, como todos os testes mentais são correlacionados positivamente, é possível descrever as intercorrelações em termos de vários fatores, de maneira que um dos fatores será evidente, na comparação com os demais. Porém, a definição exata deste varia de um conjunto de teste para outro (THURSTONE, 1934).

Segundo Thurstone, se este é o fator que Spearman afirmou na sua teoria da inteligência, então seu critério é inadequado, pois o *tetrad criterion*² somente diz se qualquer conjunto de intercorrelação pode ou não ser descrito em termos de um e somente um fator comum (THURSTONE, 1934).

Lawley foi quem fez a primeira abordagem da análise fatorial do ponto de vista da teoria estatística, tratando do problema da estimação, em 1940, quando publicou o artigo “*The estimation of factor loadings by the method of maximum likelihood*” (FACHEL, 1976).

De acordo com LAWLEY (1940), quando uma bateria de testes de inteligência é aplicada a um grupo de pessoas, é prática comum entre psicólogos explicar os escores obtidos em termos de número de fatores. Seja X_i o escore de qualquer pessoa no i -ésimo teste, e supondo que existam p testes, então assume-se que:

$$X_i = l_{i1}F_1 + l_{i2}F_2 + \dots + l_{im}F_m + \tau_i\varepsilon_i \quad , \quad (i = 1, 2, \dots, p) \quad (3.1)$$

onde $l_{i1}, l_{i2}, \dots, l_{im}, \tau_i$ representam as cargas fatoriais, F_1, F_2, \dots, F_m os fatores comuns e ε_i a habilidade específica da pessoa no i -ésimo teste. Tanto os fatores comuns quanto ε_i são independentes e normalmente distribuídos.

² O *tetrad criterion* é baseado no cálculo da diferença de produtos de quatro elementos da diagonal da matriz de correlação. Este critério é observado quando o resultado for igual ou próximo de zero.

Lawley apresenta o método da máxima verossimilhança para obtenção das cargas fatoriais fazendo comparações com outros métodos desenvolvidos por Spearman, Thurstone e Thomson (LAWLEY, 1940). Para estimar as cargas fatoriais pelo método da máxima verossimilhança, X_i deve seguir distribuição normal multivariada.

Segundo LAWLEY e MAXWELL (1962), a suposição básica na análise fatorial é:

$$X_i = \sum_{r=1}^k l_{ir} f_r + e_i, \quad (i = 1, 2, \dots, p) \quad (3.2)$$

em que:

l_{ir} é a carga do r-ésimo fator na i-ésima variável;

k é um valor especificado;

f_r é o r-ésimo fator comum;

e_i é o i-ésimo resíduo, representando fontes de variação que afetam somente X_i .

De acordo com FACHEL (1976), com a utilização dos resultados da Inferência Estatística a Análise Fatorial passou a ser considerada um método estatístico propriamente dito. Foi desenvolvido o teste sobre a adequação de um modelo com m fatores, pelo princípio da razão de verossimilhança.

3.2 MODELO FATORIAL ORTOGONAL

Seja um conjunto de p variáveis, cada um com n observações, formando o vetor de dados $\underline{X}' = [X_1, X_2, \dots, X_p]$; supondo que as p variáveis sejam correlacionadas, e que seja possível reduzir sua dimensão inicial através de novas variáveis hipotéticas que explicarão a maior parte da variação das variáveis originais. Assim, cada variável X_i , $i = 1, 2, \dots, p$ é representada como combinação linear de variáveis hipotéticas, chamadas fatores comuns (por serem comuns a várias variáveis), mais um fator residual ou específico para cada variável.

Dado o vetor aleatório $\underline{X}' = [X_1, X_2, \dots, X_p]$, com média $\underline{\mu}$ e matriz de covariância Σ , é possível escrever cada variável no modelo fatorial como segue (JOHNSON e WICHERN, 1988):

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \dots + l_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \dots + l_{2m}F_m + \varepsilon_2 \\ \dots\dots\dots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \dots + l_{pm}F_m + \varepsilon_p \end{aligned} \quad (3.3)$$

Na forma matricial, tem-se:

$$\underline{X}_{(px1)} - \underline{\mu}_{(px1)} = L_{(pxm)} \underline{F}_{(mx1)} + \underline{\varepsilon}_{(px1)} \quad (3.4)$$

em que:

L é a matriz das cargas fatoriais e l_{ij} é chamado de carga da i -ésima variável do j -ésimo fator;

\underline{F} é o vetor de fatores comuns;

$\underline{\varepsilon}$ é vetor de fatores específicos ou erros.

Os p desvios $X_1 - \mu_1$, $X_2 - \mu_2$, ..., $X_p - \mu_p$ são expressos em termos de $p + m$ variáveis aleatórias, $F_1, F_2, \dots, F_m, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_p$, que não são observáveis.

Assumindo-se que:

$$E(\underline{F}) = \underline{0} \quad \text{e} \quad \text{COV}(\underline{F}) = E(\underline{F}\underline{F}') = I \quad (3.5)$$

$$E(\underline{\varepsilon}) = \underline{0} \quad \text{e} \quad \text{COV}(\underline{\varepsilon}) = E(\underline{\varepsilon}\underline{\varepsilon}') = \Psi = \begin{bmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \psi_p \end{bmatrix} \quad (3.6)$$

e que \underline{F} e $\underline{\varepsilon}$ são independentes, então, tem-se que:

$$\text{COV}(\underline{\varepsilon}, \underline{F}) = E(\underline{\varepsilon}\underline{F}') = 0 \quad (3.7)$$

As suposições apresentadas em (3.5) e (3.6), juntamente com a expressão (3.4), constituem o modelo fatorial ortogonal, apresentado a seguir:

$$\underline{X} = \underline{\mu} + L\underline{F} + \underline{\varepsilon} \quad (3.8)$$

O modelo fatorial ortogonal sugere uma estrutura de covariância para \underline{X} . Do modelo em (3.8) tem-se:

$$(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})' = (L\underline{F} + \underline{\varepsilon})(L\underline{F} + \underline{\varepsilon})' \quad (3.9)$$

$$(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})' = (L\underline{F} + \underline{\varepsilon})[(L\underline{F})' + \underline{\varepsilon}'] \quad (3.10)$$

$$(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})' = L\underline{F}(L\underline{F})' + \underline{\varepsilon}(L\underline{F})' + L\underline{F}\underline{\varepsilon}' + \underline{\varepsilon}\underline{\varepsilon}' \quad (3.11)$$

A esperança matemática da expressão (3.11) será dada por:

$$E(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})' = E[L\underline{F}(L\underline{F})' + \underline{\varepsilon}(L\underline{F})' + L\underline{F}\underline{\varepsilon}' + \underline{\varepsilon}\underline{\varepsilon}'] \quad (3.12)$$

$$E(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})' = LE(\underline{F}\underline{F}')L' + E(\underline{\varepsilon}\underline{F}')L' + LE(\underline{F}\underline{\varepsilon}') + E(\underline{\varepsilon}\underline{\varepsilon}') \quad (3.13)$$

$$E(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})' = LIL' + 0L' + L0 + \psi \quad (3.14)$$

$$\text{Mas tem-se que } \Sigma = \text{COV}(\underline{X}) = E(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})' \quad (3.15)$$

Assim, substituindo a expressão (3.15) em (3.14) tem-se:

$$\Sigma = \text{COV}(\underline{X}) = LIL' + 0L' + L0 + \psi \quad (3.16)$$

$$\Sigma = \text{COV}(\underline{X}) = L L' + \psi \quad (3.17)$$

A expressão acima é fundamental, pois os parâmetros a serem estimados são os elementos de L e ψ .

Tem-se da expressão (3.8) que $\underline{X} - \underline{\mu} = L\underline{F} + \underline{\varepsilon}$. Pós-multiplicando ambos os membros da igualdade por \underline{F}' tem-se:

$$(\underline{X} - \underline{\mu}) \underline{F}' = (L\underline{F} + \underline{\varepsilon})\underline{F}' \quad (3.18)$$

$$(\underline{X} - \underline{\mu}) \underline{F}' = L\underline{F}\underline{F}' + \underline{\varepsilon} \underline{F}' \quad (3.19)$$

$$\text{Mas: } \text{COV}(\underline{X}, \underline{F}) = E(\underline{X} - \underline{\mu})(\underline{F} - E(\underline{F}))' = E(\underline{X} - \underline{\mu}) \underline{F}' \quad (3.20)$$

Então, obtendo a esperança matemática da expressão (3.19) e substituindo na (3.20), tem-se:

$$\text{COV}(\underline{X}, \underline{F}) = E[L\underline{F}\underline{F}' + \underline{\varepsilon} \underline{F}'] \quad (3.21)$$

$$\text{COV}(\underline{X}, \underline{F}) = LE(\underline{F}\underline{F}') + E(\underline{\varepsilon} \underline{F}') = LI + 0 = L \quad (3.22)$$

As expressões (3.17) e (3.22) podem ser escritas como segue:

$$V(X_i) = \ell_{i1}^2 + \ell_{i2}^2 + \dots + \ell_{im}^2 + \psi_i \quad (3.23)$$

$$\text{COV}(X_i, F_j) = \ell_{ij} \quad (3.24)$$

e ainda,

$$\text{COV}(X_i, X_k) = \ell_{i1}\ell_{k1} + \ell_{i2}\ell_{k2} + \dots + \ell_{im}\ell_{km} \quad (3.25)$$

A porção da variância da i -ésima variável aleatória X_i , devido à contribuição dos m fatores comuns, é chamada de i -ésima comunalidade. A porção da $V(X_i) = \sigma_i^2$, devido ao fator específico, é chamada de variância específica. Então, tem-se:

$$V(X_i) = \sigma_i^2 = \ell_{i1}^2 + \ell_{i2}^2 + \dots + \ell_{im}^2 + \psi_i \quad (3.26)$$

Chamando a i -ésima comunalidade de h_i^2 , tem-se que:

$$h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 + \dots + \ell_{im}^2 \quad (3.27)$$

Assim, a expressão (3.26) poderá ser escrita da seguinte forma:

$$\sigma_i^2 = h_i^2 + \psi_i, \quad i = 1, 2, \dots, p \quad (3.28)$$

Na prática, os parâmetros do modelo são desconhecidos e devem ser estimados a partir das observações amostrais. A utilização da Análise Fatorial é justificável quando Σ difere de uma matriz diagonal, ou quando ρ (matriz de correlações) difere da identidade (JOHNSON e WICHERN, 1988, p.384). Assim, fica claro que só se justifica o emprego da Análise Fatorial quando efetivamente existem relacionamentos entre as variáveis.

Embora outros métodos matemáticos de estimação dos parâmetros do modelo fatorial tenham sido desenvolvidos, como o de Spearman, Thurstone e Thomson (LAWLEY, 1940), não serão abordados neste trabalho. Serão apresentados dois métodos de estimação, sendo o primeiro o da Máxima Verossimilhança, aplicado quando é possível assumir que as variáveis envolvidas sejam normalmente distribuídas e, assim, possibilita testar a significância sobre a validade do modelo de m-fatores. O segundo é o Método das Componentes Principais.

3.3 MÉTODO DA MÁXIMA VEROSSIMILHANÇA

Este método de estimação foi desenvolvido por Lawley e publicado no artigo *“The estimation of factor loadings by the method of maximum likelihood”*, em 1940 (LAWLEY, 1940). Em 1943, ele publicou um outro artigo apresentando a aplicação do Método da Máxima Verossimilhança na Análise Fatorial (LAWLEY, 1943).

O modelo fatorial apresentado na seção 3.2, expressão (3.4), pode ser escrito na forma:

$$\underline{X} = \underline{\mu} + \underline{LF} + \underline{\varepsilon} \quad (3.29)$$

em que:

$\underline{\mu}' = [\mu_1, \mu_2, \dots, \mu_p]$ é o vetor médio de \underline{X} ;

$\underline{\varepsilon}' = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p]$ é o vetor dos fatores específicos ou erros;

$L = (\ell_{ij})$, $i=1,2,\dots,p$, $j=1,2,\dots,m$ é a matriz das cargas fatoriais;

$\underline{F}' = [F_1, F_2, \dots, F_m]$ é o vetor de fatores comuns.

Tem-se que:

$$E(\underline{X}) = E(\underline{\mu} + L\underline{F} + \underline{\varepsilon}) \quad (3.30)$$

$$E(\underline{X}) = E(\underline{\mu} + L\underline{F} + \underline{\varepsilon}) \quad (3.31)$$

$$E(\underline{X}) = \underline{\mu} \quad (3.32)$$

Da expressão (3.17) tem-se que:

$$\text{COV}(\underline{X}) = LL' + \Psi = \Sigma$$

Os estimadores não viesados de $\underline{\mu}$ e Σ são: o vetor médio \bar{X} e a matriz de covariância amostral S , conforme apresentados na seção 2.5.4.

A matriz de covariância amostral S segue distribuição de Wishart, cuja função densidade de probabilidade é dada pela expressão a seguir e que pode ser encontrada em MORRISON (1976, p. 308). Assim, tem-se:

$$L(S) = K |S|^{1/2(n-p-1)} |\Sigma|^{-1/2n} \exp\left(-\frac{1}{2}n \text{tr} \Sigma^{-1} S\right) \quad (3.33)$$

onde $K = \frac{1}{2^{1/2np} \pi^{p(p-1)/4} \prod_{i=1}^p \Gamma\left[\frac{1}{2}(n+1-i)\right]}$ é uma constante (ANDERSON, 1958). (3.34)

O logaritmo da função de verossimilhança apresentada acima é dado por:

$$L^* = \ln L(S) = \ln \left[K |S|^{1/2(n-p-1)} |\Sigma|^{-1/2n} \exp\left(-\frac{1}{2} n \text{tr} \Sigma^{-1} S\right) \right] \quad (3.35)$$

$$L^* = \ln L(S) = \ln K + \frac{1}{2}(n-p-1) \ln |S| - \frac{1}{2} n \ln |\Sigma| - \frac{1}{2} n \text{tr} \Sigma^{-1} S \quad (3.36)$$

Substituindo Σ por $LL' + \psi$ e omitindo as funções das observações, tem-se:

$$L^* = -\frac{1}{2} n \ln |LL' + \psi| - \frac{1}{2} n \text{tr} (LL' + \psi)^{-1} S \quad (3.37)$$

Diferenciando L^* em relação a L e efetuando algumas operações algébricas, chega-se à expressão apresentada a seguir. Pode-se obter os detalhes da diferenciação em MORRISON (1976, p. 310):

$$(\Sigma^{-1} - \Sigma^{-1} S \Sigma^{-1}) L = 0 \quad (3.38)$$

$$\Sigma^{-1} (L - S \Sigma^{-1} L) = 0 \quad (3.39)$$

E, para satisfazer à igualdade tem-se que:

$$L - S \Sigma^{-1} L = 0 \quad (3.40)$$

$$\text{logo, } S \Sigma^{-1} L = L \quad (3.41)$$

Pós-multiplicando ambos os membros da igualdade da expressão $\Sigma = LL' + \psi$ por $\psi^{-1} L$, tem-se:

$$\Sigma \psi^{-1} L = (LL' + \psi) \psi^{-1} L \quad (3.42)$$

$$\psi^{-1} L (L' \psi^{-1} L + I)^{-1} = \Sigma^{-1} L \quad (3.43)$$

e substituindo a expressão (3.43) em (3.41), resulta:

$$S \psi^{-1} L (L' \psi^{-1} L + I)^{-1} = L \quad (3.44)$$

$$S \psi^{-1} L = L (L' \psi^{-1} L + I) \quad (3.45)$$

Da mesma forma, diferenciando L^* em relação a ψ e efetuando algumas operações algébricas chega-se à expressão a seguir. Pode-se obter os detalhes da diferenciação em MORRISON (1976, p.308):

$$\text{diag} \left\{ [LL' + \psi]^{-1} (I - S[LL' + \psi]^{-1}) \right\} \quad (3.46)$$

$$\text{diag} \left\{ [LL' + \psi]^{-1} - [LL' + \psi]^{-1} S [LL' + \psi]^{-1} \right\} = 0 \quad (3.47)$$

$$\text{diag} [\Sigma]^{-1} = \text{diag} \left\{ \Sigma^{-1} S \Sigma^{-1} \right\} \quad (3.48)$$

Pré e pós-multiplicando ambos os membros da igualdade pela matriz diagonal $\Sigma - LL'$, tem-se:

$$\text{diag} [\Sigma - 2LL' + LL'\Sigma^{-1}LL'] = \text{diag} [S - LL'\Sigma^{-1}S - S\Sigma^{-1}LL' + LL'\Sigma^{-1}S\Sigma^{-1}LL'] \quad (3.49)$$

E, ainda, utilizando a expressão (3.41), resulta:

$$\text{diag} [\Sigma] = \text{diag} [S] \quad (3.50)$$

$$\text{diag} [LL' + \psi] = \text{diag} [S] \quad (3.51)$$

Assim, as equações de máxima verossimilhança são dadas por:

$$S\psi^{-1}L = L(L'\psi^{-1}L + I) \quad (3.52)$$

$$\text{diag} [LL' + \psi] = \text{diag} [S] \quad (3.53)$$

As equações anteriores podem ser resolvidas apenas por processos iterativos e com dados amostrais obtêm-se \hat{L} e $\hat{\psi}$.

Os estimadores de máxima verossimilhança das comunalidades são apresentados em JOHNSON e WICHERN (1988, p.393):

$$\hat{h}_i^2 = \hat{l}_{i1}^2 + \hat{l}_{i2}^2 + \dots + \hat{l}_{im}^2, \quad i = 1, 2, \dots, p \quad (3.54)$$

A proporção explicada pelo j-ésimo fator é obtida através das seguintes expressões:

$$\text{Var Explicada} = \frac{\sum_{i=1}^p \hat{\ell}_{ij}^2}{\text{tr}(S)} \quad (3.55)$$

para fatores estimados a partir da matriz de covariância amostral S .

$$\text{Var Explicada} = \frac{\sum_{i=1}^p \hat{\ell}_{z(ij)}^2}{p} \quad (3.56)$$

para fatores estimados a partir da matriz de correlação amostral R .

3.3.1 Teste para o Número de Fatores Comuns em Grandes Amostras

A suposição da normalidade da população conduz ao teste de ajuste do modelo. Suponha que o modelo ajustado tenha m fatores comuns. Neste caso $\Sigma = LL' + \psi$, e testar a adequação do modelo de m fatores comuns é equivalente a testar as hipóteses:

$$H_0 : \Sigma = LL' + \psi \quad (3.57)$$

$$H_1 : \Sigma = \text{qualquer outra matriz definida positiva}$$

Sob a hipótese H_0 , a estimativa de verossimilhança de Σ é $\hat{\Sigma}$, que pode ser obtida resolvendo-se as equações apresentadas em (3.52) e (3.53). A estatística da razão de verossimilhança é dada por:

$$\lambda(\underline{x}) = \frac{L_{\hat{\Sigma}}}{L_{S_n}} \quad (3.58)$$

A simplificação do cálculo exige a utilização da estatística $-2 \ln \lambda(\underline{x})$, cuja distribuição é conhecida, ou seja, é aproximadamente χ_r^2 , quando n é moderadamente grande (BARTLETT, 1950). Assim, tem-se que:

$$-2 \ln \lambda(\underline{x}) = -2 \ln \left[\frac{L_{\hat{\Sigma}}}{L_{S_n}} \right] \quad (3.59)$$

$$-2 \ln \lambda(\underline{x}) = n \left[\ln |\hat{\Sigma}| + \text{tr}(\hat{\Sigma}^{-1} S_n) - \ln |S_n| - p \right] \quad (3.60)$$

Se $\hat{\Sigma}$ é obtida com bastante precisão, tem-se que $\text{tr}(\hat{\Sigma}^{-1} S_n) = \text{tr}(I_p) = p$, pois, $S_n = ((n-1)/n)S = \hat{\Sigma}$. Assim,

$$-2 \ln \lambda(\underline{x}) = n \left[\ln |\hat{\Sigma}| - \ln |S_n| \right] \quad (3.61)$$

$$-2 \ln \lambda(\underline{x}) = n \ln \frac{|\hat{\Sigma}|}{|S_n|} \quad (3.62)$$

No entanto, se o algoritmo para obter $\hat{\Sigma}$ não convergir rapidamente para uma solução precisa, necessita-se da expressão (3.60).

BARTLETT (1950, p.82) propôs uma modificação na expressão (3.62), para que a sua convergência à distribuição limite χ^2 fosse mais rápida. Ele sugeriu que, ao invés de n , fosse utilizado o fator multiplicativo n' , dado por:

$$n' = n - 1 - \frac{1}{6}(2p + 4m + 5) \quad (3.63)$$

em que:

n é o número de observações da amostra;

p é o número de variáveis;

m é o número de fatores.

Assim, a estatística a ser usada para testar H_0 é:

$$U_m = \left[n - 1 - \frac{1}{6}(2p + 4m + 5) \right] \left[\ln |\hat{\Sigma}| - \ln |S_n| \right] \quad (3.64)$$

$$\text{sendo } U_m \sim \chi_r^2 \text{ com } r = \frac{1}{2} \{ (p-m)^2 - p - m \} \quad (3.65)$$

Dessa forma, rejeita-se a hipótese nula, ao nível de significância α , se $U_m > \chi_{\alpha,r}^2$.

Como o número de graus de liberdade precisa ser positivo, igualando a expressão $\frac{1}{2}[(p-m)^2 - p - m]$ a zero e desenvolvendo, tem-se:

$$\frac{1}{2}[p^2 - 2mp + m^2 - p - m] = 0 \quad (3.66)$$

$$\frac{1}{2}p^2 - \frac{1}{2}p - mp + \frac{m^2}{2} - \frac{m}{2} = 0 \quad (3.67)$$

Agrupando os termos semelhantes tem-se:

$$\frac{m^2}{2} - \left(\frac{1}{2} + p\right)m + \left(\frac{1}{2}p^2 - \frac{1}{2}p\right) = 0 \quad (3.68)$$

Resolvendo a equação de 2º grau em m:

$$m = \frac{1}{2} + p \pm \sqrt{\frac{8p+1}{4}} \quad (3.69)$$

$$m = \frac{1}{2}[(2p+1) \pm \sqrt{8p+1}] \quad (3.70)$$

Mas como o número de graus de liberdade precisa ser positivo, então:

$$m < \frac{1}{2}[(2p+1) - \sqrt{8p+1}] \quad (3.71)$$

3.4 MÉTODO DAS COMPONENTES PRINCIPAIS

O Método da Análise de Componentes Principais foi desenvolvido inicialmente por Karl Pearson, em 1901, e posteriormente por Hotelling, em 1933 (FACHEL, 1976).

Inicialmente, será abordado o Método das Componentes Principais para população e, posteriormente, para amostra.

3.4.1 Análise Fatorial para População

Dado o vetor aleatório $\underline{X}' = [X_1, X_2, \dots, X_p]$, com vetor médio $\underline{\mu}$ e matriz de covariância Σ , que tem os pares de autovalor-autovetor $(\lambda_1, \underline{e}_1)$, $(\lambda_2, \underline{e}_2)$, ..., $(\lambda_p, \underline{e}_p)$ com $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$. Tem-se então, pelo teorema da decomposição espectral (ver seção 2.3), que:

$$\Sigma = \lambda_1 \underline{e}_1 \underline{e}'_1 + \lambda_2 \underline{e}_2 \underline{e}'_2 + \dots + \lambda_p \underline{e}_p \underline{e}'_p$$

$$\Sigma = \left[\begin{array}{cccc} \sqrt{\lambda_1} \underline{e}_1 & & & \\ & \sqrt{\lambda_2} \underline{e}_2 & & \\ & & \dots & \\ & & & \sqrt{\lambda_p} \underline{e}_p \end{array} \right] \begin{bmatrix} \sqrt{\lambda_1} \underline{e}'_1 \\ \sqrt{\lambda_2} \underline{e}'_2 \\ \vdots \\ \sqrt{\lambda_p} \underline{e}'_p \end{bmatrix} = LL' \quad (3.72)$$

em que:

L é uma matriz $p \times p$, de cargas fatoriais.

A expressão (3.72) reflete um ajuste da estrutura de covariância através do modelo fatorial em que o número de fatores é igual ao de variáveis ($m = p$) e variâncias específicas ψ_i nulas, para todo $i = 1, 2, \dots, p$. Este modelo não é útil, pois o número de fatores é igual ao número de variáveis e a variação para os fatores específicos é nula.

É preferível um modelo que explique a estrutura de covariâncias através de poucos fatores comuns. Neste caso $(p - m)$, autovalores menores do que 1 (um) e os respectivos autovetores são desconsiderados. Assim, a contribuição de $\lambda_{m+1} \underline{e}_{m+1} \underline{e}'_{m+1} + \lambda_{m+2} \underline{e}_{m+2} \underline{e}'_{m+2} + \dots + \lambda_p \underline{e}_p \underline{e}'_p$ é descartada e é possível obter a seguinte aproximação de Σ :

$$\Sigma \cong \left[\begin{array}{cccc} \sqrt{\lambda_1} \underline{e}_1 & & & \\ & \sqrt{\lambda_2} \underline{e}_2 & & \\ & & \dots & \\ & & & \sqrt{\lambda_m} \underline{e}_m \end{array} \right] \begin{bmatrix} \sqrt{\lambda_1} \underline{e}'_1 \\ \sqrt{\lambda_2} \underline{e}'_2 \\ \dots \\ \sqrt{\lambda_m} \underline{e}'_m \end{bmatrix} = LL' \quad (3.73)$$

onde L é uma matriz $p \times m$. A expressão (3.73) não considera a contribuição dos fatores específicos, que pode ser estimada tomando-se a diagonal de $\Sigma - LL'$, sendo LL' definida em (3.73).

Da expressão (3.17) tem-se que $\Sigma = LL' + \Psi$. Então:

$$\Sigma = \left[\sqrt{\lambda_1} \underline{e}_1 \quad \sqrt{\lambda_2} \underline{e}_2 \quad \dots \quad \sqrt{\lambda_m} \underline{e}_m \right] \begin{bmatrix} \sqrt{\lambda_1} \underline{e}'_1 \\ \sqrt{\lambda_2} \underline{e}'_2 \\ \dots \\ \sqrt{\lambda_m} \underline{e}'_m \end{bmatrix} + \begin{bmatrix} \Psi_1 & 0 & \dots & 0 \\ 0 & \Psi_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \Psi_p \end{bmatrix} \quad (3.74)$$

onde $\psi_i = \sigma_i^2 - \sum_{j=1}^m \ell_{ij}^2$, da expressão (3.26).

É habitual trabalhar com as variáveis em uma escala padronizada, ou seja, fazendo a transformação, subtraindo suas respectivas médias e dividindo-as pelos seus desvios padrão. A padronização evita que uma variável com variação grande influencie indevidamente a obtenção das cargas fatoriais.

Sejam as variáveis padronizadas Z_i , com média 0 e variância 1. A padronização pode ser feita como se segue:

$$\underline{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \dots \\ Z_p \end{bmatrix} = V^{-1/2} (\underline{X} - \underline{\mu}) = \begin{bmatrix} \frac{X_1 - \mu_1}{\sqrt{\sigma_1^2}} \\ \frac{X_2 - \mu_2}{\sqrt{\sigma_2^2}} \\ \dots \\ \frac{X_p - \mu_p}{\sqrt{\sigma_p^2}} \end{bmatrix} \quad (3.75)$$

A matriz de covariância de \underline{Z} é a matriz de correlação ρ , e o processo de estimação dos parâmetros do modelo fatorial é o mesmo descrito anteriormente, substituindo Σ por ρ . A expressão $\psi_i = \sigma_i^2 - \sum_{j=1}^m \ell_{ij}^2$ ficará igual a $\psi_i = 1 - \sum_{j=1}^m \ell_{ij}^2$, uma vez que $Z_i \sim N(0,1)$.

3.4.2 Análise Fatorial para Amostra

A Análise Fatorial por Componentes Principais obtida a partir da matriz de covariância amostral S é definida em função dos pares de autovalores e autovetores estimados $(\hat{\lambda}_i, \hat{e}_i)$, $i = 1, 2, \dots, p$, em que $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_p$.

A matriz das cargas fatoriais estimadas $(\hat{\ell}_{ij})$ é dada pela expressão a seguir, quando se tem $m < p$:

$$\hat{L} = \left[\sqrt{\hat{\lambda}_1} \hat{e}_1 \sqrt{\hat{\lambda}_2} \hat{e}_2 \dots \sqrt{\hat{\lambda}_m} \hat{e}_m \right] \quad (3.76)$$

As variâncias específicas estimadas são fornecidas pelos elementos da diagonal da matriz $S - \hat{L}\hat{L}'$. Então, tem-se:

$$\tilde{\Psi} = \begin{bmatrix} \tilde{\psi}_1 & 0 & \dots & 0 \\ 0 & \tilde{\psi}_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \tilde{\psi}_p \end{bmatrix}, \text{ com } \tilde{\psi}_i = S_i^2 - \sum_{j=1}^m \hat{\ell}_{ij}^2 \quad (3.77)$$

E, as comunalidades estimadas são:

$$\hat{h}_i^2 = \hat{\ell}_{i1}^2 + \hat{\ell}_{i2}^2 + \dots + \hat{\ell}_{im}^2 \quad (3.78)$$

Quando se utiliza a matriz de correlação amostral, basta substituir S por R , a matriz de correlação estimada, nas expressões apresentadas.

3.5 NÚMERO DE FATORES

A definição do número de fatores é uma decisão importante na aplicação da Análise Fatorial. São apresentados adiante os critérios mais comuns para definir o número de fatores.

3.5.1 Número de Fatores Definido com Base no Grau de Explicação dos Autovalores Estimados

A escolha de m fatores pode ser baseada nos autovalores estimados. Escolhe-se o número de fatores que explicam uma proporção especificada da variância total (JOHNSON e WICHERN, 1998, p.387).

A proporção da variação total devido a j -ésimo fator é dada por:

$$\text{Var Explicada} = \frac{\hat{\lambda}_j}{\text{tr}(S)} \quad (3.79)$$

para fatores estimados a partir da matriz de covariância amostral S .

$$\text{Var Explicada} = \frac{\hat{\lambda}_j}{p} \quad (3.80)$$

para fatores estimados a partir da matriz de correlação amostral R .

O número de fatores comuns extraído do modelo deve aumentar até que uma proporção adequada da variação total amostral seja explicada.

3.5.2 Número de Fatores Definido com Base no Critério de Kaiser

Quando é usada a matriz de correlação, retêm-se apenas os autovalores da matriz que são maiores que a unidade. Este critério é devido a Kaiser (KAISER, 1960;1970).

3.6 ROTAÇÃO DOS FATORES

A etapa posterior à da obtenção das cargas fatoriais é a sua interpretação delas. Segundo FACHEL (1976), para uma melhor interpretação dos fatores é comum fazer uma rotação ou uma transformação neles. O objetivo da rotação dos

fatores é obter uma matriz de cargas mais fácil de ser interpretada ou que mais se identifique com a natureza das variáveis analisadas.

Diversos métodos de rotação e transformação dos fatores foram propostos. Apresenta-se, aqui, o método analítico mais utilizado. Trata-se do Critério *Varimax* e do *Varimax* Normal.

O Critério *Varimax* foi proposto por Kaiser em 1958 (KAISER, 1958). O critério dá maior ênfase à simplificação das colunas (correspondentes aos fatores) da matriz de cargas, de forma a ter uma estrutura mais simples. Assim, Kaiser definiu a simplificação de uma coluna, correspondente a um fator j , como a variância de suas cargas ao quadrado, como apresentado adiante, onde p é dimensão do vetor observado.

$$V_j = \left[p \sum_{i=1}^p (\ell_{ij}^2)^2 - \left(\sum_{i=1}^p \ell_{ij}^2 \right)^2 \right] / p^2 \quad (3.81)$$

O critério de máxima simplicidade de uma matriz fatorial completa é definido como a maximização da expressão a seguir.

$$V = \sum_{j=1}^m V_j = \sum_{j=1}^m \left\{ \left[p \sum_{i=1}^p (\ell_{ij}^2)^2 - \left(\sum_{i=1}^p \ell_{ij}^2 \right)^2 \right] / p^2 \right\} \quad (3.82)$$

em que se tem:

$i = 1, 2, \dots, p$ (variáveis);

$j = 1, 2, \dots, m$ (fatores);

ℓ_{ij} é o peso (carga) do j -ésimo fator na i -ésima variável.

Esse critério atribui pesos iguais às variáveis com comunalidades grandes ou pequenas. Assim, Kaiser sugeriu que, antes de iniciar o processo de maximização, as cargas sejam padronizadas. Desse modo, tem-se a expressão que deve ser maximizada:

$$V = \sum_{j=1}^m \left\{ \left[p \sum_{i=1}^p x_{ij}^4 - \left(\sum_{i=1}^p x_{ij}^2 \right)^2 \right] / p^2 \right\} \quad (3.83)$$

em que $x_{ij} = \frac{\ell_{ij}}{\sqrt{\sum_{j=1}^m \ell_{ij}^2}} = \frac{\ell_{ij}}{\sqrt{h_i^2}}$ é a j-ésima carga fatorial na i-ésima variável resposta

dividida pela raiz quadrada de sua comunalidade.

Após a rotação, os valores de x_{ij} devem ser multiplicados pela raiz quadrada de sua comunalidade respectiva, para que a escala original seja restaurada.

3.7 ESCORES FATORIAIS

Os fatores são variáveis latentes, ou seja, não observáveis. Porém, existem métodos de estimação indireta. São propostos dois métodos de estimação indireta, os quais são descritos adiante.

3.7.1 Método dos Mínimos Quadrados Ponderados

Este método foi desenvolvido por Bartlett, adotando o princípio de mínimos quadrados. Tendo em vista que $V(\varepsilon_i) = \psi_i$ não é necessariamente igual para todo i , ele sugeriu o uso dos mínimos quadrados ponderados, tendo como peso o inverso das variâncias específicas (JOHNSON e WICHERN, 1988, p.140).

Os escores fatoriais são obtidos de forma que a soma de quadrados dos resíduos ponderados seja mínima, em relação aos elementos de F . Assim,

$$\sum_{i=1}^p \frac{\varepsilon_i^2}{\psi_i} = \underline{\varepsilon}' \underline{\psi}^{-1} \underline{\varepsilon} = (\underline{X} - \underline{\mu} - \underline{L}\underline{F})' \underline{\psi}^{-1} (\underline{X} - \underline{\mu} - \underline{L}\underline{F}) \quad (3.84)$$

Derivando em relação a F e igualando a zero, obtém-se:

$$-2 \left\{ \underline{L}' \underline{\psi}^{-1} (\underline{X} - \underline{\mu} - \underline{L}\underline{F}') \right\} = 0 \quad (3.85)$$

$$\underline{L}' \underline{\psi}^{-1} (\underline{X} - \underline{\mu}) = \underline{\psi}^{-1} \underline{L}\underline{F}\underline{L}' \quad (3.86)$$

$$\hat{\underline{F}} = (\underline{L}' \underline{\psi}^{-1} \underline{L})^{-1} \underline{L}' \underline{\psi}^{-1} (\underline{X} - \underline{\mu}) \quad (3.87)$$

O que ocorre de fato é que $\underline{\mu}$, L e ψ são desconhecidos, portanto deve-se utilizar as respectivas estimativas. Assim, tem-se:

$$\hat{\underline{E}}_j = (\hat{L}'\hat{\psi}^{-1}\hat{L})^{-1}\hat{L}'\hat{\psi}^{-1}(\underline{X}_j - \bar{\underline{X}}) \quad , \quad j = 1, 2, \dots, n \quad (3.88)$$

E, se for utilizada a matriz de correlação, a expressão (3.88) tornar-se-á:

$$\hat{\underline{E}}_j = (\hat{L}'_z\hat{\psi}_z^{-1}\hat{L}_z)^{-1}\hat{L}'_z\hat{\psi}_z^{-1}\underline{Z}_j \quad , \quad j = 1, 2, \dots, n \quad (3.89)$$

Quando se utilizam as cargas fatoriais que sofreram rotação, $\hat{L}^* = \hat{L}T$, tem-se:

$$\hat{\underline{E}}_j^* = T'\hat{\underline{E}}_j \quad , \quad j = 1, 2, \dots, n \quad (3.90)$$

3.7.2 Método da Regressão

Este método foi desenvolvido por G. H. Thomson (LAWLEY e MAXWELL, 1962) e é também chamado de método de Thomson.

Seja $\underline{X}' = [X_1, X_2, \dots, X_p]$ o vetor das observações, $\underline{f}' = (f_1, f_2, \dots, f_m)$ o vetor dos escores fatoriais e $\underline{\varepsilon}' = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p)$ o vetor dos erros. As suposições do modelo fatorial ortogonal, apresentadas na seção 3.2, são:

$$E(\underline{F}\underline{F}') = I \quad (3.91)$$

$$E(\underline{X}\underline{F}') = L \quad (3.92)$$

$$E(\underline{X}\underline{X}') = \Sigma = LL' + \psi \quad (3.93)$$

O método de Thomson, baseado na regressão de \underline{f} sobre \underline{X} , é equivalente a encontrar para cada j , $j = 1, 2, \dots, m$, uma função linear das observações, que dará um bom preditor de f_j , dado pela expressão a seguir (FACHEL, 1976):

$$\hat{f}_j = \underline{a}'_j \underline{X} = \underline{X}' \underline{a}_j \quad (3.94)$$

onde \underline{a}_j é um vetor de ordem p , escolhido de tal forma que a variância de $(\hat{f}_j - f_j)$ é mínima. Tem-se que:

$$V(\hat{f}_j - f_j) = E(\underline{X}'\underline{a}_j - f_j)^2 \quad (3.95)$$

$$V(\hat{f}_j - f_j) = E(\underline{X}'\underline{a}_j - f_j)(\underline{X}'\underline{a}_j - f_j)' \quad \text{ou} \quad (3.96)$$

$$V(\hat{f}_j - f_j) = E(\underline{a}_j'\underline{X} - f_j)(\underline{a}_j'\underline{X} - f_j)' \quad (3.97)$$

Derivando a expressão acima em relação a \underline{a}_j' e igualando a zero, tem-se:

$$2(\underline{a}_j'\underline{\Sigma} - \underline{L}_j) = 0 \quad (3.98)$$

$$\underline{a}_j'\underline{\Sigma} = \underline{L}_j \quad (3.99)$$

E, tem-se que $\hat{f}_j = \underline{X}'\underline{a}_j$,

$$\text{logo } \underline{a}_j = \frac{\hat{f}_j}{\underline{X}'} \quad (3.100)$$

Substituindo a expressão (3.100) na (3.99), obtém-se:

$$\underline{\hat{f}} = \underline{L}'\underline{\Sigma}^{-1}\underline{X} \quad (3.101)$$

Dado que $\underline{\Sigma} = \underline{L}\underline{L}' + \underline{\Psi}$ e pré-multiplicando ambos os membros da igualdade por $\underline{L}'\underline{\Psi}^{-1}$ resulta:

$$\underline{L}'\underline{\Psi}^{-1}\underline{\Sigma} = \underline{L}'\underline{\Psi}^{-1}[\underline{L}\underline{L}' + \underline{\Psi}] \quad (3.102)$$

$$\underline{L}'\underline{\Sigma}^{-1} = [\underline{L}\underline{\Psi}^{-1}\underline{L}' + \underline{I}]^{-1}\underline{L}'\underline{\Psi}^{-1} \quad (3.103)$$

Agora, substituindo a expressão (3.103) na (3.101), tem-se o resultado final:

$$\underline{\hat{f}} = [\underline{L}\underline{\Psi}^{-1}\underline{L}' + \underline{I}]^{-1}\underline{L}'\underline{\Psi}^{-1}\underline{X} \quad (3.104)$$

3.8 SIGNIFICÂNCIA ESTATÍSTICA DA MATRIZ DE CORRELAÇÃO

Sendo o objetivo da Análise Fatorial a modelagem do relacionamento existente no conjunto de variáveis, ou seja, no vetor observado, e a redução da sua dimensão inicial, através de fatores, faz-se necessário verificar a adequação do grau de correlação, isto é, a significância da relação entre as variáveis.

A significância estatística da estrutura da matriz de correlação pode ser verificada através do Teste de Esfericidade de Bartlett. Outra estatística para avaliar a adequação da aplicação da Análise Fatorial é a medida de adequabilidade da amostra (MSA), de Kaiser-Meyer-Olkin.

3.8.1 Teste de Esfericidade de Bartlett

O Teste de Esfericidade de Bartlett testa a hipótese nula de que a matriz de correlação da população é uma matriz identidade, o que indica que as variáveis não são correlacionadas e o modelo fatorial é inadequado.

A estatística do teste segue a distribuição χ^2 com $\nu = \frac{1}{2}[p(p-1)]$ graus de liberdade e foi proposta por Bartlett, em 1950 (BARTLETT, 1950). Assim, tem-se:

$$\chi^2 = - \left[(n-1) - \frac{1}{6}(2p+5) \right] \ln |R| \sim \chi^2_{\nu} \quad (3.105)$$

em que:

n é o tamanho da amostra;

p é o número de variáveis;

|R| é o determinante da matriz de correlação.

3.8.2 Medida de Adequabilidade da Amostra de Kaiser-Meyer-Olkin

A medida de adequabilidade da amostra de Kaiser-Meyer-Olkin compara os valores dos coeficientes de correlação simples com os dos coeficientes de correlação parcial. Esta medida é um índice que varia de 0 a 1 e é calculada por:

$$MSA = \frac{\sum_{i \neq j} \sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} \sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} \sum_{i \neq j} q_{ij}^2} \quad (3.106)$$

em que:

r_{ij} é o coeficiente de correlação entre as variáveis X_i e X_j ;

q_{ij} é o elemento fora da diagonal da matriz antiimagem de correlação (que corresponde ao coeficiente de correlação parcial entre as variáveis X_i e X_j , com o sinal invertido).

Conforme KAISER (1970), a matriz antiimagem de correlação é obtida por:

$$Q = SR^{-1}S, \quad (3.107)$$

em que:

$$S = (\text{diag } R^{-1})^{-1/2};$$

R é a matriz de correlação.

Segundo HAIR et al. (1998) e MARQUES (2006), para valores de MSA entre 0,5 e 1 a Análise Fatorial é adequada; no entanto, se abaixo de 0,5, a análise fatorial não é apropriada.

4 MATERIAL E MÉTODO

Definiu-se o número de variáveis entre 5 e 20 e em 81 o número de populações normais multivariadas. Foram inicialmente geradas as amostras de variáveis normais bivariadas com 100.000 observações e, utilizando-se dos seus vetores de médias e matrizes de covariâncias, foram geradas as amostras de variáveis normais multivariadas, também compostas pelo mesmo número de observações (100.000), as quais foram consideradas como populações (universos). Dessas populações, foram retiradas 1.000 amostras aleatórias de diferentes tamanhos. Para o dimensionamento das amostras, adotou-se a expressão (2.123), que tem por finalidade a estimação do vetor médio populacional. O nível de confiança utilizado foi de 95% e os erros amostrais relativos foram fixados em 5%, 10% e 15%.

Todos os procedimentos de cálculo, neste trabalho, foram feitos utilizando-se o Sistema Computacional R, cujos *scripts* estão disponíveis no apêndice 3.

4.1 MÉTODO DE DETERMINAÇÃO DA POPULAÇÃO

Tendo em vista os objetivos do trabalho, optou-se pela utilização do Método de Simulação de Monte Carlo, para a obtenção das 81 populações (universos). De acordo com CLIFF e HAMBURGER (1967), em áreas da Estatística onde as questões de interesse prático têm sido tão complexas e tão difíceis de se especificar soluções analíticas matematicamente, é comum o uso do Método de Simulação Monte Carlo, em que amostras de alguma população específica são geradas por algum processo aleatório.

Tendo-se definido inicialmente o número de variáveis entre 5 e 20, geraram-se as variáveis normais bivariadas. Ao gerar estas variáveis, deve-se fixar o coeficiente de correlação para cada par de variáveis, bem como os outros parâmetros. O procedimento adotado para a geração destes valores é apresentado no resultado 4.1.

Resultado 4.1

Sejam U e V variáveis aleatórias independentes com distribuição normal padrão, ou seja, $U \sim N(0, 1)$ e $V \sim N(0, 1)$, e os seguintes parâmetros predefinidos: μ_1 , μ_2 , σ_1^2 , σ_2^2 e ρ . As variáveis aleatórias X_1 e X_2 são expressas por:

$$X_1 = \mu_1 + \sqrt{\sigma_1^2} U \quad (4.1)$$

$$X_2 = \mu_2 + \rho \left(\frac{\sqrt{\sigma_2^2}}{\sqrt{\sigma_1^2}} \right) (X_1 - \mu_1) + \sqrt{\sigma_2^2(1-\rho^2)} V \quad (4.2)$$

em que μ_1 e $\mu_2 \in \mathbb{R}$; σ_1^2 e $\sigma_2^2 \in \mathbb{R}^+$; $-1 \leq \rho \leq 1$

A distribuição conjunta de (X_1, X_2) tem distribuição normal bivariada com parâmetros μ_1 , μ_2 , σ_1^2 , σ_2^2 e ρ .

Prova:

Tem-se que $U \sim N(0, 1)$ e $V \sim N(0, 1)$ são independentes; logo, a distribuição conjunta é dada da seguinte forma:

$$g(u, v) = g(u)g(v) \quad (4.3)$$

$$g(u, v) = \left\{ \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}u^2\right] \right\} \left\{ \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}v^2\right] \right\} \quad (4.4)$$

que poderá ser escrita como:

$$g(u, v) = \left(\frac{1}{\sqrt{2\pi}} \right)^2 \exp\left\{-\frac{1}{2}(u^2 + v^2)\right\} \quad (4.5)$$

$$\text{Tem-se ainda que: } X_1 = \mu_1 + \sqrt{\sigma_1^2} U, \text{ logo } U = \frac{X_1 - \mu_1}{\sigma_1} \quad (4.6)$$

$$e \quad X_2 = \mu_2 + \rho \left(\frac{\sqrt{\sigma_2^2}}{\sqrt{\sigma_1^2}} \right) (X_1 - \mu_1) + \sqrt{\sigma_2^2 (1 - \rho^2)} V \quad (4.7)$$

$$\text{Portanto, } V = \frac{(X_2 - \mu_2) - \rho \frac{\sigma_2}{\sigma_1} (X_1 - \mu_1)}{\sigma_2 \sqrt{1 - \rho^2}} \quad (4.8)$$

A densidade conjunta de (X_1, X_2) é dada por:

$$f(x_1, x_2) = f(g(u, v)) J \quad (4.9)$$

Em que J é o jacobiano de transformação de U, V para X_1, X_2 e é definido como o valor absoluto de $|J|$.

Tem-se, então, que o jacobiano de transformação é igual a:

$$J = \begin{vmatrix} \frac{1}{\sigma_1} & 0 \\ \rho & 1 \end{vmatrix} = \frac{1}{\sigma_1 \sigma_2 \sqrt{1 - \rho^2}} \quad (4.10)$$

Portanto:

$$f(x_1, x_2) = f\left(\left(\frac{1}{\sqrt{2\pi}}\right)^2 \exp\left\{-\frac{1}{2}(u^2 + v^2)\right\}\right) J \quad (4.11)$$

Fazendo as substituições de u, v e J , tem-se:

$$f(x_1, x_2) = \frac{1}{\sigma_1 \sigma_2 \sqrt{1 - \rho^2}} \left(\frac{1}{2\pi} \right) \exp \left\{ -\frac{1}{2} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{(x_2 - \mu_2) - \rho \frac{\sigma_2}{\sigma_1} (x_1 - \mu_1)}{\sigma_2 \sqrt{1 - \rho^2}} \right)^2 \right] \right\} \quad (4.12)$$

Desenvolvendo a expressão anterior (4.12), é possível obter:

$$f(x_1, x_2) = \frac{1}{\sigma_1 \sigma_2 \sqrt{1-\rho^2}} \left(\frac{1}{2\pi} \right) \exp \left\{ -\frac{1}{2} \left(\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \frac{(x_2 - \mu_2)^2}{\sigma_2^2 (1-\rho^2)} - \frac{2\rho(x_2 - \mu_2)(x_1 - \mu_1)}{\sigma_1 \sigma_2 \sqrt{1-\rho^2}} + \frac{\rho^2(x_1 - \mu_1)^2}{\sigma_1^2 (1-\rho^2)} \right) \right\} \quad (4.13)$$

E, agrupando os termos semelhantes, tem-se:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left(\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right) \right\} \quad (4.14)$$

que é a função densidade de probabilidade da distribuição normal bivariada.

Tendo-se definido o número de variáveis entre 5 e 20, geraram-se tantas variáveis normais bivariadas quanto necessárias. Obteve-se, então, o vetor de médias ($\underline{\mu}$) e a matriz de covariância (Σ), que são utilizados como parâmetros para gerar as variáveis normais multivariadas, através da função³ *mvrnorm*, disponível no Sistema Computacional R. Esta função é baseada na decomposição espectral, descrita a seguir.

Tem-se, da decomposição espectral, que $\Sigma = \Gamma\Lambda\Gamma'$, onde Σ é a matriz de covariância, Λ é a matriz diagonal de autovalores e Γ é a matriz ortogonal cujas colunas são os autovetores padronizados de Σ .

Uma vez que a matriz Σ é simétrica e definida positiva, é possível decompor em $\Sigma = LL'$, onde L é a matriz raiz quadrada de Σ . Seja então $\Lambda^{1/2}$ a matriz diagonal em que $\sqrt{\lambda_i}$ é o i -ésimo elemento, sendo λ_i o i -ésimo autovalor, e tem-se que $L = \Gamma\Lambda^{1/2}\Gamma'$, da definição de matriz raiz quadrada. A demonstração poderá ser obtida no trabalho de STEINER (2000).

Finalmente, gera-se p vetores $N(0,1)$ formando $\underline{Z} \sim N(\underline{0}, I)$, e fazendo-se a transformação $\underline{X} = L\underline{Z} + \underline{\mu}$ têm-se as variáveis normais multivariadas. Uma vez que $\underline{Z} \sim N(\underline{0}, I)$, pelo resultado 2.3 da seção 2.5.2, qualquer combinação linear de variáveis normais multivariadas tem também a mesma distribuição. Foram, então, gerados vetores $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$.

³ A função *mvrnorm* é disponível na biblioteca MASS do Sistema R. A referência citada é: RIPLEY, B. D. **Stochastic Simulation**. New York: John Wiley & Sons, 1987.

4.1.1 População Normal Multivariada com 5 Variáveis

Descrevem-se, a seguir, as etapas para determinação da população normal multivariada, composta de 5 variáveis.

Sejam, então, os parâmetros definidos para população 1, que se encontram no apêndice 1. O vetor médio, a matriz de covariância e as correlações entre as variáveis foram assim definidos:

$$\underline{\mu}'_i = [70 \quad 75 \quad 80 \quad 20 \quad 10] \quad , \quad i = 1, 2, \dots, 5 \quad (4.15)$$

$$\underline{\sigma}^2'_i = [80 \quad 69 \quad 90 \quad 35 \quad 6] \quad , \quad i = 1, 2, \dots, 5 \quad (4.16)$$

Correlações: $\rho_{X_1, X_2} = 0,90$; $\rho_{X_1, X_3} = 0,30$; $\rho_{X_4, X_2} = 0,95$;

$$\rho_{X_5, X_3} = 0,70 \quad (4.17)$$

Então, seguiram-se as etapas abaixo:

- (i) Gerou-se a variável X_1 substituindo-se os valores correspondentes aos parâmetros μ_1 e σ_1^2 , dos vetores de médias e variâncias, apresentados em (4.15) e (4.16), na expressão abaixo:

$$X_1 = \mu_1 + \sqrt{\sigma_1^2} U \quad \text{onde } U \sim N(0, 1); \quad (4.18)$$

- (ii) para gerar a variável X_2 , que é correlacionada com a variável X_1 , foram substituídos os valores correspondentes aos parâmetros μ_1, σ_1^2 , μ_2 , σ_2^2 e ρ_{X_1, X_2} , apresentados em (4.15), (4.16) e (4.17) e X_1 , obtida em (i), na seguinte expressão:

$$X_2 = \mu_2 + \rho_{X_1, X_2} \left(\frac{\sqrt{\sigma_2^2}}{\sqrt{\sigma_1^2}} \right) (X_1 - \mu_1) + \sqrt{\sigma_2^2 (1 - \rho_{X_1, X_2}^2)} V \quad (4.19)$$

onde $V \sim N(0, 1)$;

O mesmo procedimento foi adotado para as variáveis X_3 , X_4 e X_5 .

(iii) as variáveis X_1 , X_2 , X_3 , X_4 e X_5 constituem a matriz de dados que fornecerão o vetor de médias e matriz de covariância, que serão os parâmetros para gerar as variáveis normais multivariadas.

O vetor de médias e a matriz de covariância das variáveis X_1 , X_2 , X_3 , X_4 e X_5 estão apresentados a seguir.

$$\underline{\mu}_p' = [69,9944 \quad 75,0105 \quad 79,9515 \quad 20,0114 \quad 9,9864]$$

$$\Sigma_p = \begin{bmatrix} 80,4732 & 67,2652 & 25,5289 & 45,4789 & 4,6300 \\ 67,2652 & 69,3827 & 21,3423 & 46,9155 & 3,8652 \\ 25,5289 & 21,3423 & 89,6475 & 14,3982 & 16,2273 \\ 45,4789 & 46,9155 & 14,3982 & 35,1576 & 2,6127 \\ 4,6300 & 3,8652 & 16,2273 & 2,6127 & 5,9995 \end{bmatrix}$$

Em que $\underline{\mu}_p$ e Σ_p são os parâmetros utilizados para gerar as variáveis normais multivariadas, sendo, neste caso, $p = 5$;

(iv) foram geradas, então, as variáveis normais multivariadas, que compõem a população a partir da qual serão retiradas amostras aleatórias. O vetor de médias e a matriz de covariância populacional são:

$$\underline{\mu}' = [69,9934 \quad 74,9868 \quad 79,9813 \quad 19,9972 \quad 9,9960]$$

$$\Sigma = \begin{bmatrix} 80,9198 & 67,6965 & 25,7340 & 45,7638 & 4,6463 \\ 67,6965 & 69,7561 & 21,5079 & 47,1444 & 3,8668 \\ 25,7340 & 21,5079 & 90,0413 & 14,5134 & 16,2767 \\ 45,7638 & 47,1444 & 14,5134 & 35,2944 & 2,6076 \\ 4,6463 & 3,8668 & 16,2767 & 2,6076 & 6,0237 \end{bmatrix}$$

Obtiveram-se, da mesma forma, as demais 80 populações normais multivariadas e os respectivos vetores de médias e as matrizes de covariâncias.

4.2 MÉTODO DE OBTENÇÃO DAS AMOSTRAS

Dado o vetor \underline{X} , com distribuição normal multivariada, ou seja, $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$, calcularam-se os tamanhos das amostras necessários para se estimar o vetor de médias populacional $\underline{\mu}$, utilizando-se da expressão (2.123), da seção 2.5.8. Adotaram-se os erros amostrais relativos de 5%, 10% e 15%, e nível de confiança de 95%. Estes valores estão apresentados no quadro 1.

QUADRO 1 - TAMANHOS DE AMOSTRAS PARA DIFERENTES ERROS RELATIVOS E NÍVEL DE CONFIANÇA DE 95%, SEGUNDO VARIÁVEIS

VARIÁVEIS	ERROS RELATIVOS		
	5%	10%	15%
X_1	44	11	5
X_2	33	9	4
X_3	38	10	5
X_4	235	59	27
X_5	160	40	18

FONTE: Dados obtidos por simulação Monte Carlo

Após o dimensionamento das amostras, mínimos necessários para cada uma das variáveis, adotou-se o maior dentre eles, assegurando, desta forma, o erro amostral relativo máximo, definido em 5%, 10% ou 15%. As amostras foram obtidas pelo processo aleatório simples, com reposição. Este procedimento foi adotado para cada uma das 81 populações apresentadas no apêndice 1.

Os tamanhos de amostras dimensionados para as 81 populações normais multivariadas variam entre 24 e 984. Assim, a razão entre o tamanho da amostra e o número de variáveis (n/p) está compreendida entre 3,7 e 49,5. Apesar de diversos autores reforçarem a necessidade de amostras grandes para aplicação da Análise Fatorial, neste trabalho foram consideradas amostras pequenas, pois o objetivo é avaliar os efeitos do tamanho da amostra, além do número de variáveis, número de fatores e explicação estimada dos fatores, nas estimativas dos parâmetros do Modelo Fatorial Ortogonal.

4.3 MÉTODO DE AVALIAÇÃO DAS ESTIMATIVAS DOS PARÂMETROS DO MODELO FATORIAL ORTOGONAL

A avaliação das estimativas dos autovalores, autovetores, cargas fatoriais e comunalidades foi feita usando-se o maior coeficiente de variação e a maior raiz quadrada do erro quadrático médio relativa.

Essa opção deve-se ao fato de que, para cada modelo fatorial estimado, têm-se diferentes números de autovalores, autovetores, cargas fatoriais e comunalidades, e cada elemento estimado possui uma precisão e erro total. Desta forma, está-se avaliando a menor precisão e o maior erro total das estimativas.

A precisão, geralmente, refere-se à variabilidade e não inclui o viés. Já, o erro total é a raiz quadrada do erro quadrático médio, incluindo o viés (KISH, 1965; SILVA, 1998). Assim, a precisão é medida pelo erro padrão, ou a precisão relativa, pelo coeficiente de variação. De acordo com KISH (1965), quando o viés não for desprezível a melhor medida para avaliar a estimativa é o erro total, ou seja, a raiz quadrada do erro quadrático médio.

O erro quadrático médio é obtido através da expressão demonstrada no resultado 2.1:

$$EQM(\hat{\theta}) = b^2(\hat{\theta}) + V(\hat{\theta}) \quad (4.20)$$

Extraindo a raiz quadrada de ambos os membros da igualdade da expressão (4.20) e dividindo por $E(\hat{\theta})$, tem-se:

$$\frac{\sqrt{EQM(\hat{\theta})}}{E(\hat{\theta})} = \frac{\sqrt{b^2(\hat{\theta}) + V(\hat{\theta})}}{E(\hat{\theta})} \quad (4.21)$$

Se $b^2(\hat{\theta})$ for igual a zero, tem-se que:

$$\frac{\sqrt{EQM(\hat{\theta})}}{E(\hat{\theta})} = \frac{\sqrt{V(\hat{\theta})}}{E(\hat{\theta})} \quad (4.22)$$

Neste caso, a raiz quadrada do erro quadrático médio relativa, ou o erro total relativo, equivale ao coeficiente de variação. A precisão é o inverso do coeficiente de variação (CV), ou seja, quanto maior CV, menor é a precisão.

O critério adotado avalia a precisão mínima e o maior erro total relativo dos autovalores, autovetores, cargas fatoriais e comunalidades estimados. As expressões do coeficiente de variação e raiz quadrada do erro quadrático médio relativa são as apresentadas a seguir:

$$CV(\hat{\theta}) = \frac{\sqrt{V(\hat{\theta})}}{E(\hat{\theta})}, \text{ que é o coeficiente de variação da distribuição amostral}$$

do estimador $\hat{\theta}$.

$$REQM(\hat{\theta}) = \frac{\sqrt{EQM(\hat{\theta})}}{E(\hat{\theta})}, \text{ que é a raiz quadrada do erro quadrático médio}$$

relativa do estimador $\hat{\theta}$.

A definição dos modelos matemáticos que explicam a precisão e o erro total das estimativas dos autovalores, autovetores, cargas fatoriais e comunalidades, em função de algumas variáveis explicativas, foi feita utilizando-se a Análise de Regressão Linear Múltipla, apresentada na seção 2.7.

5 RESULTADOS E DISCUSSÃO

5.1 APLICAÇÃO DA ANÁLISE FATORIAL EM DADOS POPULACIONAIS

Apresentam-se, a seguir, os resultados da aplicação da Análise Fatorial a partir da matriz de correlação populacional ($p = 5$ variáveis), cujo vetor de médias e matriz de covariância populacional já foram apresentados na seção 4.1.

$$\rho = \begin{bmatrix} 1,0000 & 0,9010 & 0,3015 & 0,8563 & 0,2105 \\ 0,9010 & 1,0000 & 0,2714 & 0,9501 & 0,1886 \\ 0,3015 & 0,2714 & 1,0000 & 0,2575 & 0,6989 \\ 0,8563 & 0,9501 & 0,2575 & 1,0000 & 0,1788 \\ 0,2105 & 0,1886 & 0,6989 & 0,1788 & 1,0000 \end{bmatrix}$$

Utilizou-se o Método das Componentes Principais e adotou-se o Critério de Kaiser para definir o número de fatores. Os autovalores, autovetores, cargas fatoriais e as comunalidades encontram-se no quadro 2.

QUADRO 2 - AUTOVALORES, AUTOVETORES, CARGAS FATORIAIS E COMUNALIDADES DO MODELO FATORIAL ORTOGONAL

$\lambda_1 = 3,0532$		$\lambda_2 = 1,4573$		h_i^2
e_{i1}	l_{i1}	e_{i2}	l_{i2}	
0,5266	0,9201	-0,1954	-0,2359	0,9022
0,5384	0,9408	-0,2368	-0,2858	0,9669
0,3004	0,5250	0,6261	0,7558	0,8469
0,5284	0,9233	-0,2415	-0,2915	0,9374
0,2517	0,4399	0,6749	0,8147	0,8572

FONTE: Dados obtidos a partir da simulação Monte Carlo

NOTA: i é o número de variáveis, sendo $i = 1, 2, \dots, 5$.

As variâncias explicadas pelos dois primeiros fatores são respectivamente de 61,06% e 29,15%, totalizando 90,21%.

5.2 APLICAÇÃO DA ANÁLISE FATORIAL EM DADOS AMOSTRAIS

Para a aplicação da Análise Fatorial, em dados amostrais, foram realizados os testes de Esfericidade de Bartlett e calculadas as medidas de adequabilidade da amostra (MSA), para as amostras de menor tamanho, obtidas de cada uma das populações, geradas pelo Método de Simulação Monte Carlo. Os resultados dos testes de Esfericidade de Bartlett foram todos significativos (valor-p = 0,0000). Tem-se portanto que as matrizes de correlações das populações são diferentes das matrizes identidades, e a estatística $MSA > 0,50$, indicando que as estruturas de correlações são adequadas para aplicação da Análise Fatorial. Os resultados dos testes estão disponíveis no apêndice 2.

As estimativas dos parâmetros do modelo fatorial ortogonal são os valores médios de 1.000 amostras, extraídas pelo processo aleatório simples, com reposição. O dimensionamento da amostra (n), para cada uma das 81 populações normais multivariadas obtidas pelo Método de Simulação Monte Carlo, foi feito utilizando 95% de confiança e margens de erros relativos fixadas em 5%, 10% e 15%.

No quadro 3 estão as estimativas dos autovalores e os respectivos viés, variância e o erro quadrático médio, para amostras de diferentes tamanhos, obtidas da população apresentada na seção 4.1. Os parâmetros do modelo fatorial ortogonal, que são os autovalores, autovetores, cargas fatoriais e comunalidades, já foram apresentados na seção 5.1.

QUADRO 3 - ESTIMATIVAS DOS AUTOVALORES, VIÉS, VARIÂNCIA E ERRO QUADRÁTICO MÉDIO, SEGUNDO TAMANHOS DE AMOSTRAS

TAMANHOS DE AMOSTRAS	PRIMEIRO AUTOVALOR ESTIMADO MÉDIO				SEGUNDO AUTOVALOR ESTIMADO MÉDIO			
	$\hat{\lambda}_1$	$b(\hat{\lambda}_1)$	$V(\hat{\lambda}_1)$	$EQM(\hat{\lambda}_1)$	$\hat{\lambda}_2$	$b(\hat{\lambda}_2)$	$V(\hat{\lambda}_2)$	$EQM(\hat{\lambda}_2)$
235	3,0578	0,0045	0,0102	0,0102	1,4540	-0,0033	0,0089	0,0089
59	3,0701	0,0169	0,0339	0,0342	1,4441	-0,0132	0,0293	0,0295
26	3,1133	0,0601	0,0662	0,0698	1,4144	-0,0428	0,0555	0,0573

FONTE: Dados obtidos a partir da simulação Monte Carlo

As estimativas dos autovetores, viés, variância e erro quadrático médio, segundo diferentes tamanhos de amostras, encontram-se no quadro 4.

QUADRO 4 - ESTIMATIVAS DOS AUTOVETORES, VIÉS, VARIÂNCIA E ERRO QUADRÁTICO MÉDIO, SEGUNDO TAMANHOS DE AMOSTRAS

TAMANHOS DE AMOSTRAS	PRIMEIRO AUTOVETOR ESTIMADO MÉDIO				SEGUNDO AUTOVETOR ESTIMADO MÉDIO			
	\hat{e}_{i1}	$b(\hat{e}_{i1})$	$V(\hat{e}_{i1})$	$EQM(\hat{e}_{i1})$	\hat{e}_{i2}	$b(\hat{e}_{i2})$	$V(\hat{e}_{i2})$	$EQM(\hat{e}_{i2})$
235	0,5244	-0,0021	0,0025	0,0025	-0,1907	0,0047	0,0015	0,0015
	0,5368	-0,0017	0,0027	0,0027	-0,2313	0,0055	0,0015	0,0016
	0,2947	-0,0057	0,0019	0,0020	0,6211	-0,0050	0,0076	0,0076
	0,5266	-0,0018	0,0026	0,0026	-0,2365	0,0049	0,0017	0,0017
	0,2461	-0,0056	0,0023	0,0023	0,6693	-0,0055	0,0080	0,0080
59	0,4835	-0,0431	0,0446	0,0464	-0,1720	0,0234	0,0076	0,0081
	0,4950	-0,0435	0,0476	0,0495	-0,2057	0,0311	0,0110	0,0120
	0,2740	-0,0264	0,0096	0,0103	0,5112	-0,1149	0,1371	0,1503
	0,4854	-0,0430	0,0459	0,0478	-0,2083	0,0332	0,0128	0,0139
	0,2348	-0,0170	0,0078	0,0080	0,5500	-0,1249	0,1524	0,1680
26	0,3844	-0,1422	0,1273	0,1475	-0,1679	0,0275	0,0125	0,0133
	0,3928	-0,1456	0,1358	0,1570	-0,1923	0,0445	0,0183	0,0203
	0,2611	-0,0394	0,0195	0,0210	0,3634	-0,2627	0,2632	0,3322
	0,3847	-0,1437	0,1311	0,1518	-0,1907	0,0508	0,0224	0,0250
	0,2362	-0,0155	0,0124	0,0126	0,3980	-0,2768	0,2916	0,3682

FONTE: Dados obtidos a partir da simulação Monte Carlo

NOTA: i é o número de variáveis, sendo i = 1, 2, ..., 5.

O quadro 5 apresenta as estimativas das cargas fatoriais e os respectivos viés, variância e erro quadrático médio, segundo diferentes tamanhos de amostras.

QUADRO 5 - ESTIMATIVAS DAS CARGAS FATORIAIS, VIÉS, VARIÂNCIA E ERRO QUADRÁTICO MÉDIO, SEGUNDO TAMANHOS DE AMOSTRAS

TAMANHOS DE AMOSTRAS	CARGA FATORIAL ESTIMADA MÉDIO DO PRIMEIRO FATOR				CARGA FATORIAL ESTIMADA MÉDIO DO SEGUNDO FATOR			
	$\hat{\ell}_{i1}$	$b(\hat{\ell}_{i1})$	$V(\hat{\ell}_{i1})$	$EQM(\hat{\ell}_{i1})$	$\hat{\ell}_{i2}$	$b(\hat{\ell}_{i2})$	$V(\hat{\ell}_{i2})$	$EQM(\hat{\ell}_{i2})$
235	0,9168	-0,0033	0,0070	0,0070	-0,2287	0,0072	0,0018	0,0018
	0,9383	-0,0025	0,0075	0,0075	-0,2775	0,0083	0,0017	0,0018
	0,5164	-0,0086	0,0070	0,0071	0,7488	-0,0070	0,0137	0,0137
	0,9205	-0,0028	0,0072	0,0072	-0,2839	0,0077	0,0020	0,0020
	0,4315	-0,0083	0,0079	0,0080	0,8065	-0,0082	0,0138	0,0139
59	0,8488	-0,0713	0,1302	0,1353	-0,2021	0,0338	0,0098	0,0109
	0,8689	-0,0719	0,1387	0,1439	-0,2417	0,0441	0,0146	0,0165
	0,4840	-0,0410	0,0326	0,0342	0,6065	-0,1493	0,2150	0,2373
	0,8521	-0,0712	0,1339	0,1390	-0,2446	0,0469	0,0173	0,0195
	0,4150	-0,0249	0,0268	0,0274	0,6513	-0,1634	0,2363	0,2630
26	0,6861	-0,2340	0,3793	0,4341	-0,1922	0,0437	0,0158	0,0177
	0,7010	-0,2398	0,4038	0,4613	-0,2190	0,0669	0,0236	0,0280
	0,4677	-0,0573	0,0656	0,0689	0,4118	-0,3440	0,4016	0,5200
	0,6865	-0,2368	0,3901	0,4462	-0,2165	0,0750	0,0295	0,0351
	0,4227	-0,0172	0,0437	0,0440	0,4502	-0,3644	0,4399	0,5727

FONTE: Dados obtidos a partir da simulação Monte Carlo

NOTA: i é o número de variáveis, sendo i = 1, 2, ..., 5.

As estimativas das comunalidades, os respectivos viés, variância e erro quadrático médio estão apresentados no quadro 6.

QUADRO 6 - ESTIMATIVAS DAS COMUNALIDADES, VIÉS, VARIÂNCIA E ERRO QUADRÁTICO MÉDIO, SEGUNDO TAMANHOS DE AMOSTRAS

TAMANHOS DE AMOSTRAS	\hat{h}_i^2 MÉDIAS	$b(\hat{h}_i^2)$	$V(\hat{h}_i^2)$	$EQM(\hat{h}_i^2)$
235	0,9016	-0,0006	0,0002	0,0002
	0,9667	-0,0002	0,0000	0,0000
	0,8481	0,0012	0,0003	0,0003
	0,9371	-0,0003	0,0001	0,0001
	0,8583	0,0012	0,0003	0,0003
59	0,9013	-0,0009	0,0007	0,0007
	0,9667	-0,0002	0,0001	0,0001
	0,8496	0,0028	0,0012	0,0012
	0,9371	-0,0003	0,0003	0,0003
	0,8595	0,0023	0,0010	0,0010
26	0,9027	0,0005	0,0015	0,0015
	0,9667	-0,0002	0,0002	0,0002
	0,8555	0,0087	0,0026	0,0027
	0,9378	0,0004	0,0006	0,0006
	0,8650	0,0078	0,0023	0,0023

FONTE: Dados obtidos a partir da simulação Monte Carlo

NOTA: i é o número de variáveis, sendo $i = 1, 2, \dots, 5$.

5.2.1 Coeficiente de Variação e Raiz Quadrada do Erro Quadrático Médio Relativa das Estimativas dos Autovalores

Calcularam-se o coeficiente de variação e a raiz quadrada do erro quadrático médio relativa das estimativas dos autovalores, apresentados no quadro 7.

Em se tratando de medidas relativas, tanto o coeficiente de variação (Cv_{max}) quanto a raiz quadrada do erro quadrático médio relativa ($REQM_{max}$) devem ser analisados considerando-se o valor médio da estimativa em questão e, respectivamente, a sua variância e o erro quadrático médio. Quando as estimativas médias são pequenas, como no caso dos autovetores e cargas fatoriais, as medidas relativas são consideravelmente grandes. É indicado, nestes casos, analisar junto com a variância e o erro quadrático médio.

QUADRO 7 - COEFICIENTE DE VARIAÇÃO E RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DOS AUTOVALORES, SEGUNDO TAMANHOS DE AMOSTRAS

TAMANHOS DE AMOSTRAS	PRIMEIRO AUTOVALOR ESTIMADO		SEGUNDO AUTOVALOR ESTIMADO	
	$CV(\hat{\lambda}_1)$	$\sqrt{EQM(\hat{\lambda}_1)}/\hat{\lambda}_1$	$CV(\hat{\lambda}_2)$	$\sqrt{EQM(\hat{\lambda}_2)}/\hat{\lambda}_2$
235	0,0330	0,0331	0,0647	0,0648
59	0,0600	0,0603	0,1185	0,1188
26	0,0826	0,0849	0,1666	0,1693

FONTE: Dados obtidos a partir da simulação Monte Carlo

Tem-se que o maior coeficiente de variação para amostra de tamanho 235 é o correspondente ao segundo autovalor, ou seja, 0,0647. No caso da raiz quadrada do erro quadrático médio relativa, é também correspondente ao segundo autovalor, igual a 0,0648.

Também, para as amostras de tamanho 59 e 26, tanto o coeficiente de variação quanto a raiz quadrada do erro quadrático médio relativa, os maiores valores são os correspondentes ao segundo autovalor. Para as três amostras, os autovalores correspondentes ao maior CV e maior REQM são os menores valores estimados, conforme mostra o quadro 3.

Os coeficientes de variação, bem como as raízes quadradas dos erros quadráticos médios relativos, aumentam à medida que diminuem os tamanhos de amostras.

5.2.2 Coeficiente de Variação e Raiz Quadrada do Erro Quadrático Médio Relativa das Estimativas dos Autovetores

Ressalta-se que as estimativas dos autovetores e das cargas fatoriais foram consideradas em valores absolutos. Os sinais das cargas fatoriais, nos fatores, podem ser invertidos, sem que haja alterações na análise (JOHNSON e WICHERN, 1988). Uma vez que os sinais dos autovetores é que determinam os sinais das cargas fatoriais, ambas as estimativas foram utilizadas em valores absolutos.

Os coeficientes de variação e as raízes quadradas dos erros quadráticos médios relativos das estimativas dos autovetores, para a amostra de tamanho 235, estão apresentados no quadro 8.

QUADRO 8 - COEFICIENTE DE VARIAÇÃO E RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DOS AUTOVETORES, SEGUNDO VARIÁVEIS

VARIÁVEIS	\hat{e}_1		\hat{e}_2	
	$CV(\hat{e}_{i1})$	$\sqrt{EQM(\hat{e}_{i1})}/\hat{e}_{i1}$	$CV(\hat{e}_{i2})$	$\sqrt{EQM(\hat{e}_{i2})}/\hat{e}_{i2}$
X_1	0,0961	0,0962	0,2031	0,2045
X_2	0,0974	0,0974	0,1701	0,1717
X_3	0,1491	0,1504	0,1404	0,1406
X_4	0,0973	0,0974	0,1747	0,1759
X_5	0,1942	0,1955	0,1335	0,1337

FONTE: Dados obtidos a partir da simulação Monte Carlo

NOTA: i é o número de variáveis, sendo $i = 1, 2, \dots, 5$.

Observa-se que o maior coeficiente de variação e a maior raiz quadrada do erro quadrático médio relativa são referentes à variável X_1 , correspondentes ao segundo autovetor, sendo respectivamente de 0,2031 e 0,2045. Ao observar a estimativa do autovetor correspondente, no quadro 4, nota-se que é o menor valor absoluto, sendo igual a 0,1907.

5.2.3 Coeficiente de Variação e Raiz Quadrada do Erro Quadrático Médio Relativa das Estimativas das Cargas Fatoriais

Os coeficientes de variação e raízes quadradas dos erros quadráticos médios relativos das estimativas das cargas fatoriais, para amostra de tamanho 235, estão apresentados no quadro 9. Conforme já mencionado, as estimativas das cargas fatoriais foram consideradas em valores absolutos.

QUADRO 9 - COEFICIENTE DE VARIAÇÃO E RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DAS CARGAS FATORIAIS, SEGUNDO VARIÁVEIS

VARIÁVEIS	$\hat{\ell}_{i1}$		$\hat{\ell}_{i2}$	
	$CV(\hat{\ell}_{i1})$	$\sqrt{EQM(\hat{\ell}_{i1})}/\hat{\ell}_{i1}$	$CV(\hat{\ell}_{i2})$	$\sqrt{EQM(\hat{\ell}_{i2})}/\hat{\ell}_{i2}$
X_1	0,0914	0,0915	0,1830	0,1857
X_2	0,0922	0,0923	0,1507	0,1536
X_3	0,1619	0,1627	0,1562	0,1565
X_4	0,0922	0,0923	0,1569	0,1592
X_5	0,2063	0,2072	0,1457	0,1461

FONTE: Dados obtidos a partir da simulação Monte Carlo

NOTA: i é o número de variáveis, sendo $i = 1, 2, \dots, 5$.

O maior coeficiente de variação e raiz quadrada do erro quadrático médio relativa correspondem à variável X_5 , correspondentes ao primeiro fator, sendo respectivamente de 0,2063 e 0,2072. A carga fatorial estimada correspondente, conforme mostra o quadro 5, é igual a 0,4315.

5.2.4 Coeficiente de Variação e Raiz Quadrada do Erro Quadrático Médio Relativa, das Estimativas das Comunalidades

O maior coeficiente de variação e a maior raiz quadrada do erro quadrático médio relativa das comunalidades estimadas, para a amostra de tamanho 235, referem-se à variável X_3 , sendo respectivamente de 0,0203 e 0,0204, apresentados no quadro 10.

QUADRO 10 - COEFICIENTES DE VARIAÇÃO E RAÍZES QUADRADAS DO ERRO QUADRÁTICO MÉDIO RELATIVAS DAS ESTIMATIVAS DAS COMUNALIDADES, SEGUNDO VARIÁVEIS

VARIÁVEIS	$CV(\hat{h}_i^2)$	$\sqrt{EQM(\hat{h}_i^2)}/\hat{h}_i^2$
X_1	0,0142	0,0142
X_2	0,0042	0,0042
X_3	0,0203	0,0204
X_4	0,0084	0,0084
X_5	0,0187	0,0188

FONTE: Dados obtidos a partir da simulação Monte Carlo

NOTA: i é o número de variáveis, sendo $i = 1, 2, \dots, 5$.

Ao observar a estimativa da comunalidade correspondente, no quadro 6, tem-se que é o menor valor, igual a 0,8481.

O critério descrito acima foi adotado para cada uma das 243 amostras, obtidas a partir das 81 populações. Os maiores coeficientes de variação e maiores raízes quadradas do erro quadrático médio relativas e as respectivas estimativas dos autovalores, autovetores, cargas fatoriais e comunalidades, bem como as variâncias e os erros quadráticos médios, encontram-se nos quadros A.5.1 e A.5.2, do apêndice 5.

Apresentam-se, a seguir, as descrições das variáveis utilizadas nas próximas seções.

QUADRO 11 - DESCRIÇÃO DAS VARIÁVEIS

VARIÁVEIS	DESCRIÇÃO
n	Tamanho da amostra
p	Número de variáveis
f	Número de fatores (autovalores > 1)
n/p	Razão entre o tamanho da amostra e o número de variáveis
p/f	Razão entre o número de variáveis e o de fatores
expl	Estimativa da explicação total dos fatores
CVmax	Maior coeficiente de variação
REQMmax	Maior raiz quadrada do erro quadrático médio relativa
autoval	Estimativa do autovalor, correspondente ao maior coeficiente de variação e à maior raiz quadrada do erro quadrático médio relativa
autovet	Estimativa do autovetor em valor absoluto, correspondente ao maior coeficiente de variação e à maior raiz quadrada do erro quadrático médio relativa
car	Estimativa da carga fatorial, em valor absoluto, correspondente ao maior coeficiente de variação e à maior raiz quadrada do erro quadrático médio relativa
comun	Estimativa da comunalidade, correspondente ao maior coeficiente de variação e à maior raiz quadrada do erro quadrático médio relativa
CVmaxautoval	Maior coeficiente de variação das estimativas dos autovalores
REQMmaxautoval	Maior raiz quadrada do erro quadrático médio relativa das estimativas dos autovalores
CVmaxautovet	Maior coeficiente de variação das estimativas dos autovetores
REQMmaxautovet	Maior raiz quadrada do erro quadrático médio relativa das estimativas dos autovetores
CVmaxcar	Maior coeficiente de variação das estimativas das cargas fatoriais
REQMmaxcar	Maior raiz quadrada do erro quadrático médio relativa das estimativas das cargas fatoriais
CVmaxcomun	Maior coeficiente de variação das estimativas das comunalidades
REQMmaxcomun	Maior raiz quadrada do erro quadrático médio relativa das estimativas das comunalidades

FONTE: A autora

Foram excluídas duas amostras, obtidas das populações 27 e 51, por apresentarem autovalores médios menores do que a unidade (respectivamente, o quarto e o quinto autovalor), embora os parâmetros sejam superiores a um. Isto ocorre devido às estimativas dos coeficientes de correlação linear de Pearson ($\hat{\rho}$). Um dos fatores que afetam a intensidade, bem como a precisão, ao se estimar a correlação populacional (ρ), é o tamanho da amostra, principalmente quando ele é pequeno. Nesses casos, os tamanhos de amostras e números de variáveis são respectivamente de: 46 e 10; 76 e 14. As matrizes de correlações populacionais e as amostrais, bem como os respectivos autovalores, encontram-se no apêndice 4.

Assim, foram consideradas 241 amostras, para as análises posteriores. Ressalta-se que cada uma delas representa o valor médio de 1.000 amostras (NA=1.000) de mesmo tamanho.

O quadro 12 apresenta os valores mínimo, percentil 25, mediana, percentil 75 e máximo, das variáveis utilizadas na definição dos modelos matemáticos, conforme as descrições já apresentadas no quadro 11.

QUADRO 12 - VALORES MÍNIMO, PERCENTIL 25, MEDIANA, PERCENTIL 75 E MÁXIMO, SEGUNDO VARIÁVEIS

VARIÁVEIS	MÍNIMO	PERCENTIL 25	MEDIANA	PERCENTIL 75	MÁXIMO
n	24	67	124	396	984
p	5	8	12	16	20
f	2	3	4	5	6
n/p	3,7	5,4	10,8	37,2	49,5
p/f	2,0	3,0	3,2	3,5	4,5
expl	0,6363	0,7543	0,7838	0,8070	0,9055
autoval	1,0034	1,1404	1,3729	1,6974	3,8291
autovet (valores absolutos)	0,0001	0,0008	0,0026	0,0143	0,5112
car (valores absolutos)	0,0001	0,0010	0,0034	0,0181	0,6065
comun	0,0241	0,3857	0,5609	0,6644	0,8589
CVmaxautoval	0,0384	0,0607	0,0952	0,1246	0,2117
REQMmaxautoval	0,0384	0,0627	0,0997	0,1311	0,2282
CVmaxautovet	0,1682	8,5410	44,9792	153,2397	791,0738
REQMmaxautovet	0,1684	9,0849	46,7847	178,2858	899,4263
CVmaxcar	0,1752	8,5280	44,6841	147,6012	795,2769
REQMmaxcar	0,1753	9,1716	46,9831	170,1066	1.037,2921
CVmaxcomun	0,0203	0,0937	0,1769	0,3260	1,0107
REQMmaxcomun	0,0204	0,0992	0,1923	0,4142	1,3793

FONTE: Dados obtidos por simulação Monte Carlo

Observa-se que o maior coeficiente de variação (CVmax) e a maior raiz quadrada do erro quadrático médio relativa (REQMmax) dos autovalores e das comunalidades são pequenos, quando comparados aos dos autovetores e das cargas fatoriais. Isso ocorre devido ao fato de os valores médios das estimativas das duas últimas componentes, correspondentes às medidas analisadas (CVmax e REQMmax), serem muito pequenos.

Os maiores coeficientes de variação (CVmax) e as maiores raízes quadradas dos erros quadráticos médios relativas (REQMmax), dos autovalores,

autovetores, cargas fatoriais e comunalidades, bem como as estimativas correspondentes, estão apresentados nos quadros A.5.1 e A.5.2 do apêndice 5.

Para analisar as relações existentes entre o maior coeficiente de variação e a maior raiz quadrada do erro quadrático médio relativa, com as variáveis explicativas: estimativas dos parâmetros do modelo fatorial ortogonal, tamanho da amostra, número de variáveis, estimativa da explicação dos fatores e número de fatores, utilizou-se o modelo de regressão linear múltipla, apresentado nas próximas seções.

Para facilitar a apresentação dos resultados e tendo em vista que estão sendo avaliados os maiores valores, tanto do coeficiente de variação quanto da raiz quadrada do erro quadrático médio relativa, serão utilizados os termos coeficiente de variação e raiz quadrada do erro quadrático médio relativa.

5.3 ANÁLISE DO COEFICIENTE DE VARIAÇÃO E RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DOS AUTOVALORES

5.3.1 Coeficiente de Variação das Estimativas dos Autovalores

As estimativas dos coeficientes de regressão, seus respectivos erros padrão, estatísticas t e F, erro padrão da estimativa e o coeficiente de determinação do modelo, que relaciona o coeficiente de variação das estimativas dos autovalores com as variáveis explicativas: autovalor estimado, estimativa da explicação média dos fatores e tamanho da amostra, encontram-se no quadro 13. Os testes aplicados para avaliar a multicolinearidade, homogeneidade da variância, Gussianidade dos resíduos e $E(\underline{\varepsilon}) = \underline{0}$ estão apresentados no apêndice 7 e as médias e os desvios padrão das variáveis do modelo, no apêndice 6.

QUADRO 13 - ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO, ERRO PADRÃO, ESTATÍSTICAS t E F, VALOR-p E COEFICIENTE DE DETERMINAÇÃO DO MODELO AJUSTADO PARA O MAIOR COEFICIENTE DE VARIAÇÃO DOS AUTOVALORES ESTIMADOS

VARIÁVEIS EXPLICATIVAS	ESTIMATIVA DOS COEFICIENTES	ERRO PADRÃO	ESTATÍSTICA t	VALOR-p
rautoval	-0,4696	0,0193	-24,3594	0,0000
r(n)	-0,0289	0,0010	-29,3419	0,0000
expm	-0,3329	0,0682	-4,8819	0,0000
S = 0,1051				
Coeficiente de Determinação (R^2) = 0,9903				
Estatística $F_{3, 238} = 8.226,0200$				0,0000

FONTE: Dados obtidos por simulação Monte Carlo

O modelo ajustado é:

$$\log CV_{\max \text{ autoval}} = -0,4696 \text{ rautoval} - 0,0289 r(n) - 0,3329 \text{ expm} \quad (5.4)$$

em que:

$\log CV_{\max \text{ autoval}}$ é o logaritmo decimal do maior coeficiente de variação da estimativa do autovalor;

rautoval é a raiz quadrada da estimativa do autovalor, correspondente ao maior coeficiente de variação;

$r(n)$ é a raiz quadrada do tamanho da amostra;

expm é a estimativa da explicação média dos fatores (razão entre a explicação total estimada e o número de fatores).

O modelo anterior poderá ser escrito na forma exponencial, ou seja:

$$CV_{\max \text{ autoval}} = 10^{-0,4696 \text{ rautoval} - 0,0289 r(n) - 0,3329 \text{ expm}} \quad (5.5)$$

O coeficiente de determinação do modelo ajustado é $R^2 = 99,03\%$, indicando que a qualidade do ajuste é muito boa. Para a verificação do modelo ajustado, identificaram-se os pontos de *leverages* e influentes, que se encontram no apêndice 7. Após a análise desses pontos, optou-se pela não exclusão. Conclui-se, pelo modelo, que o aumento em uma ou mais variáveis explicativas: estimativa do

autovalor; estimativa da explicação média dos fatores e tamanho da amostra diminui o coeficiente de variação.

5.3.2 Raiz Quadrada do Erro Quadrático Médio Relativa das Estimativas dos Autovalores

As estimativas dos coeficientes de regressão, seus respectivos erros padrão, estatísticas t e F, erro padrão da estimativa e o coeficiente de determinação do modelo que relaciona a raiz quadrada do erro quadrático médio relativa das estimativas dos autovalores com as variáveis explicativas: autovalor estimado, estimativa da explicação média dos fatores e tamanho da amostra, encontram-se no quadro 14.

QUADRO 14 - ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO, ERRO PADRÃO, ESTATÍSTICAS t E F, VALOR-p E COEFICIENTE DE DETERMINAÇÃO DO MODELO AJUSTADO PARA A MAIOR RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DOS AUTOVALORES ESTIMADOS

VARIÁVEIS EXPLICATIVAS	ESTIMATIVA DOS COEFICIENTES	ERRO PADRÃO	ESTATÍSTICA t	VALOR-p
rautoval	-0,4202	0,0180	-23,3912	0,0000
r(n)	-0,0306	0,0009	-33,2804	0,0000
expm	-0,4125	0,0635	-6,4922	0,0000
S = 0,0979				
Coeficiente de Determinação (R^2) = 0,9913				
Estatística $F_{3, 238} = 9.164,2600$				0,0000

FONTE: Dados obtidos por simulação Monte Carlo

O modelo ajustado é:

$$\log\text{REQMmax autoval} = -0,4202\text{rautoval} - 0,0306\text{r}(n) - 0,4125\text{expm} \quad (5.6)$$

em que:

$\log\text{REQMmax autoval}$ é o logaritmo decimal da maior raiz quadrada do erro quadrático médio relativa da estimativa do autovalor;

rautoval é a raiz quadrada da estimativa do autovalor, correspondente à maior raiz quadrada do erro quadrático médio relativa;

$r(n)$ é a raiz quadrada do tamanho da amostra;

$expm$ é a estimativa da explicação média dos fatores (razão entre a explicação total e o número de fatores).

O modelo anterior pode ser escrito na forma exponencial, como segue:

$$REQM_{\max} \text{ autoval} = 10^{-0,4202 \text{ rautoval} - 0,0306 \text{ r}(n) - 0,4125 \text{ expm}} \quad (5.7)$$

O coeficiente de determinação do modelo ajustado é $R^2 = 99,13\%$, indicando que a qualidade do ajuste é muito boa. Identificaram-se os pontos de *leverages* e influentes para verificar o modelo, e os resultados se encontram no apêndice 7. Após a análise de cada um deles, optou-se por não excluí-los.

Conclui-se, então, pelo modelo, que o aumento em uma ou mais variáveis explicativas: estimativa do autovalor; estimativa da explicação média dos fatores e tamanho da amostra diminui a raiz quadrada do erro quadrático médio relativa.

5.4 ANÁLISE DO COEFICIENTE DE VARIAÇÃO E RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DOS AUTOVETORES

5.4.1 Coeficiente de Variação das Estimativas dos Autovetores

As estimativas dos coeficientes de regressão, seus respectivos erros padrão, estatísticas t e F , erro padrão da estimativa e o coeficiente de determinação do modelo ajustado para explicar o comportamento do coeficiente de variação das estimativas dos autovetores, em função das variáveis explicativas: autovetor estimado, tamanho da amostra, razão entre o número de fatores e o de variáveis e estimativa da explicação total dos fatores, encontram-se no quadro 15.

QUADRO 15 - ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO, ERRO PADRÃO, ESTATÍSTICAS t E F, VALOR-p E COEFICIENTE DE DETERMINAÇÃO DO MODELO AJUSTADO PARA O MAIOR COEFICIENTE DE VARIAÇÃO DOS AUTOVETORES ESTIMADOS

VARIÁVEIS EXPLICATIVAS	ESTIMATIVA DOS COEFICIENTES	ERRO PADRÃO	ESTATÍSTICA t	VALOR-p
logautovet	-0,9282	0,0167	-55,4830	0,0000
n	-0,0009	0,0001	-12,7093	0,0000
r(f/p)	2,3360	0,4382	5,3310	0,0000
expl	-2,3731	0,3199	-7,4189	0,0000
S = 0,2381				
Coeficiente de Determinação (R^2) = 0,9808				
Estatística $F_{4, 237} = 3.076,6690$				0,0000

FONTE: Dados obtidos por simulação Monte Carlo

O modelo ajustado é:

$$\log CV \max \text{ autovet} = -0,9282 \log \text{ autovet} - 0,0009 n + 2,3360 r(f/p) - 2,3731 \text{ expl} \quad (5.8)$$

em que:

$\log CV \max \text{ autovet}$ é o logaritmo decimal do maior coeficiente de variação da estimativa do autovetor;

$\log \text{ autovet}$ é o logaritmo decimal do valor absoluto da estimativa do autovetor, correspondente ao maior coeficiente de variação;

n é o tamanho da amostra;

r(f/p) é a raiz quadrada da razão entre o número de fatores e o de variáveis;

expl é a estimativa da explicação total dos fatores.

O modelo anterior pode ser escrito na forma exponencial, ou seja:

$$CV \max \text{ autovet} = 10^{-0,9282 \log \text{ autovet} - 0,0009 n + 2,3360 r(f/p) - 2,3731 \text{ expl}} \quad (5.9)$$

O coeficiente de determinação do modelo ajustado é $R^2 = 98,08\%$, indicando um bom ajuste. Foram identificados os pontos influentes e *leverages*, que se encontram no apêndice 7. Após a análise, não se excluiu nenhum ponto.

Conclui-se, pelo modelo, que o aumento em uma ou mais variáveis explicativas: estimativa, considerando em valor absoluto, do autovetor; estimativa da explicação total dos fatores e tamanho da amostra, reduz o coeficiente de variação.

No entanto, o aumento na razão entre o número de fatores e o de variáveis aumenta o coeficiente de variação.

5.4.2 Raiz Quadrada do Erro Quadrático Médio Relativa das Estimativas dos Autovetores

Apresentam-se, a seguir, no quadro 16, as estimativas dos coeficientes de regressão, seus respectivos erros padrão, estatísticas t e F, erro padrão da estimativa e o coeficiente de determinação do modelo ajustado para explicar a relação entre a variável resposta raiz quadrada do erro quadrático médio relativa, em função das variáveis explicativas: autovetor estimado, explicação total dos fatores, tamanho da amostra e razão entre o número de fatores e o de variáveis.

QUADRO 16 - ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO, ERRO PADRÃO, ESTATÍSTICAS t E F, VALOR-p E COEFICIENTE DE DETERMINAÇÃO DO MODELO AJUSTADO PARA A MAIOR RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DOS AUTOVETORES ESTIMADOS

VARIÁVEIS EXPLICATIVAS	ESTIMATIVA DOS COEFICIENTES	ERRO PADRÃO	ESTATÍSTICA t	VALOR-p
logautovet	-0,9336	0,0187	-49,8876	0,0000
n	-0,0009	0,0001	-11,8609	0,0000
r(f/p)	2,7525	0,4902	5,6154	0,0000
expl	-2,6261	0,3578	-7,3390	0,0000
S = 0,2664				
Coeficiente de Determinação (R^2) = 0,9771				
Estatística $F_{4, 237} = 2.571,0810$				0,0000

FONTE: Dados obtidos por simulação Monte Carlo

O modelo ajustado é:

$$\log\text{REQMmax autovet} = -0,9336 \log\text{autovet} - 0,0009 n + 2,7525r(f/p) - 2,6261\text{expl} \quad (5.10)$$

em que:

$\log\text{REQMmax autovet}$ é o logaritmo decimal da maior raiz quadrada do erro quadrático médio relativa da estimativa do autovetor;

$\log\text{autovet}$ é o logaritmo decimal do valor absoluto da estimativa do autovetor correspondente à maior raiz quadrada do erro quadrático médio relativa;

n é o tamanho da amostra;

$r(f/p)$ é a raiz quadrada da razão entre o número de fatores e o de variáveis; $expl$ é a estimativa da explicação total dos fatores.

O modelo anterior na forma exponencial é:

$$REQM_{\max} \text{ autovet} = 10^{-0,9336 \log \text{ autovet} - 0,0009n + 2,7525 r(f/p) - 2,6261 \text{ expl}} \quad (5.11)$$

O coeficiente de determinação do modelo ajustado é $R^2 = 97,71\%$, indicando que é um bom ajuste. Após a análise dos pontos de *leverages* e influentes, apresentados no apêndice 7, não se excluiu nenhum ponto.

É possível concluir, através do modelo ajustado, que o aumento em uma ou mais variáveis explicativas: estimativa – considerando em valor absoluto – do autovetor; tamanho da amostra e estimativa da explicação total dos fatores, reduz a raiz quadrada do erro quadrático médio relativa. Por outro lado, o aumento na razão entre o número de fatores e o de variáveis aumenta a raiz quadrada do erro quadrático médio relativa.

5.5 ANÁLISE DO COEFICIENTE DE VARIAÇÃO E RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DAS CARGAS FATORIAIS

5.5.1 Coeficiente de Variação das Cargas Fatoriais Estimadas

O quadro 17 apresenta as estimativas dos coeficientes de regressão, seus respectivos erros padrão, estatísticas t e F, erro padrão da estimativa e o coeficiente de determinação do modelo ajustado para explicar a relação entre a variável coeficiente de variação das cargas fatoriais e as variáveis explicativas: estimativa das cargas fatoriais, tamanho da amostra, estimativa da explicação total dos fatores e razão entre o número de fatores e o de variáveis.

QUADRO 17 - ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO, ERRO PADRÃO, ESTATÍSTICAS t E F, VALOR-p E COEFICIENTE DE DETERMINAÇÃO DO MODELO AJUSTADO PARA O MAIOR COEFICIENTE DE VARIAÇÃO DAS CARGAS FATORIAIS ESTIMADAS

VARIÁVEIS EXPLICATIVAS	ESTIMATIVA DOS COEFICIENTES	ERRO PADRÃO	ESTATÍSTICA t	VALOR-p
logcar	-0,9183	0,0165	-55,5270	0,0000
n	-0,0009	0,0001	-12,6867	0,0000
r(f/p)	2,4407	0,4333	5,6334	0,0000
expl	-2,3103	0,3158	-7,3166	0,0000
S = 0,2355				
Coeficiente de Determinação (R^2) = 0,9810				
Estatística $F_{4, 237} = 3.112,2030$				0,0000

FONTE: Dados obtidos por simulação Monte Carlo

O modelo ajustado é:

$$\log CV \max car = -0,9183 \log car - 0,0009 n + 2,4407 r(f/p) - 2,3103 expl \quad (5.12)$$

em que:

logCV max car é o logaritmo decimal do maior coeficiente de variação da estimativa da carga fatorial;

logcar é o logaritmo decimal do valor absoluto da estimativa da carga fatorial, correspondente ao maior coeficiente de variação;

n é o tamanho da amostra;

r(f/p) é a raiz quadrada da razão entre o número de fatores e o de variáveis;

expl é a estimativa da explicação total dos fatores.

O modelo anterior na forma exponencial é:

$$CV \max car = 10^{-0,9183 \log car - 0,0009 n + 2,4407 r(f/p) - 2,3103 expl} \quad (5.13)$$

O coeficiente de determinação do modelo é $R^2 = 98,10\%$, indicando que o ajuste do modelo é muito bom. Após a análise dos pontos de *leverages* e influentes, que se encontram no apêndice 7, optou-se pela não exclusão deles.

Conclui-se, pelo modelo ajustado, que o aumento em uma ou mais variáveis explicativas: estimativa (em valor absoluto) das cargas fatoriais, tamanho da amostra e explicação total dos fatores reduz o coeficiente de variação. Por outro lado, o aumento na razão entre o número de fatores e o de variáveis aumenta o coeficiente de variação.

5.5.2 Raiz Quadrada do Erro Quadrático Médio Relativa das Cargas Fatoriais

Estimadas

Encontram-se, no quadro 18, as estimativas dos coeficientes de regressão, seus respectivos erros padrão, estatísticas t e F, erro padrão da estimativa e o coeficiente de determinação do modelo para explicar a variável raiz quadrada do erro quadrático médio relativa das cargas fatoriais, em função das variáveis explicativas: estimativas das cargas fatoriais, tamanho da amostra, estimativa da explicação total dos fatores e razão entre o número de fatores e o de variáveis.

QUADRO 18 - ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO, ERRO PADRÃO, ESTATÍSTICAS t E F, VALOR-p E COEFICIENTE DE DETERMINAÇÃO DO MODELO AJUSTADO PARA A MAIOR RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS CARGAS FATORIAIS ESTIMADAS

VARIÁVEIS EXPLICATIVAS	ESTIMATIVA DOS COEFICIENTES	ERRO PADRÃO	ESTATÍSTICA t	VALOR-p
logcar	-0,9216	0,0187	-49,1767	0,0000
n	-0,0009	0,0001	-11,6754	0,0000
r(f/p)	2,8681	0,4910	5,8417	0,0000
expl	-2,5651	0,3578	-7,1687	0,0000
S = 0,2668				
Coeficiente de Determinação (R^2) = 0,9767				
Estatística $F_{4, 237} = 2.531,7660$				0,0000

FONTE: Dados obtidos por simulação Monte Carlo

O modelo ajustado é:

$$\log \text{REQMmax car} = -0,9216 \log \text{car} - 0,0009 n + 2,8681 r(f/p) - 2,5651 \text{expl} \quad (5.14)$$

em que:

$\log \text{REQM}_{\max \text{car}}$ é o logaritmo decimal da maior raiz quadrada do erro quadrático médio relativa da estimativa da carga fatorial;

$\log \text{car}$ é o logaritmo decimal do valor absoluto da estimativa da carga fatorial, correspondente à maior raiz quadrada do erro quadrático médio relativa;

n é o tamanho da amostra;

$r(f/p)$ é a raiz quadrada da razão entre o número de fatores e o de variáveis;

expl é a estimativa da explicação total dos fatores.

O modelo anterior na forma exponencial é dado por:

$$\text{REQM}_{\max \text{car}} = 10^{-0,9216 \log \text{car} - 0,0009n + 2,8681r(f/p) - 2,5651 \text{expl}} \quad (5.15)$$

O coeficiente de determinação do modelo é $R^2 = 97,67\%$, indicando que é bom ajuste. Após a análise dos pontos de *leverages* e influentes, optou-se pela não exclusão deles.

Conclui-se, pelo modelo ajustado, que o aumento em uma ou mais variáveis explicativas: estimativa (em valor absoluto) das cargas fatoriais; tamanho da amostra e estimativa da explicação total dos fatores reduz o valor da raiz quadrada do erro quadrático médio relativa. No entanto, o aumento na razão entre o número de fatores e o de variáveis aumenta a raiz quadrada do erro quadrático médio relativa.

5.6 ANÁLISE DO COEFICIENTE DE VARIAÇÃO E RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DAS COMUNALIDADES

5.6.1 Coeficiente de Variação das Estimativas das Comunalidades

As estimativas dos coeficientes de regressão, erros padrão, estatísticas t e F e o coeficiente de determinação do modelo ajustado para explicar a relação existente entre o coeficiente de variação, das comunalidades estimadas e as variáveis

explicativas: estimativas das comunalidades, razão entre o tamanho da amostra e o número de variáveis e número de fatores, estão apresentados no quadro 19.

QUADRO 19 - ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO, ERRO PADRÃO, ESTATÍSTICAS t E F, VALOR-p E COEFICIENTE DE DETERMINAÇÃO DO MODELO AJUSTADO PARA O MAIOR COEFICIENTE DE VARIAÇÃO DAS COMUNALIDADES ESTIMADAS

VARIÁVEIS EXPLICATIVAS	ESTIMATIVA DOS COEFICIENTES	ERRO PADRÃO	ESTATÍSTICA t	VALOR-p
comun	-1,3902	0,0617	-22,5424	0,0000
r(n/p)	-0,0743	0,0065	-11,4623	0,0000
r(f)	0,1261	0,0216	5,8516	0,0000
S = 0,2025				
Coeficiente de Determinação (R^2) = 0,9440				
Estatística $F_{3, 238} = 1.355,0600$				0,0000

FONTE: Dados obtidos por simulação Monte Carlo

O modelo ajustado é:

$$\log CV \max \text{comun} = -1,3902 \text{comun} - 0,0743 r(n/p) + 0,1261 r(f) \quad (5.16)$$

em que:

$\log CV \max \text{comun}$ é o logaritmo decimal do maior coeficiente de variação da estimativa da comunalidade;

comun é a estimativa da comunalidade, correspondente ao maior coeficiente de variação;

$r(n/p)$ é a raiz quadrada da razão entre o tamanho da amostra e o número de variáveis;

$r(f)$ é a raiz quadrada do número de fatores.

O modelo anterior pode ser escrito na forma exponencial, como segue:

$$CV \max \text{comun} = 10^{-1,3902 \text{comun} - 0,0743 r(n/p) + 0,1261 r(f)} \quad (5.17)$$

O coeficiente de determinação do modelo ajustado é $R^2 = 94,40\%$, indicando que é um bom ajuste. Após a análise dos pontos de *leverages* e influentes, optou-se pela não exclusão deles.

O modelo ajustado indica que o aumento na comunalidade estimada e/ou na razão entre o tamanho da amostra e o número de variáveis reduzem o coeficiente de variação. Por outro lado, o aumento no número de fatores aumenta o coeficiente de variação.

5.6.2 Raiz Quadrada do Erro Quadrático Médio Relativa das Estimativas das Comunalidades

As estimativas dos coeficientes de regressão, seus respectivos erros padrão, estatísticas t e F, erro padrão da estimativa e o coeficiente de determinação do modelo para explicar as variações ocorridas na variável resposta, raiz quadrada do erro quadrático médio relativa, das comunalidades estimadas, em função das variáveis explicativas: estimativas das comunalidades, razão entre o tamanho da amostra e o número de variáveis e número de fatores, encontram-se no quadro 20.

QUADRO 20 - ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO, ERRO PADRÃO, ESTATÍSTICAS t E F, VALOR-p E COEFICIENTE DE DETERMINAÇÃO DO MODELO AJUSTADO PARA A MAIOR RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS COMUNALIDADES ESTIMADAS

VARIÁVEIS EXPLICATIVAS	ESTIMATIVA DOS COEFICIENTES	ERRO PADRÃO	ESTATÍSTICA t	VALOR-p
comun	-1,5276	0,0696	-21,9498	0,0000
r(n/p)	-0,0901	0,0073	-12,3290	0,0000
r(f)	0,2262	0,0243	9,3010	0,0000
S = 0,2285				
Coeficiente de Determinação (R^2) = 0,9232				
Estatística $F_{3, 238} = 966,6200$				0,0000

FONTE: Dados obtidos por simulação Monte Carlo

O modelo ajustado é:

$$\log\text{REQMmax comun} = -1,5276 \text{ comun} - 0,0901 r(n/p) + 0,2262 r(f) \quad (5.18)$$

em que:

$\log\text{REQMmax comun}$ é o logaritmo decimal da maior raiz quadrada do erro quadrático médio relativa da estimativa da comunalidade;

comun é a estimativa da comunalidade, correspondente à maior raiz quadrada do erro quadrático médio relativa;

$r(n/p)$ é a raiz quadrada da razão entre o tamanho da amostra e o número de variáveis;

$r(f)$ é a raiz quadrada do número de fatores.

O modelo anterior pode ser escrito na forma exponencial, ou seja:

$$\text{REQMmax comun} = 10^{-1,5276 \text{ comun} - 0,0901r(n/p) + 0,2262r(f)} \quad (5.19)$$

O modelo ajustado apresenta coeficiente de determinação $R^2 = 92,32\%$, indicando um bom ajuste. Identificados os pontos de *leverages* e influentes, optou-se pela não eliminação.

É possível concluir que o aumento na comunalidade estimada e/ou na razão entre tamanho da amostra e o número de variáveis reduzem a REQMmax. Por outro lado, o aumento no número de fatores aumenta o valor da raiz quadrada do erro quadrático médio relativa.

5.7 MODELOS AJUSTADOS PARA O MAIOR COEFICIENTE DE VARIAÇÃO E MAIOR RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DOS AUTOVALORES, AUTOVETORES, CARGAS FATORIAIS E COMUNALIDADES

É importante ressaltar que a utilização da Análise Fatorial em dados amostrais requer, inicialmente, a avaliação da estrutura de correlação do vetor de dados. Deve-se, portanto, aplicar o Teste de Esfericidade de Bartlett para a comprovação de que a matriz de correlação da população difere de uma matriz identidade. Outra medida a ser calculada é a Medida de Adequabilidade da Amostra de Kaiser-Meyer-Olkin (MSA), cujo valor deve ser superior a 0,5.

O quadro adiante apresenta os modelos ajustados para o maior coeficiente de variação e a maior raiz quadrada do erro quadrático médio relativa, das estimativas dos autovalores, autovetores, cargas fatoriais e comunalidades. Também estão apresentados os coeficientes de determinação, R^2 , indicando que as qualidades dos ajustes são muito boas. Assim, os modelos possibilitam estimar a precisão relativa e o erro total relativo das estimativas dos parâmetros do modelo fatorial ortogonal, estimado pelo Método das Componentes Principais.

QUADRO 21 - MODELOS AJUSTADOS PARA O MAIOR COEFICIENTE DE VARIAÇÃO E A MAIOR RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA E OS COEFICIENTES DE DETERMINAÇÃO, SEGUNDO OS ESTIMADORES DO MODELO FATORIAL ORTOGONAL

ESTIMADORES	MODELOS AJUSTADOS	R^2 (%)
AUTOVALOR	1 Coeficiente de Variação: $\log CV_{\max \text{ autoval}} = -0,4696 \text{ rautoval} - 0,0289 \text{ r(n)} - 0,3329 \text{ expm}$ Na forma exponencial: $CV_{\max \text{ autoval}} = 10^{-0,4696 \text{ rautoval} - 0,0289 \text{ r(n)} - 0,3329 \text{ expm}}$	99,03
	2 Raiz Quadrada do Erro Quadrático Médio Relativa: $\log REQM_{\max \text{ autoval}} = -0,4202 \text{ rautoval} - 0,0306 \text{ r(n)} - 0,4125 \text{ expm}$ Na forma exponencial: $REQM_{\max \text{ autoval}} = 10^{-0,4202 \text{ rautoval} - 0,0306 \text{ r(n)} - 0,4125 \text{ expm}}$	99,13
AUTOVETOR	3 Coeficiente de Variação: $\log CV_{\max \text{ autovet}} = -0,9282 \log \text{ autovet} - 0,0009 \text{ n} + 2,3360 \text{ r (f/p)} - 2,3731 \text{ expl}$ Na forma exponencial: $CV_{\max \text{ autovet}} = 10^{-0,9282 \log \text{ autovet} - 0,0009 \text{ n} + 2,3360 \text{ r (f/p)} - 2,3731 \text{ expl}}$	98,08
	4 Raiz Quadrada do Erro Quadrático Médio Relativa: $\log REQM_{\max \text{ autovet}} = -0,9336 \log \text{ autovet} - 0,0009 \text{ n} + 2,7525 \text{ r (f/p)} - 2,6261 \text{ expl}$ Na forma exponencial: $REQM_{\max \text{ autovet}} = 10^{-0,9336 \log \text{ autovet} - 0,0009 \text{ n} + 2,7525 \text{ r (f/p)} - 2,6261 \text{ expl}}$	97,71
CARGA FATORIAL	5 Coeficiente de Variação: $\log CV_{\max \text{ car}} = -0,9183 \log \text{ car} - 0,0009 \text{ n} + 2,4407 \text{ r (f/p)} - 2,3103 \text{ expl}$ Na forma exponencial: $CV_{\max \text{ car}} = 10^{-0,9183 \log \text{ car} - 0,0009 \text{ n} + 2,4407 \text{ r (f/p)} - 2,3103 \text{ expl}}$	98,10
	6 Raiz Quadrada do Erro Quadrático Médio Relativa: $\log REQM_{\max \text{ car}} = -0,9216 \log \text{ car} - 0,0009 \text{ n} + 2,8681 \text{ r (f/p)} - 2,5651 \text{ expl}$ Na forma exponencial: $REQM_{\max \text{ car}} = 10^{-0,9216 \log \text{ car} - 0,0009 \text{ n} + 2,8681 \text{ r (f/p)} - 2,5651 \text{ expl}}$	97,67
COMUNALIDADE	7 Coeficiente de Variação: $\log CV_{\max \text{ comun}} = -1,3902 \text{ comun} - 0,0743 \text{ r(n/p)} + 0,1261 \text{ r (f)}$ Na forma exponencial: $CV_{\max \text{ comun}} = 10^{-1,3902 \text{ comun} - 0,0743 \text{ r(n/p)} + 0,1261 \text{ r (f)}}$	94,40
	8 Raiz Quadrada do Erro Quadrático Médio Relativa: $\log REQM_{\max \text{ comun}} = -1,5276 \text{ comun} - 0,0901 \text{ r(n/p)} + 0,2262 \text{ r (f)}$ Na forma exponencial: $REQM_{\max \text{ comun}} = 10^{-1,5276 \text{ comun} - 0,0901 \text{ r(n/p)} + 0,2262 \text{ r (f)}}$	92,32

Baseado no quadro 21, apresentado anteriormente, onde se encontram os modelos ajustados para estimar a precisão e o erro total relativos das estimativas dos parâmetros do modelo fatorial ortogonal, bem como os respectivos coeficientes de determinação (R^2), tem-se que:

- a) o coeficiente de variação e a raiz quadrada do erro quadrático médio relativa das estimativas dos autovalores (modelos 1 e 2) são funções das variáveis explicativas: raiz quadrada da estimativa do autovalor (r_{autoval}), explicação média dos fatores estimada (expm) e raiz quadrada do tamanho da amostra ($r(n)$). O aumento em pelo menos uma dessas variáveis explicativas contribui para a redução no coeficiente de variação e na raiz quadrada do erro quadrático médio relativa da estimativa do autovalor;
- b) o coeficiente de variação e a raiz quadrada do erro quadrático médio relativa das estimativas dos autovetores (modelos 3 e 4) são funções das variáveis explicativas: logaritmo decimal do valor absoluto da estimativa do autovetor (\log_{autovet}), explicação total dos fatores estimada, tamanho da amostra e raiz quadrada da razão entre o número de fatores e o de variáveis ($r(f/p)$). Quanto maior o valor de pelo menos uma das três primeiras variáveis explicativas, menores serão o coeficiente de variação e a raiz quadrada do erro quadrático médio relativa da estimativa do autovetor. Por outro lado, quanto maior a razão ($r(f/p)$) maiores serão o coeficiente de variação e a raiz quadrada do erro quadrático médio relativa.
- c) o coeficiente de variação e a raiz quadrada do erro quadrático médio relativa das estimativas das cargas fatoriais (modelos 5 e 6) indicam que são funções das variáveis explicativas: logaritmo decimal do valor absoluto da estimativa da carga fatorial (\log_{car}), explicação total dos fatores estimada, tamanho da amostra e raiz quadrada da razão entre o número de fatores e o de variáveis ($r(f/p)$). As variáveis explicativas:

logaritmo decimal do valor absoluto da estimativa da carga fatorial ($\log_{10} car$); explicação total dos fatores estimada e tamanho da amostra contribuem negativamente nas variáveis respostas, ou seja, quanto maiores, menores serão o coeficiente de variação e a raiz quadrada do erro quadrático médio relativa. No entanto, quanto maior a razão ($r(f/p)$), maiores serão o CVmax e REQMmax das estimativas das cargas fatoriais.

- d) o coeficiente de variação e a raiz quadrada do erro quadrático médio das estimativas das comunalidades (modelos 7 e 8) são funções das variáveis explicativas: estimativa da comunalidade ($comun$), raiz quadrada da razão entre o tamanho da amostra e o número de variáveis ($r(n/p)$) e raiz quadrada do número de fatores ($r(f)$). Quanto maiores as duas primeiras variáveis explicativas, menores serão o CVmax e REQMmax das estimativas das comunalidades. Por outro lado, quanto maior o número de fatores maiores serão o CVmax e REQMmax.

CONCLUSÕES E RECOMENDAÇÕES

Apresentam-se, a seguir, os modelos ajustados para estimar a precisão e o erro total relativos das estimativas dos parâmetros do Modelo Fatorial Ortogonal.

a) precisão relativa das estimativas dos autovalores, autovetores, cargas fatoriais e comunalidades:

$$\log CV_{\max \text{ autoval}} = -0,4696 r_{\text{autoval}} - 0,0289 r(n) - 0,3329 \text{ expm};$$

$$\log CV_{\max \text{ autovet}} = -0,9282 \log \text{ autovet} - 0,0009 n + 2,3360 r(f/p) - 2,3731 \text{ expl};$$

$$\log CV_{\max \text{ car}} = -0,9183 \log \text{ car} - 0,0009 n + 2,4407 r(f/p) - 2,3103 \text{ expl};$$

$$\log CV_{\max \text{ comun}} = -1,3902 \text{ comun} - 0,0743 r(n/p) + 0,1261 r(f).$$

b) erro total relativo das estimativas dos autovalores, autovetores, cargas fatoriais e comunalidades:

$$\log REQ_{\max \text{ autoval}} = -0,4202 r_{\text{autoval}} - 0,0306 r(n) - 0,4125 \text{ expm};$$

$$\log REQ_{\max \text{ autovet}} = -0,9336 \log \text{ autovet} - 0,0009 n + 2,7525 r(f/p) - 2,6261 \text{ expl};$$

$$\log REQ_{\max \text{ car}} = -0,9216 \log \text{ car} - 0,0009 n + 2,8681 r(f/p) - 2,5651 \text{ expl};$$

$$\log REQ_{\max \text{ comun}} = -1,5276 \text{ comun} - 0,0901 r(n/p) + 0,2262 r(f).$$

c) como medida de avaliação da qualidade das estimativas, propõe-se o erro total relativo, pois este considera, além da variância, o viés. Constatou-se que o viés não é desprezível, principalmente nas estimativas dos autovetores e das cargas fatoriais. Quando o tamanho da amostra é pequeno e o número de variáveis é grande, o viés tende a aumentar.

Como conclusão final tem-se que a grandeza das estimativas dos autovalores, dos autovetores, das cargas fatoriais e das comunalidades é determinante, juntamente com o tamanho da amostra, o número de variáveis e de fatores, na precisão e no erro total relativos. A precisão e o erro total relativos das estimativas dos autovalores, autovetores e cargas fatoriais, além das variáveis explicativas citadas anteriormente, dependem também da estimativa da explicação dos fatores. Estimativas pequenas, em valor, tendem a apresentar grandes coeficientes de variação e, conseqüentemente, precisões baixas.

Espera-se que os resultados e as conclusões deste trabalho possam, de alguma forma, contribuir na utilização da Análise Fatorial em dados amostrais.

Recomendam-se outras pesquisas para avaliar a precisão e o erro total das estimativas dos autovalores, autovetores, cargas fatoriais e comunalidades, utilizando-se de:

- (i) amostras de tamanhos maiores, já que n (tamanho da amostra) é importante na qualidade das estimativas;
- (ii) amostras com maior número de variáveis (p);
- (iii) testes de significância para os coeficientes de correlação, principalmente para amostras pequenas;
- (iv) rotação ortogonal dos fatores e comparação da precisão e do erro total relativos das novas estimativas das cargas fatoriais, com as obtidas inicialmente.

REFERÊNCIAS

- ANDERSON, T. W. **An introduction to multivariate statistical analysis**. New York: John Wiley & Sons, 1958. 375 p.
- BARTLETT, M. S. Tests of significance in factor analysis. **British Journal of Psychology**, n. 3, p. 77-85, 1950.
- BRITO, Luiza T. de L. et al. Uso de análise multivariada na classificação das fontes hídricas subterrâneas da bacia hidrográfica do salitre. **Eng. Agric.**, v. 26, n. 1, p.36-44, jan-abr 2006.
- CLIFF, Norman; HAMBURGER, Charles D. The study of sampling errors in factor analysis by means of artificial experiments. **Psychological Bulletin**, v. 68, n. 6, p. 430-445, 1967.
- COCHRAN, William G. **Sampling techniques**. 3.ed. New York: John Wiley & Sons, 1977. 428 p.
- COSTA, Giovani G. O. **Um procedimento inferencial para análise fatorial utilizando as técnicas *bootstrap* e *jackknife***: construção de intervalos de confiança de testes de hipóteses. Rio de Janeiro, 2006. 196 p. Tese (doutorado). Departamento de Engenharia Elétrica, PUC-RIO.
- EHLERS, Ricardo S. **Métodos computacionalmente intensivos em estatística**. Curitiba: UFPR. Junho de 2003. Notas de aula.
- FABRIGAR, Leandre R. et al. Evaluating the use of exploratory factor analysis in psychological research. **Psychological Methods**, v. 4, n. 3, p. 272-299, 1999.
- FACHEL, Jandira M. G. **Análise fatorial**. São Paulo, 1976. Dissertação (Mestrado) – IME, USP.
- FURTADO, Emerson Marcos et al. Ranqueamento de faxinais do Estado do Paraná. **Revista de Ciências Exatas e Naturais**, v. 5, n. 1, jan.-jun. 2003.
- GIRÃO, Enio G. et al. Seleção dos indicadores da qualidade de água no Rio Jaibaras pelo emprego da análise da componente principal. **Revista Ciência Agronômica**, v. 38, n. 1, p.17-24, 2007.
- GUJARATI, Domador N. **Econometria básica**. São Paulo: Pearson Education do Brasil, 2000, 846 p.
- HAIR, Joseph F. et al. **Multivariate data analysis**. 5. ed. New Jersey: Prentice Hall, 1998. 745 p.
- JOHNSON, Richard A.; WICHERN, Dean W. **Applied multivariate statistical analysis**. 2. ed. New Jersey: Prentice Hall International, 1988. 607 p.
- KAISER, Henry F. The varimax criterion for analytic rotation in factor analysis. **Psychometrika**, v. 33, n. 3, p. 187-200, 1958.

KAISER, Henry F. The application of electronic computers to factor analysis. **Educational and Psychological Measurement**, v. 20, n. 1, 1960.

KAISER, Henry F. A second generation little jiffy. **Psychometrika**, v. 35, n. 4, dez. 1970.

KISH, Leslie. **Survey sampling**. New York: John Wiley & Sons, 1965, 643 p.

KURTZ, Fabio C. et al. Zoneamento ambiental em pantanais (banhados). **Rev. Bras. Eng. Agric. Ambient.** v. 5, n. 2, maio-ago 2001.

LAWLEY, D. N. The estimation of factor loadings by the method of maximum likelihood. **Proceedings of the Royal Society of Edinburg**, v. 60, p. 64-82, 1940.

LAWLEY, D. N. The application of the maximum likelihood method for factor analysis. **British Journal of Psychology**, v. 33, p. 172-175, 1943.

LAWLEY, D. N.; MAXWELL, A. E. Factor analysis as a statistical method. **The Statistician**, v. 12, n. 3, p. 209-229, 1962.

MARDIA, K. V. Measures of multivariate skewness and kurtosis with applications. **Biometrika**, v. 57, n. 3, p. 519-530, 1970.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. **Multivariate analysis**. New York: Academic Press, 1982, 521 p.

MARQUES, Marcos A. Mendes. **Aplicação da análise multivariada no estudo da infraestrutura dos serviços de saúde dos municípios paranaenses**. Curitiba, 2006. 133 p. Dissertação (mestrado). Setores de Ciências Exatas e de Tecnologia, UFPR.

MONTEGOMERY, Douglas C.; PECK, Elizabeth A. **Introduction to linear regression analysis**. New York: John Wiley & Sons, 1982, 504 p.

MORRISON, Donald F. **Multivariate statistical methods**. New York: McGraw-Hill, 1976, 451 p.

MÜLLER, Sonia I. M. G; CHAVES NETO, Anselmo. Programa de técnicas integradas para análise e avaliação de fornecedores. **RNTI – Revista Negócios e Tecnologia da Informação**, v.2, p.8, 2007.

NETER, John et al. **Applied linear statistical models**. New York: McGraw-Hill, 1996, 1.408p.

OSBORNE, Jason W.; COSTELLO, Anna B. Sample size and subject to item ratio in principal components analysis. **Practical Assessment, Research and Evaluation**, v. 9, n.11, 2004. Disponível em: <http://PAREonline.net/getvn.asp?v=98&n=11>. Acess em: 12/05/2007.

PRESS, S. James. **Applied multivariate analysis: using bayesian and frequentist methods of inference**. Florida: Robert E. Krieger Publishing, 1982. 600 p.

SILVA, Nilza Nunes da. **Amostragem Probabilística: um curso introdutório**, São Paulo: Editora da Universidade de São Paulo, 1998, 125 p.

SPEARMAN, C. H. General intelligence objectively determined and measured. **American Journal of Psychology**, v. 15, p. 201-293, 1904.

SRIVASTAVA, M. S.; CARTER, E. M. **An introduction to applied multivariate statistics**. New York: Elsevier Science Publishing, 1983. 394 p.

STEINER, M. T. A.; SOUZA, R. C. Testing statistical methods in pattern recognition via simulacón. **Investigación Operativa**. Rio de Janeiro, v. 9, n. 1, 2, 3, p. 49-70, 2000.

THURSTONE, L. L. The vectors of mind. Psychological Review. v. 41, p. 1-32, 1934. In: **Classics in the History of Psychology**. Disponível em: <<http://psychclassics.yorku.ca/thurstone>. Acesso em: 28/02/2007>.

VILLWOCK, Rosangela et al. **Análise multivariada aplicada à análise de dados de instrumentação de barragens**. In: SIMPÓSIO DE ENGENHERIA DE PRODUÇÃO, 14, 2007, Bauru. Anais... Bauru: UNESP, 2007.

ZANELLA, Andreia et al. Identificação de fatores que influenciam na satisfação dos clientes de um clube recreativo por meio da análise fatorial. **GEPROS**. Ano 2, v. 3, maio-jun. 2007.

BIBLIOGRAFIAS CONSULTADAS

An Introduction to R. Notes on R: A programming for data analysis and graphics. v. 2.3.1, 2006.

CHATFIELD, Christopher; COLLINS, A. J. **Introduction to multivariate analysis.** London: Chapman & Hall, 1980. 246 p.

CHAVES NETO, Anselmo. **Probabilidade e estatística matemática II.** Curitiba: UFPR, 1º semestre de 2002. Notas de aula.

CHAVES NETO, Anselmo. **Análise multivariada aplicada à pesquisa.** Curitiba: UFPR, 2º semestre de 2002. Notas de aula.

CHAVES NETO, Anselmo. **Probabilidade e estatística matemática I.** Curitiba: UFPR, 1º semestre de 2003. Notas de aula.

Comandos do R para simulações Monte Carlo, disponível em:
<http://www.est.ufpr.br/~ehlers/ce718/praticas>

KENDALL, Maurice. **Multivariate analysis.** 2. ed. London: Charles Griffin, 1980. 210 p.

LAWLEY, D. N. Tests of significance for the latent roots of covariance and correlation matrices. **Biometrika**, v. 43, p. 128-136, 1956.

MOOD, Alexander M.; GRAYBILL, Franklin A.; BOES, Duane C. **Introduction to the theory of statistics.** 3. ed. Singapore: McGraw-Hill Book, 1974, 564 p.

SRIVASTAVA, M. S. On fixed-width confidence bounds for regression parameters and mean vector. **Journal of the Royal Statistical Society. B**, v. 29, p. 132-140, 1967.

APÊNDICE 1 - PARÂMETROS PARA SIMULAÇÃO MONTE CARLO

População 1 - 5 variáveis

Médias:

70,00 75,00 80,00 20,00 10,00

Variâncias:

80,00 69,00 90,00 35,00 6,00

Correlações:

X_1 e $X_2 = 0,90$ X_1 e $X_3 = 0,30$ X_2 e $X_4 = 0,95$ X_3 e $X_5 = 0,70$

Semente: 30 (para gerar as variáveis normais bivariadas)

30 (para gerar as variáveis normais multivariadas)

População 2 - 5 variáveis

Médias:

70,00 75,00 80,00 20,00 10,00

Variâncias:

80,00 65,00 90,00 35,00 8,00

Correlações:

X_1 e $X_2 = 0,90$ X_1 e $X_3 = 0,30$ X_2 e $X_4 = 0,95$ X_3 e $X_5 = 0,20$

Semente: 31 (para gerar as variáveis normais bivariadas)

31 (para gerar as variáveis normais multivariadas)

População 3 - 5 variáveis

Médias e variâncias iguais às da população 2, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,20$ X_1 e $X_3 = 0,95$ X_2 e $X_4 = 0,90$ X_3 e $X_5 = 0,20$

Semente: 31 (para gerar as variáveis normais bivariadas)

31 (para gerar as variáveis normais multivariadas)

População 4 - 5 variáveis

Médias:

70,00 75,00 80,00 20,00 10,00

Variâncias:

80,00 79,00 90,00 25,00 8,00

Correlações:

 X_1 e $X_2 = 0,90$ X_1 e $X_3 = 0,30$ X_2 e $X_4 = 0,95$ X_3 e $X_5 = 0,70$

Semente: 40 (para gerar as variáveis normais bivariadas)

40 (para gerar as variáveis normais multivariadas)

População 5 - 5 variáveis

Médias:

16,00 5,00 155,00 89,00 2.676,00

Variâncias:

6,00 2,00 1.370,00 600,00 365.713,00

Correlações:

 X_1 e $X_2 = 0,21$ X_1 e $X_3 = -0,21$ X_2 e $X_4 = 0,78$ X_3 e $X_5 = 0,91$

Semente: 30 (para gerar as variáveis normais bivariadas)

40 (para gerar as variáveis normais multivariadas)

População 6 - 5 variáveis

Médias e variâncias iguais às da população 5, sendo algumas correlações diferentes.

Correlações:

 X_1 e $X_2 = 0,80$ X_1 e $X_3 = -0,21$ X_2 e $X_4 = 0,78$ X_3 e $X_5 = 0,91$

Semente: 30 (para gerar as variáveis normais bivariadas)

40 (para gerar as variáveis normais multivariadas)

População 7 - 6 variáveis

Médias:

3,99 3,44 2,86 1,44 3,04 2,26

Variâncias:

0,37 0,20 0,79 0,17 0,39 0,24

Correlações:

X_1 e $X_2 = 0,60$ X_1 e $X_3 = 0,12$ X_2 e $X_4 = 0,28$ X_3 e $X_5 = 0,91$
 X_3 e $X_6 = 0,85$

Semente: 30 (para gerar as variáveis normais bivariadas)

40 (para gerar as variáveis normais multivariadas)

População 8 - 6 variáveis

Médias e variâncias iguais às da população 7, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,70$ X_2 e $X_3 = 0,55$ X_3 e $X_4 = -0,92$ X_4 e $X_5 = -0,65$
 X_5 e $X_6 = 0,95$

Semente: 39 (para gerar as variáveis normais bivariadas)

39 (para gerar as variáveis normais multivariadas)

População 9 - 6 variáveis

Médias e variâncias iguais às da população 7, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,76$ X_2 e $X_3 = 0,30$ X_3 e $X_4 = 0,62$ X_4 e $X_5 = -0,25$
 X_5 e $X_6 = 0,89$

Semente: 39 (para gerar as variáveis normais bivariadas)

39 (para gerar as variáveis normais multivariadas)

População 10 - 6 variáveis

Médias e variâncias iguais às da população 7, sendo algumas correlações diferentes.

Correlações:

$$X_1 \text{ e } X_2 = 0,76$$

$$X_2 \text{ e } X_3 = 0,90$$

$$X_3 \text{ e } X_4 = 0,42$$

$$X_4 \text{ e } X_5 = 0,85$$

$$X_5 \text{ e } X_6 = 0,25$$

Semente: 39 (para gerar as variáveis normais bivariadas)

39 (para gerar as variáveis normais multivariadas)

População 11 - 6 variáveis

Médias e variâncias iguais às da população 7, sendo algumas correlações diferentes.

Correlações:

$$X_1 \text{ e } X_2 = 0,05$$

$$X_1 \text{ e } X_3 = 0,90$$

$$X_2 \text{ e } X_4 = 0,10$$

$$X_3 \text{ e } X_5 = 0,70$$

$$X_3 \text{ e } X_6 = 0,85$$

Semente: 30 (para gerar as variáveis normais bivariadas)

40 (para gerar as variáveis normais multivariadas)

População 12 - 7 variáveis

Médias:

1.467,00 294,00 582,00 578,00 62,00 87,00 69,00

Variâncias:

128.098,00 8.000,00 40.412,00 34.787,00 385,00 20,00 360,00

Correlações:

$$X_1 \text{ e } X_2 = -0,80$$

$$X_1 \text{ e } X_3 = 0,30$$

$$X_2 \text{ e } X_4 = 0,85$$

$$X_3 \text{ e } X_5 = 0,98$$

$$X_3 \text{ e } X_6 = 0,60$$

$$X_4 \text{ e } X_7 = 0,20$$

Semente: 34 (para gerar as variáveis normais bivariadas)

34 (para gerar as variáveis normais multivariadas)

População 13 - 7 variáveis

Médias e variâncias iguais às da população 12, sendo algumas correlações diferentes.

Correlações:

$$X_1 \text{ e } X_2 = -0,80$$

$$X_2 \text{ e } X_3 = 0,50$$

$$X_3 \text{ e } X_4 = -0,95$$

$$X_4 \text{ e } X_5 = 0,68$$

$$X_5 \text{ e } X_6 = 0,75$$

$$X_6 \text{ e } X_7 = 0,25$$

Semente: 34 (para gerar as variáveis normais bivariadas)

34 (para gerar as variáveis normais multivariadas)

População 14 - 7 variáveis

Médias e variâncias iguais às da população 12, sendo algumas correlações diferentes.

Correlações:

$$X_1 \text{ e } X_2 = 0,80$$

$$X_2 \text{ e } X_3 = 0,50$$

$$X_3 \text{ e } X_4 = -0,79$$

$$X_4 \text{ e } X_5 = 0,88$$

$$X_5 \text{ e } X_6 = 0,30$$

$$X_6 \text{ e } X_7 = 0,82$$

Semente: 34 (para gerar as variáveis normais bivariadas)

34 (para gerar as variáveis normais multivariadas)

População 15 - 7 variáveis

Médias e variâncias iguais às da população 12, sendo algumas correlações diferentes.

Correlações:

$$X_1 \text{ e } X_2 = 0,80$$

$$X_2 \text{ e } X_3 = 0,62$$

$$X_3 \text{ e } X_4 = -0,75$$

$$X_4 \text{ e } X_5 = 0,88$$

$$X_5 \text{ e } X_6 = 0,92$$

$$X_6 \text{ e } X_7 = 0,65$$

Semente: 34 (para gerar as variáveis normais bivariadas)

34 (para gerar as variáveis normais multivariadas)

População 16 - 7 variáveis

Médias e variâncias iguais às da população 12, sendo algumas correlações diferentes.

Correlações:

$$X_1 \text{ e } X_2 = -0,80$$

$$X_2 \text{ e } X_3 = 0,50$$

$$X_3 \text{ e } X_4 = -0,75$$

$$X_4 \text{ e } X_5 = 0,88$$

$$X_5 \text{ e } X_6 = 0,70$$

$$X_6 \text{ e } X_7 = 0,92$$

Semente: 34 (para gerar as variáveis normais bivariadas)

34 (para gerar as variáveis normais multivariadas)

População 17 - 8 variáveis

Médias:

105,00 0,83 0,75 139,00 5,66 79,00 6,75 6,92

Variâncias:

1.199,00 0,09 0,05 1.064,00 4,22 820,00 5,00 3,39

Correlações:

$$X_1 \text{ e } X_2 = 0,80$$

$$X_1 \text{ e } X_3 = 0,70$$

$$X_2 \text{ e } X_4 = 0,89$$

$$X_3 \text{ e } X_5 = 0,78$$

$$X_3 \text{ e } X_6 = 0,85$$

$$X_4 \text{ e } X_7 = 0,87$$

$$X_5 \text{ e } X_8 = 0,88$$

Semente: 86 (para gerar as variáveis normais bivariadas)

86 (para gerar as variáveis normais multivariadas)

População 18 - 8 variáveis

Médias e variâncias iguais às da população 17, sendo algumas correlações diferentes.

Correlações:

$$X_1 \text{ e } X_2 = 0,80$$

$$X_1 \text{ e } X_3 = 0,70$$

$$X_2 \text{ e } X_4 = 0,89$$

$$X_3 \text{ e } X_5 = 0,48$$

$$X_3 \text{ e } X_6 = 0,85$$

$$X_4 \text{ e } X_7 = 0,87$$

$$X_5 \text{ e } X_8 = 0,88$$

Semente: 36 (para gerar as variáveis normais bivariadas)

36 (para gerar as variáveis normais multivariadas)

População 19 - 8 variáveis

Médias e variâncias iguais às da população 17, sendo algumas correlações diferentes.

Correlações:

$$X_1 \text{ e } X_2 = 0,80$$

$$X_1 \text{ e } X_3 = 0,70$$

$$X_2 \text{ e } X_4 = 0,58$$

$$X_3 \text{ e } X_5 = 0,42$$

$$X_3 \text{ e } X_6 = 0,88$$

$$X_4 \text{ e } X_7 = 0,50$$

$$X_5 \text{ e } X_8 = 0,90$$

Semente: 36 (para gerar as variáveis normais bivariadas)

36 (para gerar as variáveis normais multivariadas)

População 20 - 8 variáveis

Médias e variâncias iguais às da população 17, sendo algumas correlações diferentes.

Correlações:

$$X_1 \text{ e } X_2 = 0,80$$

$$X_1 \text{ e } X_3 = 0,70$$

$$X_2 \text{ e } X_4 = 0,58$$

$$X_3 \text{ e } X_5 = 0,42$$

$$X_3 \text{ e } X_6 = 0,88$$

$$X_4 \text{ e } X_7 = 0,60$$

$$X_5 \text{ e } X_8 = 0,85$$

Semente: 36 (para gerar as variáveis normais bivariadas)

36 (para gerar as variáveis normais multivariadas)

População 21 - 8 variáveis

Médias e variâncias iguais às da população 17, sendo algumas correlações diferentes.

Correlações:

$$X_1 \text{ e } X_2 = 0,80$$

$$X_1 \text{ e } X_3 = 0,70$$

$$X_2 \text{ e } X_4 = 0,28$$

$$X_3 \text{ e } X_5 = 0,62$$

$$X_3 \text{ e } X_6 = 0,78$$

$$X_4 \text{ e } X_7 = 0,60$$

$$X_5 \text{ e } X_8 = 0,85$$

Semente: 36 (para gerar as variáveis normais bivariadas)

36 (para gerar as variáveis normais multivariadas)

População 22 - 9 variáveis

Médias:

104,60 0,83 0,75 139,00 5,66 79,00 6,75 6,92 105,35

Variâncias:

1.199,00 0,09 0,05 1.064,00 4,22 820,00 5,00 3,39 802,00

Correlações:

X_1 e $X_2 = 0,80$	X_1 e $X_3 = 0,75$	X_2 e $X_4 = 0,65$	X_3 e $X_5 = 0,32$
X_3 e $X_6 = 0,68$	X_4 e $X_7 = 0,55$	X_5 e $X_8 = 0,85$	X_6 e $X_9 = 0,78$

Semente: 76 (para gerar as variáveis normais bivariadas)

76 (para gerar as variáveis normais multivariadas)

População 23 - 9 variáveis

Médias e variâncias iguais às da população 22, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,86$	X_1 e $X_3 = 0,50$	X_2 e $X_4 = 0,28$	X_3 e $X_5 = -0,92$
X_3 e $X_6 = 0,48$	X_4 e $X_7 = -0,85$	X_5 e $X_8 = 0,73$	X_6 e $X_9 = 0,45$

Semente: 76 (para gerar as variáveis normais bivariadas)

76 (para gerar as variáveis normais multivariadas)

População 24 - 9 variáveis

Médias e variâncias iguais às da população 22, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,80$	X_1 e $X_3 = 0,55$	X_2 e $X_4 = 0,38$	X_3 e $X_5 = -0,32$
X_3 e $X_6 = 0,48$	X_4 e $X_7 = -0,85$	X_5 e $X_8 = 0,73$	X_6 e $X_9 = 0,75$

Semente: 76 (para gerar as variáveis normais bivariadas)

76 (para gerar as variáveis normais multivariadas)

População 25 - 9 variáveis

Médias e variâncias iguais às da população 22, sendo algumas correlações diferentes.

Correlações:

$$\begin{array}{llll} X_1 \text{ e } X_2 = 0,80 & X_1 \text{ e } X_3 = -0,20 & X_2 \text{ e } X_4 = 0,79 & X_3 \text{ e } X_5 = -0,32 \\ X_3 \text{ e } X_6 = 0,48 & X_4 \text{ e } X_7 = -0,85 & X_5 \text{ e } X_8 = 0,73 & X_6 \text{ e } X_9 = 0,75 \end{array}$$

Semente: 76 (para gerar as variáveis normais bivariadas)

76 (para gerar as variáveis normais multivariadas)

População 26 - 9 variáveis

Médias e variâncias iguais às da população 22, sendo algumas correlações diferentes.

Correlações:

$$\begin{array}{llll} X_1 \text{ e } X_2 = 0,80 & X_1 \text{ e } X_3 = -0,20 & X_2 \text{ e } X_4 = 0,79 & X_3 \text{ e } X_5 = -0,62 \\ X_3 \text{ e } X_6 = 0,48 & X_4 \text{ e } X_7 = -0,85 & X_5 \text{ e } X_8 = 0,73 & X_6 \text{ e } X_9 = 0,75 \end{array}$$

Semente: 76 (para gerar as variáveis normais bivariadas)

76 (para gerar as variáveis normais multivariadas)

População 27 - 10 variáveis

Médias:

104,60 0,83 0,75 139,00 9,66 79,00 6,75 6,92 139,00 4,91

Variâncias:

1.199,00 0,09 0,07 964,00 6,22 229,00 5,25 4,39 830,00 1,30

Correlações:

$$\begin{array}{llll} X_1 \text{ e } X_2 = 0,80 & X_1 \text{ e } X_3 = 0,70 & X_2 \text{ e } X_4 = 0,38 & X_3 \text{ e } X_5 = 0,42 \\ X_3 \text{ e } X_6 = 0,68 & X_4 \text{ e } X_7 = 0,60 & X_5 \text{ e } X_8 = 0,55 & X_6 \text{ e } X_9 = 0,78 \\ X_9 \text{ e } X_{10} = 0,89 & & & \end{array}$$

Semente: 69 (para gerar as variáveis normais bivariadas)

69 (para gerar as variáveis normais multivariadas)

População 28 - 10 variáveis

Médias e variâncias iguais às da população 27, sendo algumas correlações diferentes.

Correlações:

$$X_1 \text{ e } X_2 = 0,80$$

$$X_1 \text{ e } X_3 = 0,70$$

$$X_2 \text{ e } X_4 = 0,38$$

$$X_3 \text{ e } X_5 = 0,42$$

$$X_3 \text{ e } X_6 = 0,48$$

$$X_4 \text{ e } X_7 = 0,60$$

$$X_5 \text{ e } X_8 = 0,55$$

$$X_6 \text{ e } X_9 = 0,75$$

$$X_9 \text{ e } X_{10} = 0,82$$

Semente: 69 (para gerar as variáveis normais bivariadas)

69 (para gerar as variáveis normais multivariadas)

População 29 - 10 variáveis

Médias e variâncias iguais às da população 27, sendo algumas correlações diferentes.

Correlações:

$$X_1 \text{ e } X_2 = 0,80$$

$$X_1 \text{ e } X_3 = 0,70$$

$$X_2 \text{ e } X_4 = 0,68$$

$$X_3 \text{ e } X_5 = 0,42$$

$$X_3 \text{ e } X_6 = 0,88$$

$$X_4 \text{ e } X_7 = 0,60$$

$$X_5 \text{ e } X_8 = 0,55$$

$$X_6 \text{ e } X_9 = 0,75$$

$$X_9 \text{ e } X_{10} = 0,82$$

Semente: 69 (para gerar as variáveis normais bivariadas)

69 (para gerar as variáveis normais multivariadas)

População 30 - 10 variáveis

Médias e variâncias iguais às da população 27, sendo algumas correlações diferentes.

Correlações:

$$X_1 \text{ e } X_2 = 0,90$$

$$X_1 \text{ e } X_3 = 0,80$$

$$X_2 \text{ e } X_4 = 0,58$$

$$X_3 \text{ e } X_5 = 0,42$$

$$X_3 \text{ e } X_6 = 0,88$$

$$X_4 \text{ e } X_7 = 0,60$$

$$X_5 \text{ e } X_8 = 0,35$$

$$X_6 \text{ e } X_9 = 0,75$$

$$X_9 \text{ e } X_{10} = 0,82$$

Semente: 69 (para gerar as variáveis normais bivariadas)

69 (para gerar as variáveis normais multivariadas)

População 31 - 10 variáveis

Médias e variâncias iguais às da população 27, sendo algumas correlações diferentes.

Correlações:

$$\begin{array}{llll} X_1 \text{ e } X_2 = 0,30 & X_1 \text{ e } X_3 = 0,80 & X_2 \text{ e } X_4 = 0,58 & X_3 \text{ e } X_5 = 0,42 \\ X_3 \text{ e } X_6 = 0,88 & X_4 \text{ e } X_7 = 0,60 & X_5 \text{ e } X_8 = 0,35 & X_6 \text{ e } X_9 = 0,75 \\ X_9 \text{ e } X_{10} = 0,82 & & & \end{array}$$

Semente: 69 (para gerar as variáveis normais bivariadas)

69 (para gerar as variáveis normais multivariadas)

População 32 - 11 variáveis

Médias:

3,97 0,83 0,75 18,99 5,66 79,27 6,75 53,17 138,99 4,91
10,46

Variâncias:

1,12 0,02 0,03 24,20 4,22 828,87 6,99 112,26 1.664,20 1,30
15,02

Correlações:

$$\begin{array}{llll} X_1 \text{ e } X_2 = 0,80 & X_1 \text{ e } X_3 = 0,70 & X_2 \text{ e } X_4 = 0,38 & X_3 \text{ e } X_5 = 0,42 \\ X_3 \text{ e } X_6 = 0,68 & X_4 \text{ e } X_7 = 0,60 & X_5 \text{ e } X_8 = 0,55 & X_6 \text{ e } X_9 = 0,78 \\ X_9 \text{ e } X_{10} = 0,89 & X_{10} \text{ e } X_{11} = 0,30 & & \end{array}$$

Semente: 69 (para gerar as variáveis normais bivariadas)

69 (para gerar as variáveis normais multivariadas)

População 33 - 11 variáveis

Médias e variâncias iguais às da população 32, sendo algumas correlações diferentes.

Correlações:

$$\begin{array}{llll} X_1 \text{ e } X_2 = 0,80 & X_1 \text{ e } X_3 = 0,82 & X_2 \text{ e } X_4 = 0,38 & X_3 \text{ e } X_5 = 0,42 \\ X_3 \text{ e } X_6 = 0,40 & X_4 \text{ e } X_7 = 0,60 & X_5 \text{ e } X_8 = 0,55 & X_6 \text{ e } X_9 = 0,78 \\ X_9 \text{ e } X_{10} = 0,89 & X_{10} \text{ e } X_{11} = 0,30 & & \end{array}$$

Semente: 78 (para gerar as variáveis normais bivariadas)

78 (para gerar as variáveis normais multivariadas)

População 34 - 11 variáveis

Médias e variâncias iguais às da população 32, sendo algumas correlações diferentes.

Correlações:

$$X_1 \text{ e } X_2 = 0,60$$

$$X_1 \text{ e } X_3 = 0,82$$

$$X_2 \text{ e } X_4 = 0,38$$

$$X_3 \text{ e } X_5 = 0,42$$

$$X_3 \text{ e } X_6 = 0,25$$

$$X_4 \text{ e } X_7 = 0,88$$

$$X_5 \text{ e } X_8 = 0,55$$

$$X_6 \text{ e } X_9 = 0,78$$

$$X_9 \text{ e } X_{10} = 0,89$$

$$X_{10} \text{ e } X_{11} = 0,30$$

Semente: 78 (para gerar as variáveis normais bivariadas)

78 (para gerar as variáveis normais multivariadas)

População 35 - 11 variáveis

Médias e variâncias iguais às da população 32, sendo algumas correlações diferentes.

Correlações:

$$X_1 \text{ e } X_2 = 0,70$$

$$X_1 \text{ e } X_3 = 0,82$$

$$X_2 \text{ e } X_4 = 0,38$$

$$X_3 \text{ e } X_5 = 0,85$$

$$X_3 \text{ e } X_6 = 0,25$$

$$X_4 \text{ e } X_7 = 0,88$$

$$X_5 \text{ e } X_8 = 0,55$$

$$X_6 \text{ e } X_9 = 0,78$$

$$X_9 \text{ e } X_{10} = 0,89$$

$$X_{10} \text{ e } X_{11} = 0,30$$

Semente: 78 (para gerar as variáveis normais bivariadas)

78 (para gerar as variáveis normais multivariadas)

População 36 - 11 variáveis

Médias e variâncias iguais às da população 32, sendo algumas correlações diferentes.

Correlações:

$$X_1 \text{ e } X_2 = 0,70$$

$$X_1 \text{ e } X_3 = 0,82$$

$$X_2 \text{ e } X_4 = 0,38$$

$$X_3 \text{ e } X_5 = 0,85$$

$$X_3 \text{ e } X_6 = 0,65$$

$$X_4 \text{ e } X_7 = 0,88$$

$$X_5 \text{ e } X_8 = 0,55$$

$$X_6 \text{ e } X_9 = 0,78$$

$$X_9 \text{ e } X_{10} = 0,89$$

$$X_{10} \text{ e } X_{11} = 0,30$$

Semente: 78 (para gerar as variáveis normais bivariadas)

78 (para gerar as variáveis normais multivariadas)

População 37 - 12 variáveis

Médias:

3,97	0,83	0,75	18,99	5,66	79,27	6,75	53,17	138,99	4,91
10,46	15,76								

Variâncias:

1,12	0,02	0,03	24,20	4,22	828,87	5,99	112,26	1.664,20	1,30
15,02	5,40								

Correlações:

X_1 e $X_2 = 0,70$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,38$	X_3 e $X_5 = 0,85$
X_3 e $X_6 = 0,25$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,85$	

Semente: 48 (para gerar as variáveis normais bivariadas)

48 (para gerar as variáveis normais multivariadas)

População 38 - 12 variáveis

Médias e variâncias iguais às da população 37, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,50$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,38$	X_3 e $X_5 = 0,85$
X_3 e $X_6 = 0,25$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,45$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,85$	

Semente: 48 (para gerar as variáveis normais bivariadas)

48 (para gerar as variáveis normais multivariadas)

População 39 - 12 variáveis

Médias e variâncias iguais às da população 37, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,40$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,85$
X_3 e $X_6 = 0,75$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,25$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,90$	

Semente: 48 (para gerar as variáveis normais bivariadas)

48 (para gerar as variáveis normais multivariadas)

População 40 - 12 variáveis

Médias e variâncias iguais às da população 37, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = -0,82$	X_1 e $X_3 = 0,20$	X_2 e $X_4 = 0,78$	X_3 e $X_5 = 0,85$
X_3 e $X_6 = 0,75$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,15$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,34$	

Semente: 48 (para gerar as variáveis normais bivariadas)

48 (para gerar as variáveis normais multivariadas)

População 41 - 12 variáveis

Médias e variâncias iguais às da população 37, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = -0,82$	X_1 e $X_3 = 0,20$	X_2 e $X_4 = 0,78$	X_3 e $X_5 = 0,85$
X_3 e $X_6 = 0,45$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,34$	

Semente: 48 (para gerar as variáveis normais bivariadas)

48 (para gerar as variáveis normais multivariadas)

População 42 - 13 variáveis

Médias:

3,97	0,83	0,75	18,99	5,66	79,27	6,75	53,17	138,99	4,91
10,46	15,76	55,54							

Variâncias:

1,12	0,02	0,03	24,20	4,22	628,87	4,99	112,26	1.664,20	1,30
10,02	5,40	320,20							

Correlações:

X_1 e $X_2 = 0,70$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,38$	X_3 e $X_5 = 0,85$
X_3 e $X_6 = 0,25$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,85$	X_{12} e $X_{13} = 0,70$

Semente: 18 (para gerar as variáveis normais bivariadas)

18 (para gerar as variáveis normais multivariadas)

População 43 - 13 variáveis

Médias e variâncias iguais às da população 42, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,70$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,38$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,25$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$

Semente: 18 (para gerar as variáveis normais bivariadas)

18 (para gerar as variáveis normais multivariadas)

População 44 - 13 variáveis

Médias e variâncias iguais às da população 42, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,70$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$

Semente: 18 (para gerar as variáveis normais bivariadas)

18 (para gerar as variáveis normais multivariadas)

População 45 - 13 variáveis

Médias e variâncias iguais às da população 42, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,70$	X_1 e $X_3 = 0,42$	X_2 e $X_4 = 0,86$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,95$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,85$	X_{12} e $X_{13} = 0,70$

Semente: 18 (para gerar as variáveis normais bivariadas)

18 (para gerar as variáveis normais multivariadas)

População 46 - 13 variáveis

Médias e variâncias iguais às da população 42, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,90$	X_1 e $X_3 = 0,62$	X_2 e $X_4 = 0,86$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,95$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,40$	X_{11} e $X_{12} = 0,85$	X_{12} e $X_{13} = 0,30$

Semente: 18 (para gerar as variáveis normais bivariadas)

18 (para gerar as variáveis normais multivariadas)

População 47 - 14 variáveis

Médias:

104,66	0,83	0,75	138,90	5,66	79,27	6,75	6,92	98,99	4,91
10,46	14,76	55,54	19,62						

Variâncias:

2.198,99	0,02	0,003	960,00	6,22	1.128,87	6,99	5,39	890,00	1,30
18,02	6,39	320,20	66,53						

Correlações:

X_1 e $X_2 = 0,70$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$			

Semente: 19 (para gerar as variáveis normais bivariadas)

19 (para gerar as variáveis normais multivariadas)

População 48 - 14 variáveis

Médias e variâncias iguais às da população 47, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = -0,35$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,48$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$			

Semente: 19 (para gerar as variáveis normais bivariadas)

19 (para gerar as variáveis normais multivariadas)

População 49 - 14 variáveis

Médias e variâncias iguais às da população 47, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = -0,35$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,98$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,86$	X_4 e $X_7 = 0,48$	X_5 e $X_8 = 0,25$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,90$
X_{13} e $X_{14} = 0,86$			

Semente: 19 (para gerar as variáveis normais bivariadas)

19 (para gerar as variáveis normais multivariadas)

População 50 - 14 variáveis

Médias e variâncias iguais às da população 47, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = -0,95$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,70$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,86$	X_4 e $X_7 = 0,48$	X_5 e $X_8 = 0,85$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$			

Semente: 99 (para gerar as variáveis normais bivariadas)

99 (para gerar as variáveis normais multivariadas)

População 51 - 14 variáveis

Médias e variâncias iguais às da população 47, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = -0,65$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,70$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,86$	X_4 e $X_7 = 0,48$	X_5 e $X_8 = 0,25$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,40$
X_{13} e $X_{14} = 0,86$			

Semente: 99 (para gerar as variáveis normais bivariadas)

99 (para gerar as variáveis normais multivariadas)

População 52 - 15 variáveis

Médias:

16,27	4,66	11,93	86,33	2.547,36	2,95	0,83	0,75	138,99	1,45
8,71	1,58	6,92	10,07	4,91					

Variâncias:

8,13	1,66	7,58	434,72	359.158,00	0,29	0,10	0,05	1.664,20	0,13
3,46	0,20	5,38	4,01	6,90					

Correlações:

X_1 e $X_2 = 0,70$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,15$		

Semente: 56 (para gerar as variáveis normais bivariadas)

56 (para gerar as variáveis normais multivariadas)

População 53 - 15 variáveis

Médias e variâncias iguais às da população 52, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,40$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,98$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,86$	X_4 e $X_7 = 0,28$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,95$		

Semente: 56 (para gerar as variáveis normais bivariadas)

56 (para gerar as variáveis normais multivariadas)

População 54 - 15 variáveis

Médias e variâncias iguais às da população 52, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,30$	X_1 e $X_3 = 0,72$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,25$
X_3 e $X_6 = 0,86$	X_4 e $X_7 = 0,28$	X_5 e $X_8 = 0,65$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,29$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,15$		

Semente: 56 (para gerar as variáveis normais bivariadas)

56 (para gerar as variáveis normais multivariadas)

População 55 - 15 variáveis

Médias e variâncias iguais às da população 52, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,80$	X_1 e $X_3 = 0,72$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,25$
X_3 e $X_6 = 0,86$	X_4 e $X_7 = 0,68$	X_5 e $X_8 = 0,85$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,29$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,15$		

Semente: 56 (para gerar as variáveis normais bivariadas)

56 (para gerar as variáveis normais multivariadas)

População 56 - 15 variáveis

Médias e variâncias iguais às da população 52, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,80$	X_1 e $X_3 = 0,72$	X_2 e $X_4 = -0,68$	X_3 e $X_5 = 0,70$
X_3 e $X_6 = 0,86$	X_4 e $X_7 = 0,38$	X_5 e $X_8 = 0,85$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,69$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,55$		

Semente: 56 (para gerar as variáveis normais bivariadas)

56 (para gerar as variáveis normais multivariadas)

População 57 - 16 variáveis

Médias:

104,64	0,84	0,75	168,90	5,66	79,27	6,75	6,91	138,99	4,91
10,46	14,76	55,53	19,62	45,97	29,68				

Variâncias:

1.698,90	0,095	0,027	3.642,00	3,22	828,87	6,99	5,39	1.664,00	1,31
18,02	6,39	520,19	66,54	61,85	54,46				

Correlações:

X_1 e $X_2 = 0,70$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,15$	X_{15} e $X_{16} = 0,70$	

Semente: 96 (para gerar as variáveis normais bivariadas)

96 (para gerar as variáveis normais multivariadas)

População 58 - 16 variáveis

Médias e variâncias iguais às da população 57, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,40$	X_1 e $X_3 = 0,89$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,29$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,75$	X_{15} e $X_{16} = 0,70$	

Semente: 96 (para gerar as variáveis normais bivariadas)

96 (para gerar as variáveis normais multivariadas)

População 59 - 16 variáveis

Médias e variâncias iguais às da população 57, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,40$	X_1 e $X_3 = 0,89$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,92$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,15$	X_{15} e $X_{16} = 0,60$	

Semente: 96 (para gerar as variáveis normais bivariadas)

96 (para gerar as variáveis normais multivariadas)

População 60 - 16 variáveis

Médias e variâncias iguais às da população 57, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,40$	X_1 e $X_3 = 0,89$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,26$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,92$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,15$	X_{15} e $X_{16} = 0,75$	

Semente: 96 (para gerar as variáveis normais bivariadas)

96 (para gerar as variáveis normais multivariadas)

População 61 - 16 variáveis

Médias e variâncias iguais às da população 57, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,90$	X_1 e $X_3 = 0,49$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,26$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,82$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,80$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,65$	X_{15} e $X_{16} = 0,75$	

Semente: 96 (para gerar as variáveis normais bivariadas)

96 (para gerar as variáveis normais multivariadas)

População 62 - 17 variáveis

Médias:

5,66	79,27	6,75	6,92	105,35	4,91	10,46	14,76	55,54	4,31
45,97	29,68	-1,08	19,85	2,52	15,00	8,04			

Variâncias:

6,22	1.128,87	6,99	5,39	1.702,13	1,30	18,02	6,39	420,20	1,03
21,85	22,46	0,08	22,73	1,08	25,03	2,33			

Correlações:

X_1 e $X_2 = 0,70$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,25$	X_{15} e $X_{16} = 0,70$	X_{16} e $X_{17} = -0,60$

Semente: 96 (para gerar as variáveis normais bivariadas)

96 (para gerar as variáveis normais multivariadas)

População 63 - 17 variáveis

Médias e variâncias iguais às da população 62, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,20$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,38$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,95$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,25$	X_{15} e $X_{16} = 0,70$	X_{16} e $X_{17} = -0,80$

Semente: 96 (para gerar as variáveis normais bivariadas)

96 (para gerar as variáveis normais multivariadas)

População 64 - 17 variáveis

Médias e variâncias iguais às da população 62, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,50$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,38$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,96$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,85$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,15$	X_{15} e $X_{16} = 0,30$	X_{16} e $X_{17} = -0,80$

Semente: 96 (para gerar as variáveis normais bivariadas)

96 (para gerar as variáveis normais multivariadas)

População 65 - 17 variáveis

Médias e variâncias iguais às da população 62, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,50$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,96$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,82$
X_9 e $X_{10} = 0,80$	X_{10} e $X_{11} = 0,73$	X_{11} e $X_{12} = 0,85$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,65$	X_{15} e $X_{16} = 0,30$	X_{16} e $X_{17} = -0,80$

Semente: 96 (para gerar as variáveis normais bivariadas)

96 (para gerar as variáveis normais multivariadas)

População 66 - 17 variáveis

Médias e variâncias iguais às da população 62, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,80$	X_1 e $X_3 = 0,52$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,90$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,82$
X_9 e $X_{10} = 0,80$	X_{10} e $X_{11} = 0,73$	X_{11} e $X_{12} = 0,85$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,65$	X_{15} e $X_{16} = 0,30$	X_{16} e $X_{17} = -0,70$

Semente: 96 (para gerar as variáveis normais bivariadas)

96 (para gerar as variáveis normais multivariadas)

População 67 - 18 variáveis

Médias:

1,45	8,71	6,75	6,92	105,35	4,91	10,46	14,76	7,24	4,31
45,97	29,68	0,54	19,85	1,54	15,00	1,37	8,04		

Variâncias:

0,13	3,46	6,99	5,38	1.702,13	1,30	18,02	6,39	3,92	1,03
81,85	42,46	0,01	22,73	0,16	25,03	0,18	2,33		

Correlações:

X_1 e $X_2 = 0,70$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,25$	X_{15} e $X_{16} = 0,70$	X_{16} e $X_{17} = -0,60$
X_{17} e $X_{18} = 0,40$			

Semente: 56 (para gerar as variáveis normais bivariadas)

56 (para gerar as variáveis normais multivariadas)

População 68 - 18 variáveis

Médias e variâncias iguais às da população 67, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,30$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,96$	X_4 e $X_7 = 0,68$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,25$	X_{15} e $X_{16} = 0,80$	X_{16} e $X_{17} = -0,60$
X_{17} e $X_{18} = 0,20$			

Semente: 56 (para gerar as variáveis normais bivariadas)

56 (para gerar as variáveis normais multivariadas)

População 69 - 18 variáveis

Médias e variâncias iguais às da população 67, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,65$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,85$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,28$	X_5 e $X_8 = 0,85$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,70$	X_{11} e $X_{12} = 0,85$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,75$	X_{15} e $X_{16} = 0,80$	X_{16} e $X_{17} = 0,40$
X_{17} e $X_{18} = 0,60$			

Semente: 56 (para gerar as variáveis normais bivariadas)

56 (para gerar as variáveis normais multivariadas)

População 70 - 18 variáveis

Médias e variâncias iguais às da população 67, apenas algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,85$	X_1 e $X_3 = 0,72$	X_2 e $X_4 = 0,88$	X_3 e $X_5 = 0,75$
X_3 e $X_6 = 0,80$	X_4 e $X_7 = 0,68$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,80$	X_{10} e $X_{11} = 0,90$	X_{11} e $X_{12} = 0,69$	X_{12} e $X_{13} = 0,87$
X_{13} e $X_{14} = 0,68$	X_{14} e $X_{15} = 0,75$	X_{15} e $X_{16} = 0,80$	X_{16} e $X_{17} = 0,40$
X_{17} e $X_{18} = 0,95$			

Semente: 56 (para gerar as variáveis normais bivariadas)

56 (para gerar as variáveis normais multivariadas)

População 71 - 18 variáveis

Médias e variâncias iguais às da população 67, apenas algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,40$	X_1 e $X_3 = 0,72$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,75$
X_3 e $X_6 = 0,86$	X_4 e $X_7 = 0,38$	X_5 e $X_8 = 0,65$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,25$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,26$	X_{14} e $X_{15} = 0,25$	X_{15} e $X_{16} = 0,60$	X_{16} e $X_{17} = -0,70$
X_{17} e $X_{18} = 0,30$			

Semente: 56 (para gerar as variáveis normais bivariadas)

56 (para gerar as variáveis normais multivariadas)

População 72 - 19 variáveis

Médias:

9,13	0,83	0,74	11,68	5,66	79,27	6,75	7,23	4,31	6,77
29,68	0,30	4,42	2,51	1,94	111,67	1,37	8,04	6,78	

Variâncias:

4,03	0,02	0,03	4,92	6,22	1.128,87	6,99	3,91	1,02	3,11
42,45	0,008	0,91	1,08	0,50	2.304,98	0,20	2,33	3,98	

Correlações:

X_1 e $X_2 = 0,70$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,25$	X_{15} e $X_{16} = 0,70$	X_{16} e $X_{17} = -0,60$
X_{17} e $X_{18} = 0,40$	X_{18} e $X_{19} = 0,87$		

Semente: 66 (para gerar as variáveis normais bivariadas)

66 (para gerar as variáveis normais multivariadas)

População 73 - 19 variáveis

Médias e variâncias iguais às da população 72, sendo algumas correlações diferentes.

Correlações:

$X_1 \text{ e } X_2 = 0,30$

$X_1 \text{ e } X_3 = 0,82$

$X_2 \text{ e } X_4 = 0,68$

$X_3 \text{ e } X_5 = 0,45$

$X_3 \text{ e } X_6 = 0,76$

$X_4 \text{ e } X_7 = 0,88$

$X_5 \text{ e } X_8 = 0,55$

$X_6 \text{ e } X_9 = 0,78$

$X_9 \text{ e } X_{10} = 0,92$

$X_{10} \text{ e } X_{11} = 0,30$

$X_{11} \text{ e } X_{12} = 0,65$

$X_{12} \text{ e } X_{13} = 0,80$

$X_{13} \text{ e } X_{14} = 0,86$

$X_{14} \text{ e } X_{15} = 0,25$

$X_{15} \text{ e } X_{16} = 0,50$

$X_{16} \text{ e } X_{17} = -0,60$

$X_{17} \text{ e } X_{18} = 0,20$

$X_{18} \text{ e } X_{19} = 0,87$

Semente: 66 (para gerar as variáveis normais bivariadas)

66 (para gerar as variáveis normais multivariadas)

População 74 - 19 variáveis

Médias e variâncias iguais às da população 72, sendo algumas correlações diferentes.

Correlações:

$X_1 \text{ e } X_2 = 0,50$

$X_1 \text{ e } X_3 = 0,82$

$X_2 \text{ e } X_4 = 0,78$

$X_3 \text{ e } X_5 = 0,45$

$X_3 \text{ e } X_6 = 0,76$

$X_4 \text{ e } X_7 = 0,88$

$X_5 \text{ e } X_8 = 0,55$

$X_6 \text{ e } X_9 = 0,78$

$X_9 \text{ e } X_{10} = 0,92$

$X_{10} \text{ e } X_{11} = 0,30$

$X_{11} \text{ e } X_{12} = 0,65$

$X_{12} \text{ e } X_{13} = 0,80$

$X_{13} \text{ e } X_{14} = 0,86$

$X_{14} \text{ e } X_{15} = 0,95$

$X_{15} \text{ e } X_{16} = 0,70$

$X_{16} \text{ e } X_{17} = -0,30$

$X_{17} \text{ e } X_{18} = 0,20$

$X_{18} \text{ e } X_{19} = 0,57$

Semente: 66 (para gerar as variáveis normais bivariadas)

66 (para gerar as variáveis normais multivariadas)

População 75 - 19 variáveis

Médias e variâncias iguais às da população 72, sendo algumas correlações diferentes.

Correlações:

$X_1 \text{ e } X_2 = 0,45$

$X_1 \text{ e } X_3 = 0,82$

$X_2 \text{ e } X_4 = 0,78$

$X_3 \text{ e } X_5 = 0,25$

$X_3 \text{ e } X_6 = 0,86$

$X_4 \text{ e } X_7 = 0,68$

$X_5 \text{ e } X_8 = 0,55$

$X_6 \text{ e } X_9 = 0,78$

$X_9 \text{ e } X_{10} = 0,87$

$X_{10} \text{ e } X_{11} = 0,30$

$X_{11} \text{ e } X_{12} = 0,65$

$X_{12} \text{ e } X_{13} = 0,80$

$X_{13} \text{ e } X_{14} = 0,86$

$X_{14} \text{ e } X_{15} = 0,65$

$X_{15} \text{ e } X_{16} = 0,70$

$X_{16} \text{ e } X_{17} = -0,30$

$X_{17} \text{ e } X_{18} = 0,60$

$X_{18} \text{ e } X_{19} = 0,27$

Semente: 66 (para gerar as variáveis normais bivariadas)

66 (para gerar as variáveis normais multivariadas)

População 76 - 19 variáveis

Médias e variâncias iguais às da população 72, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,75$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,78$	X_3 e $X_5 = 0,25$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,87$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,80$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,65$	X_{15} e $X_{16} = 0,70$	X_{16} e $X_{17} = -0,30$
X_{17} e $X_{18} = 0,60$	X_{18} e $X_{19} = 0,77$		

Semente: 76 (para gerar as variáveis normais bivariadas)

76 (para gerar as variáveis normais multivariadas)

População 77 - 20 variáveis

Médias:

1,45	8,71	6,75	1,63	105,35	4,91	10,46	14,76	55,54	19,62
45,97	29,68	0,30	19,85	0,76	15,00	111,67	1,37	8,04	6,78

Variâncias:

0,13	3,46	6,99	0,11	1.072,13	1,30	18,02	6,39	820,19	66,53
61,85	52,46	0,01	22,73	0,12	25,04	2.304,98	0,19	2,33	3,99

Correlações:

X_1 e $X_2 = 0,70$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,68$	X_3 e $X_5 = 0,45$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,30$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,25$	X_{15} e $X_{16} = 0,70$	X_{16} e $X_{17} = -0,60$
X_{17} e $X_{18} = 0,40$	X_{18} e $X_{19} = 0,87$	X_{19} e $X_{20} = 0,10$	

Semente: 06 (para gerar as variáveis normais bivariadas)

06 (para gerar as variáveis normais multivariadas)

População 78 - 20 variáveis

Médias e variâncias iguais às da população 77, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,50$	X_1 e $X_3 = 0,82$	X_2 e $X_4 = 0,28$	X_3 e $X_5 = 0,85$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,88$	X_5 e $X_8 = 0,55$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,49$	X_{10} e $X_{11} = 0,90$	X_{11} e $X_{12} = 0,65$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,25$	X_{15} e $X_{16} = 0,70$	X_{16} e $X_{17} = -0,60$
X_{17} e $X_{18} = 0,40$	X_{18} e $X_{19} = 0,87$	X_{19} e $X_{20} = 0,70$	

Semente: 06 (para gerar as variáveis normais bivariadas)

06 (para gerar as variáveis normais multivariadas)

População 79 - 20 variáveis

Médias e variâncias iguais às da população 77, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,90$	X_1 e $X_3 = 0,72$	X_2 e $X_4 = 0,88$	X_3 e $X_5 = 0,85$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,80$	X_5 e $X_8 = 0,65$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,89$	X_{10} e $X_{11} = 0,60$	X_{11} e $X_{12} = 0,50$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,75$	X_{15} e $X_{16} = 0,85$	X_{16} e $X_{17} = -0,70$
X_{17} e $X_{18} = 0,86$	X_{18} e $X_{19} = 0,47$	X_{19} e $X_{20} = 0,80$	

Semente: 06 (para gerar as variáveis normais bivariadas)

06 (para gerar as variáveis normais multivariadas)

População 80 - 20 variáveis

Médias e variâncias iguais às da população 77, sendo algumas correlações diferentes.

Correlações:

X_1 e $X_2 = 0,86$	X_1 e $X_3 = 0,72$	X_2 e $X_4 = 0,80$	X_3 e $X_5 = 0,85$
X_3 e $X_6 = 0,76$	X_4 e $X_7 = 0,80$	X_5 e $X_8 = 0,65$	X_6 e $X_9 = 0,78$
X_9 e $X_{10} = 0,80$	X_{10} e $X_{11} = 0,78$	X_{11} e $X_{12} = 0,55$	X_{12} e $X_{13} = 0,70$
X_{13} e $X_{14} = 0,86$	X_{14} e $X_{15} = 0,75$	X_{15} e $X_{16} = 0,75$	X_{16} e $X_{17} = -0,80$
X_{17} e $X_{18} = 0,60$	X_{18} e $X_{19} = 0,87$	X_{19} e $X_{20} = 0,70$	

Semente: 06 (para gerar as variáveis normais bivariadas)

06 (para gerar as variáveis normais multivariadas)

População 81 - 20 variáveis

Médias e variâncias iguais às da população 77, sendo algumas correlações diferentes.

Correlações:

$$X_1 \text{ e } X_2 = 0,30$$

$$X_3 \text{ e } X_6 = 0,86$$

$$X_9 \text{ e } X_{10} = 0,95$$

$$X_{13} \text{ e } X_{14} = 0,86$$

$$X_{17} \text{ e } X_{18} = 0,40$$

$$X_1 \text{ e } X_3 = 0,72$$

$$X_4 \text{ e } X_7 = 0,70$$

$$X_{10} \text{ e } X_{11} = 0,60$$

$$X_{14} \text{ e } X_{15} = 0,75$$

$$X_{18} \text{ e } X_{19} = 0,77$$

$$X_2 \text{ e } X_4 = 0,28$$

$$X_5 \text{ e } X_8 = 0,55$$

$$X_{11} \text{ e } X_{12} = 0,65$$

$$X_{15} \text{ e } X_{16} = 0,25$$

$$X_{19} \text{ e } X_{20} = 0,60$$

$$X_3 \text{ e } X_5 = 0,65$$

$$X_6 \text{ e } X_9 = 0,78$$

$$X_{12} \text{ e } X_{13} = 0,40$$

$$X_{16} \text{ e } X_{17} = -0,80$$

Semente: 06 (para gerar as variáveis normais bivariadas)

06 (para gerar as variáveis normais multivariadas)

APÊNDICE 2 - TESTE DE ESFERICIDADE DE BARTLETT E ESTATÍSTICA DE ADEQUABILIDADE DA AMOSTRA (MSA)

QUADRO A.2.1 - TAMANHOS DE AMOSTRA, NÚMERO DE VARIÁVEIS, TESTE DE ESFERICIDADE DE BARTLETT E ESTATÍSTICA
MSA, SEGUNDO POPULAÇÃO AMOSTRADA

continua

POPULAÇÃO AMOSTRADA	TAMANHOS DE AMOSTRA (n)	NÚMERO DE VARIÁVEIS (p)	TESTE DE ESFERICIDADE DE BARTLETT			ESTATÍSTICA MSA
			Estatística	G. L.	Valor-p	
1	27	5	109,4056	10	0,0000	0,6890
2	26	5	97,2277	10	0,0000	0,5271
3	26	5	105,7238	10	0,0000	0,5403
3	59	5	236,2248	10	0,0000	0,5262
4	24	5	128,2171	10	0,0000	0,5590
4	54	5	253,8502	10	0,0000	0,6766
5	24	5	56,7742	10	0,0000	0,5151
6	24	5	103,5423	10	0,0000	0,5241
7	30	6	110,9042	15	0,0000	0,6450
8	30	6	180,0705	15	0,0000	0,6646
9	30	6	101,0484	15	0,0000	0,5252
10	30	6	133,3047	15	0,0000	0,7052
11	31	6	106,5103	15	0,0000	0,7325
12	39	7	219,2075	21	0,0000	0,6791
13	39	7	184,2687	21	0,0000	0,6122
14	39	7	184,0143	21	0,0000	0,6396
15	39	7	214,3797	21	0,0000	0,7430
16	39	7	225,9117	21	0,0000	0,7201
17	44	8	384,2807	28	0,0000	0,8237
18	44	8	312,1709	28	0,0000	0,6796
19	44	8	267,8601	28	0,0000	0,5815
20	44	8	234,3451	28	0,0000	0,5939
21	44	8	218,9010	28	0,0000	0,6142
22	46	9	272,4551	36	0,0000	0,7254
23	46	9	278,8216	36	0,0000	0,6843
24	46	9	217,9345	36	0,0000	0,6456
25	46	9	201,0899	36	0,0000	0,6210
26	46	9	234,6816	36	0,0000	0,6394
27	46	10	376,0085	45	0,0000	0,7924
28	46	10	306,7559	45	0,0000	0,7338
29	47	10	378,3792	45	0,0000	0,8149
30	47	10	435,0023	45	0,0000	0,8273
31	47	10	354,7565	45	0,0000	0,7674
32	56	11	361,2456	55	0,0000	0,7246
33	55	11	382,8642	55	0,0000	0,7092
34	55	11	407,9247	55	0,0000	0,6699
35	55	11	444,1666	55	0,0000	0,6820
36	55	11	526,3939	55	0,0000	0,7350
37	50	12	432,4518	66	0,0000	0,6667
38	50	12	366,0033	66	0,0000	0,6551
39	50	12	468,5575	66	0,0000	0,7427
40	50	12	496,8471	66	0,0000	0,7303
41	50	12	476,5809	66	0,0000	0,7138
42	49	13	475,1039	78	0,0000	0,6576
43	49	13	381,6958	78	0,0000	0,6191

QUADRO A.2.1 - TAMANHOS DE AMOSTRA, NÚMERO DE VARIÁVEIS, TESTE DE ESFERICIDADE DE BARTLETT E ESTATÍSTICA MSA, SEGUNDO POPULAÇÃO AMOSTRADA

conclusão

POPULAÇÃO AMOSTRADA	TAMANHOS DE AMOSTRA (n)	NÚMERO DE VARIÁVEIS (p)	TESTE DE ESFERICIDADE DE BARTLETT			ESTATÍSTICA MSA
44	49	13	454,9799	78	0,0000	0,7236
45	49	13	581,4072	78	0,0000	0,7493
46	50	13	623,8040	78	0,0000	0,7812
47	76	14	901,3803	91	0,0000	0,6924
48	76	14	756,0524	91	0,0000	0,6662
49	76	14	1.025,2970	91	0,0000	0,7035
50	76	14	937,7240	91	0,0000	0,7856
51	76	14	716,5425	91	0,0000	0,7411
52	56	15	740,8990	105	0,0000	0,7577
53	56	15	898,0560	105	0,0000	0,6825
54	56	15	488,9463	105	0,0000	0,6283
55	56	15	667,2090	105	0,0000	0,6961
56	56	15	674,2128	105	0,0000	0,7361
57	67	16	795,6372	120	0,0000	0,7014
58	67	16	763,9589	120	0,0000	0,7118
59	67	16	807,6650	120	0,0000	0,6680
60	67	16	832,2709	120	0,0000	0,6774
61	67	16	753,6208	120	0,0000	0,7080
62	78	17	973,4270	136	0,0000	0,6961
63	77	17	1.086,4310	136	0,0000	0,7085
64	77	17	1.123,2860	136	0,0000	0,6970
66	78	17	1.151,3290	136	0,0000	0,7503
67	67	18	866,3990	153	0,0000	0,6667
68	66	18	884,9783	153	0,0000	0,6474
69	66	18	1.059,4500	153	0,0000	0,7771
70	66	18	1.177,8040	153	0,0000	0,7785
71	66	18	696,2197	153	0,0000	0,6648
72	79	19	1.109,9900	171	0,0000	0,6561
73	78	19	1.130,4670	171	0,0000	0,6642
74	78	19	1.272,6440	171	0,0000	0,7467
75	79	19	1.080,8460	171	0,0000	0,7234
76	79	19	1.196,6330	171	0,0000	0,6995
77	109	20	1.548,5490	190	0,0000	0,7066
78	110	20	1.701,5970	190	0,0000	0,7345
79	110	20	2.165,9150	190	0,0000	0,8072
80	110	20	2.033,8800	190	0,0000	0,8151
81	110	20	1.662,7950	190	0,0000	0,7666

FONTE: Dados obtidos por simulação Monte Carlo

APÊNDICE 3 - SCRIPTS DO SISTEMA R

SCRIPT 1

Este *scrip*, constrói a função densidade de probabilidade, da distribuição normal bivariada.

```
d2normal<-function(x1,x2,mi1,mi2,sig1,sig2,r){
  z1<-(x1-mi1)/sig1
  z2<-(x2-mi2)/sig2
  return(exp(-0.5/(1-r^2)*(z1^2-2*r*z1*z2+z2^2))/
    (2*pi*sqrt(1-r^2)*sig1*sig2))
}
grafico<-function(mi1,mi2,sig1,sig2,r){
  M=50
  x1<-seq(mi1-5*sig1,mi1+5*sig1,length=M)
  x2<-seq(mi2-5*sig2,mi2+5*sig2,length=M)
  f<-matrix(0,M,M)
  for (i in 1:M)
    for (j in 1:M)
      f[i,j]<-d2normal(x1[i],x2[j],mi1,mi2,sig1,sig2,r)
  persp(x1,x2,f,theta=30,phi=30,expand=0.5,col="lightgreen", ltheta=120,
    shade=1,xlab="X1",ylab="X2",zlab="densidade")
}
grafico(0,0,1,1,0) # correlação zero
```

SCRIPT 2

Este script executa as seguintes funções:

- gera as variáveis normais bivariadas e, a partir delas, as normais multivariadas;
- calcula os tamanhos de amostras para os erros relativos definidos, para nível de 95% de confiança;
- aplica a análise fatorial para os dados populacionais;
- aplica a análise fatorial para os dados amostrais e calcula as estimativas médias;
- calcula o viés, variância e erro quadrático médio das estimativas do modelo fatorial ortogonal.

```
tempo1=Sys.time() # início da contagem do tempo

#===== gera variáveis normais bivariadas =====#
require(MASS)

p<-20 # definindo o número de variáveis

# definindo as médias
m<-c(1.45,8.71,6.75,1.63,105.35,4.91,10.46,14.76,55.54,19.62,45.97,
  29.68,0.30,19.85,0.76,15.00,111.67,1.37,8.04,6.78)
```

```

# definindo as variâncias
v<-c(0.13,3.46,6.99,0.11,1702.13,1.30,18.02,6.39,650.19,66.53,61.85,52.46,
0.01,22.73,0.09,25.04,2304.98,0.19,2.33,3.99)

# definindo os coeficientes de correlação

r<-c(0.70,0.82,0.68,0.45,0.76,0.88,0.55,0.78,0.89,0.30,0.65,0.70,0.86,0.25,
0.70,-0.60,0.40,0.87,0.10)

n=100000 # número de observações

set.seed(6)
x1<-m[1]+sqrt(v[1])*rnorm(n,0,1)
# correlação entre x1 e x2
x2<-m[2]+r[1]*(sqrt(v[2])/sqrt(v[1]))*(x1-m[1])+sqrt(v[2]*(1-
r[1]**2))*rnorm(n,0,1)
# correlação entre x1 e x3
x3<-m[3]+r[2]*(sqrt(v[3])/sqrt(v[1]))*(x1-m[1])+sqrt(v[3]*(1-
r[2]**2))*rnorm(n,0,1)
# correlação entre x2 e x4
x4<-m[4]+r[3]*(sqrt(v[4])/sqrt(v[2]))*(x2-m[2])+sqrt(v[4]*(1-
r[3]**2))*rnorm(n,0,1)
# correlação entre x3 e x5
x5<-m[5]+r[4]*(sqrt(v[5])/sqrt(v[3]))*(x3-m[3])+sqrt(v[5]*(1-
r[4]**2))*rnorm(n,0,1)
# correlação entre x3 e x6
x6<-m[6]+r[5]*(sqrt(v[6])/sqrt(v[3]))*(x3-m[3])+sqrt(v[6]*(1-
r[5]**2))*rnorm(n,0,1)
# correlação entre x4 e x7
x7<-m[7]+r[6]*(sqrt(v[7])/sqrt(v[4]))*(x4-m[4])+sqrt(v[7]*(1-
r[6]**2))*rnorm(n,0,1)
# correlação entre x5 e x8
x8<-m[8]+r[7]*(sqrt(v[8])/sqrt(v[5]))*(x5-m[5])+sqrt(v[8]*(1-
r[7]**2))*rnorm(n,0,1)
# correlação entre x6 e x9
x9<-m[9]+r[8]*(sqrt(v[9])/sqrt(v[6]))*(x6-m[6])+sqrt(v[9]*(1-
r[8]**2))*rnorm(n,0,1)
# correlação entre x9 e x10
x10<-m[10]+r[9]*(sqrt(v[10])/sqrt(v[9]))*(x9-m[9])+sqrt(v[10]*(1-
r[9]**2))*rnorm(n,0,1)
# correlação entre x10 e x11
x11<-m[11]+r[10]*(sqrt(v[11])/sqrt(v[10]))*(x10-m[10])+sqrt(v[11]*(1-r[10]**2))*
rnorm(n,0,1)
# correlação entre x11 e x12
x12<-m[12]+r[11]*(sqrt(v[12])/sqrt(v[11]))*(x11-m[11])+sqrt(v[12]*(1-r[11]**2))*
rnorm(n,0,1)
# correlação entre x12 e x13
x13<-m[13]+r[12]*(sqrt(v[13])/sqrt(v[12]))*(x12-m[12])+sqrt(v[13]*(1-r[12]**2))*
rnorm(n,0,1)
# correlação entre x13 e x14
x14<-m[14]+r[13]*(sqrt(v[14])/sqrt(v[13]))*(x13-m[13])+sqrt(v[14]*(1-r[13]**2))*
rnorm(n,0,1)
# correlação entre x14 e x15
x15<-m[15]+r[14]*(sqrt(v[15])/sqrt(v[14]))*(x14-m[14])+sqrt(v[15]*(1-r[14]**2))*
rnorm(n,0,1)
# correlação entre x15 e x16
x16<-m[16]+r[15]*(sqrt(v[16])/sqrt(v[15]))*(x15-m[15])+sqrt(v[16]*(1-r[15]**2))*
rnorm(n,0,1)
# correlação entre x16 e x17
x17<-m[17]+r[16]*(sqrt(v[17])/sqrt(v[16]))*(x16-m[16])+sqrt(v[17]*(1-r[16]**2))*
rnorm(n,0,1)
# correlação entre x17 e x18
x18<-m[18]+r[17]*(sqrt(v[18])/sqrt(v[17]))*(x17-m[17])+sqrt(v[18]*(1-r[17]**2))*
rnorm(n,0,1)
# correlação entre x18 e x19
x19<-m[19]+r[18]*(sqrt(v[19])/sqrt(v[18]))*(x18-m[18])+sqrt(v[19]*(1-r[18]**2))*
rnorm(n,0,1)

```

```

# correlação entre x19 e x20
x20<-m[20]+r[19]*(sqrt(v[20])/sqrt(v[19]))*(x19-m[19])+sqrt(v[20]*(1-r[19]**2))*
rnorm(n,0,1)

dados<-cbind(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12,x13,x14,x15,x16,x17,
  x18,x19,x20)
mi1<-mean(x1)
mi2<-mean(x2)
mi3<-mean(x3)
mi4<-mean(x4)
mi5<-mean(x5)
mi6<-mean(x6)
mi7<-mean(x7)
mi8<-mean(x8)
mi9<-mean(x9)
mi10<-mean(x10)
mi11<-mean(x11)
mi12<-mean(x12)
mi13<-mean(x13)
mi14<-mean(x14)
mi15<-mean(x15)
mi16<-mean(x16)
mi17<-mean(x17)
mi18<-mean(x18)
mi19<-mean(x19)
mi20<-mean(x20)

si<-cov(dados)
ri<-cor(dados)
mi<-cbind(mi1,mi2,mi3,mi4,mi5,mi6,mi7,mi8,mi9,mi10,mi11,mi12,mi13,mi14,
  mi15,mi16,mi17,mi18,mi19,mi20)

#mi      # vetor de médias
#si      # matriz de covariâncias
#ri      # matriz de correlação

p<-ncol(dados)      # número de variáveis

#==== gera a distribuição normal multivariada a partir do vetor de médias
#==== e matriz de covariância, obtida acima (mi e si)

set.seed(6)
y<-mvrnorm(100000,mi,si)
y1<-(y[,1])
y2<-(y[,2])
y3<-(y[,3])
y4<-(y[,4])
y5<-(y[,5])
y6<-(y[,6])
y7<-(y[,7])
y8<-(y[,8])
y9<-(y[,9])
y10<-(y[,10])
y11<-(y[,11])
y12<-(y[,12])
y13<-(y[,13])
y14<-(y[,14])
y15<-(y[,15])
y16<-(y[,16])
y17<-(y[,17])
y18<-(y[,18])
y19<-(y[,19])
y20<-(y[,20])

dados1<-cbind(y1,y2,y3,y4,y5,y6,y7,y8,y9,y10,y11,y12,y13,y14,y15,y16,y17,
  y18,y19,y20)

```

```

s<-cov(dados1) # calcula a matriz de covariância

write.table(s,"c:/sp_1.csv",dec="," , sep=";", row.names=FALSE) # grava a matriz de
covariância

y1m<-mean(y1)
y2m<-mean(y2)
y3m<-mean(y3)
y4m<-mean(y4)
y5m<-mean(y5)
y6m<-mean(y6)
y7m<-mean(y7)
y8m<-mean(y8)
y9m<-mean(y9)
y10m<-mean(y10)
y11m<-mean(y11)
y12m<-mean(y12)
y13m<-mean(y13)
y14m<-mean(y14)
y15m<-mean(y15)
y16m<-mean(y16)
y17m<-mean(y17)
y18m<-mean(y18)
y19m<-mean(y19)
y20m<-mean(y20)

m<-cbind(y1m,y2m,y3m,y4m,y5m,y6m,y7m,y8m,y9m,y10m,y11m,y12m,
y13m,y14m,y15m,y16m,y17m,y18m,y19m,y20m) # cria o vetor de médias

# grava o vetor de médias
write.table(m,"c:/mp_1.csv",dec="," , sep=";", row.names=FALSE)
errorel<-0.05 # Definindo o erro relativo

erro<-matrix(c(m*errorel),p,1) # Calculando os erros absolutos

conf<-0.95 #Definindo o nível de confiança

alfa<-(1-conf)

z<-qnorm(1-(alfa/(2*p)))
z2<-z*z
z2

# Calculando os tamanhos de amostras
n<-matrix(0,p,1)
for(i in 1:p){
n[i,1]<-(z2*s[i,i])/(erro[i,1]*erro[i,1])
}
for (i in 1:p){
n[i,1]}
write.table(n,"c:/n_1.csv",dec="," , sep=";", row.names=FALSE) # grava os
# tamanhos de amostras
n

# Análise Fatorial para dados populacional

p<-ncol(dados1) # lê o número de variáveis do arquivo dados1
N<-nrow(dados1) # lê o número de observações do arquivo dados1
r<-cor(dados1) # calcula a matriz de correlação do arquivo dados1
write.table(r,"c:/rp_1.csv",dec="," , sep=";", row.names=FALSE) # grava a matriz
# de correlação

ee<-eigen(r,EISPACK=TRUE) # calcula os autovalores e autovetores

autoval<-ee$values # separa os autovalores

autoval1<-matrix(0,p,1)

```

```

for (i in 1:p){
  autoval1[i,]<-autoval[i]
  autovalp<-t(autoval1)
#autoval1      # autovalores da matriz de correl. pop.

# grava os autovalores
write.table(autovalp,"c:/autovalp_1.csv",dec="," ,sep=";",row.names=FALSE)

autovet<-ee$vector      # separa os autovetores
#autovet

# grava os autovetores
write.table(autovet,"c:/autovetp_1.csv",dec="," ,sep=";",row.names=FALSE)

autovet1<-autovet

auto<-matrix(0,1,p)      # seleciona os autovalores populacional >1
  f<-0
  for (i in 1:p){
    if (autoval1[i,]>1) {(auto[1,i]=autoval1[i,])
    f<-(f+1)}
  }

autol<-matrix(c(auto[auto>=1]),1,f)
#autol # mostra os autovalores >1

exp<-autoval1/(sum(ee$values)) # calcula as explicações da cada autovalor

exp<-t(exp)
#exp

# grava as explicações dos autovalores
write.table(exp,"c:/exp_1.csv",dec="," ,sep=";",row.names=FALSE)

expacum<-cumsum(exp)      # explicação acumulada
#expacum

l<-matrix(0,p,ncol(auto1))
  for (k in 1:ncol(auto1)){
    for (i in 1:p){
      l[i,k]<-round((sqrt(auto1[1,k])*autovet1[i,k]),5)
    }
  }
lp<-t(l)
#l      # matriz dos carregamentos

# grava os carregamentos
write.table(lp,"c:/l_1.csv",dec="," ,sep=";",row.names=FALSE)

comun<-matrix(0,p,ncol(auto1))
comunt<-matrix(0,p,1)
  for (i in 1:p){
    for (k in 1:ncol(l)){
      comun[i,k]<-l[i,k]*l[i,k]
    }
  }

comunp<-t(comun)
#comun # calcula as comunalidades parciais

# grava as comunalidades
write.table(comunp,"c:/comun_1.csv",dec="," ,sep=";",row.names=FALSE)

for (i in 1:p){
  comunt[i,]<-sum(comun[i,])
}
#comunt # calcula a comunalidade total

varesp<-matrix(0,i,1)

```



```

for (i in 1:p){
  varesp[i,]<-(1-comunt[i,])} # calcula as variâncias específicas

varespp<-t(varesp)
# varesp
# grava as var. específicas
write.table(varespp,"c:/varesp_1.csv",dec="," , sep=";", row.names=FALSE)

#===== Obtém amostras de tamanhos definidos =====#

N<-nrow(dados1) # tamanho da população
n<-776 # tamanho da amostra
niter<-1000 # número de amostras
p<-ncol(dados1) # número de variáveis

autovalla<-matrix(0,niter,p)
vet<-array(0,dim=c(p,p,niter))
expacum1<-matrix(0,niter,p)
expla<-matrix(0,niter,p)
nlidos<-matrix(0,niter)

t1<-format(Sys.time(), "%X")

{windows(width = 7, height = 2)

for (j in 1:niter){

  i<-round(runif(n,1,N),0)
  amostral<-dados1[i,] # seleciona amostras

  barplot(j, width=.8, horiz = T, xlim=c(0,niter), ylim=c(0,1),
    col="red", main=paste(t1,"PROCESSANDO...", "AMOSTRA: ",j,
    " - ",format(Sys.time(), "%X")))

r1<-cor(amostral) # calcula a matriz de correlação amostral

eel<-eigen(r1,EISPACK=TRUE) # calcula os autovalores e autovetores
# amostral
autovala<-round(eel$values,5) # separa os autovalores

for (i in 1:p){
  autovalla[j,i]<-autovala[i]} # cria a matriz de autovalores para as
# iterações

autoveta<-eel$vector # separa os autovetores

# cria a matriz de autovetores para as iterações
for (i in 1:p){
  for (k in 1:p){
    vet[i,k,j]<-autoveta[i,k]}}

autovetla<-vet

expla=autovalla/sum(eel$values) # calcula as explicações da cada autovalor
#expla

la<-array(0,dim=c(p,f,niter))
for (j in 1:niter){
  for (k in 1:f){
    for (i in 1:p){
      la[i,k,j]<-round((sqrt(autovalla[j,k])*autovetla[i,k,j]),5)
    }}}
#la # matriz dos carregamentos

comun1<-array(0,dim=c(p,f,niter))

```

```

for (j in 1:niter){
  for (i in 1:p){
    for (k in 1:f){
      comunl[i,k,j]<-la[i,k,j]*la[i,k,j]
    }}
}}

#comunl # calcula as comunalidades parciais

comuntl<-array(0,dim=c(p,1,niter))
for (j in 1:niter){
  for (i in 1:p){
    comuntl[i,,j]<-sum(comunl[i,,j])
  }}
#comuntl # comunalidades totais

varespl<-array(0,dim=c(p,1,niter))
for (j in 1:niter){
  for (i in 1:p){
    varespl[i,,j]<-(1-comuntl[i,,j])
  }}
#varespl # variâncias específicas
}}

# Cálculo dos valores médios

autovallam<-matrix(0,1,f)
for (i in 1:f){
  autovallam[,i]<-mean(autovalla[,i]) # autovalores médios
}
autovallam

# grava os autovalores médios
write.table(autovallam,"c:/autovallam_1.csv",dec="," , sep=";" , row.names=FALSE)

vetm<-matrix(0,p,f)
for (i in 1:p){
  for (k in 1:f){
    vetm[i,k]<-mean(vet[i,k,]) }}
#vetm # autovetores médios
write.table(vetm,"c:/vetm_1.csv",dec="," , sep=";" , row.names=FALSE) # grava os
autovetores médios

explam<-matrix(0,1,p)
for (i in 1:f){
  explam[,i]<-mean(expla[,i]) # explicação média dos autovalores
}
explam

# grava as expl. médias
write.table(explam,"c:/expalam_1.csv",dec="," , sep=";" , row.names=FALSE)

lam<-matrix(0,p,f)
for (i in 1:p){
  for (k in 1:f){
    lam[i,k]<-mean(la[i,k,]) }} # matriz de carregamentos médios
lamp<-t(lam)
#lam

# grava os carregamentos médios
write.table(lamp,"c:/lam_1.csv",dec="," , sep=";" , row.names=FALSE)

comuntlm<-matrix(0,p,1)
for (i in 1:p){
  comuntlm[i,]<-mean(comuntl[i,,]) # comunalidades médias
}
comuntlp<-t(comuntlm)
#comuntlm

# grava as comunalidades médias
write.table(comuntlp,"c:/comuntlm_1.csv",dec="," , sep=";" , row.names=FALSE)

```

```

# Cálculo do viés, variância e erro quadrático médio dos autovalores
autovallt<-t(autovall)

viesauto<-matrix(0,1,f)
for (k in 1:f){
viesauto[1,k]<-(mean(autovalla[,k])-autovallt[1,k])}
#viesauto      # vies do autovalor

# grava o viés dos autovalores
write.table(viesauto,"c:/viesauto_1.csv",dec=" ",sep=";",row.names=FALSE)

Vauto<-matrix(0,1,f)
  for (k in 1:f){
  Vauto[1,k]<-((niter-1)/niter)*var(autovalla[,k])}
#Vauto      # variância do autovalor
# grava a variância dos autovalores
write.table(Vauto,"c:/Vauto_1.csv",dec=" ",sep=";",row.names=FALSE)

EQMauto<-Vauto+(viesauto^2)
#EQMauto      # EQM do autovalor

# grava o EQM dos autovalores
write.table(EQMauto,"c:/EQMauto_1.csv",dec=" ",sep=";",row.names=FALSE)

# cálculo do viés, variância e erro quadrático dos autovetores
viesvet<-matrix(0,p,f)
for (i in 1:p){
  for (k in 1:f){
  viesvet[i,k]<-mean(vet[i,k,]-autovet[i,k])}
#viesvet      # viés do autovetor
# grava o viés dos autovetores
write.table(viesvet,"c:/viesvet_1.csv",dec=" ",sep=";",row.names=FALSE)

Vvet<-matrix(0,p,f)
  for (i in 1:p){
  for (k in 1:f){
  Vvet[i,k]<-((niter-1)/niter)*var(vet[i,k,])}
#Vvet      # variância do autovetor

# grava a variância dos autovetores
write.table(Vvet,"c:/Vvet_1.csv",dec=" ",sep=";",row.names=FALSE)

EQMvet<-Vvet+(viesvet^2)
#EQMvet      # EQM do autovetor

# grava o EQM dos autovetores
write.table(EQMvet,"c:/EQMvet_1.csv",dec=" ",sep=";",row.names=FALSE)

# cálculo do viés, variância e erro quadrático médio dos carregamentos
viescar<-matrix(0,p,f)
  for (i in 1:p){
  for (k in 1:f){
  viescar[i,k]=mean(la[i,k,])-l[i,k]}
viescarp<-t(viescar)
#viescar      # viés do carregamento

# grava o viés dos carregamentos
write.table(viescarp,"c:/viescar_1.csv",dec=" ",sep=";",row.names=FALSE)

Vcar<-matrix(0,p,f)
  for (i in 1:p){
  for (k in 1:f){

```

```

    Vcar[i,k]<-((niter-1)/niter)*var(la[i,k,])}}
Vcarp<-t(Vcar)
#Vcar                # variância do carregamento

# grava a variância dos carregamentos
write.table(Vcarp,"c:/Vcar_1.csv",dec="," , sep=";" , row.names=FALSE)

EQMcar<-Vcar+(viescar^2)
EQMcarp<-t(EQMcar)
#EQMcar                # EQM do carregamento

# grava o EQM dos carregamentos
write.table(EQMcarp,"c:/EQMcar_1.csv",dec="," , sep=";" , row.names=FALSE)

# cálculo do viés, variância e erro quadrático médio das comunalidades
viescom<-matrix(0,p,1)
  for (i in 1:p){
    viescom[i,1]<-mean(comunt1[i,1,])-comunt[i,1]}
viescom<-round(viescom,7)
viescomp<-t(viescom)
#viescom                # viés das comunalidades

# grava o viés das comunalidades
write.table(viescomp,"c:/viescom_1.csv",dec="," , sep=";" , row.names=FALSE)

Vcom<-matrix(0,p,1)
  for (i in 1:p){
    Vcom[i,1]<-((niter-1)/niter)*var(comunt1[i,1,])}
Vcom<-round(Vcom,7)
Vcomp<-t(Vcom)
#Vcom                # variância das comunalidades

# grava a variância das comunalidades
write.table(Vcomp,"c:/Vcom_1.csv",dec="," , sep=";" , row.names=FALSE)

EQMcom<- Vcom +(viescom^2)
EQMcom<-round(EQMcom,7)
EQMcomp<-t(EQMcom)
#EQMcom                # EQM das comunalidades

# grava o EQM das comunalidades
write.table(EQMcomp,"c:/EQMcom_1.csv",dec="," , sep=";" , row.names=FALSE)

#===== mostra barra de contagem =====#
tempo2<-Sys.time()
A<-round(print(tempo2-tempo1),2)
{windows(width = 7, height = 2)
  barplot(j, width=.8, horiz = T, xlim=c(0,niter), ylim=c(0,1),
    col="red", main=paste("TEMPO DE PROCESSAMENTO:",A,"MINUTOS"))}

```

SCRIPT 3

Este *script* calcula a estatística MSA e teste de esfericidade de Bartlett.

```

require(MASS)

amostral<-read.table("c:/amostral.csv", sep=";" , header=T)

p<-ncol(amostral) # número de variáveis
n<-nrow(amostral) # tamanho da amostra
p;n

```

```

R<-cor(amostral)          # calcula a matriz de correlação

cat("Matriz de Correlação");R

qui2<-(-(n-1)-(2*p+5)/6)*log(det(R)) # estatística de Bartlett
gl<-p*(p-1)/2 # graus de liberdade
pvalor<-dchisq(qui2,gl) # p-valor
cat ("Estatística do teste :");qui2
cat ("Graus de liberdade :");gl
cat ("valor-p :");pvalor

Ri<-ginv(R) # calcula o inverso da matriz de correlação
digRi<-diag(Ri) # diagonal de Ri
S<-(digRi)^(-1/2) # diagonal de Ri elevado a (-1/2)
#S
S1<-matrix(diag(S),p,p)
S1
Q<-S1%*%Ri%*%S1 # matriz com as correlações parciais
#Q
t<-matrix(0,p,p)
for (i in 1:p) {
  for (j in 1:p) {
    if (i!=j) (t[i,j]=Q[i,j])
  }
}
#t
Q2<-matrix(0,p,1)
for (i in 1:p){
  Q2[i,]<-sum(Q[i,]^2)-1}
R2=matrix(0,p,1)
for (i in 1:p){
  R2[i,]<-sum(R[i,]^2)-1
}
diagQ<-matrix(0,p,1)
for (i in 1:p) {
  diagQ[i,]<-R2[i,]/(R2[i,]+Q2[i,])
}
#diagQ

#===== Measure Sample Adequacy =====#
Manti<-matrix(0,p,p)
for (i in 1:p) {
  for ( j in 1:p) {
    if (i!=j) (Manti[i,j]<-t[i,j]) else (Manti[i,j]<-diagQ[j,1])
  }
}

cat ("Manti :")
Manti # Measure Sample Adequacy na diagonal (matriz anti-imagem)

#===== Medida KMO =====#
SomaQ<-sum(Q2[,1])
#SomaQ
SomaR<-sum(R2[,1])
#SomaR

MSA<-SomaR/(SomaR+SomaQ) # estatística MSA
cat ("MSA :")
MSA

```

SCRIPT 4

Esta *script* executa as seguintes funções:

- a) ajusta os modelos de regressão linear múltipla;
- b) calcula a medida K para testar a multicolinearidade;
- c) realiza o teste de Golfeld-Quandt para verificar a homogeneidade da variância dos resíduos;
- d) realiza o teste de Kolmogorov-Smirnov para verificar a Gaussianidade dos resíduos;
- e) realiza o teste t de Student para verificar se a média do resíduo é igual a zero.

```
##### AJUSTA MODELO DE REGRESSÃO LINEAR MULTIPLA #####

##### LEITURA DO ARQUIVO DE DADOS #####
library(MASS) # ativa a biblioteca MASS
library(nortest) # ativa a biblioteca nortest
library(lmtest) # ativa a biblioteca lmtest

dados<-read.csv("c:/dados.csv", sep=";", dec=".", header=T)

##### CRIA NOVAS VARIÁVEIS #####

attach(dados)
n<-dados$n
p<-dados$p
f<-dados$f
errorel<-dados$errorel
expl<-dados$exptot
autoval<-dados$autoval
cvautoval<-dados$cvautoval
eqmautoval<-dados$eqmautoval
autovet<-dados$autovet
cvautovet<-dados$cvautovet
eqmautovet<-dados$eqmautovet
car<-dados$car
cvcar<-dados$cvcar
eqmcar<-dados$eqmcar
comu<-dados$comu
cvcomu<-dados$cvcomu
eqmcomu<-dados$eqmcomu
rautoval<-dados$rautoval
logcvval<-dados$logcvval
logeqmval<-dados$logeqmval
logcvvet<-dados$logcvvet
logeqmvvet<-dados$logeqmvvet
logcvcar<-dados$logcvcar
logeqmcar<-dados$logeqmcar
logcvcomu<-dados$logcvcomu
logeqmcomu<-dados$logeqmcomu
n_p<-dados$n_p
p_f<-dados$p_f
raizn<-sqrt(n)
raizn_p<-dados$raizn_p
expm<- (expl/f)
raizf<-sqrt(f)
```

```

#===== CÁLCULO DA ESTATÍSTICA k =====#

dados1<-cbind(rautoval,raizn,expm)
dados1<-as.data.frame(dados1)
matriz.corr<-round(cor(dados1),5)
matriz.corr
ee<-eigen(matriz.corr) #obtem os autovalores e autovetores
autoval=ee$values # separa os autovalores

i<-nrow(matriz.corr)
k<-autoval[1]/autoval[i]
k

#===== modelo 1 - CV AUTOVALORES =====#

modelo<-lm(logcvval ~(rautoval+raizn+expm)-1) # ajusta o modelo de regressão

out<-summary(modelo)
#out      # mostra o resumo do modelo ajustado

coef<-round(out$coeff,4)
coef      # mostra os coeficientes estimados

s<-out$sigma
s         # mostra o erro padrão da estimativa

R2<-out$r.squared # obtém o R2
R2         # mostra o R2

F<-out$fstatistic
F

#===== TESTE DE HOMOGENEIDADE DA VARIÂNCIA =====#

homocedast<-gqtest(modelo,alternative="two.sided") # teste de Goldfeld-
Quandt
#homocedast

#===== TESTE DE NORMALIDADE DOS RESÍDUOS =====#

residuo<-modelo$residuals # obtém os resíduos
#residuo      # mostra os resíduos

lillie.test(residuo) # teste Kolmogorov-Smirnov para normalidade

qqnorm(residuo,main="QQ-PLOT - residuo",xlab="Quantis Teórica",ylab="Quantis
Amostral")
qqline(residuo)      # gráfico QQ-PLOT

#===== TESTE DE HIPÓTESE DE QUE MÉDIA DO RESÍDUO=0 =====#

teste<-t.test(residuo) # testa a hipótese de que média=0
teste

```

SCRIPT 5

Este script identifica os *outliers*, pontos de *leverages* e DFFITS.

```

#####      RESÍDUOS STUDENTIZADOS      #####
library(DAAG)
library(tseries)
library(car)
residuo<-rstudent(modelo) # obtém os resíduos studentizados
#sort(residuo,decreasing = TRUE)
plot(residuo,xlab="Índice", ylab="Resíduos Studentizados")
abline(h=c(-3,3),lty=2) # mostra a linha em + ou - 3
identify(residuo,y=NULL,cex=0.6,offset = 0.5,
         col="red") # mostra o número da observação

outlier.test(modelo) # teste de outlier de Bonferroni

#####      LEVERAGES      #####

hat<-hatvalues(modelo) # leverages
media<-mean(hat)
plot(hat,xlab="Índice",ylab="Leverage")
abline(h=3*media,lty=2)
identify(hat,y=NULL,cex=0.7,offset = 0.5,
        col="red") # mostra o número da observação

#####      DFFITS      #####

difs<-round(dffits(modelo),4)
p<-3 #*** mudar o número de parâmetros estimados
n<-241 #*** mudar o total de observações
plot(difs,xlab="Observação", ylab="DIFS",main="EQM COMUNALIDADES",
     cex.main=0.8,font=1)
vdif<-2*sqrt(p/n)
abline(h=c(-vdif,vdif),lty=2)
#identify(difs,y=NULL,cex=0.6, offset = 0.5, col="red",
#        label=as.character(difs)) # mostra o valor do DFFITS
identify(difs,y=NULL,cex=0.6,offset = 0.5,
        col="red") # mostra o número da observação

```


APÊNDICE 4 - MATRIZES DE CORRELAÇÃO DAS POPULAÇÕES 27 E 51 E DAS RESPECTIVAS AMOSTRAS

1 MATRIZ DE CORRELAÇÃO DA POPULAÇÃO 27

A matriz de correlação apresentada abaixo é proveniente da população normal multivariada ($\rho=10$), gerada a partir dos parâmetros definidos no apêndice 1, correspondente à população 27.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1,0000	0,7976	0,6992	0,3053	0,2964	0,4693	0,1775	0,1723	0,3659	0,3256
X_2	0,7976	1,0000	0,5591	0,3827	0,2374	0,3766	0,2249	0,1390	0,2923	0,2604
X_3	0,6992	0,5591	1,0000	0,2136	0,4249	0,6777	0,1241	0,2385	0,5252	0,4660
X_4	0,3053	0,3827	0,2136	1,0000	0,0855	0,1475	0,5975	0,0494	0,1151	0,1014
X_5	0,2964	0,2374	0,4249	0,0855	1,0000	0,2876	0,0495	0,5558	0,2205	0,1941
X_6	0,4693	0,3766	0,6777	0,1475	0,2876	1,0000	0,0843	0,1614	0,7768	0,6895
X_7	0,1775	0,2249	0,1241	0,5975	0,0495	0,0843	1,0000	0,0278	0,0669	0,0593
X_8	0,1723	0,1390	0,2385	0,0494	0,5558	0,1614	0,0278	1,0000	0,1238	0,1094
X_9	0,3659	0,2923	0,5252	0,1151	0,2205	0,7768	0,0669	0,1238	1,0000	0,8890
X_{10}	0,3256	0,2604	0,4660	0,1014	0,1941	0,6895	0,0593	0,1094	0,8890	1,0000

Os autovalores da matriz de correlação populacional são:

$$\lambda_1 = 4,1256 ; \lambda_2 = 1,6974 ; \lambda_3 = 1,4068 ; \lambda_4 = 1,0482 ; \lambda_5 = 0,4706 ; \lambda_6 = 0,3926 ;$$

$$\lambda_7 = 0,3649 ; \lambda_8 = 0,2310 ; \lambda_9 = 0,1651 ; \lambda_{10} = 0,0978.$$

2 MATRIZ DE CORRELAÇÃO AMOSTRAL PROVENIENTE DA POPULAÇÃO 27

A matriz de correlação apresentada abaixo é proveniente de uma amostra de tamanho igual a 46 observações, retirada da população 27.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1,0000	0,8833	0,7371	0,3327	0,2963	0,6982	0,2496	0,3971	0,5859	0,6111
X_2	0,8833	1,0000	0,6544	0,4045	0,2816	0,6383	0,3267	0,3470	0,4741	0,4844
X_3	0,7371	0,6544	1,0000	0,4500	0,4685	0,8192	0,2267	0,3922	0,7319	0,7035
X_4	0,3327	0,4045	0,4500	1,0000	0,1689	0,4817	0,6787	0,2937	0,2994	0,2025
X_5	0,2963	0,2816	0,4685	0,1689	1,0000	0,4202	-0,0195	0,4584	0,3586	0,3817
X_6	0,6982	0,6383	0,8192	0,4817	0,4202	1,0000	0,2246	0,3712	0,8762	0,8359
X_7	0,2496	0,3267	0,2267	0,6787	-0,0195	0,2246	1,0000	0,1823	0,1895	0,1082
X_8	0,3971	0,3470	0,3922	0,2937	0,4584	0,3712	0,1823	1,0000	0,2248	0,2593
X_9	0,5859	0,4741	0,7319	0,2994	0,3586	0,8762	0,1895	0,2248	1,0000	0,9352
X_{10}	0,6111	0,4844	0,7035	0,2025	0,3817	0,8359	0,1082	0,2593	0,9352	1,0000

Os autovalores da matriz de correlação amostral são:

$$\hat{\lambda}_1 = 5,3102 ; \quad \hat{\lambda}_2 = 1,5044 ; \quad \hat{\lambda}_3 = 1,0823 ; \quad \hat{\lambda}_4 = 0,8213 ; \quad \hat{\lambda}_5 = 0,4898 ; \quad \hat{\lambda}_6 = 0,3535 ;$$

$$\hat{\lambda}_7 = 0,2097 ; \quad \hat{\lambda}_8 = 0,1023 ; \quad \hat{\lambda}_9 = 0,0775 ; \quad \hat{\lambda}_{10} = 0,0490.$$

3 MATRIZ DE CORRELAÇÃO DA POPULAÇÃO 51

A matriz de correlação apresentada abaixo é proveniente da população normal multivariada ($p=14$), gerada a partir dos parâmetros definidos no apêndice 1, correspondente à população 51.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}
X_1	1,0000	-0,6474	0,8188	-0,4534	0,3654	0,7040	-0,2097	0,0940	0,5496	0,4882	0,1427	0,0941	0,0361	0,0252
X_2	-0,6474	1,0000	-0,5293	0,6993	-0,2391	-0,4539	0,3282	-0,0665	-0,3547	-0,3123	-0,0993	-0,0592	-0,0238	-0,0144
X_3	0,8188	-0,5293	1,0000	-0,3714	0,4445	0,8597	-0,1705	0,1144	0,6709	0,5975	0,1757	0,1142	0,0470	0,0362
X_4	-0,4534	0,6993	-0,3714	1,0000	-0,1679	-0,3195	0,4774	-0,0418	-0,2532	-0,2215	-0,0661	-0,0389	-0,0171	-0,0088
X_5	0,3654	-0,2391	0,4445	-0,1679	1,0000	0,3837	-0,0754	0,2494	0,2970	0,2646	0,0797	0,0544	0,0237	0,0163
X_6	0,7040	-0,4539	0,8597	-0,3195	0,3837	1,0000	-0,1464	0,0984	0,7786	0,6946	0,2088	0,1392	0,0542	0,0419
X_7	-0,2097	0,3282	-0,1705	0,4774	-0,0754	-0,1464	1,0000	-0,0187	-0,1150	-0,0996	-0,0250	-0,0152	-0,0017	0,0006
X_8	0,0940	-0,0665	0,1144	-0,0418	0,2494	0,0984	-0,0187	1,0000	0,0776	0,0709	0,0147	0,0115	0,0070	0,0052
X_9	0,5496	-0,3547	0,6709	-0,2532	0,2970	0,7786	-0,1150	0,0776	1,0000	0,8902	0,2674	0,1785	0,0694	0,0558
X_{10}	0,4882	-0,3123	0,5975	-0,2215	0,2646	0,6946	-0,0996	0,0709	0,8902	1,0000	0,2997	0,2004	0,0810	0,0658
X_{11}	0,1427	-0,0993	0,1757	-0,0661	0,0797	0,2088	-0,0250	0,0147	0,2674	0,2997	1,0000	0,6510	0,2603	0,2205
X_{12}	0,0941	-0,0592	0,1142	-0,0389	0,0544	0,1392	-0,0152	0,0115	0,1785	0,2004	0,6510	1,0000	0,3986	0,3398
X_{13}	0,0361	-0,0238	0,0470	-0,0171	0,0237	0,0542	-0,0017	0,0070	0,0694	0,0810	0,2603	0,3986	1,0000	0,8604
X_{14}	0,0252	-0,0144	0,0362	-0,0088	0,0163	0,0419	0,0006	0,0052	0,0558	0,0658	0,2205	0,3398	0,8604	1,0000

Os autovalores da matriz de correlação populacional são:

$$\begin{aligned} \lambda_1 &= 4,8360 ; \lambda_2 = 2,3244 ; \lambda_3 = 1,5378 ; \lambda_4 = 1,1637 ; \lambda_5 = 1,0528 ; \lambda_6 = 0,7694 ; \\ \lambda_7 &= 0,6763 ; \lambda_8 = 0,4941 ; \lambda_9 = 0,3309 ; \lambda_{10} = 0,2715 ; \lambda_{11} = 0,2044 ; \lambda_{12} = 0,1370 ; \\ \lambda_{13} &= 0,1096 ; \lambda_{14} = 0,0925. \end{aligned}$$

4 MATRIZ DE CORRELAÇÃO AMOSTRAL PROVENIENTE DA POPULAÇÃO 51

A matriz de correlação apresentada abaixo é proveniente de uma amostra de tamanho igual a 76 observações, retirada da população 51.

	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	X ₁₁	X ₁₂	X ₁₃	X ₁₄
X ₁	1,0000	-0,6200	0,8507	-0,3676	0,4183	0,7449	-0,2842	-0,0094	0,6080	0,5251	0,0707	0,0474	0,1810	0,1423
X ₂	-0,6200	1,0000	-0,5341	0,6621	-0,2382	-0,4408	0,4178	0,1472	-0,4208	-0,2601	0,0048	-0,0595	-0,1769	-0,1936
X ₃	0,8507	-0,5341	1,0000	-0,3429	0,4761	0,8700	-0,2673	0,0926	0,7034	0,6100	0,2123	0,0867	0,2194	0,1695
X ₄	-0,3676	0,6621	-0,3429	1,0000	-0,1397	-0,2423	0,5192	0,0814	-0,3010	-0,2032	-0,0349	0,0415	-0,0631	-0,1591
X ₅	0,4183	-0,2382	0,4761	-0,1397	1,0000	0,4716	-0,0207	0,3950	0,3352	0,1843	0,0963	0,0119	0,0518	-0,1027
X ₆	0,7449	-0,4408	0,8700	-0,2423	0,4716	1,0000	-0,1954	0,0939	0,7563	0,6092	0,2131	0,0995	0,1638	0,1195
X ₇	-0,2842	0,4178	-0,2673	0,5192	-0,0207	-0,1954	1,0000	0,0256	-0,2837	-0,1694	0,0861	0,1945	0,0132	-0,0904
X ₈	-0,0094	0,1472	0,0926	0,0814	0,3950	0,0939	0,0256	1,0000	0,0515	-0,0215	-0,0466	-0,0503	0,1015	0,0136
X ₉	0,6080	-0,4208	0,7034	-0,3010	0,3352	0,7563	-0,2837	0,0515	1,0000	0,8538	0,2315	0,0389	0,1703	0,1864
X ₁₀	0,5251	-0,2601	0,6100	-0,2032	0,1843	0,6092	-0,1694	-0,0215	0,8538	1,0000	0,2669	0,1518	0,1472	0,1512
X ₁₁	0,0707	0,0048	0,2123	-0,0349	0,0963	0,2131	0,0861	-0,0466	0,2315	0,2669	1,0000	0,5549	0,2340	0,2102
X ₁₂	0,0474	-0,0595	0,0867	0,0415	0,0119	0,0995	0,1945	-0,0503	0,0389	0,1518	0,5549	1,0000	0,6346	0,5308
X ₁₃	0,1810	-0,1769	0,2194	-0,0631	0,0518	0,1638	0,0132	0,1015	0,1703	0,1472	0,2340	0,6346	1,0000	0,8378
X ₁₄	0,1423	-0,1936	0,1695	-0,1591	-0,1027	0,1195	-0,0904	0,0136	0,1864	0,1512	0,2102	0,5308	0,8378	1,0000

Os autovalores da matriz de correlação populacional são:

$$\begin{aligned} \hat{\lambda}_1 &= 4,9601 ; \hat{\lambda}_2 = 2,4365 ; \hat{\lambda}_3 = 1,6945 ; \hat{\lambda}_4 = 1,2595 ; \hat{\lambda}_5 = 0,8951 ; \hat{\lambda}_6 = 0,8139 ; \\ \hat{\lambda}_7 &= 0,4829 ; \hat{\lambda}_8 = 0,3803 ; \hat{\lambda}_9 = 0,3173 ; \hat{\lambda}_{10} = 0,2459 ; \hat{\lambda}_{11} = 0,2201 ; \hat{\lambda}_{12} = 0,1292 ; \\ \hat{\lambda}_{13} &= 0,1016 ; \hat{\lambda}_{14} = 0,0631. \end{aligned}$$

APÊNDICE 6 - MÉDIA E DESVIO PADRÃO DAS VARIÁVEIS DOS MODELOS AJUSTADOS

QUADRO A.6.1 - MÉDIA E DESVIO PADRÃO DAS VARIÁVEIS DOS MODELOS AJUSTADOS PARA O MAIOR COEFICIENTE DE VARIAÇÃO E MAIOR RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DOS AUTOVALORES, AUTOVETORES, CARGAS FATORIAIS E COMUNALIDADES ESTIMADOS

VARIÁVEIS	MÉDIA	DESVIO PARÃO
logCVmaxautoval (logaritmo decimal do maior coeficiente de variação da estimativa do autovalor)	-1,0527	0,1842
logREQMmaxautoval (logaritmo decimal da maior raiz quadrada do erro quadrático médio relativa da estimativa do autovalor)	-1,0319	0,1976
rautoval (raiz quadrada do autovalor, correspondente ao CVmax e à REQMmax)	1,2147	0,1799
r(n) (raiz quadrada do tamanho da amostra)	13,6292	6,8632
expm (explicação média dos fatores)	0,2331	0,0952
logCVmaxautovet (logaritmo decimal do maior coeficiente de variação da estimativa do autovetor)	1,4898	0,8575
logREQMmaxautovet (logaritmo decimal da maior raiz quadrada do erro quadrático médio relativa da estimativa do autovetor)	1,5311	0,8701
logautovet (logaritmo decimal do valor absoluto da estimativa do autovetor, correspondente ao Cvmax e à REQMmax)	-2,4077	0,885
n (tamanho da amostra)	232,6639	228,2989
r(f/p) (raiz quadrada da razão entre o número de fatores e o de variáveis)	0,5637	0,0417
expl (explicação total dos fatores)	0,7837	0,0461
logCVmaxcar (logaritmo decimal do maior coeficiente de variação da estimativa da carga fatorial)	1,4832	0,8494
logREQMmaxcar (logaritmo decimal da maior raiz quadrada do erro quadrático médio relativa da estimativa da carga fatorial)	1,5240	0,8614
logCVmaxcomun (logaritmo decimal do maior coeficiente de variação da estimativa da comunalidade)	-0,7685	0,3767
logREQMmaxcomun (logaritmo decimal da maior raiz quadrada do erro quadrático médio relativa da estimativa da comunalidade)	-0,7071	0,4248
comun (estimativa da comunalidade, correspondente ao Cvmax e à REQMmax)	0,5261	0,1913
r(n/p) (raiz quadrada da razão entre o tamanho da amostra e número de variáveis)	3,971	1,854
r(f) (raiz quadrada do número de fatores)	1,9322	0,3371

FONTE: Dados obtidos por simulação Monte Carlo

APÊNDICE 7 - AVALIAÇÃO DAS SUPOSIÇÕES DO MODELO DE REGRESSÃO LINEAR MÚLTIPLA E IDENTIFICAÇÃO DE *OUTLIERS* E PONTOS INFLUENTES

Os testes aplicados para avaliar a multicolinearidade, homogeneidade da variância, Gaussianidade dos resíduos e $E(\underline{\varepsilon}) = \underline{0}$ dos modelos ajustados encontram-se apresentados a seguir, bem como a identificação dos *outliers* e pontos influentes.

1 MODELO AJUSTADO PARA ESTIMAR O COEFICIENTE DE VARIAÇÃO DAS ESTIMATIVAS DOS AUTOVALORES, EM FUNÇÃO DAS VARIÁVEIS EXPLICATIVAS: $r_{autoval}$, $r(n)$ e $expm$

1.1 Testes para Verificar Multicolinearidade, Homogeneidade da Variância, Gaussianidade dos Resíduos e $E(\underline{\varepsilon}) = \underline{0}$

QUADRO A.7.1 - ESTATÍSTICAS E VALOR-p SEGUNDO TESTES APLICADOS

TESTES	ESTATÍSTICA	VALOR-p
Multicolinearidade	$k = 1,9884$	
Homogeneidade da variância (Goldfeld-Quandt)	$GQ = 1,2669$	0,2014
Gaussianidade dos resíduos (Kolmogorov-Smirnov com correção de Lilliefors)	$D = 0,0509$	0,1333
Teste t (média do resíduo=0)	$t = -1,5680$	0,1182

FONTE: Dados obtidos por simulação Monte Carlo

O valor de k , para avaliar a multicolinearidade, é menor que 100. Portanto, não existe problema de multicolinearidade entre as variáveis explicativas. Quanto aos demais testes, todos os valores-p são maiores do que 0,01, portanto as suposições de que os resíduos devem distribuir-se normalmente com média zero e variâncias constantes estão atendidas.

1.2 Teste para Identificar *Outlier* (Resíduo Studentizado Externamente)

$$\max |r_{student}| = 3,5585, \text{ g.l.} = 237$$

$$\text{Bonferroni } p = 0,1086$$

Observação : 220

Não é ponto de *outlier*, pois o valor-p de Bonferroni é maior do que 0,01.

1.3 Pontos de Leverages

Foram identificados 3 pontos de *leverages*. São eles: 207, 210, 225.

1.4 Pontos Influentes

Foram identificados 15 pontos influentes. São eles: 4, 31, 32, 49, 51, 120, 138, 187, 193, 207, 210, 214, 220, 225, 229.

2 MODELO AJUSTADO PARA ESTIMAR A RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DOS AUTOVALORES, EM FUNÇÃO DAS VARIÁVEIS EXPLICATIVAS: $r_{autoval}$, $r(n)$ e $expm$

2.1 Testes para Verificar Multicolinearidade, Homogeneidade da Variância, Gaussianidade dos Resíduos e $E(\varepsilon) = 0$

QUADRO A.7.2 - ESTATÍSTICAS E VALOR-p SEGUNDO TESTES APLICADOS

TESTES	ESTATÍSTICA	VALOR-p
Multicolinearidade	$k = 1,9884$	
Homogeneidade da variância (Goldfeld-Quandt)	$GQ = 1,2058$	0,3119
Gaussianidade dos resíduos (Kolmogorov-Smirnov com correção de Lilliefors)	$D = 0,0488$	0,1747
Teste t (média do resíduo=0)	$t = -1,3877$	0,1665

FONTE: Dados obtidos por simulação Monte Carlo

O valor de k é menor que 100, portanto não existe problema de multicolinearidade entre as variáveis explicativas. Quanto aos demais testes, todos os valores-p são maiores do que 0,01, atendendo portanto às suposições de que os resíduos distribuem-se normalmente com média zero e variâncias constantes.

2.2 Teste para Identificar *Outlier* (Resíduo *Studentizado* Externamente)

$$\max |r_{student}| = 3,4106, \text{ g.l} = 237$$

$$\text{Bonferroni } p = 0,1836$$

Observação: 220

Não é ponto de *outlier*, pois o valor-p de Bonferroni é maior do que 0,01.

2.3 Pontos de Leverages

Foram identificados 3 pontos de *leverages*. São eles: 207, 210, 225.

2.4 Pontos Influentes

Foram identificados 16 pontos influentes. São eles: 4, 5, 24, 31, 32, 49, 51, 138, 187, 193, 207, 210, 214, 220, 225, 229.

O quadro A.7.3 apresenta os tamanhos de amostra, números de variáveis e de fatores, estimativa da explicação total dos fatores e do autovalor, CVmax e REQMax, correspondentes aos pontos de *leverages* e influentes.

QUADRO A.7.3 - TAMANHOS DE AMOSTRA, NÚMEROS DE VARIÁVEIS E DE FATORES, ESTIMATIVAS DA EXPLICAÇÃO TOTAL DOS FATORES E DO AUTOVALOR, CVmax E REQMax, SEGUNDO O NÚMERO DA AMOSTRA

NÚMERO DA AMOSTRA	TAMANHOS DE AMOSTRA	NÚMERO DE VARIÁVEIS	NÚMERO DE FATORES	EXPLICAÇÃO TOTAL DOS FATORES	ESTIMATIVA DO AUTOVALOR	CVmax	REQMax
4	233	5	2	0,8118	1,1134	0,0571	0,0573
5	59	5	2	0,8166	1,1192	0,0976	0,0976
24	30	6	2	0,8335	1,3288	0,2965	0,2967
31	272	6	2	0,7257	1,1027	0,0514	0,0514
32	68	6	2	0,7349	1,1293	0,0814	0,0850
49	396	8	2	0,8349	1,5605	0,0708	0,0708
51	44	8	2	0,8378	1,5991	0,2117	0,2133
120	50	12	4	0,8006	2,9228	0,1147	0,1148
138	50	13	4	0,8278	2,3836	0,1519	0,1697
187	690	17	5	0,7670	2,9976	0,0436	0,0440
193	695	17	5	0,8023	2,7289	0,0472	0,0475
207	66	18	5	0,8017	3,3683	0,1321	0,1366
210	66	18	5	0,8275	3,4538	0,1322	0,1384
214	704	19	6	0,7878	2,3849	0,0463	0,0467
220	702	19	5	0,7369	3,8291	0,0422	0,0425
225	79	19	6	0,7828	3,5413	0,1000	0,1023
229	978	20	6	0,7486	1,7399	0,0395	0,0398

FONTE: Dados obtidos por simulação Monte Carlo

NOTA: As estimativas da explicação total dos fatores e dos autovalores são médias de 1.000 amostras.

3 MODELO AJUSTADO PARA ESTIMAR O COEFICIENTE DE VARIAÇÃO DAS ESTIMATIVAS DOS AUTOVETORES, EM FUNÇÃO DAS VARIÁVEIS EXPLICATIVAS: \log_{autovet} , n , $r(f/p)$ e expl

3.1 Testes para Verificar Multicolinearidade, Homogeneidade da Variância, Gaussianidade dos Resíduos e $E(\underline{\varepsilon}) = \underline{0}$

QUADRO A.7.4 - ESTATÍSTICAS E VALOR-p SEGUNDO TESTES APLICADOS

TESTES	ESTATÍSTICA	VALOR-p
Multicolinearidade	$k = 4,6134$	
Homogeneidade da variância (Goldfeld-Quandt)	$GQ = 1,0302$	0,8731
Gaussianidade dos resíduos (Kolmogorov-Smirnov com correção de Lilliefors)	$D = 0,0344$	0,6944
Teste t (média do resíduo=0)	$t = -0,0507$	0,9596

FONTE: Dados obtidos por simulação Monte Carlo

O valor de k é menor que 100, portanto não existe problema de multicolinearidade entre as variáveis explicativas. Quanto aos demais testes, todos os valores-p são maiores do que 0,01, atendendo portanto às suposições de que os resíduos distribuem-se normalmente com média zero e variâncias constantes.

3.2 Teste para identificar *Outlier* (Resíduo *Studentizado* Externamente)

Não foi identificado nenhum ponto de *outlier*.

3.3 Pontos de *Leverages*

Foram identificados 7 pontos de *leverages*. São eles: 49, 50, 51, 232, 235, 238, 241.

3.4 Pontos Influentes

Foram identificados 17 pontos influentes. São eles 1, 3, 10, 11, 12, 18, 28, 43, 46, 49, 50, 67, 160, 163, 235, 238, 241.

4 MODELO AJUSTADO PARA ESTIMAR A RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DOS AUTOVETORES, EM FUNÇÃO DAS VARIÁVEIS EXPLICATIVAS: \log_{autovet} , n , $r(f/p)$ e expl

4.1 Testes para Verificar Multicolinearidade, Homogeneidade da Variância, Gaussianidade dos Resíduos e $E(\underline{\varepsilon}) = \underline{0}$

QUADRO A.7.5 - ESTATÍSTICAS E VALOR-p SEGUNDO TESTES APLICADOS

TESTES	ESTATÍSTICA	VALOR-p
Multicolinearidade	$k = 4,6134$	
Homogeneidade da variância (Goldfeld-Quandt)	$GQ = 1,0149$	0,9367
Normalidade dos resíduos (Kolmogorov-Smirnov com correção de Lilliefors)	$D = 0,0564$	0,0610
Teste t (média do resíduo=0)	$t = -0,0143$	0,9886

FONTE: Dados obtidos por simulação Monte Carlo

O valor de k é menor que 100, portanto não existe multicolinearidade entre as variáveis explicativas. Quanto aos demais testes, todos os valores-p são maiores do que 0,01, atendendo portanto às suposições de que os resíduos distribuem-se normalmente com média zero e variâncias constantes.

4.2 Teste para Identificar *Outlier* (Resíduo *Studentizado* Externamente)

Nenhum ponto de *outlier* foi identificado.

4.3 Pontos de *Leverages*

Foram identificador 7 pontos de *leverages*, que são: 49, 50, 51, 232, 235, 238, 241.

4.4 Pontos Influentes

Foram identificados 17 pontos influentes. São eles: 1, 3, 10, 12, 18, 25, 28, 43, 46, 49, 50, 76, 160, 163, 190, 238, 241.

O quadro A.7.6 apresenta os tamanhos de amostra, números de variáveis e de fatores, estimativas da explicação total dos fatores e do autovetor, CV_{max} e $REQM_{\text{max}}$, correspondentes aos pontos (amostras) influentes e de *leverages*.

QUADRO A.7.6 - TAMANHOS DE AMOSTRA, NÚMEROS DE VARIÁVEIS E DE FATORES, ESTIMATIVAS DA EXPLICAÇÃO TOTAL DOS FATORES E DO AUTOVETOR, C_{vmax} E $REQM_{max}$, SEGUNDO O NÚMERO DA AMOSTRA

NÚMERO DA AMOSTRA	TAMANHOS DE AMOSTRA	NÚMERO DE VARIÁVEIS	NÚMERO DE FATORES	EXPLICAÇÃO TOTAL DOS FATORES	ESTIMATIVA DO AUTOVETOR	C_{vmax}	$REQM_{max}$
1	235	5	2	0,9023	-0,1907	0,2031	0,2045
3	27	5	2	0,9055	0,3634	1,4116	1,5859
10	214	5	2	0,9019	-0,1877	0,2256	0,2287
11	54	5	2	0,9021	0,4803	0,8593	0,9135
12	24	5	2	0,9032	0,3338	1,6146	1,8411
18	24	5	2	0,8817	-0,3185	1,6504	1,9195
25	269	6	3	0,8811	0,0158	4,5595	4,5626
28	268	6	2	0,7614	0,1288	0,3668	0,3676
43	344	7	2	0,8138	0,1358	0,2419	0,2419
46	344	7	2	0,7984	0,0876	0,3534	0,3534
49	396	8	2	0,8349	-0,1311	0,1682	0,1684
50	99	8	2	0,8367	-0,1285	0,3628	0,3639
51	44	8	2	0,8378	-0,1227	0,6512	1,3059
67	406	9	3	0,7351	-0,0372	10,6519	11,2211
76	414	9	2	0,6363	0,2013	2,1071	2,4900
160	502	15	5	0,7145	0,0026	112,3834	151,0975
163	501	15	5	0,7680	0,0004	634,8399	899,4263
190	691	17	5	0,7603	0,0021	53,1637	83,1882
232	982	20	6	0,7763	-0,0042	10,9573	10,9588
235	984	20	5	0,7846	0,0170	2,1127	2,1130
238	984	20	5	0,7871	0,0053	8,2316	8,2318
241	983	20	6	0,7432	0,0018	32,6302	35,5179

FONTE: Dados obtidos por simulação Monte Carlo

NOTA: As estimativas da explicação total dos fatores e dos autovetores são médias de 1.000 amostras.

5 MODELO AJUSTADO PARA ESTIMAR O COEFICIENTE DE VARIAÇÃO DAS ESTIMATIVAS DAS CARGAS FATORIAIS, EM FUNÇÃO DAS VARIÁVEIS EXPLICATIVAS: $\log car$, n , $r(f/p)$ e $expl$

5.1 Testes para Verificar Multicolinearidade, Homogeneidade da Variância, Gaussianidade dos Resíduos e $E(\varepsilon) = 0$

QUADRO A.7.7 - ESTATÍSTICAS E VALOR-p SEGUNDO TESTES APLICADOS

TESTES	ESTATÍSTICA	VALOR-p
Multicolinearidade	$k = 4,5721$	
Homogeneidade da variância (Goldfeld-Quandt)	$GQ = 0,9757$	0,8944
Gaussianidade dos resíduos (Kolmogorov-Smirnov com correção de Lilliefors)	$D = 0,0547$	0,0774
Teste t (média do resíduo=0)	$t = -0,0008$	0,9994

FONTE: Dados obtidos por simulação Monte Carlo

O valor de k é menor que 100, portanto não existe problema de multicolinearidade entre as variáveis explicativas. Quanto aos demais testes, todos os valores- p são maiores do que 0,01, atendendo portanto às suposições de que os resíduos distribuem-se normalmente com média zero e variâncias constantes.

5.2 Teste para Identificar *Outlier* (Resíduo *Studentizado* Externamente)

Não foi identificado ponto de *outlier*.

5.3 Pontos de *Leverages*

Foram identificados 8 pontos de *leverages*: São eles: 49, 50, 51, 229, 232, 235, 238, 241.

5.4 Pontos Influentes

Foram identificados 16 pontos influentes. São eles: 3, 12, 18, 28, 43, 46, 49, 50, 70, 73, 76, 160, 163, 235, 238, 241.

6 MODELO AJUSTADO PARA ESTIMAR A RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DAS CARGAS FATORIAIS, EM FUNÇÃO DAS VARIÁVEIS EXPLICATIVAS: $\log car$, n , $r(f/p)$ e $expl$

6.1 Testes para Verificar Multicolinearidade, Homogeneidade da Variância, Gaussianidade dos Resíduos e $E(\varepsilon) = 0$

QUADRO A.7.8 - ESTATÍSTICAS E VALOR- p SEGUNDO TESTES APLICADOS

TESTES	ESTATÍSTICA	VALOR- p
Multicolinearidade	$k = 4,5721$	
Homogeneidade da variância (Goldfeld-Quandt)	$GQ = 0,9654$	0,8496
Gaussianidade dos resíduos (Kolmogorov-Smirnov com correção de Lilliefors)	$D = 0,0478$	0,1984
Teste t (média do resíduo=0)	$t = 0,0295$	0,9765

FONTE: Dados obtidos por simulação Monte Carlo

O valor de k para avaliar a multicolinearidade é menor que 100, portanto não existe problema de multicolinearidade entre as variáveis explicativas. Quanto aos demais

testes, todos os valores-p são maiores do que 0,01, atendendo portanto às suposições de que os resíduos distribuem-se normalmente com média zero e variâncias constantes.

6.2 Teste para Identificar *Outlier* (Resíduo *Studentizado* Externamente)

Nenhum ponto de *outlier* foi identificado.

6.3 Pontos de *Leverages*

Foram identificados 8 pontos de *leverages*: São eles: 49, 50, 51, 229, 232, 235, 238, 241.

6.4 Pontos Influentes

Foram identificados 18 pontos influentes. São eles: 3, 12, 18, 25, 43, 46, 49, 50, 70, 73, 76, 125, 160, 163, 190, 235, 238, 241.

O quadro A.7.9, apresenta os tamanhos de amostra, números de variáveis e de fatores, estimativas da explicação total dos fatores e da carga fatorial, CVmax e REQmmax, correspondentes às amostras identificadas como pontos de *leverages* e influentes.

QUADRO A.7.9 - TAMANHOS DE AMOSTRA, NÚMEROS DE VARIÁVEIS E DE FATORES, ESTIMATIVAS DA EXPLICAÇÃO TOTAL DOS FATORES E DAS CARGAS FATORIAIS, CVmax E REQmmax, SEGUNDO O NÚMERO DA AMOSTRA

NÚMERO DA AMOSTRA	TAMANHOS DE AMOSTRA	NÚMERO DE VARIÁVEIS	NÚMERO DE FATORES	EXPLICAÇÃO TOTAL DOS FATORES	ESTIMATIVA DAS CARGAS FATORIAIS	CVmax	REQmmax
3	27	5	2	0,9055	0,4118	1,5390	1,7512
12	24	5	2	0,9032	0,3732	1,7697	2,0525
18	24	5	2	0,8817	-0,3872	1,7760	2,0922
25	269	6	3	0,8811	0,0215	4,5036	4,5075
28	268	6	2	0,7614	0,2294	0,3712	0,3719
43	344	7	2	0,8138	0,1589	0,2551	0,2552
46	344	7	2	0,7984	0,1082	0,3422	0,3422
49	396	8	2	0,8349	-0,1638	0,1752	0,1753
50	99	8	2	0,8367	-0,1614	0,3714	0,3720
51	44	8	2	0,8378	-0,1535	0,6755	0,6795
70	414	9	3	0,6888	0,0032	145,6121	207,6058
73	414	9	3	0,7155	0,0004	150,3994	156,7841
76	414	9	2	0,6363	0,3701	2,0949	2,4723
125	110	13	4	0,8025	0,0011	528,1218	825,7367
160	502	15	5	0,7145	-0,0032	122,0632	163,6294
163	501	15	5	0,7680	0,0007	470,9324	665,0234
190	691	17	5	0,7603	0,0027	56,6755	89,6575
229	978	20	6	0,7486	-0,0004	123,4965	123,7059
232	982	20	6	0,7763	-0,0056	10,6936	10,6954
235	984	20	5	0,7846	0,0242	2,1120	2,1123
238	984	20	5	0,7871	0,0056	8,2600	8,2602
241	983	20	6	0,7432	0,0024	32,1502	34,9836

FONTE: Dados obtidos por simulação Monte Carlo

NOTA: As estimativas da explicação total dos fatores e das cargas fatoriais são médias de 1.000 amostras.

7 MODELO AJUSTADO PARA ESTIMAR O COEFICIENTE DE VARIAÇÃO DAS ESTIMATIVAS DAS COMUNALIDADES, EM FUNÇÃO DAS VARIÁVEIS EXPLICATIVAS: comun, $r(n/p)$ e $r(f)$

7.1 Testes para Verificar Multicolinearidade, Homogeneidade da Variância, Gaussianidade dos Resíduos e $E(\underline{\varepsilon}) = \underline{0}$

QUADRO A.7.10 - ESTATÍSTICAS E VALOR-p SEGUNDO TESTES APLICADOS

TESTES	ESTATÍSTICA	VALOR-p
Multicolinearidade	$k = 1,4721$	
Homogeneidade da variância (Goldfeld-Quandt)	$GQ = 1,2668$	0,2016
Gaussianidade dos resíduos (Kolmogorov-Smirnov com correção de Lilliefors)	$D = 0,0554$	0,0703
Teste t (média do resíduo=0)	$t = 1,0858$	0,2787

FONTE: Dados obtidos por simulação Monte Carlo

O valor de k para avaliar a multicolinearidade é menor que 100, portanto não existe problema de multicolinearidade entre as variáveis explicativas. Quanto aos demais testes, todos os valores-p são maiores do que 0,01, atendendo portanto às suposições de que os resíduos distribuem-se normalmente com média zero e variâncias constantes.

7.2 Teste para Identificar *Outlier* (Resíduo *Studentizado* Externamente)

$$\max |r_{student}| = 3,0411, \text{ g.l} = 237$$

$$\text{Bonferroni } p = 0,6319$$

Observação: 163

Não é ponto de *outlier*, pois o valor-p de Bonferroni é maior do que 0,01.

7.3 Pontos de *Leverages*

Foram identificados 02 pontos de *leverages*: 1, 229.

7.4 Pontos Influentes

Foram identificados 10 pontos influentes. São eles: 37, 64, 67, 70, 73, 76, 115, 118, 160, 163.

8 MODELO AJUSTADO PARA ESTIMAR A RAIZ QUADRADA DO ERRO QUADRÁTICO MÉDIO RELATIVA DAS ESTIMATIVAS DAS COMUNALIDADES, EM FUNÇÃO DAS VARIÁVEIS EXPLICATIVAS: comun , $r(n/p)$ e $r(f)$

8.1 Testes para Verificar Multicolinearidade, Homogeneidade da Variância, Gaussianidade dos Resíduos e $E(\varepsilon) = 0$

QUADRO A.7.11 - ESTATÍSTICAS E VALOR-p SEGUNDO TESTES APLICADOS

TESTES	ESTATÍSTICA	VALOR-p
Multicolinearidade	$k = 1,4721$	
Homogeneidade da variância (Goldfeld-Quandt)	$GQ = 1,0988$	0,6108
Gaussianidade dos resíduos (Kolmogorov-Smirnov com correção de Lilliefors)	$D = 0,0430$	0,3415
Teste t (média do resíduo=0)	$t = 1,1881$	0,2360

FONTE: Dados obtidos por simulação Monte Carlo

O valor de k , para avaliar a multicolinearidade é menor que 100, portanto, não existe problema de multicolinearidade entre as variáveis explicativas. Quanto aos demais testes, todos os valores-p são maiores do que 0,01, atendendo portanto as suposições de que o resíduos distribuem-se normalmente com média zero e variâncias constantes.

8.2 Teste para Identificar *Outlier* (Resíduo *Studentizado* Externamente)

Não foi identificado nenhum ponto de *outlier*.

8.3 Pontos de *Leverages*

Foram identificados 2 pontos de *leverages*, que são: 1, 229.

8.4 Pontos Influentes

Foram identificados 6 pontos influentes. São eles: 64, 67, 70, 73, 76, 118.

O quadro A.7.12 apresenta os tamanhos de amostra, números de variáveis e de fatores, estimativas da explicação total dos fatores e da comunalidade, CV_{\max} e $REQM_{\max}$, correspondentes às amostras identificadas como pontos de *leverages* e influentes.

QUADRO A.7.12 - TAMANHOS DE AMOSTRA, NÚMEROS DE VARIÁVEIS E DE FATORES, ESTIMATIVAS DA EXPLICAÇÃO TOTAL DOS FATORES E DA COMUNALIDADE, C_{vmax} E $REQM_{max}$, SEGUNDO O NÚMERO DA AMOSTRA

NÚMERO DA AMOSTRA	TAMANHOS DE AMOSTRA	NÚMERO DE VARIÁVEIS	NÚMERO DE FATORES	EXPLICAÇÃO TOTAL DOS FATORES	ESTIMATIVA DA COMUNALIDADE	C_{vmax}	$REQM_{max}$
1	235	5	2	0,9023	0,8481	0,0203	0,0204
37	345	7	2	0,7121	0,2002	0,4686	0,4701
64	408	9	3	0,7248	0,2097	0,8676	1,1171
67	406	9	3	0,7351	0,1903	1,0107	1,3793
70	414	9	3	0,6888	0,0819	0,7566	0,9626
73	414	9	3	0,7155	0,1255	0,8911	1,0787
76	414	9	2	0,6363	0,0241	0,8551	0,9625
115	446	12	4	0,7996	0,6174	0,2963	0,2963
118	445	12	4	0,7764	0,8312	0,1441	0,1468
160	502	15	5	0,7145	0,5293	0,4439	0,4474
163	501	15	5	0,7680	0,6529	0,3534	0,3565
229	978	20	6	0,7486	0,0892	0,4842	0,6023

FONTE: Dados obtidos por simulação Monte Carlo

NOTA: As estimativas da explicação total dos fatores e das comunalidades, são médias de 1.000 amostras.

APÊNDICE 8 - AUTOVALOR, FORMA QUADRÁTICA E PROPRIEDADE DOS DETERMINANTES

1 AUTOVALOR

Seja $C_{p \times p}$ uma matriz não-singular. Então

$$|A - \lambda I| = |C| |A - \lambda C^{-1} C| |C^{-1}| = |CAC^{-1} - \lambda I| \quad (\text{A8.1})$$

assim, tem-se que A e CAC^{-1} têm o mesmo autovalor (MARDIA, KENT e BIBBY, 1982).

2 FORMA QUADRÁTICA

A forma quadrática nas variáveis X_1, X_2, \dots, X_n é a expressão do tipo:

$$f(x_1, x_2, \dots, x_n) = a_{11}x_1^2 + a_{22}x_2^2 + \dots + a_{nn}x_n^2 + 2a_{12}x_1x_2 + \dots + 2a_{(n-1)(n)}x_{n-1}x_n = \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_i x_j \quad (\text{A8.2})$$

A forma quadrática pode ser escrita na notação matricial como se segue:

$$Q = \underline{x}' A \underline{x} \quad (\text{A8.3})$$

onde $\underline{x}' = [x_1, x_2, \dots, x_n]$ e A é uma matriz simétrica $n \times n$ de coeficientes.

2.1 Maximização de Forma Quadrática

Uma alternativa conveniente é considerar uma solução normalizada de \underline{x} , isto é, uma solução que tenha comprimento unitário. Portanto, a maximização da forma quadrática Q pode ser transformada na maximização da razão:

$$\lambda = \frac{\underline{x}' A \underline{x}}{\underline{x}' \underline{x}} \quad (\text{A8.4})$$

Para a maximização deve-se derivar a expressão (A8.4) em relação a \underline{x} e igualar a zero, conforme demonstrado a seguir.

$$\frac{\partial \lambda}{\partial \underline{x}} = \frac{2A\underline{x}(\underline{x}'\underline{x}) - 2(\underline{x}'A\underline{x})\underline{x}}{(\underline{x}'\underline{x})^2} = \frac{2}{\underline{x}'\underline{x}} \left(A - \frac{\underline{x}'A\underline{x}}{\underline{x}'\underline{x}} I \right) \underline{x} \quad (\text{A8.5})$$

Igualando a expressão (A8.5) a zero e dividindo por $\frac{2}{\underline{x}'\underline{x}}$, obtém-se o sistema homogêneo de equações:

$$\left(A - \frac{\underline{x}'A\underline{x}}{\underline{x}'\underline{x}} I \right) \underline{x} = 0 \quad (\text{A8.6})$$

Sustituindo a expressão (A8.4) em (A8.6) tem-se:

$$(A - \lambda I)\underline{x} = 0 \quad (\text{A8.7})$$

Para que o sistema acima não possua apenas a solução trivial, $(A - \lambda I)$ não pode ter posto completo, o que significa que seu determinante deve ser igual a zero:

$$|A - \lambda I| = 0 \quad (\text{A8.8})$$

3 PROPRIEDADE DOS DETERMINANTES

Seja uma matriz A com as seguintes partições:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

O determinante de A , se A_{11} e A_{22} são não singulares, é dado por:

$$|A| = |A_{11}| |A_{22} - A_{21}A_{11}^{-1}A_{12}|, \text{ ou} \\ |A| = |A_{22}| |A_{11} - A_{12}A_{22}^{-1}A_{21}| \quad (\text{A8.9})$$