UNIVERSIDADE FEDERAL DO PARANÁ

CARLOS SMAKA

APLICAÇÃO DA ANÁLISE MULTIVARIADA NA IDENTIFICAÇÃO DE FATORES QUE INFLUENCIAM NO CUSTO DE UM PLANO DE SAÚDE

Dissertação apresentada como requisito parcial a obtenção de grau de mestre, do curso de pós-graduação em Métodos Numéricos de Engenharia do Departamento de Engenharia Civil e Setor de Ciências da Universidade Federal do Paraná.
Orientador: Jair Mendes Marques

Curitiba

2010

TERMO DE APROVAÇÃO

CARLOS SMAKA

APLICAÇÃO DA ANÁLISE MULTIVARIADA NA IDENTIFICAÇÃO DE FATORES QUE INFLUENCIAM NO CUSTO DE UM PLANO DE SAÚDE

Dissertação aprovada como requisito parcial à obtenção de grau de Mestre em Ciências, Programa de Pós-Graduação em Métodos Numéricos de Engenharia, área de concentração em Programação Matemática, Setor de Tecnologia, Departamento de Construção Civil e Setor de Ciências Exatas, Departamento de Matemática da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientador: Prof. Dr. Jair Mendes Marques
Programa de Pós – Graduação em Métodos Numéricos em Engenharia - UFPR.

Prof. Dr. Anselmo Chaves Neto
Programa de Pós-Graduação em Métodos Numéricos em Engenharia - UFPR

Prof. Dr. Mário Romero Pelegrini de Souza
FAE - Centro Universitário

Curitiba

2010

Ao meu filho Eduardo Smaka.

Agradecimentos

Primeiro a Deus .....

À Universidade Federal do Paraná que através do Programa de Pós Graduação em Engenharia de Métodos Numéricos, viabilizou o curso;

Ao meu orientador Prof Dr. Jair Mendes Marques, pelo apoio, tolerância, dedicação e competência com que conduziu o desenvolvimento deste trabalho, oferecendo toda contribuição necessária à realização do mesmo;

Aos colegas do curso pelo apoio e amizade, especialmente Edson Rovina, Thober Coradi Detofeno**,** Clodoaldo José Figueiredo, meu grupo de estudos;

Ao Departamento de Matemática;

À Maristela Bandil secretária do CESEC, pelas ajudas e pela sua competência no que faz;

Ao meu filho Eduardo Smaka pela compreensão e apoio;

A todos que direta ou indiretamente contribuíram para a realização

desta pesquisa.

RESUMO

O custo com saúde representa uma parcela significativa dos gastos gerais de várias empresas. Embora maioria das empresas ofereçam um plano de benefícios, muitas vezes o uso inadequado contribui para o aumento dos custos das empresas. O presente trabalho tem como objetivo a identificação das variáveis que mais influenciam no alto custo de um plano de saúde privado. Metodologia: A presente pesquisa envolve um plano de saúde composto por 5008 usuários,contendo inicialmente 20 variáveis. A análise utilizada envolve o método multivariado conhecido como Análise de Componentes principais (ACP) e a Análise de Regressão Múltipla (ARM). Resultados: Usando a metodologia das componentes principais obtém uma redução da dimensionalidade dos dados, com isso, somente aquelas que influenciam expressivamente o conjunto de dados são preservados. Tão importante afirmação que na análise da correlação variáveis originais versus componentes principais, destacam-se as (07) sete primeiras componentes representando 71% de todas as informações iniciais, evidenciando entre 20 variáveis as que contribuem no aumento do custo do plano de saúde. O ajuste do modelo de regressão linear múltipla aconteceu a partir do uso dos escores das componentes principais (ACP), que verificou que essas (07) sete variáveis estão significativamente relacionadas com o alto custo anual do plano de saúde, com uma representação de 60% entre todas as variáveis. Considerações finais: Pode a empresa estudada ter uma previsão de quanto cada titular do plano de saúde custa anualmente para a empresa, através de suas informações e de seus dependentes, permitindo a mesma proporcionar políticas de intervenção junto a esses em relação ao alto custo do plano de saúde e seu aumento.

ABSTRACT

The cost of healthcare represents a significant portion of overall costs for many companies. While most companies offer a benefit plan, often misuse contributes to increased business costs. This study aims to identify the variables that most influence the high cost of a private health plan. Methodology: This research involves a health plan consisting of 5,008 users, initially containing 20 variables. The analysis used involves the multivariate method known as principal component analysis (PCA) and Multiple Regression Analysis (MRA). Results: Using the method of principal components obtained a reduction of dimensionality and, thus, only those that significantly influence the data set are preserved. As important affirmation that the original variables correlation analysis versus principal components stand out (07) the first seven principal components representing 71% of all variables, showing among the 20 variables that contribute to the rising cost of health insurance. The model fit multiple regression occurred from the use of scores of principal components (PCA), which verified that these (07) seven variables are significantly related to the high annual cost of health insurance, with a representation of 60% between all variables. Conclusion: Can the company studied has a preview of how each holder of the health plan costs for the company, through its information and their dependents, enabling it to provide policy intervention with respect to those in high-cost plan health and its increase.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE ABREVIATURAS

ABNT – Associação Brasileira de Normas Técnicas

IBGE – Instituto Brasileiro de Geografia e Estatística

IDH – Índices de Desenvolvimento Humano

NBR – Norma Brasileira de Regulamentação

(ACP) - Análise de componentes principais (ACP).

(RH) --Recursos Humanos RH

(Pnda) - Pesquisa Nacional por Amostra de Domicílios Pnda

(ANS) - Agencia Nacional de Saúde

(SUS) - Sistema Único de Saúde

SUMÁRIO

INTRODUÇÃO

Investir em prevenção a fim de melhorar a saúde de seus funcionários e baixar os custos com assistência médica: este é um dos focos das corporações nos dias de hoje para reduzir as despesas. É cada vez mais freqüente empresas oferecerem aos seus trabalhadores opções que vão desde palestras sobre o tema, academias de ginástica, ginásticas laboral no inicio de turno, revezamento de operadores em máquinas onde só se trabalha uma hora por dia evitando assim doenças por movimentos repetitivos, até a elaboração de cardápio balanceado feito por nutricionistas e oferecido, em alguns casos, no refeitório da empresa.

E isso tem uma explicação lógica. O benefício é valorizadíssimo pelos funcionários e é o mínimo que eles esperam de uma empresa com uma política moderna de recursos humanos. O Brasil, segundo o IBGE (1998), país onde a classe média gasta 10% do orçamento familiar com saúde e o sistema de atendimento público é absolutamente precário, vincular-se a um plano empresarial sempre foi uma alternativa bem mais atraente do que pagar um plano privado do próprio bolso.

Sem contar que um funcionário com boa saúde tem melhor desempenho, além disso, a empresa consegue reduzir os índices de absenteísmo e os gastos com saúde. Este último item é o que mais pesa numa organização EMPRESATIVA (2008). O problema que assombra as empresas aqui e lá fora é que a assistência médica está se tornando um benefício financeiramente insustentável. ROSENBURG (2005), estima que os gastos anuais das companhias americanas com esse item alcancem 389 bilhões de dólares. Se a economia dos Estados Unidos mantiver o ritmo atual de crescimento, em 2008 uma grande corporação americana típica gastará com o plano de saúde o equivalente a todo o seu lucro, segundo estimativas da consultoria MCKINSEY (2008). No Brasil, dos 40,7 milhões de beneficiários de planos de saúde, quase 30 milhões mais de 70% do total são vinculados a planos contratados por empresas segundo a revista portal Exame.

O custo com saúde é de fato alto e preocupante para as corporações, tanto para aquelas que trabalham com auto-gestão quanto para as que contratam serviços de uma operadora de saúde. E essa despesa aumenta muitas vezes por conta do

uso inadequado do plano de saúde. Embora a empresa ofereça o plano como benefício, falta, talvez, a consciência adequada de seu uso.

É muito comum pessoas fazerem verdadeiras peregrinações em médicos credenciados, ou por não confiarem na avaliação de algum profissional, ou porque são hipocondríacos, ou, ainda, são portadores de alguma doença crônica, estes últimos acabam resultando nos casos mais alarmantes.

Em situações como essas, segundo a REVISTA DIGITAL (2008), uma grande contribuição pode ser dada pela área de Recursos Humanos das corporações que é ajudar o funcionário a usufruir seu plano de saúde da melhor forma possível. No atual momento de crise financeira do estado, torna-se importante o uso adequado dos recursos econômicos disponíveis, gastar menos e melhor deve ser um dos objetivos a ser seguido pelo setor da saúde.

A reflexão sobre o tema, também o estudo aqui das razões importantes envolvidas no crescimento dos gastos em saúde sendo um dos grandes desafios na gestão de um plano privado de assistência médico-hospitalar e o manejo entendimento como se correlacionam que resultam em gastos elevados. Os gastos com saúde não afetam as pessoas de forma uniforme, e é quase intuitivo que uma minoria gasta muito, enquanto a grande maioria gasta pouco, visto que em muitas vezes dentro do mesmo plano de saúde já existe hierarquia quanto ao tipo e número de dependentes que titulares podem indicar dentro de um mesmo plano de saúde.

A presente pesquisa tem por objetivo identificar quais as variáveis que influenciam no alto custo de um plano de saúde privado. Essa identificação pode proporcionar a empresa melhor controle e entendimento das características de seus colaboradores e dependentes quanto ao uso e consumo do plano de saúde, e o que esses influenciam realmente no alto custo deste benefício, para isso tornar-se mais visível e palpável, utiliza-se as técnicas estatísticas de análise de componentes principais (ACP) e regressão linear múltipla. Pois, a possibilidade de se obter uma ferramenta poderosa como uma equação que possa correlacionar essas variáveis, junto com as principais características de cada colaborador e seu custo, sem mencionar que é possível uma análise individual de cada titular, bastando com isso algumas poucas informações relacionadas ao titular do plano.

1.1 OBJETIVOS DO TRABALHO

1.1.1 Objetivo geral

Estratificar dados obtidos através de relatórios de gastos mensais, apresentado pela empresa Operadora do Plano de Saúde. Identificar quais as variáveis que tem maior influência no alto custo de um plano de saúde privado.

1.1.2 Objetivos específicos

- Estratificar as informações dos relatórios mensais fornecidos pela Operadora do Plano de Saúde à empresa.

- Agrupar essa informações mensais transformando-as em anuais, formando assim uma matriz de informações que sejam relevante para este estudo de análise das informações.

- Aplicar a técnica da análise de componentes principais, reduzindo segundo critérios as variáveis originais em um grupo menor de componentes principais.

- Correlacionar as variáveis originais com essas componentes principais verificando – se a porcentagem de explicação de cada componente através de seus autovalores.

- Aplicar a técnica de regressão múltipla nos escorres fatoriais analisando a real importância de cada informação obtida acima representa no custo do plano de saúde.

1.2 DELIMITAÇÃO DO TEMA

O universo do estudo foi uma amostra dos relatórios mensais que a empresa em questão recebe da operadora do seu plano de saúde, que contem informações dos titulares do plano de saúde e de seus dependentes. O acesso à base de dados foi autorizado pela empresa, salvaguardado o aspecto sigiloso da informação. Portanto, o nome da empresa e a dos seus colaboradores não são citados, havendo a existência de um contrato de sigilo assinado pelo pesquisador junto a empresa.

1.3 JUSTIFICATIVA DO TRABALHO

Os crescentes gastos com o benefício do plano de saúde vêm preocupando os diretores e acionistas da empresa, pois o seu crescimento anual é de 10% a 15% conforme levantamento efetuado internamente. Analisar e compreender as variáveis que influenciam os custos permite traçar políticas racionais para a redução.

Os gastos com a assistência médica dos funcionários não param de crescer e ameaçam comprometer o resultado da empresa e a continuação do benefício pois a uma temeridade do mesmo se tornar inacessível financeiramente.

1.4 ESTRUTURA DO TRABALHO

O trabalho está estruturada em cinco capítulos, construídos de forma a facilitar a leitura e compreensão das metodologias e técnicas aqui descritas e aplicadas.

O primeiro capítulo que comporta a introdução do trabalho, contém uma contextualização do assunto, objetivos, métodos de desenvolvimento, justificativa, delimitação e estrutura da pesquisa.

O segundo capítulo contém uma revisão de literatura abordando o conteúdos utilizados neste trabalho e a descrição do problema onde temos uma

prévia do que estamos a pesquisar. Este capítulo dedica-se a descrever o cenário atual no qual se insere este problema de pesquisa.

O capítulo 3 e 4 trata da parte central da pesquisa, com a descrição detalhada do material e metodologia aplicada, e da análise e dos resultados obtidos, propostos para esta pesquisa.

No capítulo cinco temos os resultados e análises deste trabalho.

REVISÃO DE LITERATURA

2.1 PLANOS DE SAÚDE

No Brasil, o sistema público de saúde foi regulamentado em 1988 determinando acesso universal, integral e gratuito para toda a população e permitindo a livre atuação do setor privado. O sistema de saúde suplementar cobre cerca de 30% da população e essa participação tem se mantido praticamente estável nos últimos 10 anos segundo (IBGE 1998). Essa opção de sistema institucional embora seja democrática gera iniqüidades no acesso aos serviços de saúde. Os grupos de status sócio-econômico mais elevado têm duplo acesso ao sistema. Apesar dessa iniqüidade a ampliação da população coberta por seguro privado é uma alternativa interessante do ponto de vista de bem estar social na medida em que pode minorar o problema de congestão no provimento dos serviços públicos de saúde, fazendo com que cada vez mais empresas privadas possam oferecer a seus colaboradores o acesso a Planos de Saúde privado. Mas gerenciar o benefício saúde nas empresas continua como um dos grandes desafios do Recurso Humano, nem toda a criatividade do mercado é suficiente para conter os aumentos de custos e manter a qualidade da assistência médica no padrão exigido pelos funcionários.

A despeito do número de beneficiários de plano de saúde privado, cerca de quarenta milhões, que corresponde ao segundo maior mercado de planos de saúde privado mundial, não existem estudos que proponham modelos de estimação de demanda. Como mencionado, em um sistema como o nosso, no qual coexiste o financiamento público e privado, conhecer os fatores que contribuem e interferem para o aumento exorbitante dos valores do plano de saúde é importante para o estabelecimento de políticas públicas e privadas que visem a ampliação da cobertura ou a sua continuação. Um maior grau de cobertura gera ganhos de bem estar social uma vez que reduz a incerteza associada ao estado de saúde, aumenta o acesso aos serviços preventivos e por conseqüência pode melhorar o estado de saúde médio da população, o que reflete em níveis maiores de produtividade. Além disso, o sistema privado é um sistema alternativo ao sistema público e a ampliação

da cobertura pode resultar em uma redução do tempo de espera e melhora da qualidade de vida de seus usuários.

2.1.1 Plano de Saúde Empresarial

Os planos de saúde empresariais representam uma importante parcela do segmento de assistência médica suplementar. Segundo informações da Pesquisa Nacional por Amostra de Domicílios realizada pelo IBGE (Pnad 98), entre os cerca de 29 milhões de titulares de planos de saúde, pelo menos 75% estão diretamente vinculados aos planos privados (operadoras comerciais e empresas com plano de auto-gestão). Esta proporção é um pouco menor daquela encontrada por uma pesquisa do Ibope sobre assistência à saúde realizada em 1998 (65% de contratos empresariais). Estes últimos dados sugerem certa associação entre a forma de contratação (individual ou empresarial) e a renda. Segundo BAHIA (2008), de fato, os planos empresariais representam o principal eixo sobre o qual se movem a grande maioria das empresas médicas, as seguradoras, inúmeros estabelecimentos de saúde e profissionais médicos. A lógica que preside os contratos formais ou não dos planos empresariais no Brasil é bastante distinta da que orienta os planos individuais. Os planos empresa são mediados por contratos que pressupõem um risco homogêneo para os integrantes de apólices coletivas. Os contratos entre empresas empregadoras e operadoras de planos são formais, mas o mesmo não ocorre na relação entre o estipulante (empregador) e empregado. Os clientes de planos coletivos de saúde costumam dispor de garantia efetiva, mas não contratual de coberturas e preços, mas isso não impede que as coberturas dos planos empresariais sejam mais amplas do que as dos individuais.

Ainda BAHIA (1999), em sua tese descreve o poder de barganha das empresas empregadoras e o de vocalização de demandas e pressão conjunta dos empregados aliado a lógica da homogeneização e diluição dos riscos de grandes grupos resulta em padrões assistenciais mais pródigos e de menor custo (pelo menos 60% mais baixos) do que os contratos individuais. Daí a ênfase dos instrumentos legais nos planos individuais. Contudo a restrita formalização dos

contratos no âmbito das empresas empregadoras e, sobretudo a influência destes grandes compradores de planos e serviços privados de saúde e as possibilidades de relação entre a ANS e os estabelecimentos empresariais deve ser examinada, considerando-os como um dos elementos centrais deste mercado de assistência médica suplementar. Por outro lado as transformações no mercado de trabalho - redução das grandes plantas industriais, precarização e informalização do trabalho - afetam o mercado de assistência médica suplementar remetendo à ANS mais um desafio: o de regulação, monitoramento da viabilidade assistencial e econômico-financeira e estímulo aos contratos para trabalhadores de empresas de menor porte e autônomos.

Em artigo onde SILVA (2003), já escreve que os empresários possuem poder sobre as Operadoras de Planos de Saúde, uma vez que decidem, agregam grupos de usuários e, em definitivo, pagam ou viabilizam a maior parcela dos recursos. Entretanto, com exceções, boa parte dos empresários promove e patrocina planos de saúde para seus funcionários como mais um benefício, desconectado de um programa estruturado de qualidade de vida de seus colaboradores e de suas famílias. Muitas vezes, o plano de saúde significa ficar livre de um problema, que traz incomodações e tira o foco principal dos empresários.

Nem os empresários, nem as Operadoras e tampouco os Prestadores de Serviços tem investido na avaliação sistemática dos resultados obtidos com a decisão de contratar um plano de saúde. Os atrativos continuam sendo a rede credenciada e as tecnologias colocadas a disposição que, na hora da venda são exaltadas e na utilização consideradas como vilões do sistema. Seria importante mudanças que progressivamente e mudassem completamente essas relações. Os empresários passarão a ser os principais motivadores de um cenário mais coerente para a área da saúde. O primeiro será o reposicionamento dos empresários que passarão a comprar a saúde de seus funcionários e dependentes e não mais um plano de saúde que só pode ser utilizado nas doenças. Os empresários estarão monitorando e capitalizando os ganhos da decisão para o seu negócio. O segundo será a necessidade de reduzir custos e, portanto, estarão abertos para mudanças do modelo vigente. Forma-se aqui um cenário propício para uma remodelagem do modelo atual, onde Operadoras e Prestadores, juntos, estarão compartilhando riscos. De um lado, significa que os custos têm aumentado pela utilização dos serviços, de outro, cria a perspectiva de que, operadoras, prestadores e

empresários, através de suas políticas de recursos humanos e qualidade de vida, juntos, possam obter grandes resultados, tanto econômicos quanto clínicos.

2.1.2 Tipos de Planos de Saúde Empresarial

Os planos empresa contratados das operadoras de planos de saúde são em via de regra integralmente financiados pelas empresas empregadoras. Uma parcela das empresas com planos próprios, especialmente as do setor privado também custeiam totalmente os planos de saúde dos seus trabalhadores. No Brasil as empresas realizam, quase sempre, contratos com única operadora. Exceções são encontradas para empresas que possuem atividades em várias localidades, nas quais a operadora selecionada preferencialmente não dispõe rede assistencial.

Segundo Pesquisa de Benefícios realizada pela TOWERS PERRIN (2004), o plano de saúde é o benefício mais comumente oferecido no Brasil. Pelo sexto ano consecutivo, praticamente 100% das empresas informaram que oferecem o benefício. Os planos segurados com pré-pagamento continuam como os mais freqüentemente adotados.

TABELA 1 - TIPOS DE SISTEMAS PLANOS DE SAUDE

TIPOS DE SISTEMA	% TOTAL
Segurado (pré-pagamento)	57%
Auto-segurado	29%
Misto	14%

FONTE: PESQUISA TOWERS PERRIN (2004)

Porém, a variação entre os padrões de planos oferecidos por uma mesma operadora ou ainda organizados pelas próprias empresas empregadoras vem se

ampliando. A padronização dos planos está associada com uma hierarquização das demandas segundo status sócio-econômico. A Towers Perrin identifica 4 padrões de planos segundo os níveis hierárquicos das empresas e registra que em 83% das empresas pesquisas os quadros de diretoria estavam cobertos por planos executivos e 73% das empresas oferecem o plano básico para o pessoal operacional tabela (02). Mas também existe uma quantidade de grandes empresas estatais e multinacionais que mantém um único padrão assistencial para todos os empregados.

TABELA 2 - TIPOS DE PLANOS DE SAÚDE OFERECIDOS PELAS EMPRESAS

Tipos de Plano	Total de Usuários	% Total
Básico	1.131.215	56%
Intermediário	559.943	27%
Superior	247.523	12%
Executivo	101.683	5%
Total	2.040.364	100%

FONTE:PESQUISA TOWERS PERRIN (2004)

Os planos empresariais podem ser operados por empresas comerciais (medicinas de grupo, cooperativas médicas ou seguradoras) ou ainda organizados pelas próprias empresas empregadoras (autogestões). Ambos sub-segmentos o comercial e o não lucrativo ocupam importantes posições no mercado de planos empresariais. Estes segmentos tipicamente se diferenciam pela forma de gestão do risco. Segundo ANS (2006), as autogestões são formas de retenção do risco pelas empresas empregadoras, enquanto que as empresas comerciais configuram alternativas de transferência dos riscos das despesas assistenciais para operadoras especializadas.

As empresas a contratarem uma operadora podem optar por várias formas e proposta de Plano de Saúde:

- Coletivo Empresarial: Planos de Saúde em que o contrato é assinado entre uma pessoa jurídica, tal como uma empresa, associação, fundação ou

sindicato, e uma Operadora de Planos de Saúde para a assistência a grupos determinados de pessoas, vinculados a esta pessoa jurídica, podendo prever a inclusão ou não de dependentes. Tais Planos de Saúde regem-se por regras diferentes dos contratos individuais, no que diz respeito, por exemplo, a reajustes e possibilidade de rescisão contratual. A adesão é automática para no mínimo a maioria absoluta dos funcionários ou membros da contratante.

- Coletivo por Adesão: Planos de Saúde em que o contrato é assinado entre uma pessoa jurídica, tal como uma empresa, associação, fundação ou sindicato, e uma Operada de Planos de Saúde. A adesão a este tipo de Plano de Saúde por parte dos funcionários ou membros da contratante é espontânea e opcional.

Os Planos Coletivos, empresariais ou por adesão, podem ainda ser classificados de acordo com a existência ou não de um patrocinador:

- Coletivo com patrocinador: Planos contratados com mensalidade total ou parcialmente paga à Operadora de Planos de Saúde pela pessoa jurídica contratante.
- Coletivo sem patrocinador: Planos contratados por pessoa jurídica com mensalidade integralmente paga pelo Beneficiário diretamente à Operadora de Planos de Saúde.

Dentro de Planos Coletivos há ainda uma segmentação para distinguir os tipos de contratos de acordo com o número de beneficiários. Essa segmentação define critérios para a fixação de carências para atendimento a determinadas patologias e procedimentos: até 50 Beneficiários está prevista a possibilidade de carências contratuais; acima de 50 Beneficiários não é possível a negociação de carências.

Salvo exceções legais, segundo MIRANDA (2003), os Planos Individuais não podem ser rescindidos ou suspensos de forma unilateral pela Operadora de Planos de Saúde, ao passo que os benefícios concedidos pelos Planos Coletivos podem ser rescindidos unilateralmente por motivos de inelegibilidade, perda de direitos de titularidade ou dependência, desde que previstos por regulamento e contrato.

A maioria dos Planos de Saúde brasileiros são Planos Coletivos. Estes Planos de Saúde representam 76% dos Beneficiários dos Planos de Saúde comercializados após a regulamentação do setor em 1998, conforme Caderno de Informações de Saúde Suplementar ANS – Março 2006.

2.2 ANÁLISE MULTIVARIADA

2.2.1 Introdução

Com os avanços tecnológicos ocorridos nos últimos tempos em termos computacionais, quase que imaginável a cinco décadas atrás, tem-se tornado possível os avanços extraordinários na análise de dados, onde os computadores analisam uma grande quantidade de dados complexos. Para essa análise dos dados o emprego de técnicas estatísticas multidimensionais torna-se, então, uma ferramenta fundamental. Com a possibilidade dos programas estarem hoje diretamente associados à rede mundial de computadores tem-se uma grande facilidade no acesso à grandes base de dados. Basta ver os relatórios de pesquisa e mesmo os bancos de dados, com um grande número de matrizes de informações não trabalhadas.

A pura utilização de técnicas estatísticas hoje em dia é bastante facilitada graças à vasta disposição de programas computacionais, mas não é condição suficiente se o estudo não for embasado num sólido conhecimento de uma estatística mais refinada podemos assim dizer. Pois se até então os dados eram trabalhados de uma forma mais simples como a análise unívariada, ou seja, com uma estatística de cálculo mais fácil, mas devido a grande revolução computacional começou uma grande expansão no conhecimento de técnicas estatística conhecidas como análise multivariada.

Essas técnicas da análise multivariada estão sendo amplamente utilizadas em indústrias, em centros de pesquisas e universidades.

2.2.2 Característica Da Análise Multivariada

Análise Multivariada é a parte da estatística e da análise de dados que estuda, interpreta e elabora o material estatístico sobre a base de um conjunto de

n>1 variáveis, geralmente do tipo quantitativo, qualitativo ou uma mescla de ambos, a informação em Análise Multivariada é, portanto, de caráter multidimensional.

Para FERREIRA (1996), a dificuldade e a complexidade em se obter as informações sobre as várias variáveis torna necessário o uso da multivariada tornando as informações mais claras e mais simples, sem sacrificar informações valiosas. Para isso desenvolve-se noções de estimativas e de testes fundados em hipóteses muito restritivas. Entretanto, na prática, os indivíduos observados são freqüentemente caracterizados por um grande número de caracteres (ou variáveis). Os métodos de análise de dados permitem um estudo global dessas variáveis, pondo em evidência ligações, semelhanças ou diferenças entre si.

A análise multivariada é a rigor qualquer abordagem analítica que considere o comportamento de duas ou mais variáveis simultaneamente, num vasto campo do conhecimento que envolve uma grande multiplicidade de conceitos estatísticos e matemáticos PEREIRA (1999).

A proposta multivariada é agrupar dados em conjuntos que têm atributos similares, apresentando os de uma maneira que enfatize os agrupamentos naturais, analogamente à classificação para tanto, as distâncias entre pares de amostras são calculadas e comparadas. Quando as distâncias entre as amostras são relativamente pequenas, isto implica que as mesmas são similares; já amostras diferente serão separadas por distâncias relativamente grandes, de uma forma geral PEREIRA (1999), os objetivos gerais para quais a análise multivariada é conduzida são:

a) Redução de dados ou simplificação estrutural, entretanto o fenômeno estudado deve ser representado da maneira mais simples possível, sem sacrificar valiosas informações

b) Ordenação e agrupamento

Agrupamento de objetos, tratamentos, ou variáveis similares baseados em dados amostrais ou experimentais.

c) Investigação da dependência entre variáveis

O estudo das relações estruturais entre variáveis muitas vezes é de interesse do pesquisador.

d) Predição

Relação entre variáveis devem ser determinadas para o propósito de predição de uma ou mais variáveis com base na observação de outras variáveis.

e) Teste de hipóteses.

Os modelos multivariados possuem em geral, um propósito através do qual o pesquisador pode testar ou inferir a respeito de uma hipótese sobre um determinado fenômeno. No entanto a sua utilização adequada depende do conhecimento das técnicas e das suas limitações.

2.3 ESTATÍSTICAS DESCRITIVAS

Quando se trabalha com um número muito grande de variáveis é matematicamente impossível visualizar ou observar qualquer relação entre as mesmas, nesse aspecto é importantíssimo o uso de medidas que forneçam certas informações como por exemplo média aritmética ou média amostral, é uma estatística descritiva que fornece informação de posição,ou seja representa o centro do conjunto de dados ou sua dispersão. Outro exemplo a ser citado é uma medida que mede a dispersão e a variabilidade dos dados. Procura-se descrever a amostra, pondo em evidência as características principais e as suas propriedades. As descrições formais destas medidas estão apresentadas a seguir.

Seguindo material apresentado em aula por MARQUES (2006), dada uma amostra de tamanho n com p variáveis estes valores podem ser representados em um arranjo retangular, denominado de X**,** com n linhas e p colunas, da seguinte forma :

$$X_{np} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix} \quad (1)$$

Da matriz de dados (amostra) X, a média amostral da variável X_k, simbolizada por $\overline{X}$, é dada por:

$$\overline{X}_k = \frac{1}{n}\sum_{j=1}^{n} X_{jk} \qquad k = 1,2,\dots p \tag{2}$$

E o vetor médio amostral é $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ que estima o parâmetro $\underline{\mu}$, $E(X)$.

Uma medida de variação é fornecida pela variância amostral, definida para as n observações de k-ésima variável por:

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_{jk} - \overline{X}_k)^2 \quad k = 1,2,3,\dots,p \tag{3}$$

e a matriz de covariância amostral do vetor $\overline{X}$ é dado por $S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})$ $(X_i - \overline{X})'$

Então a matriz de covariância amostral que estima a verdadeira matriz de covariância populacional $\sum = E\left[\left(\underline{\mathbf{X}} - \underline{\boldsymbol{\mu}}\right)\left(\underline{\mathbf{X}} - \underline{\boldsymbol{\mu}}\right)'\right]$ tem por expressão.

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})\,(X_i - \overline{X})' \Rightarrow \text{cov} = \begin{bmatrix} s_1^2 & \cdots & s_{1k} & \cdots & s_{1p} \\ s_{21} & s_2^2 & s_{2k} & \ddots & s_{2p} \\ s_{31} & \cdots & s_k^2 & \ddots & s_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & \cdots & s_p^2 \end{bmatrix} \tag{4}$$

onde, S_i^2 representa a variância amostral de X_i.

Apesar da covariância ser uma estatística adequada para medir a relação entre duas variáveis ela é complicada para comparar graus de relação entre variáveis, e isto devido à que está influenciada pelas unidades de medida de cada variável. Para evitar a influência da ordem de grandeza e unidades de cada variável, dividi-se a covariância pelo desvio padrão de X e de Y, dando origem ao coeficiente

de correlação, pois esta mede o grau e o tipo do relacionamento entre as variáveis estudadas, esta medida e também chamada de coeficiente de correlação de Pearson, em homenagem ao seu criador.

Simbologia: r (coeficiente amostral) ou ρ (coeficiente populacional).

O valor de r, pode ser calculado por:

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n(\sum X^2) - (\sum X)^2}\sqrt{n(\sum Y^2) - (\sum Y)^2}} \quad (5)$$

ou

$$\hat{\rho} = r = \frac{\sigma_{xk}}{\sigma_x \sigma_k} \quad (6)$$

Onde n é o número de pares de valores (X,Y) observados.Pode-se maior que -1,0 $\leq \rho \leq$ 1,0. O mesmo ocorre com o valor de r. A partir dos valores de r ou ρ, pode-se verificar o tipo da correlação existente entre as variáveis estudadas, segundo MARQUES (2005,p.207), o coeficiente de correlação linear pode ser avaliado qualitativamente da seguinte forma:

Valor de **r** ou correlação	
0,0 ------| 0,30	fraca correlação linear
0,3 ------| 0,60	moderada correlação linear
0,60-----| 0,90	forte correlação linear
0,90 ----| 1,00	correlação linear muito forte

A matriz de ρ é dada por:

$$r = \hat{\rho} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \quad (7)$$

A matriz de correlação mostra a correlação entre todas as variáveis, logo é uma matriz simétrica e na diagonal sempre terá o valor 1, uma vez que se trata da correlação da variável com ela mesma.

2.4 ANÁLISE DAS COMPONENTES PRINCIPAIS

2.4.1 Introdução

A análise de componentes principais (principal component analysis –PCA) é uma das mais antigas técnicas multivariadas e seu tratamento matemático já é bem difundido, não sendo necessária nenhuma nova discussão a respeito desse tema específico. A análise de componentes principais é uma técnica que tem o propósito de analisar estruturas de covariâncias e correlações, baseada nas raízes (ou valores) características e nos vetores gerados a partir delas, em matrizes simétricas positivas definidas.

No sentido mais geral, MOITA (2004), a técnica de componentes principais é um método para transformar variáveis correlacionadas em outro grupo de variáveis não correlacionadas, servindo ainda para a obtenção de combinações lineares das variáveis originais com variabilidade relativamente grande (ou pequena, dependendo do propósito). Além de ser uma ferramenta para a redução da dimensionalidade dos dados, pode-se ver que a análise de componentes principais tem um fim por si só ou como um passo intermediário para análises subseqüentes dos dados. Segundo FERREIRA (2006), apesar das técnicas de análise multivariada terem sido desenvolvidas para resolver problemas específicos, principalmente de Medicina, Sociologia e Biologia, podem ser também utilizadas para resolver outros tipos de problemas em diversas áreas do conhecimento.

A análise de componentes principais é uma das técnicas mais conhecidas, contudo é importante ter uma visão conjunta de todas ou quase todas as técnicas para resolver a maioria dos problema práticos.

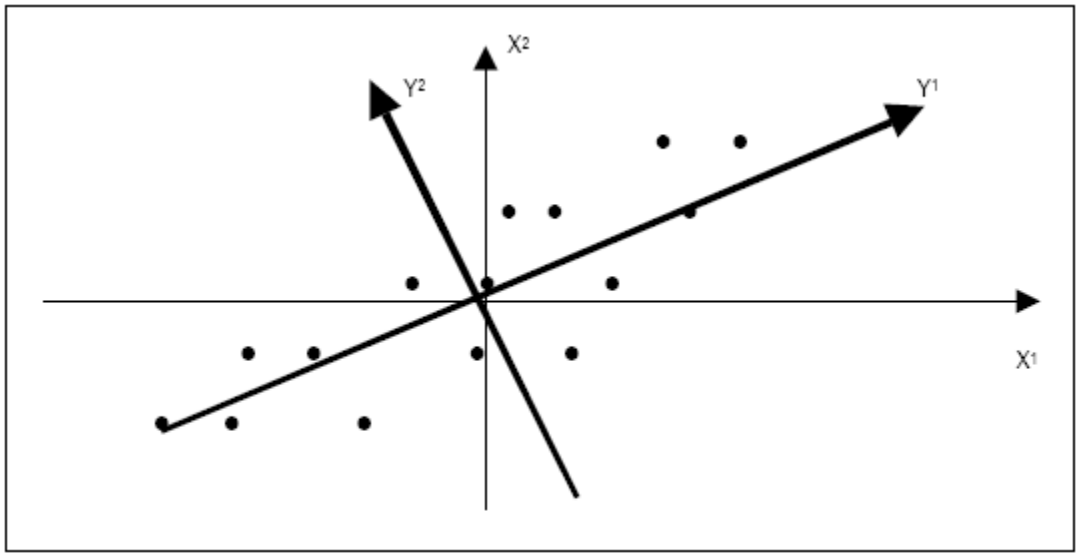
2.4.2 Componentes principais populacionais

Em breve resumo introdutório, seguindo o material de MARQUES (2006), algebricamente, as componentes principais são combinações lineares de p variáveis originais: $X_1, X_2, \ldots\ldots\ldots\ldots, X_p$

Geometricamente, as combinações lineares representam a seleção de um novo sistema de coordenadas, obtido por rotação do sistema original com $X_1, X_2, \ldots\ldots\ldots\ldots, X_p$ como eixos. Os novos eixos, $(Y_1, Y_2, \ldots\ldots, Y_p)$, representam as direções com variabilidade máxima, permitindo uma interpretação mais simples da estrutura da matriz de covariância.

Por exemplo, para p = 2.

FIGURA 1 - SIGNIFICADO GEOMÉTRICO DAS COMPONENTES PRINCIPAIS PARA P=2



FONTE: JOHNSON e WICHERN (2002)

De acordo com FONSECA (1999), algebricamente a ACP é uma combinação linear particular de p variáveis originais $X_1, X_2, \ldots\ldots\ldots\ldots, X_p$ considere $[X_1, X_2, \ldots\ldots\ldots\ldots, X_p]$ um vetor aleatório p dimensional com vetor de médias $\boldsymbol{\mu}$, matriz de covariância $\sum$ e autovalores $\lambda_1 \geq \lambda_2 \geq, \ldots\ldots\ldots, \geq \lambda_p$. Então, seja $Y_1, Y_2, \ldots\ldots, Y_p$ as combinações lineares abaixo:

$$\begin{aligned} Y_1 &= \underline{\mathbf{e}}'_1\underline{\mathbf{X}} = e_{11}X_1 + e_{21}X_2 + .. e_{p1}X_p \\ Y_1 &= \underline{\mathbf{e}}'_2\underline{\mathbf{X}} = e_{12}X_1 + e_{22}X_2 + .. e_{p2}X_p \\ &\vdots \\ Y_p &= \underline{\mathbf{e}}'_p\underline{\mathbf{X}} = e_{1p}X_1 + e_{2p}X_2 + .. e_{pp}X_p \end{aligned} \tag{8}$$

Considere-se ainda que uma variável aleatória simples, X_i, seja multiplicada por uma constante c_i. Então, o valor esperado e a variância de X_i, são dados, respectivamente por :

$$E(c_iX_i) = c_i.E(X_i) = c_i\mu_i \tag{9}$$

$$Var(c_iX_i) = E(c_iX_i - c_i\mu_i)^2 = {c_i}^2 Var(X_i) = {c_i}^2\sigma_{i1} \tag{10}$$

Se $\mathbf{X_2}$ é uma segunda variável aleatória e se a e b são constantes então, usando a propriedade da adição na expectância, vem que :

$$\begin{aligned} cov(aX_1, bX_2) &= E\big(aX_1 - a\mu_1\big)\big(bX_2 - a\mu_2\big) \\ &= abE\big(X_1 - \mu_1\big)\big(X_2 - \mu_2\big) \\ &= abCov(X_1, X_2) \\ &= ab\sigma_{12} \end{aligned} \tag{11}$$

Como:

$$E(aX_1) = aE(X_1) = a\mu_1 \tag{12}$$

$$E(bX_2) = bE(X_2) = b\mu_2 \tag{13}$$

Então, pode se escrever para a combinação linear $aX_1 + bX_2$, que:

$$E(aX_1 + bX_2) = aE(X_1) + b(X_2) = a\mu_1 + b\mu_2 \tag{14}$$

$$\begin{aligned} Var\,(aX_1 + bX_2) &= E[(aX_1 + bX_2) - (a\mu_1 + b\mu_2)]^2 \\ &= E[(aX_1 - a\mu_1) + (bX_2 - b\mu_2)]^2 \\ &= E[a^2(X_1 - \mu_1)^2 + b^2(X_2 - \mu_2)^2 + 2ab(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= [a^2 Var(X_1) + b^2 Var(X_2) + 2abCov(X_1, X_2)] \\ Var\,(aX_1 + bX_2) &= a^2\sigma_{11} + b^2\sigma_{22} + 2ab\sigma_{12} \end{aligned} \tag{15}$$

Agora, consideramos o vetor $[X_1, X_2]$ e $\underline{\mathbf{c}}' = [a, b]$, $(aX_1 + bX_2)$ pode ser escrito como:

$$[a \quad b]\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \underline{\mathbf{c}}'\underline{\mathbf{X}} \tag{16}$$

Analogamente, $E(aX_1 + bX_2) = a\mu_1 + b\mu_2$, se torna :

$$[a \quad b]\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \underline{\mathbf{c}}'\underline{\boldsymbol{\mu}} \tag{17}$$

E considerando-se a matriz de variância-covariância de X igual a :

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \tag{18}$$

Então, a variância da combinação linear poderá ser escrita como:

$$Var(aX_1 + bX_2) = Var\left(\underline{\mathbf{c}}'\underline{\mathbf{X}}\right) = \underline{\mathbf{c}}'\Sigma\underline{\mathbf{c}} \tag{19}$$

Desde (16) vem:

$$[a \quad b]\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}\begin{bmatrix}a\\ b\end{bmatrix} = a^2\sigma_{11} + 2ab\sigma_{12} + b^2 \tag{20}$$

Os resultados anteriores podem ser estendidos para uma combinação linear de p variáveis aleatórias. Assim, para uma dada combinação linear, pode se escrever:

$$\begin{gathered}\underline{\mathbf{c}}'\underline{\mathbf{X}} = c_1X_1 + c_2X_2 + \ldots c_pX_p\\ \text{média} = E\left(\underline{\mathbf{c}}'\underline{\mathbf{X}}\right) = \underline{\mathbf{c}}'\boldsymbol{\mu}\\ \text{Variância} = \text{Var}\left(\underline{\mathbf{c}}'\underline{\mathbf{X}}\right) = \underline{\mathbf{c}}'\Sigma\underline{\mathbf{c}}\end{gathered} \tag{21}$$

Usando os resultados de (21) em (11), vem que :

$$\begin{aligned}&\text{Var}(Y_i) = \underline{\mathbf{e}}'_i\Sigma\underline{\mathbf{e}}_i && i = 1,2,\ldots,p\\ &\text{Cov}\left(Y_i, Y_j\right) = \underline{\mathbf{e}}'_i\,\Sigma\underline{\mathbf{e}}_j && i,j = 1,2,\ldots p\end{aligned} \tag{22}$$

Isto implica que as componentes principais são, portanto, todas as combinações lineares não correlacionadas $Y_1, Y_2, \ldots\ldots, Y_p$ cujas as variâncias em (22) sejam tão grande quanto possível.

A primeira componente principal segundo a definição de JOHNSON e WICHERN (2002), é a combinação linear que possui a máxima variância, isto é, aquela combinação que maximizar a variância, de acordo com a equação (21). Parece evidente que a expressão $V(Y_i) = \underline{\mathbf{e}}'_i\Sigma\underline{\mathbf{e}}_i$ pode ser aumentada pela multiplicação de qualquer $\underline{\mathbf{e}}_i$ por uma dada constante. Para eliminar esta indeterminação, é conveniente restringir os vetores coeficientes ao comprimento unitário. Deste modo, pode se escrever as definições do primeiro e do segundo

componente principais, respectivamente Pc_1 e Pc_2 na forma de função objetivo com restrições.

$$\begin{gathered}\text{Maximizar Var}(\underline{\mathbf{e}}'_{\mathbf{1}}\underline{\mathbf{X}}) \\ \text{Sujeito a: } \underline{\mathbf{e}}'_{\mathbf{1}}\underline{\mathbf{e}}_{\mathbf{1}} = 1\end{gathered} \tag{23}$$

$$\begin{gathered}\text{Maximizar Var}(\underline{\mathbf{e}}'_{\mathbf{2}}\underline{\mathbf{X}}) \\ \text{Sujeito a: } \underline{\mathbf{e}}'_{\mathbf{2}}\underline{\mathbf{e}}_{\mathbf{2}} = 1 \\ \text{Cov}\left(\underline{\mathbf{e}}'_{\mathbf{1}}\underline{\mathbf{X}}, \underline{\mathbf{e}}'_{\mathbf{2}}\underline{\mathbf{X}}\right) = 0\end{gathered} \tag{24}$$

As soluções dos sistemas (23) e (24) conduzem, respectivamente, ao valor do primeiro e do segundo componentes principais.

Desta forma, o i-ésimo componente principal será a combinação linear $\mathbf{e}'\mathrm{X}$ que for solução da expressão (25) a seguir.

$$\begin{gathered}\text{Maximizar Var}(\underline{\mathbf{e}}'_{\mathbf{i}}\underline{\mathbf{X}}) \\ \text{Sujeito a: } \underline{\mathbf{e}}'_{\mathbf{i}}\underline{\mathbf{e}}_{\mathbf{i}} = 1 \\ \text{Cov}\left(\underline{\mathbf{e}}'_{\mathbf{i}}\underline{\mathbf{X}}, \underline{\mathbf{e}}'_{\mathbf{j}}\underline{\mathbf{X}}\right) = 0 \text{ para j} < i\end{gathered} \tag{25}$$

2.4.3 Propriedades das componentes principais

(1) Seja o vetor aleatório $\underline{\mathbf{X}}' = [\mathrm{X}_1, \mathrm{X}_2, \ldots\ldots, \mathrm{X}_\mathrm{p}]$ com matriz de covariância Σ e pares de autovalores e autvetores$(\lambda_1, \underline{\mathbf{e}}_1), (\lambda_2, \underline{\mathbf{e}}_2), \ldots, (\lambda_\mathrm{p}, \underline{\mathbf{e}}_\mathrm{p})$, onde $\lambda_1 \geq \lambda_2 \geq \cdots, \geq \lambda_\mathrm{p} \geq 0$. A i-ésima componente principal é dada por:

$$\mathrm{Y_j} = \underline{\mathbf{e}}'_{\mathrm{j}}\underline{\mathbf{X}} = \mathrm{e_1X_1} + \mathrm{e_2X_2} + .. \mathrm{e_pX_p}, \qquad \mathrm{j} = 1,2,\ldots,\mathrm{p} \tag{26}$$

Onde:

$$\mathrm{V(Y_j)} = \underline{\mathbf{e}}'_{\mathrm{j}}\Sigma\underline{\mathbf{e}}_{\mathrm{j}} = \lambda_\mathrm{j} \text{ e } \mathrm{Cov(Y_i, Y_j)} = \underline{\mathbf{e}}'_{\mathrm{j}}\Sigma\underline{\mathbf{e}}_{\mathrm{j}} = 0 \text{ , i} \neq \mathrm{j} \tag{27}$$

(2) Variância total.

$$V(X_1) + V(X_2) + \ \ldots + V(X_p) = \sum_{i=1}^{p} V(X_i) = \sigma_1^2 + \sigma_2^2 + \ \ldots + \sigma_p^2 = \lambda_1 + \lambda_2 + \ \ldots + \lambda_p = \sum_{j=1}^{p} V(Y_j). \quad (28)$$

3) Se $Y_1 = \underline{\mathbf{e}}'_1\underline{\mathbf{X}}$, $Y_2 = \underline{\mathbf{e}}'_2\underline{\mathbf{X}}, \ldots, Y_p = \underline{\mathbf{e}}'_p\underline{\mathbf{X}}$ são as componentes principais de Σ então:

$$\rho_{Y_i X_j} = \frac{e_{ij}\sqrt{\lambda_j}}{\sigma_i} \qquad i, j = 1,2, \ldots, p \quad (29)$$

São os coeficientes de correlação entre as componentes principais Y_j e as variáveis X_i, onde $(\lambda_1, \underline{\mathbf{e}}_1), (\lambda_2, \underline{\mathbf{e}}_2), \ldots, (\lambda_p, \underline{\mathbf{e}}_p)$, são os pares autovalores- autovetores de Σ.

(4) A proporção da variância total devida à j-ésima componente principal é.

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}, \qquad j = 1,2, \ldots, p \quad (30)$$

Vale destacar que a ACP deve ser realizada apena nos casos em que haja correlação suficientes entre as diferentes variáveis da matriz original de dados e, quanto aos resultados da metodologia, a matriz fatorial de carga dos fatores obtidos segundo JOHNSON e WICHERN (2002), em certos problemas onde se aplicam as componentes principais, se uma porcentagem de 70% ou mais for atribuída às primeiras r componentes principais, então, esses podem substituir as p variáveis originais sem significativa perda de informações.

Para que o pesquisador possa compreender a definição conceitual de cada fator.

2.4.4 Componentes principais obtidas pela padronização das variáveis

Normalmente as características são observadas em unidades de medidas diferentes entre si, e neste caso, segundo JOHNSON E WICHERN (2002), é conveniente padronizar as variáveis $X_i (i = 1, 2, 3, \dots\dots, p)$. A padronização pode ser feita pela padronização das variáveis originais por.

$$Z_i = \frac{X_i - \mu_i}{\sqrt{\sigma_i}} \tag{31}$$

Que em notação matricial é:

$$Z = V^{-\frac{1}{2}} \left(\underline{\mathbf{X}} - \underline{\boldsymbol{\mu}} \right) \tag{32}$$

Em que:

$$V^{-\frac{1}{2}} = \begin{bmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 & \cdots & 0 \\ & \frac{1}{\sqrt{\sigma_{22}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & \frac{1}{\sqrt{\sigma_{pp}}} \end{bmatrix} \tag{33}$$

Pode-se então, verificar que:

$$E(Z) = 0 \text{ e } \mathrm{Cov}\,(Z) = V^{-\frac{1}{2}} \Sigma V^{-\frac{1}{2}} = \rho \tag{34}$$

Então, os componentes principais de Z são dados pelos autovalores e autovetores de ρ, matriz de correlação de X. Os autovalores e autovetores de Σ são em geral, diferentes daqueles derivados de ρ.

O i-ésimo componente principal das variáveis padronizadas $\underline{\mathbf{Z}}' = [Z_1, Z_2, \dots\dots\dots Z_p]$ com $\mathrm{Cov}(Z) = \rho$, é dado por :

$$Y_i = \underline{\mathbf{e}}' V^{-\frac{1}{2}} \left(\underline{\mathbf{X}} - \underline{\boldsymbol{\mu}} \right), \quad i = 1,2, \dots, p \tag{35}$$

Verifica – se ainda, que:

$$\sum_{i=1}^{p} \mathrm{Var}(Y_i) = \sum_{i=1}^{p} \mathrm{Var}(Z_i) = \rho$$
$$\sum_{i=1}^{p} \mathrm{Var}(\lambda_i) = \rho \tag{36}$$

e verifica-se também que :

$$P_{yi} Z_k = \underline{\mathbf{e}}_{ik} \sqrt{\lambda_i}, i, k = 1,2, \dots, p \tag{37}$$

Onde a proporção da variância total explicada pela i-ésima componentes principal de Z é dada por $\frac{\lambda_i}{p}$ que é a porcentagem da variação total explicada por ele.

2.4.5 Componentes principais amostrais

Seja uma amostra aleatória $X_1, X_2, \dots, X_p$ retirada de uma população ρ variada com, com média $\underline{\boldsymbol{\mu}}$ e covariância Σ. Com o vetor de médias amostrais $\underline{\overline{\mathbf{X}}}$, matriz de covariância S e a matriz de correlação r. Componentes principais amostrais são as combinações lineares das variáveis mensuradas que maximizam a variação total da amostra e que são mutuamente ortogonais.

Na aplicação pratica das componentes principais os parâmetros ρ e Σ são desconhecidos, obtêm-se as componentes principais através de seus estimadores, que são a matriz de covariância amostral S ou a matriz de correlação amostral R, que são definidas segundo FERREIRA (2006), através de

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{\mathbf{x}}_i - \underline{\bar{\mathbf{x}}})(\underline{\mathbf{x}}_i - \underline{\bar{\mathbf{x}}})' \tag{38}$$

$$r = V^{-\frac{1}{2}}\ S\ V^{-\frac{1}{2}} \tag{39}$$

Onde V é a matriz desvio padrão amostral, e $\underline{\overline{\mathbf{X}}}$ é o vetor médio mostrados em (40) e (41).

$$V = \begin{bmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_p \end{bmatrix} \tag{40}$$

$$\underline{\overline{\mathbf{X}}} = \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \vdots \\ \overline{X}_p \end{bmatrix} \tag{41}$$

Calculam-se os autovalores $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p$ e os respectivos autovetores $\underline{\hat{\mathbf{e}}}_1, \underline{\hat{\mathbf{e}}}_2, \ldots, \underline{\hat{\mathbf{e}}}_p$ para depois construir as componentes principais amostrais.

$$\begin{aligned} \widehat{\mathbf{Y}}_\mathbf{1} &= \underline{\hat{\mathbf{e}}}'_\mathbf{1}\underline{\mathbf{X}} = \hat{e}_{11}X_1 + \hat{e}_{21}X_2 + \ldots \hat{e}_{p1}X_p \\ \widehat{\mathbf{Y}}_\mathbf{1} &= \underline{\hat{\mathbf{e}}}'_\mathbf{2}\underline{\mathbf{X}} = \hat{e}_{12}X_1 + \hat{e}_{22}X_2 + \ldots \hat{e}_{p2}X_p \\ &\vdots \\ \widehat{\mathbf{Y}}_\mathbf{p} &= \underline{\hat{\mathbf{e}}}'_\mathbf{p}\underline{\mathbf{X}} = \hat{e}_{1p}X_1 + \hat{e}_{2p}X_2 + \ldots \hat{e}_{pp}X_p \end{aligned} \tag{42}$$

Todas as propriedades das componentes principais se mantêm e são obtidas com base em estimadores. MARQUES (2006).

2.4.6 Propriedades das componentes principais amostrais.

$$1) V(\widehat{Y}_j) = \hat{\lambda}_j \quad j = 2, \dots, p \tag{43}$$

$$2)\ Cov(\widehat{Y}_i, \widehat{Y}_j) = 0, \qquad i \neq j. \tag{44}$$

$$3) \sum_{i=1}^{p} S_i^2 = s_1^2 + s_2^2 + \ \dots + s_p^2 = \sum_{j=1}^{p} \hat{\lambda}_1 + \hat{\lambda}_2 + \ \dots + \hat{\lambda}_p \tag{45}$$

4) A proporção da variância total explicada pela j-ésima componente principal estimada é :

$$\frac{\hat{\lambda}_j}{\hat{\lambda}_1 + \hat{\lambda}_2 + \dots + \hat{\lambda}_p}, \ j = 1,2, \dots, p \tag{46}$$

5) Correlação amostral entre $\widehat{Y}_j$ e X_i é:

$$r_{\widehat{Y}_j X_i} = \frac{\hat{e}_{ij}\sqrt{\hat{\lambda}_j}}{s_i} \quad i, j = 1,2, \dots, p \tag{47}$$

2.4.7 Análise dos Autovalores

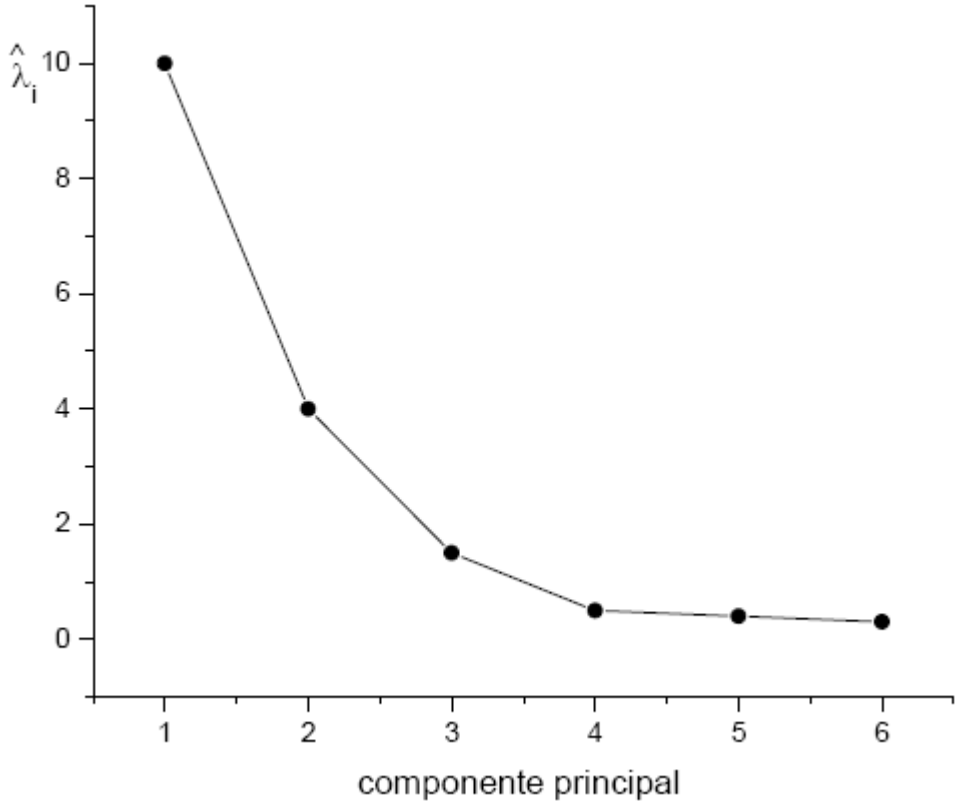
FERREIRA (2006)

> *Existe sempre a questão importante de o número de componentes a ser retido. Não existe uma resposta definitiva para essa questão. Os aspectos que devem ser considerados incluem a quantidade da variação amostral explicada, o tamanho relativo dos autovalores e a interpretação subjetiva dos componentes.*

Uma Importante questão a se considerar é o número de componentes a ser retido. Não existe uma resposta definitiva para esta questão. Os aspectos a serem considerados incluem a quantidade de variação amostral explicada, o tamanho relativo dos autovalores e a interpretação subjetiva dos componentes. Tornando muito grande a dificuldade frente a problemas práticos no cotidiano experimental. Uma das dificuldades encontrada para a utilização das componentes principais consiste na falta de critérios objetivos para determinar qual o número de componentes principais que deve ser utilizado na análise dos dados. Dentre os critérios indicados na literatura que auxiliam nessa tomada de decisão, destacam – se: *"Scree Plot"* ou gráfico de autovalores, também conhecido como "gráfico do cotovelo" é um gráfico onde são plotados $\hat{\lambda}\mathrm{X\,i}$, magnitude de um autovalor versus seu número. O teste scree é comparado a análise de separação da base de uma montanha e o acúmulo de restos de rochas caídos dela, a análise para no ponto onde começa o resto.

Para ilustrar, um exemplo retirado FERREIRA (2006), onde mostra um gráfico com 06 componentes principais e a formação de um cotovelo em i = 4 significando que todos os componentes acima de $\hat{\lambda}_3$ servem suficientemente para resumir a variação amostral total.

FIGURA 2 - EXEMPLO GRÁFICO SCREE PLOT



FONTE: FURTADO (2006) p. 256

Outro método para a escolha do número de componentes principais mais comumente usado é o Critério de Kaiser, pois trata-se de um método exploratório para indicar a redução do espaço paramétrico, o qual sugere considerar apenas os componentes com autovalor superior a 1, o que significa que o componente contabiliza mais variância do que uma variável. Critério da proporção: observa-se a proporção de variância acumulada e um nível de corte é estabelecido, representando o total da variância contabilizado pelos componentes selecionados seu uso fica limitado às condições práticas de estudo e ao bom senso do pesquisador.

Na prática a maioria dos pesquisadores raramente usa um único critério para determinar quantos fatores serão extraídos, mas HAIR (2007, p. 103), adverte das conseqüências quanto às escolhas dos números de fatores, pois se em excesso a interpretação se torna difícil e se de menos a estrutura correta não é revelada e dimensões importantes podem ser omitidas, portanto como já citado anteriormente o pesquisador deve se empenhar em ter um conjunto de fatores mais representativos e parcimonioso possível.

2.5 ANÁLISE DE REGRESSÃO

2.5.1 Introdução

A análise de regressão é uma técnica de modelagem utilizada para analisar a relação entre uma variável dependente (Y) e uma ou mais variáveis independentes X_1, X_2, X_3,..., X_p. Segundo HAIR (2005,p.136), o objetivo dessa técnica é identificar (estimar) uma função que descreva, o mais próximo possível, a relação entre essas variáveis e assim poder predizer o valor que a variável dependente (Y) irá assumir para um determinado valor da variável independente X. São técnicas utilizadas para estimar uma relação que possa existir na população pois essa correlação mede a força, ou grau, de relacionamento entre duas variáveis; a regressão dá a equação que descreve o relacionamento em termos matemáticos. Os dados para análise de regressão e correlação provêm de observações de variáveis emparelhadas. Na regressão pressupõe-se alguma relação de causa e efeito, de explanação do comportamento entre as variáveis.

Exemplos de relação entre variáveis são o consumo em relação à taxa de inflação; a produção de leite e temperatura ambiente; a resistência de um material e sua composição química; o número de peças com defeitos e a experiência; receita e gasto com publicidade e etc...

2.5.2 Modelo De Regressão

Modelos de regressão são modelos matemáticos que relacionam o comportamento de uma variável Y com outra X. Quando a função "f"' que relaciona duas variáveis é do tipo $f(x) = \beta_1 + \beta_2 x + \varepsilon$ tem-se o modelo de regressão simples. A variável X é a variável independente do modelo enquanto $Y = f(x) + \varepsilon$ é a variável dependente das variações de X. O modelo de regressão é chamado de simples quando a relação causal envolve apenas duas variáveis. O modelo de regressão é múltiplo quando o comportamento de Y é explicado por mais de uma variável independe X1, X2, ..., Xp. Os modelos citados (simples ou múltiplo) simulam

relacionamentos entre as variáveis do tipo linear (equação da reta ou do plano) ou não linear (equação exponencial, geométrica, etc.). Segundo HAIR (2005, p.136), para que serve determinar a relação entre duas variáveis? Para realizar previsões futuras sobre algum fenômeno da realidade. Neste caso extrapola-se para o futuro as relações de causa-efeito já observadas no passado entre as variáveis.

Pesquisadores interessados em "simular" os efeitos causados sobre uma variável Y em decorrência de alterações introduzidas nos valores de uma variável X também usam este modelo. Nem todas as situações são bem aproximadas por uma equação linear. Quando os dados não podem ser aproximados por um modelo linear, a alternativa é procurar um modelo não-linear conveniente, ou transformar os dados para a forma linear. Por exemplo, a conversão de uma ou de ambas escalas em logaritmos dá por vezes um modelo linear.

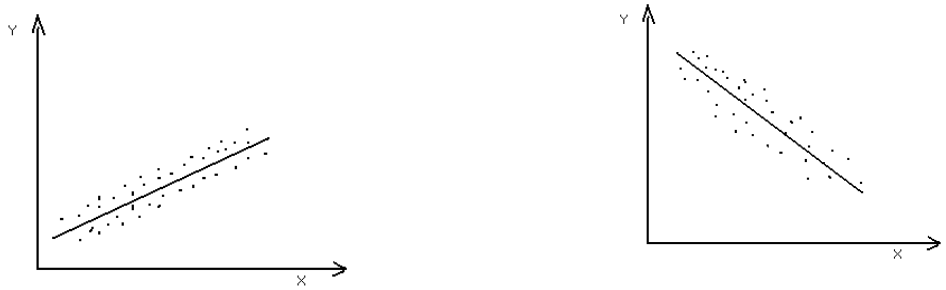
2.5.3 Diagrama De Dispersão

É um gráfico no qual cada ponto plotado representa um par observado de valores para as variáveis estudadas (X ,Y), num sistema de eixos cartesianos.

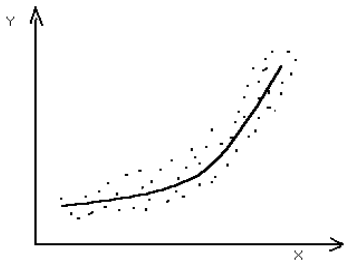
Através do diagrama de dispersão podemos ter uma idéia do tipo de relação entre as variáveis estudadas.

A seguir tem-se alguns exemplos de diagramas de dispersão.

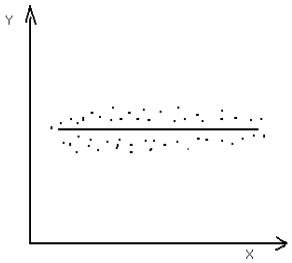
FIGURA 3 - EXEMPLOS DE DIAGRAMAS DE DISPERSÃO



(a) Relação linear direta (b) Relação linear inversa



(c) Relação curvilínea direta



(d) Não há relação

FONTE:MONTEGOMERY (2003, p. 213, 214)

2.5.4 Significância Do Coeficiente De Correlação Linear

De modo geral, muitas vezes a hipótese nula de interesse é que o coeficiente de correlação populacional seja igual a zero, pois se essa hipótese for rejeitada ao nível de significância α estipulado, pode se concluir que efetivamente existe uma relação entre as variáveis estudadas.

Na maioria das vezes utiliza-se a distribuição de “t” Student, para testar a significância do coeficiente de correlação linear populacional, sendo que a estatística de teste será calculada por:

$$t_t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \tag{48}$$

onde r é o coeficiente de correlação amostral.

O valor t (student) pode ser interpretado como o número de desvios padrões que o estimador b dista do ponto zero. Quanto maior for essa distância , menor será a chance de $\beta = 0$.

2.6 REGRESSÃO LINEAR SIMPLES

Este modelo é utilizado quando existe uma relação linear entre a variável independente e a variável dependente (neste caso apenas uma). A função que expressa esse modelo é dada pela forma seguinte:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{49}$$

Onde ε_i é a componente aleatória.

Uma vez escolhido o modelo de regressão, deve-se estimar seus parâmetros, neste caso os coeficientes da equação da reta, β_0 e β_1. Isso pode ser aplicando o Método dos Mínimos Quadrados.

Tirando a média sobre a equação acima, temos:

$$\overline{Y} = \beta_0 + \beta_1 \overline{X} \tag{50}$$

uma vez que a média dos erros é zero.
Subtraindo as duas equações temos:

$$Y_i - \overline{Y} = (\beta_0 - \beta_0) + (\beta_1)\left(X_i - \overline{X}\right) + \varepsilon_i \tag{51}$$

Chamando de Y e X as diferenças centradas nas médias, $(Y_i - \overline{Y})$ e $\left(X_i - \overline{X}\right)$ respectivamente, temos que:

$$Y_i = \beta_1 X_i + \varepsilon_i \tag{52}$$

ou ainda,

$$\varepsilon_i = Y_i - \beta_i X_i \tag{53}$$

Fazendo a soma dos quadrados dos erros,

$$\sum(\varepsilon_i)^2 = \sum(Y_i - \beta_1 X_i)^2 \tag{54}$$

$$\sum(\varepsilon_i)^2 = \sum Y_i^2 - \sum 2\beta_1 X_i Y_i + \sum \beta_i^2 X_i^2 \tag{55}$$

como β_1 é uma constante,temos:

$$\sum(\varepsilon_i)^2 = \sum Y_i^2 - 2\beta_1 \sum X_i Y_i + \beta_i^2 \sum X_i^2 \tag{56}$$

Como o objetivo é estimar uma equação que minimize os erros, devemos então derivar a equação acima em relação a β_1 e igualar a zero. E como não se tem os verdadeiros valores e sim uma amostra , ou seja o valor a ser determinado é um estimador do verdadeiro valor populacional, a nova nomenclatura para β_1 será $\widehat{\beta}_1$. Com isso temos:

$$0 = -2\sum X_i Y_i + 2\hat{\beta}_1 \sum X_i^2 \tag{57}$$

Que pode ser reescrita como:

$$\hat{\beta}_1 = \frac{\sum X_i Y_i}{\sum X_i^2} \tag{58}$$

E o estimador β_0, pode ser calculado a partir de:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \tag{59}$$

Sendo que a equação de estimativa será dada por:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \tag{60}$$

Mas será que a equação (60) foi bem estimada, ou melhor, será que ela representa bem a relação entre as variáveis? Uma maneira de avaliar é através da diferença entre os valores amostrais reais (Y) e os valores estimados ($\hat{Y}$), essa diferença é chamado de resíduo.

É preciso observar se as diferenças $(Y - \hat{Y}) = \hat{\varepsilon}$ são relativamente pequenas. Uma análise mais cuidadosa pode ser feita através da aplicação de testes estatísticos, nesse caso a ANOVA e o teste t-Student.

QUADRO 1 - TABELA ANOVA

Tabela ANOVA			
Soma dos Quadrados	Graus de Liberdade (g.l.)	Quadrados Médios (QM)	Teste F
SQE= $\hat{b}_1^{\,2} \sum x_i^2$ SQR= $\sum \left(Y - \hat{Y}\right)^2$	1 n-2	SQE/g.l. SQR/g.l.	SQEmed/SQRmed
SQT= $\sum y_i^2$	n-1	SQE/g.l + SQR/g.l.	

FONTE: AUTOR 2010

Obs: O grau de liberdade em relação ao SQE é devido a termos apenas uma variável independente; Em relação a SQT, os graus devem ser iguais a variância amostral, ou seja, n-1 (onde n é o número da elementos da amostra).

Onde:
Soma dos quadrados dos totais de Y centrado

$$SQT = \sum Y_i^2 \tag{61}$$

Soma dos quadrados explicados:

$$\mathrm{SQE} = \sum \hat{\mathrm{Y}}_{\mathrm{i}}^2 = \sum \hat{\beta}_{\mathrm{i}}^2 \, \mathrm{X}_{\mathrm{i}}^2 = \hat{\beta}_{\mathrm{i}}^2 \sum \mathrm{X}_{\mathrm{i}}^2 \tag{62}$$

Soma dos quadrados dos resíduos:

$$\mathrm{SQR} = \sum (\mathrm{Y} - \hat{\mathrm{Y}})^2 \tag{63}$$

Um outro parâmetro utilizado constantemente é o coeficiente de determinação, R^2, que explica percentualmente a relação entre as variáveis do problema.

$$\mathrm{R}^2 = \frac{\mathrm{SQE}}{\mathrm{SQT}} \tag{64}$$

2.7 CONSIDERAÇÕES RELEVANTES SOBRE A RETA DE REGRESSÃO

A previsão da variável dependente resultará sempre em um valor médio, pois, a relação entre X e Y é média. Para fazermos previsões acerca da variável dependente Y, não devemos utilizar valores da variável independente X que extrapolem o intervalo de valores utilizados no modelo de regressão.

Os pares de valores (X, Y) estão dispersos em relação a reta estimada. Isso ocorre entre outras razões, porque existem inúmeras outras variáveis externas, não consideradas no modelo que influenciam Y. Assim, não basta apenas calcularmos os coeficientes β_0 e β_1 da equação da reta de regressão pelo MQ. Precisamos verificar até que ponto tais estimativas são suficientes para explicar o relacionamento entre as variáveis X e Y. O erro padrão da estimativa S_e mede o

desvio médio entre os valores reais de Y e os valores estimados $\hat{Y}$. Ele informa de modo aproximado o quanto grande são os erros de estimativa em relação aos dados da amostra. S_e é medido na unidade de Y. O que se busca é conseguir o menor valor possível de S_e.

Assumindo que os desvios são "normalmente distribuídos", pode-se dizer então que 68% dos pontos (plotados) encontram-se dentro de 1 desvio padrão:

$$-1 \leq S_e \leq 1 \tag{65}$$

Sendo os desvios normalmente distribuídos a fórmula de Se é obtida da definição da variância da amostra, com n-1 graus de liberdade:

$$S_e = \frac{\sum(Y - \overline{Y})^2}{N - 1} \tag{66}$$

$$S_e = \sqrt{\frac{\sum(Y - \overline{Y})^2}{N - 1}} \tag{67}$$

O erro padrão existirá sempre que o poder de explicação da reta não for completo. O valor do erro significa então que existem outros fatores que interferem no comportamento de Y além da variável X.

2.8 ANÁLISE RESIDUAL

Toda a análise anterior foi desenvolvida supondo-se que os erros fossem variáveis aleatórias não correlacionadas, normal e independente distribuídas, com média zero e variância constante. Com o objetivo de validar tudo que foi visto, essas suposições necessitam ser verificadas.

Chama-se resíduo de um modelo de regressão a diferença entre o valor observado e o valor estimado da variável dependente.

$$e_i = Y_i - \hat{Y}_i \tag{68}$$

Através da análise dos resíduos se consegue verificar se os erros têm distribuição aproximadamente normal, com variância constante, e se a adição de termos adicionais ao modelo é útil.

A primeira coisa a ser verificada é a suposição de normalidade dos erros. Para isso, um histograma da freqüência dos resíduos pode ser construído. Como usualmente, as amostras têm tamanho pequeno, resultando em um histograma pouco significativo, prefere-se construir um gráfico conhecido como gráfico de probabilidade normal dos resíduos.

2.9 REGRESSÃO LINEAR MÚLTIPLA

2.9.1 Introdução

No capítulo anterior, estudou-se o modelo mais simples de relacionar uma variável dependente com apenas uma variável independente. Segundo Hair (2005), existem inúmeros fenômenos que envolvem muitas variáveis independentes. Neste capítulo, estudar-se-á ainda uma relação linear entre as variáveis independentes e a variável dependente, como a mostrada a seguir:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots\ldots\ldots + \beta_n x_n + \varepsilon \tag{69}$$

Essa equação é conhecida como modelo de regressão linear múltipla. Como antes, o parâmetro β_o é conhecido como a interseção do plano ou coeficiente linear. Os outros parâmetros são conhecidos como coeficientes parciais de regressão, porque (no caso de duas variáveis independentes) β_1 mede a variação esperada em Y por unidade de variação em x_1, quando x_2 for constante, e β_2 mede a variação esperada em Y por unidade de variação em x_2, quando x_1 for constante. No caso geral, o parâmetro β_j representa a variação esperada na resposta Y por unidade de variação unitária em x_j, quando todos os outros regressores (ou variáveis independentes) x_i ($i \neq j$) forem mantidos constantes.

Modelos que sejam mais complexos na estrutura do que a equação (69) podem freqüentemente ainda ser analisados por técnicas de regressão linear múltipla. Por exemplo, considere o modelo polinomial cúbico com um regressor.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x^2 + \beta_3 x^3 \ \ldots\ldots\ldots + \beta_n x_n + \varepsilon \tag{70}$$

Se for o caso de $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$, então a equação (70) assume à forma apresentada na equação (71), com três regressores.

Modelos que incluem efeitos de interação podem ser analisados pelos métodos de regressão linear múltipla. Uma interação entre duas variáveis pode ser representada por um termo cruzado no modelo, tal como

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon \quad (71)$$

Se fizer $x_3 = x_1 x_2$ e $\beta_3 = \beta_{12}$, então a Equação (70) pode ser escrita na forma da Equação (71). No caso de se considerar o termo de interação, a superfície gerada pelo modelo não é linear e as curvas de nível são curvadas,. Em geral, qualquer modelo de regressão que seja linear nos parâmetros (os β's) é um modelo de regressão linear, independente da forma da superfície que ele gere.

2.9.2 Regressão Linear Múltipla

Em algumas situações mais do que uma variável independente ($X_1, X_2, ..., X_n$) pode ser necessária para predizer o valor da variável dependente (Y). O modelo matemático para esse caso é dado abaixo:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ \ldots\ldots\ldots + \beta_n x_n + \varepsilon \quad (72)$$

Que para as n observações poderá se escrito da forma:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_1 + \beta_2 x_{21} + \ \ldots\ldots\ldots\ldots + \beta_k x_{k1} + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 x_2 + \beta_2 x_{22} + \ \ldots\ldots\ldots\ldots + \beta_k x_{k2} + \varepsilon_2 \\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ Y_n &= \beta_0 + \beta_1 x_n + \beta_2 x_{n2} + \ \ldots\ldots\ldots\ldots + \beta_n x_{kn} + \varepsilon_n \end{aligned} \quad (73)$$

Que forma na realidade um sistema linear, que podermos escrever na forma de matriz como:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_{21} & x_1 \\ 1 & x_2 & x_{22} & x_1 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_n & x_{2n} & x_1 \end{bmatrix} . \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_k \end{bmatrix} \tag{74}$$

O modelo de regressão nesse caso é dado por:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ \ldots\ldots\ldots + \beta_n x_n + \varepsilon \tag{75}$$

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i \qquad i = 1,2, \ldots\ldots, n \tag{76}$$

Aplicando o princípio dos mínimos quadrados à equação (76), obtém-se:

$$S = \sum_{n}^{i} (Y_i - Y_i)^2 \tag{77}$$

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = 2 \sum_{i=1}^{n} (Y_i - Y_i) = 0$$

$$\left.\frac{\partial S}{\partial \beta_j}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = 2 \sum_{i=1}^{n} (Y_i - Y_i) X_{ij} = 0 \qquad \text{j=1,2,3,...,k} \tag{78}$$

Após manipulações algébricas, as equações normais de mínimos quadrados são obtidas:

$$\begin{matrix} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \ \ldots. + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} Y_i \\ n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \ \ldots. + \hat{\beta}_k \sum_{i=1}^{n} x_{i1} x_{ik} = \sum_{i=1}^{n} x_{i1} Y_i \\ \vdots \qquad\qquad \vdots \qquad\qquad\qquad\qquad \vdots \qquad\qquad \vdots \\ n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik} x_{i2} + \ \ldots. + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik} Y_i \end{matrix} \tag{79}$$

O sistema representado na equação (79) pode ser resolvido por qualquer método de resolução de equações algébricas lineares. Existem m = k + 1 equações, que podem ser expressas na forma matricial:

$$\underline{\mathbf{Y}} = \mathrm{X}\underline{\boldsymbol{\beta}} + \underline{\boldsymbol{\varepsilon}} \tag{80}$$

em que

$$\underline{\mathbf{Y}} = \begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n1} \end{bmatrix} \qquad \mathrm{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \tag{81}$$

$$\underline{\boldsymbol{\beta}} = \begin{bmatrix} \beta_{11} \\ \beta_{21} \\ \vdots \\ \beta_{n1} \end{bmatrix} \qquad \underline{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{n1} \end{bmatrix} \tag{82}$$

Com Y sendo um vetor (*n* x 1) das observações, X sendo uma matriz (*n* x *m*) contendo os valores das variáveis independentes, $\underline{\boldsymbol{\beta}}$ sendo um vetor (*m* x 1) dos parâmetros (ou coeficientes de regressão) e $\underline{\boldsymbol{\varepsilon}}$ sendo um vetor (*n* x 1) dos erros aleatórios.

Uma outra forma de representar a equação (80) é:

$$\mathrm{X}'\mathrm{X}\underline{\widehat{\boldsymbol{\beta}}} = \mathrm{X}'\underline{\mathbf{y}} \tag{83}$$

em que **X'** é a matriz transposta de **X**. Como o objetivo é determinar os parâmetros, a equação (83) deve ser resolvida, resultando em:

$$\underline{\widehat{\boldsymbol{\beta}}} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\underline{\mathbf{y}} \tag{84}$$

em que $(X'X)^{-1}$ a matriz inversa de $X'X$. Qualquer método pode ser utilizado para fazer a inversão da matriz. Uma vez determinados os parâmetros, o modelo proposto toma a forma:

$$\hat{Y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \tag{85}$$

$$\underline{\hat{\mathbf{Y}}} = X\underline{\hat{\boldsymbol{\beta}}} \tag{86}$$

O vetor dos resíduos, $e_i = y_i - \hat{y}_i$ é dado por $\underline{\mathbf{e}} = \underline{\mathbf{Y}} - \underline{\hat{\mathbf{Y}}}$

2.9.3 Variância Dos Parâmetros

As variâncias dos $\hat{\beta}'s$ são expressas em termos dos elementos da inversa da matriz. $X'X$ A inversa de $X'X$ vezes a constante σ^2 representa a matriz de covariância dos coeficientes de regressão $\hat{\beta}$. Os elementos da diagonal de $\sigma^2(X'X)^{-1}$ são as variâncias de $\hat{\beta}_0, \hat{\beta}_1, \ldots\ldots\ldots\ldots\ldots, \hat{\beta}_k$ e os elementos fora da diagonal dessa matriz são as covariâncias.

$$C = (X'X)^{-1} \tag{87}$$

$$C = \begin{bmatrix} c_{00} & c_{01} & \cdots & c_{0k} \\ c_{10} & c_{11} & \cdots & c_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k0} & c_{k1} & \cdots & c_{kk} \end{bmatrix} \tag{88}$$

que é simétrica $(c_{10} = c_{01}, c_{20} = c_{02}, c_{k0} = c_{0k})$ porque $(X'X)^{-1}$ é simétrica, tendo-se

$$\mathrm{Var}(\hat{\beta}_j) = \sigma^2 c_{ij} \qquad j = 0,1,2,\ldots\ldots\ldots\ldots k$$
$$\mathrm{Covar}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 c_{ij} \qquad i \neq j \tag{89}$$

Em geral, a matriz de covariância de $\hat{\beta}$ é uma matriz simétrica (*m* x *m*), cujo *jj*-ésimo elemento é a variância de $\hat{\beta}_j$ e cujo *i,j*-ésimo elemento é a covariância entre $\hat{\beta}_i$ e $\hat{\beta}_j$; ou seja:

$$\mathrm{Cov}(\hat{\beta}) = \sigma^2 (X'X)^{-1} = \sigma^2 C \tag{90}$$

As estimativas das variâncias desses coeficientes de regressão são obtidas trocando σ^2 pela estimativa apropriada. Quando σ^2 for trocado pela estimativa, $\hat{\sigma}^2$ a raiz quadrada da variância estimada do j-ésimo coeficiente de regressão é chamada de erro-padrão estimado (ou desvio-padrão) de $\hat{\beta}_j$ ou $\exp(\hat{\beta}_j) = \sqrt{\sigma^2 c_{ij}}$. Da mesma forma que na regressão linear simples, a estimativa de σ^2é definida em termos da soma quadrática dos resíduos

$$SQ_E = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \underline{\mathbf{e}}'\underline{\mathbf{e}}$$
$$SQ_E = \underline{\mathbf{Y}}'\underline{\mathbf{Y}} - \underline{\hat{\boldsymbol{\beta}}}'X'\underline{\mathbf{Y}} \tag{91}$$

A média quadrática residual ou do erro é expressa por:

$$\hat{\sigma}^2 = MQ_E = \frac{SQ_E}{n-m} = \frac{\underline{\mathbf{Y}}'\underline{\mathbf{Y}} - \underline{\hat{\boldsymbol{\beta}}}'X'Y}{n-m} \tag{92}$$

2.9.4 Testes Nos Coeficientes Individuais De Regressão e nos Subconjuntos De Coeficientes

Está-se freqüentemente interessados em testar hipóteses para os coeficientes individuais de regressão. Tais testes são úteis na determinação do valor potencial de cada um dos regressores no modelo de regressão. Por exemplo, o

modelo pode ser mais efetivo com a inclusão de variáveis adicionais ou talvez com a retirada de um ou mais regressores atualmente no modelo.

A adição de uma variável ao modelo de regressão sempre aumenta a soma quadrática da regressão e sempre diminui a soma quadrática do erro. Tem-se de decidir se o aumento na soma quadrática da regressão é grande o suficiente para justificar o uso de uma variável adicional no modelo. Além disso, a adição de uma variável não importante ao modelo pode na verdade aumentar a soma quadrática do erro, indicando que a adição de tal variável fez realmente o modelo apresentar um ajuste mais pobre dos dados.

MONTGOMERY E RUNGER (2003), as hipóteses para testar a significância de qualquer coeficiente de regressão, como β_j, são:

$$H_0: \beta_j = 0$$
$$H_1: \beta_j \neq 0 \tag{93}$$

Se $H_0: \beta_j = 0$ não for rejeitada, então isso indica que o regressor x_j poderá ser retirado do modelo. A estatística de teste para essa hipótese é

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\sigma^2 c_{ij}}} \tag{94}$$

em que C_{jj} é o elemento da diagonal de $(X'X)^{-1}$ correspondente a $\hat{\beta}_j$. Observe que o denominador da equação (94) é o erro-padrão do coeficiente. $\hat{\beta}_j$ A hipótese nula $H_0: \beta_j = 0$ será rejeitada se $|t_0| > t_{\alpha/2n-m}$ isso é chamado de teste parcial ou marginal, porque o coeficiente de regressão β_j depende de todos os outros regressores X_i ($i \neq j$) que estão no modelo. O regressor associado ao parâmetro β_j, X_j, contribui então significativamente para o modelo.

2.9.5 Intervalos De Confiança Na Regressão Linear Múltipla

O intervalo de confiança para os parâmetros é calculado pela equação (95).

$$\beta_j - t_{\alpha/2n-m}\sqrt{\sigma^2 c_{ij}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2n-m}\sqrt{\sigma^2 c_{ij}} \quad (95)$$

Enquanto que para a resposta média, ele é dado pela Equação (96).

$$\hat{\mu}_{Y/x_0} - t_{\alpha/2n-m}\sqrt{\hat{\sigma}^2 x_0\,(X'X)^{-1}x_0} \leq \hat{\mu}_{Y/x_0} \leq \hat{\mu}_{Y/x_0} + t_{\alpha/2n-m}\sqrt{\hat{\sigma}^2 x_0\,(X'X)^{-1}x_0} \quad (96)$$

Sendo $\underline{\mathbf{X}}_0$ o vetor das variáveis independentes em um determinado ponto.

Coeficiente de Determinação Múltipla e de Correlação Múltipla.

Esses coeficientes são definidos da mesma forma de antes.

2.9.6 Análise Residual

Seguindo MONTGOMERY E RUNGER (2003), na regressão linear simples, os resíduos foram padronizados na forma $e_i = Y_i - \hat{Y}_i$ ou seja de modo que seus desvios-padrão sejam aproximadamente unitários. Isso possibilita a descoberta fácil de um possível *outlier*.

Uma outra forma de escalonar os resíduos é usá-los na forma de Student.

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}} \quad (97)$$

em que h_{ii} é o *i*-ésimo elemento diagonal da matriz H, conhecida como matriz "chapéu".

$$\mathrm{H} = \mathrm{X}\,(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' \tag{98}$$

Como

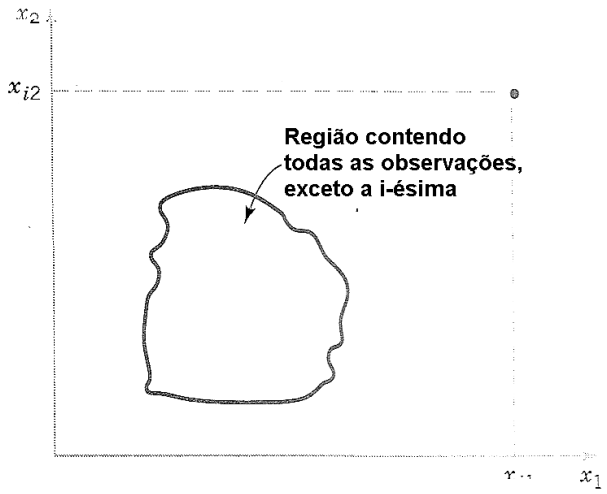
$$\underline{\hat{\mathbf{Y}}} = \mathrm{X}\underline{\hat{\boldsymbol{\beta}}} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\underline{\mathbf{Y}} = \mathrm{H}\underline{\mathbf{Y}} \tag{99}$$

conclui-se que a matriz H transforma os valores observados, Y, em valores ajustados, $\hat{\mathbf{Y}}$.

2.9.7 Observações Influentes

Algumas vezes, essas observações que influenciam estão relativamente longe da vizinhança onde o resto dos dados foi coletado. Uma situação hipotética para duas variáveis é mostrada na figura (4), onde uma observação no espaço x está distante do resto dos dados. A disposição dos pontos no espaço x é importante na determinação das propriedades do modelo. Por exemplo, o ponto (x_{i1},x_{i2}) na figura (4) pode exercer muita influência na determinação de R^2, nas estimativas dos coeficientes de regressão e na magnitude da média quadrática dos erros.

FIGURA 4 - EXEMPLO DE UM *OUTLIER*



Um ponto que está longe do espaço x.

FONTE: MONTGOMORY(2005,p.249)

Quer-se examinar os pontos que influenciam de modo a determinar se eles controlam muitas propriedades do modelo. Se esses pontos que influenciam forem pontos ″ruins″, ou errôneos de algum modo, então eles devem ser eliminados. Por outro lado, pode não haver algo errado com esses pontos, porém, no mínimo, quer-se determinar se eles produzem ou não resultados consistentes com o resto dos dados. Em qualquer evento, mesmo se um ponto de influência for válido, se ele controlar importantes propriedades do modelo, quer-se saber isso, uma vez que ele poderia ter um impacto no uso do modelo, é necessário decidir o que fazer com as observações discordantes.

A maneira mais simples de lidar com essas observações é eliminá-las. Como já foi dito, esta abordagem, apesar de ser muito utilizada, não é aconselhável. Ela só se justifica no caso de os outliers serem devidos a erros cuja correção é inviável. Caso contrário, as observações consideradas como outliers devem ser tratadas cuidadosamente pois contêm informação relevante sobre características subjacentes aos dados e poderão ser decisivas no conhecimento da população à qual pertence a amostra em estudo.

MONTGOMERY e RUNGER (2003), descrevem vários métodos de detecção de observações que influenciam.

Um excelente diagnóstico é a medida da distância, desenvolvido por Dennis R. Cook. Essa é uma medida da distância ao quadrado entre a estimativa usual de mínimos quadrados de β, baseada em todas *n* observações, e a estimativa obtida quando o *i*-ésimo ponto for removido, como $\hat{\beta}_{(i)}$. A medida da distância Cook é:

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})X'X(\hat{\beta}_{(i)} - \hat{\beta})}{m\hat{\sigma}^2} \quad i = 1,2,\dots,n \tag{100}$$

Claramente, se o *i*-ésimo ponto exercer influência, sua remoção resultará em $\hat{\beta}_{(i)}$ variando consideravelmente do valor $\hat{\beta}$. Logo, um grande valor de D_i implica que o *i*-ésimo ponto exerce influência. A estatística D_i é realmente calculada usando.

$$D_i = \frac{r_i^2}{m}\frac{h_{ii}}{(1-h_{ii})} \qquad i = 1,2,\dots,n \tag{101}$$

Da equação (99), vemos que D_i consiste do quadrado do resíduo na forma de Student, que reflete quão bem o modelo ajusta a *i*-ésima observação Y_i [lembre-se que $r_i = e_i/\sqrt{\hat{\sigma}^2(1-h_{ii})}$] e um componente que mede quão longe aquele ponto está do resto dos dados [$h_{ii}/(1 - h_{ii})$ é uma medida da distância do *i*-ésimo ponto do centróide dos n - 1 pontos restantes]. Um valor de $D_i > 1$ indicaria que o ponto exerce influência. Cada componente de D_i (ou ambos) pode contribuir para um grande valor. Outro método a ser utilizado é o Modelos de discordância:

Num modelo de discordância considera-se que num dado conjunto de dados,se existirem observações aberrantes elas têm distribuição diferente das restantes observações ou distribuições idênticas mas com parâmetros diferentes.

H0: a amostra foi retirada de uma população com distribuição específica que pode ou não ser conhecida e ser especificada completamente ou não, e onde não existem observações “anormais”

H1: todas as observações ou apenas as “anormais” têm distribuição diferente da da hipótese nula. hipótese nula será rejeitada a favor da hipótese alternativa se existirem observações aberrantes.

Para decidir pela aceitação ou rejeição da hipótese nula, da não existência de *outliers* é necessário utilizar testes de discordância que tenham distribuição desconhecida ou valores críticos tabelados. Na utilização de testes formais de *outliers* deve ter-se em conta que eles dividem-se em duas classes:

- aqueles em que as observações discordantes da amostra são identificadas como sendo *outliers*.
- aqueles que testam a presença de *outliers* mas não identificam observações particulares.

Um método citado e muito utilizado é o Z-Scores que conssiste em:
Calcular os *z-scores*, isto é, os valores z-standardizados dos dados.

- Se o conjunto dos dados é pequeno (inferior a 50), valores que tenham *z-scores*
- inferiores a -2.5 ou superiores a 2.5 devem ser considerados *outliers.*
- Se o conjunto dos dados é grande, valores que tenham z-socres inferiores a
- -3.3 ou superiores a 3.3 são tipicamente considerados *outliers.*
- Se o conjunto dos dados é muito grande (1000 ou mais), também valores
- mais extremos do que +-3.3 podem ser considerados dados normais e não *outliers.*

2.9.8 Multicolinearidade

Em problemas de regressão múltipla, espera-se encontrar dependências entre a variável de resposta Y e os regressores X_j. Na maioria dos problemas de regressão, no entanto, encontra-se que há também dependências entre os regressores X_j. Em situações onde essas dependências forem fortes, dizemos que existe multicolinearidade. A multicolinearidade pode ter sérios efeitos nas estimativas dos coeficientes de regressão e na aplicabilidade geral do modelo estimado.

Algumas indicações da presença de multicolinearidade são:

1. Valores altos do coeficiente de correlação.

2. Grandes alterações nas estimativas dos coeficientes de regressão, quando uma variável independente for adicionada ou retirada do modelo, ou quando uma observação for alterada ou eliminada.
3. A rejeição da hipótese $H_0: \beta_1 = \beta_2 = \cdots \ldots . = \beta_k = 0$ por meio da realização do teste F, mas nenhuma rejeição das hipóteses $H_0: \beta_i = 0, i = 1,2, \ldots \ldots . . k$, por meio da realização dos testes *t* sobre os coeficientes individuais de regresssão.
4. Obtenção de estimativas para os coeficientes de regressão com sinais algébricos contrários àqueles que seriam esperados a partir de conhecimentos teóricos disponíveis ou de experiências anteriores sobre o fenômeno estudado.
5. Obtenção de intervalos de confiança com elevadas amplitudes para os coeficientes de regressão, associados a variáveis independentes importantes.

Os efeitos de multicolinearidade podem ser facilmente demonstrados. Os elementos da diagonal da matriz $C = ((X'X)^{-1}$ podem ser escritos como.

$$C_{IJ=} \frac{1}{(1-R_J^2)} \qquad J = 1,2, \ldots \ldots \ldots, K \tag{102}$$

Sendo R_J^2 o coeficiente de determinação múltipla, resultante da regressão de X_j nos outros k - 1 regressores. Claramente, quanto mais forte for a dependência linear de X_j nos regressores restantes, e por conseguinte mais forte a colinearidade, maior será o valor de R_J^2. Lembre-se de que $V\hat{\beta}_J = \sigma^2 C_{IJ}$. Logo, diz-se que a variância de $\hat{\beta}_J$ é "inflacionada" pela quantidade$(1 - R_J^2)^{-1}$. Dessa maneira, define-se o fator de inflação da variância para $\hat{\beta}_J$ como:

$$FIV(\hat{\beta}_J) = \frac{1}{(1-R_J^2)} \quad J = 1,2, \ldots \ldots, K \tag{103}$$

Esses fatores são uma importante medida da extensão da presença de multicolinearidade.

Embora as estimativas dos coeficientes de regressão sejam muitas imprecisas quando a multicolinearidade está presente, a equação do modelo ajustado pode ainda ser útil. Por exemplo, suponha que se deseje prever as novas

observações para a resposta. Se essas previsões forem interpolações na região original do espaço X onde a multicolinearidade existe, então previsões satisfatórias serão freqüentemente obtidas, porque, enquanto os $\hat{\beta}_J$ qais podem ser pobremente estimados, a função $\sum_{j=1}^{k} \beta_j x_{ij}$ pode ser bem estimada. Por outro lado, se a previsão das novas observações requerer extrapolação além da região original do espaço X onde os dados foram coletados, geralmente então esperaríamos obter resultados pobres. Extrapolação requer geralmente boas estimativas dos parâmetros individuais do modelo.

Multicolinearidade aparece devido a várias razões. Ela ocorrerá quando o analista coletar dados, tal que uma restrição linear se mantenha aproximadamente entre as colunas da matriz X. Por exemplo, se quatro regressores forem os componentes de uma mistura, então tal restrição sempre existirá porque a soma dos componentes sempre é constante. Geralmente, essas restrições não se mantêm exatamente e o analista pode não saber que elas existem.

A presença de multicolinearidade pode ser detectada de várias maneiras. Duas das mais fáceis de se entender são:

1. Os fatores de inflação da variância, definidos na equação (103), são medidas muito úteis de multicolinearidade. Quanto maior for o fator de inflação da variância, mais severa será a multicolinearidade. Alguns autores sugeriram que se qualquer fator de inflação da variância exceder 10, então a multicolinearidade será um problema.
2. Outros autores consideram esse valor muito liberal e sugerem que os fatores de inflação da variância não devem exceder 4 ou 5.
3. Se o teste F para significância da regressão for significante, mas os testes para os coeficientes individuais de regressão não forem significantes, então a multicolinearidade pode estar presente.

Várias medidas têm sido propostas para resolver o problema de multicolinearidade. Freqüentemente, sugere-se aumentar os dados com novas observações especificamente projetadas para fragmentar as dependências lineares aproximadas que existem correntemente. Entretanto, isso é algumas vezes impossível por causa de razões econômicas ou por causa de restrições físicas que relacionam o X_j. Uma outra possibilidade é remover certas variáveis do modelo, porém essa abordagem tem a desvantagem de descartar a informação contida nas variáveis removidas.

3. MATERIAL E MÉTODO

3.1 LEVANTAMENTO DOS DADOS

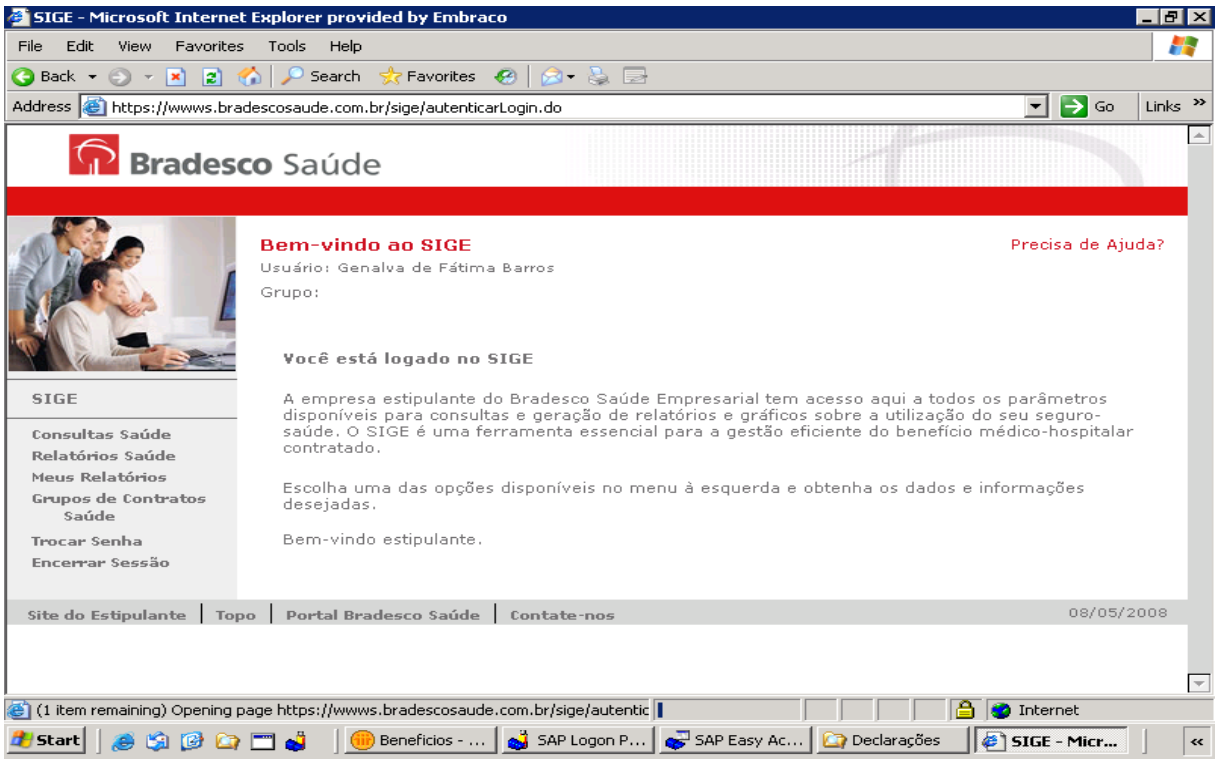
Empresa situada na região do norte de Santa Catarina emprega aproximadamente 3.500 funcionários, oferece aos colaboradores e seus dependentes um plano de saúde que chega a um total aproximado de 10.000 usuários, no qual esta na forma de pós-pagamento, isto significa que o faturamento é realizado após análise periódica dos custos referentes aos serviços utilizados pelos segurados, acrescido de uma taxa de administração previamente acordada com a empresa contratante do seguro e de tributos. Portanto a empresa não possui ferramentas e argumentos para formar estratégia para tentar estabelecer um critério do sinistro pago a operadora e para com isso tentar diminuir valor do custo do plano pago à mesma.

O presente trabalho tem por objetivo identificar quais fatores interferem no auto custo deste benefício, relacionando o perfil desses beneficiários e de seus dependentes, e analisar o impacto desses gastos nas finanças do plano e de seus beneficiários. Este plano de saúde é mantido pela empresa como um benefício social junto a seus funcionários e dependentes, na busca de um maior conforto e segurança a todos devido à grande precariedade que se encontra a saúde publica no Brasil, mas crescentes gastos com o benefício vêm preocupando os diretores e acionistas da empresa, pois o seu crescimento anual é cerca 10% a 15% conforme levantamento efetuado internamente e repassado pelo Gerente de RH.

Analisar e compreender as variáveis que influenciam os custos permite traçar políticas racionais para a redução de seu custo. O uso de ferramentas matemáticas, tais como estatística de análise multivariada, oferece a empresas a possibilidade de analisar e estratificar grande quantidade de dados possibilitando encontrar padrões ou correlações entre essas informações de modo que possa levar a empresa a criar programas de incentivo ou de melhorias para que se possa diminuir o uso do plano de saúde de forma indevida ou incorreta, conseqüentemente acarretando a redução no custo.

Mensalmente a operadora informa as quantidades e os tipos de serviços utilizados, bem como, os custos inerentes. O acesso as informações é feito através de sistema on-line, via internet, restrito ao gerente de Recursos Humanos (RH) e seus subordinados por ele autorizados. Na figura (05) tem-se um exemplo da página de acesso ao sistema.

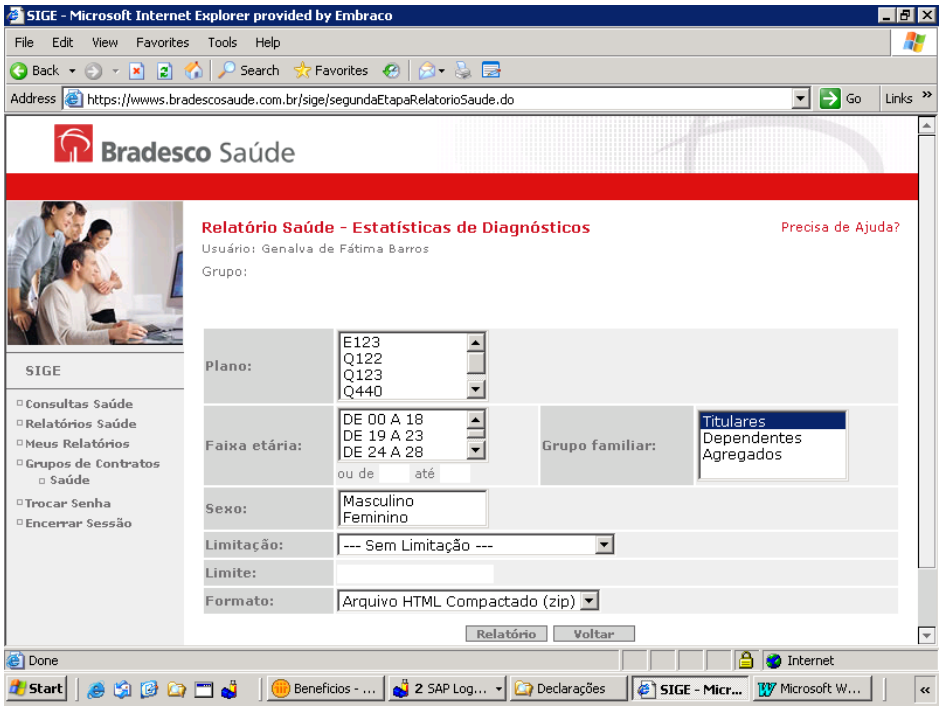
FIGURA 5 - PAGINA INICIAL DA OPERADORA NA INTERNET



FONTE: EMPRESA (2009)

Ao acessar o sistema é fornecido todas as informações referentes a utilização do Plano de Saúde, conforme pode-se visualizar na figura (06).

FIGURA 6 - PAGINA DE ACESSO DA EMPRESA JUNTO A OPERADORA DO PLANO DE SAÚDE VIA INTERNET



FONTE: EMPRESA (2009)

Os códigos listados no campo “*Plano*” estão referenciados na TABELA 03. Os campos “*Faixa etária*”, “*Grupo familiar*”, “*sexo*” e “*Formato*” são autos explicativos, não sendo necessário maiores esclarecimentos. O campo “*Limitação*” não utilizado pela empresa.

TABELA 3 - CÓDIGO DOS PLANOS DE SAÚDE

Código	Colaborador
E 123	Básico
Q 122	Líderes
Q440	Gestores
Q 990	Diretores

FONTE: EMPRESA (2009)

Os dados coletados para este trabalho não configuraram deste serviço já que o mesmo demonstrou-se muito limitado quanto a captação dos dados, ou seja informações relativas aos usuários do plano de saúde e de seus dependentes o qual não foi eficiente pois dos dados são agrupados e não individuais, portanto os dados são coletados e estratificados através dos relatórios mensais repassados pela Operadora do Plano de Saúde a empresa, essas informações são individuais de cada usuário onde contém informações de cada procedimento realizado pelo titular ou pelo dependente no ano de 2008.

Conforme já citado acima, dentro, das informações oferecidas pelo *softwer* optou-se em não fazer uso do mesmo, pois as informações estão todas agrupadas, com isso não oferece condições para alcançar o objetivo proposto neste trabalho, portanto a necessidade de se obter uma nova fonte de dados quais foram passados pelo RH da empresa, em arquivos de textos com relatórios mensais de todo o período de 2008 compactados, quais houve um exaustivo trabalho de decodificação do mesmo, essas informações só foram possível graças a uma planilha (quadro 2) na qual descreve todas as informações referente ao arquivo.

QUADRO 2 - DECODIFICAÇÃO DOS DADOS

Continua

DESCRIÇÃO DO REGISTRO	
CAMPO	CONTEÚDO
TIPO DE REGISTRO	M (fixo)
NÚMERO DA SUBFATURA	
NOME DA SUBFATURA	Nome da empresa contratante do Plano de Saúde
TIPO DA SUBFATURA	01-Técnica 02-Cancelada 03-Administrativa
NÚMERO DO CERTIFICADO	
MATRÍCULA	Matricula do titular do Plano de Saúde
NOME DO SEGURADO	Nome do titular do Plano de Saúde
CÓDIGO DO PACIENTE	1- Cônjuge 2- Filho 3- Mãe 4- Pai 5- Sogro 6- Sogra

Continua

	7- Tutelado 8- Outros
NOME DO PACIENTE	Nome do paciente
NOME DO BENEFICIÁRIO	Nome do hospital clinica ou prestador de serviço
TIPO DE EVENTO	1-Consulta 2- Evento 3- Desp. Hosp. 4- Hon. Médicos 5- Exames Simples 6- Exames Especiais 7- Clinica Especializada 8- Farmácia 9- Ter.Inf.
NÚMERO DO DOCUMENTO	Número do documento ou senha, liberação dos procedimentos
CÓDIGO DO PROCEDIMENTO	De acordo com a tabela AMB (arquivo de procedimentos)
QUANTIDADE DE PROCEDIMENTOS	Quantidade de procedimentos efetuada pelo paciente
DATA DO PAGAMENTO	Data de pagamento dos procedimentos
VALOR PAGO	Em FAJ-TR (Vigente na data do pagamento)
DATA DO EVENTO	Dia, Mês e Ano que foi realizado o procedimento
NÚMERO DO COTRATO	Número do contrato da prestadora junto a Operadora
CÓDIGO DO REFERENCIADO	Nome do Médico
DATA DE NASCIMENTO	Data de nascimento do Paciente
SEXO	Sexo do paciente 1- Masculino 2-Feminíno
GRAU DE PARENTESCO	0-Titular 1- Conjuge 2- Filho 3- Mãe 4- Pai 5- Sogro 6- Sogra 7- Tutelado

Conclusão

	8- Outros
ESPECIALIDADE	Código da especialidade do prestador de serviço
VALOR DO SINISTRO	Valor Real pago em moeda vigente R$
CÓDIGO DA AUTORIZAÇÃO	1- Segurado Novo 2- Extravio ou perda do cartão; 3- Admissional 4-Demissional 5-Periódico 6-Acidente de Trabalho 7-Outros 8-Assistência a demitidos 9-Medicina Social.
CPF/CGC DO REFERENCIADO	CPF (para Pessoas Físicas) ou CGC (para Pessoas Jurídicas) Zeros, nos casos de reembolso.
TIPO DO REFERENCIADO	F= Pessoa Física J= Pessoa Jurídica Brancos, nos casos de reembolso
VALOR DE INSS OU ISS R$	Em R$ (ISS para Pessoa Jurídica / INSS para Pessoa Fisíca)
VALOR TOTAL EM R$	Em FAJ-TR (Do último dia útil do mês) (ISS para Pessoa Jurídica / INSS para Pessoa Fisíca)
VALOR DE INSS OU ISS FAJ TR	Em FAJ-TR (ISS para Pessoa Jrurídica / INSS para Pessoa Fisíca)
CARGO DO SEGURADO	Código de hierarquia dentro da empresa
DATA DE ADMISSÃO	Data de admissão do titular do plano junto a empresa
PLANO DO SEGURADO	E123 - Básico Q122 - Líderes Q440- Gestores Q990 - Diretores EA23 -Aposentados
CDB	Código de Doença Bradesco
TROCA DE ACOMODAÇÃO	Se houve ou não troca de acomodações

FONTE:EMPRESA (2010)

Através das informações do quadro 02, e análise das mesmas, optou-se em nas focar nas variáveis que estão diretamente relacionadas ao perfil de consumo dos colaboradores e de seus dependentes, através destes identificar quais influenciam no custo do plano de saúde junto a operadora, que é o objetivo deste trabalho.

Para se obter a matriz de dados com as variáveis selecionadas, houve-se a necessidade de alteração na codificação de alguns dados principalmente, tipos de planos de saúde onde receberam nova codificação, E123=1; Q122=2; Q440=3; Q990=4; EA123=5; A planilha dos dados geral somente foi concluída após ser compactada utilizando planilha de cálculo do *EXCEL* precisamente uma tabela dinâmica qual gerou a planilha com os dados. Já que os dados são individuais para cada paciente, num total de 5.009 titulares de plano de saúde, cada um acrescido de seus respectivos dependentes.

3.2 METODOLOGIA

A metodologia aqui proposta procura determinar quais variáveis influenciam no custo de um plano de saúde, utilizando o *softwer* MATLAB .V5.3.1, a matriz de dados foi transformada através de algoritmo criado por MARQUES (2005), que decompõe as variáveis originais, usando a metodologia das componentes principais.

Procurando obter uma redução da dimensionalidade dos dados, para somente aquelas que influenciam expressivamente o conjunto de dados. Assim conservou-se as primeiras componentes e as variáveis que constituem um resumo de informações mais importantes da estrutura de covariância. Com a finalidade de alcançar o objetivo deste, que é a obtenção do modelo de precisão, fazendo uso da técnica de regressão linear múltipla foi desenvolvido um estudo, no *software MINITAB.15*, para prever qual o custo do plano de saúde para a contratação de um novo colaborador através de seu perfil.

Obtemos através da tabela de ANOVA e seus índices de regressão, a verificação em %(porcentagem) do quanto essas variáveis (componentes principais) representam de informação entre as variáveis originais.

4. RESULTADO E ANÁLISE

A análise foi efetuada com base em 20 variáveis constituídos a partir dos dados originalmente avaliados e a decisão de agrupamento destas é o escopo desse trabalho. Os principais fatores que contribuem na elevação dos custos operacionais incorridos pela empresa estudada e o pagamento do sinistro mensal a empresa Operadora do Plano de Saúde influenciaram decisivamente nos critérios de agrupamento e estudo.

Dos dados fornecidos pela empresa elaborou-se o quadro (3), que apresenta as variáveis estudadas. Este quadro descreve a nomenclatura e simbologia que será adotada no presente estudo de caso.

QUADRO 3 - VARIÁVEIS DA MATRIZ DE DADOS

X_1	Nº de dependentes do Titular do plano de saúde	X_2	Tempo de serviço na empresa do titular do plano
X_3	Tipo de Plano do titular	X_4	Sexo do titular do Plano
X_5	Idade do paciente titular até 31 / 12 /2008	X_6	Quantidade de procedimentos realizados pelo titular em 2008
X_7	Quantidade de procedimentos realizados pelo cônjuge do titular em 2008	X_8	Idade do paciente cônjuge até 31 / 12 /2008
X_9	Quantidade de procedimentos realizados pelos filho(a) em 2008	X_{10}	Idade do paciente filho(a) até 31 / 12 /2008
X_{11}	Quantidade de procedimentos realizados pelo tutelado em 2008	X_{12}	Idade do paciente tutelado até 31 / 12 /2008
X_{13}	Quantidade de procedimentos realizados pela mãe em 2008	X_{14}	Idade do paciente mãe até 31 / 12 /2008
X_{15}	Quantidade de procedimentos realizados pelo pai em 2008	X_{16}	Idade do paciente pai até 31 / 12 /2008
X_{17}	Quantidade de procedimentos realizados pelo sogro em 2008	X_{18}	Idade do paciente sogro até 31 / 12 /2008
X_{19}	Quantidade de procedimentos realizados pela sogra em 2008	X_{20}	Idade do paciente sogra até 31 / 12 /2008

FONTE: AUTOR (2010)

4.1 CÁLCULO DOS AUTOVALORES E AUTOVETORES

A partir das variáveis originais desenvolveu-se no ambiente de programação *MatLab V5.3.1* a análise de componentes principais, gerando-se 20 vetores ortonormais (componentes principais) e, 20 autovalores (diagonal da matriz de correlação), que são equivalentes às variâncias apresentadas em cada componente. Utilizando-se o Critério de Kaiser obteve-se oito autovalores que representam 73,89% dos dados e, estes autovalores foram utilizados na construção das componentes principais.

Torna-se interessante observar os valores dos coeficientes encontrados nas combinações lineares que representam os componentes principais e das correlações entre as variáveis padronizadas e os componentes principais, pois estes valores são imprescindíveis para a discussão dos resultados, os autovalores e sua proporção estão apresentados na tabela (04):

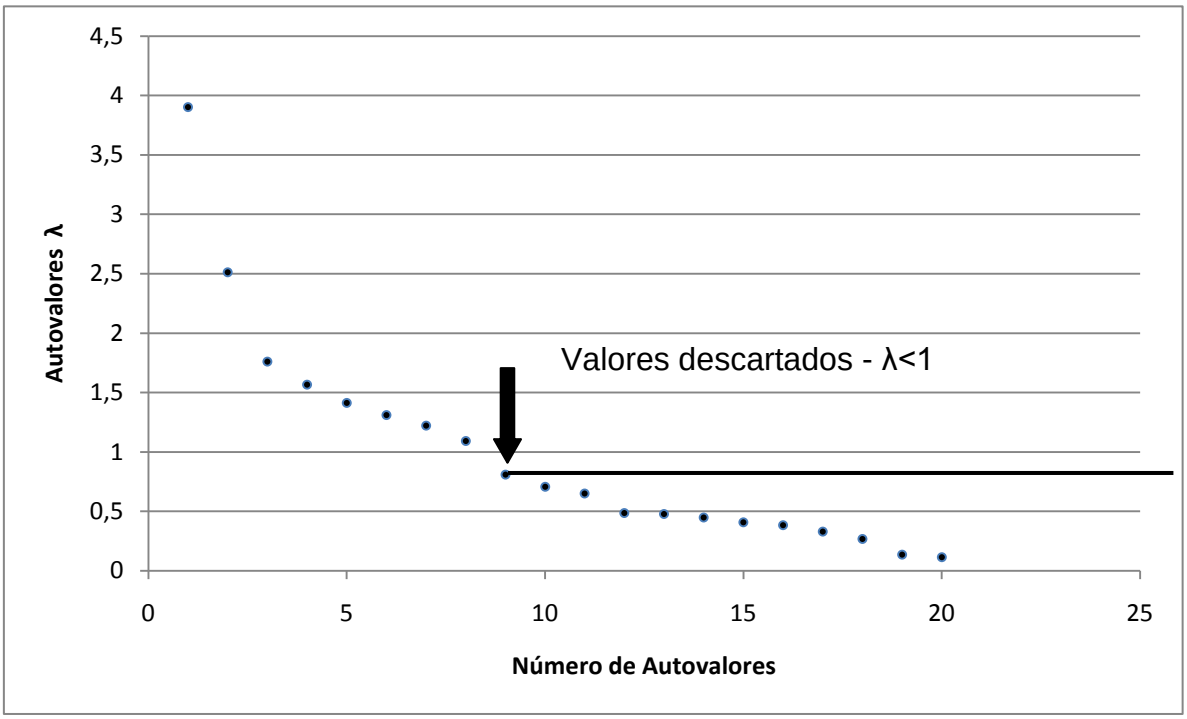
TABELA 4 - AUTO VALORES E VARIÂNCIA EXPLICADA %

Ordem	Autovalores	Variância explicada (%)	Variância explicada acumulada (%)
1	3.9017	19.51	19.51
2	2.5113	12.56	32.07
3	1.7604	8.8	40.87
4	1.5652	7.83	48.69
5	1.4132	7.07	55.76
6	1.3099	6.55	62.31
7	1.2228	6.11	68.42
8	1.0925	5.46	73.89
9	0.81	4.05	77.94
10	0.7074	3.54	81.47
11	0.6496	3.25	84.72
12	0.4866	2.43	87.15
13	0.4776	2.39	89.54
14	0.4498	2.25	91.79
15	0.4094	2.05	93.84
16	0.3834	1.92	95.75
17	0.3306	1.65	97.41
18	0.2677	1.34	98.75
19	0.1374	0.69	99.43
20	0.1135	0.57	100

FONTE: AUTOR (2010)

Para facilitar a compreensão construiu-se o gráfico 1, onde no eixo das abscissas estão representados a referencia do autovalor , de 1 a 20, e no eixo das ordenadas estão representados os valores dos autovalores obtidos. Observa-se que somente os oito primeiros autovalores são superiores a 01, conforme estabelece o critério de Kaiser.

GRÁFICO 1 - AUTOVALORES DA MATRIZ CORRELAÇÃO



FONTE: AUTOR (2010)

As oito componentes principais determinados na (tabela 04), podem ser utilizados para avaliar os fatores que interferem no auto custo do Plano de Saúde da empresa ora estudada.

4.2 CORRELAÇÃO

A estrutura dos resultados obtidos após calculo das correlações entre as variáveis padronizadas e as componentes principais se mostrou simples de ser interpretada e bastante útil, pois as correlações dos fatores, revelaram-se em oito componentes principais como sendo correlação forte ($0{,}60 < |r| < 0{,}90$).

Foram considerados os coeficientes de correlação com valores absolutos entre (0,60; 0,90) considerados como forte correlação. Tais valores encontram-se em destaque (negrito), no quadro 04.

QUADRO 4 - CORRELAÇÃO COMPONENTES PRINCIPAIS *versus* VARIÁVEIS ORIGINAL

CORRELAÇÃO COMPONENTES PRINCIPAIS *versus* VARIÁVEIS ORIGINAL																				
var	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP8 |	CP9	CP10	CP11	CP12	CP13	CP14	CP15	CP16 |	CP17	CP18	CP19	CP20
1	**0.8583**	-0.062	0.0698	0.2773	0.0918	0.1197	0.0623	0.0021	-0.0549	-0.0161	0.0939	-0.0053	-0.0293	0.0041	0.168	0.1748	0.0434	-0.0071	-0.2778	-0.0139
2	**0.6103**	0.0053	-0.0681	-0.325	-0.094	-0.1532	-0.0894	-0.0347	0.4875	0.0456	-0.0289	0.3722	0.2785	-0.0292	-0.065	-0.1104	-0.0019	0.0056	-0.043	-0.0073
3	0.4768	-0.0779	0.1333	-0.399	-0.0492	-0.1076	-0.1075	0.1043	-0.1386	**0.6901**	-0.1909	-0.0979	-0.1072	0.0038	-0.001	0.014	-0.0037	-0.0032	0.0013	0
4	**0.6712**	-0.1148	0.2191	0.0167	0.1095	0.2009	0.1005	-0.0007	-0.0077	0.0517	0.5102	-0.1655	-0.0158	-0.0001	-0.275	-0.2347	-0.0606	0.0054	0.02	-0.0003
5	0.4425	0.0571	-0.1409	-0.4697	-0.1693	-0.3238	-0.1838	0.0482	0.2995	-0.3193	-0.0994	-0.3773	-0.1877	0.104	-0.002	0.0049	-0.0127	0.0042	-0.0266	-0.0001
6	0.1686	0.0215	-0.0993	-0.5018	-0.157	-0.3708	-0.2281	0.1996	-0.5599	-0.1862	0.259	0.1588	0.099	-0.0296	0.05	0.039	0.0158	0.0065	0	0.0021
7	0.4416	-0.1887	0.376	-0.2826	0.0756	0.3	0.2047	-0.1005	-0.2165	-0.2693	-0.3715	0.1829	-0.2341	-0.1411	-0.156	-0.1015	-0.0085	-0.0004	-0.0118	-0.0019
8	**0.611**	-0.2162	0.3673	-0.2642	0.072	0.3146	0.2115	-0.148	0.0477	-0.0727	0.0598	-0.0584	0.1764	0.1165	0.274	0.1963	0.0301	-0.0022	0.175	0.0122
9	0.4316	-0.1938	0.2028	0.4086	0.1261	-0.1088	-0.2209	0.4453	-0.1015	-0.0987	-0.2806	-0.2039	0.344	-0.1602	-0.037	-0.0795	-0.0165	0.0003	0.0309	0.0009
10	0.5044	-0.1728	0.0906	0.4437	0.0793	-0.1979	-0.2476	0.4087	0.1342	-0.0209	0.0787	0.2471	-0.3424	0.1125	0.03	0.072	0.0279	0.0006	0.1199	0.0097
11	0.0842	**0.6515**	0.4226	0.1513	-0.5405	-0.0552	0.1154	0.036	-0.0059	-0.0007	0.0049	-0.0027	-0.0007	-0.0067	-0.005	0.0059	0.0358	-0.0235	0.0234	-0.2338
12	0.0804	**0.6816**	0.4454	0.1449	-0.4483	-0.1314	0.1793	0.0247	0.0007	0.0043	-0.0046	0.0051	0.009	0.0008	-0.009	-0.0043	0.0068	0.0407	-0.0119	0.2375
13	0.5023	-0.0763	-0.2159	0.3358	-0.0963	-0.2197	-0.0678	-0.4853	-0.2333	-0.0108	-0.1688	0.0191	0.1038	0.3817	-0.207	0.0373	-0.0059	-0.0027	0.0261	-0.0014
14	0.5141	-0.0318	-0.2706	0.2625	-0.1327	-0.2623	-0.104	-0.526	-0.0198	0.0264	0.0476	-0.0355	-0.0939	-0.4155	0.153	-0.0536	-0.0192	0.0085	0.0769	0.006
15	0.3506	0.1224	-0.5842	0.061	-0.1215	0.0779	0.4765	0.2419	-0.1046	0.0085	-0.0806	0.0083	-0.023	0.1393	0.241	-0.3367	0.037	-0.0016	0.0168	-0.0011
16	0.3451	0.1794	-0.6022	-0.0275	-0.0925	0.0674	0.4582	0.2492	0.0223	0.0137	-0.0076	-0.0136	0.0243	-0.1601	-0.254	0.3241	-0.0275	-0.0068	0.0497	0.0036
17	0.216	0.5851	-0.2212	-0.0113	0.0002	0.4687	-0.4134	0.0237	-0.0474	-0.0135	-0.0406	0.0352	-0.0092	0.0362	0.063	0.013	-0.382	0.1182	0.0105	-0.0034
18	0.2097	0.5576	-0.2542	-0.0473	0.0549	0.4649	-0.4475	-0.0258	-0.0251	-0.0061	-0.0102	-0.022	0.0025	-0.0234	-0.054	-0.0217	0.3645	-0.137	0.0265	0.0276
19	0.0985	**0.6029**	0.0579	-0.0532	**0.6359**	-0.2484	0.1237	-0.0544	-0.0113	0.0037	-0.0173	-0.0074	-0.0003	-0.001	-0.01	-0.0041	0.1374	0.3398	0.0147	-0.033
20	0.0963	**0.6124**	0.1101	-0.0459	0.5869	-0.3038	0.1503	-0.0453	-0.0123	-0.009	-0.0139	0.0169	-0.0025	0.0102	0.033	-0.008	-0.1451	-0.342	0.0095	-0.0004

FONTE : AUTOR (2010)

A análise é feita verificando-se o grau de influência que cada variável original X_j tem sobre o componente calculadas Y_i. O grau de influência é dado pela correlação entre cada X_j e o componente Yi que está sendo interpretado,essa análise das correlações das componentes principais versus variáveis original, fornece uma interpretação específica para cada componente.

Inicialmente neste caso se destaca que há prevalência de uma forte correlação principalmente nas duas primeiras componentes que juntas representam 32% de variação total, entre todas as variáveis.

A primeira componente agrupa as variáveis X_1, X_2, X_4 e X_8 explicando simultaneamente o número de dependentes do titular do plano, o tempo de serviço na empresa do titular do plano, o sexo do titular do plano e a idade do cônjuge, estes representam 19,51% das variáveis.

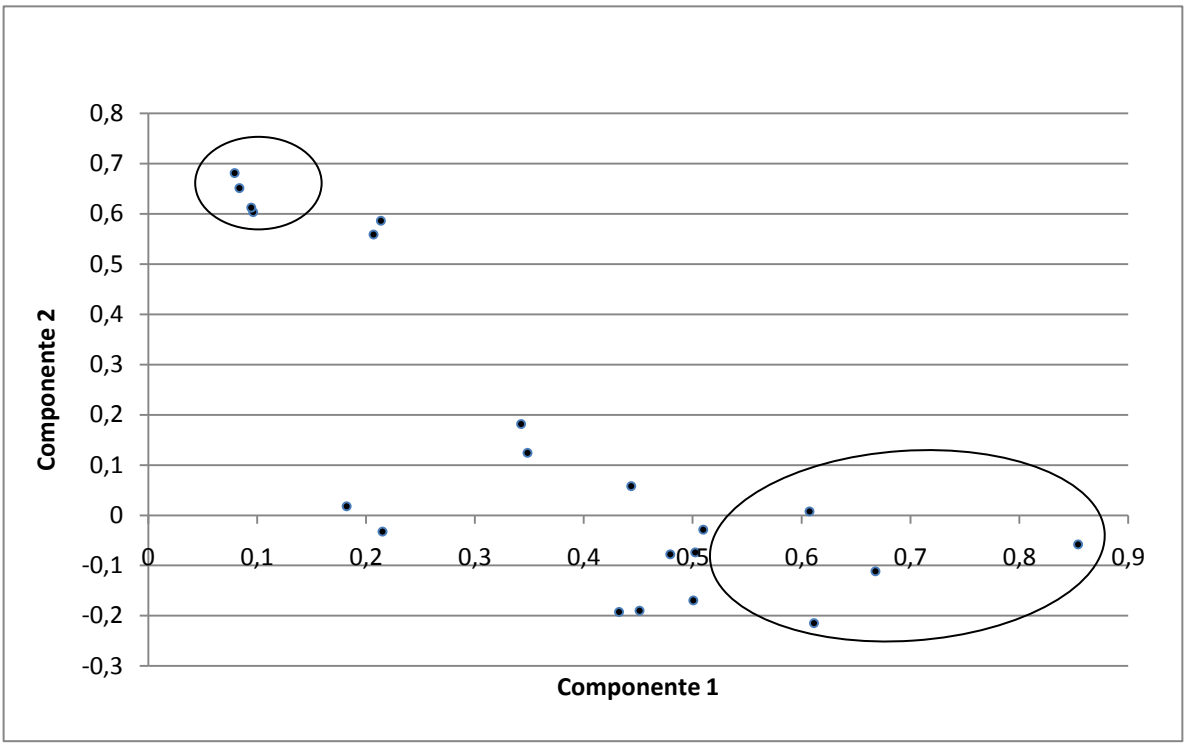
De um modo geral a primeira componente possui uma alta correlação focado diretamente ao titular do plano. Seja visto que a maior correlação existente prevalece da quantidade de dependentes, isso demonstra que quanto melhor o tipo de plano dentro da hierarquia estabelecida pela empresa, influencia no seu custo.

A segunda componente agrupa as variáveis X_{11}, X_{12}, X_{19} e X_{20} explicando simultaneamente o quantidade de procedimentos realizados pelos tutelado em 2008, idade do paciente tutelado até 31/12/08, a quantidade de procedimentos realizados pela sogra em 2008 e a idade do paciente sogra até 31/12/2008 e estes itens, representam 12.56%.

Já á segunda componente com uma alta correlação tem seu grupo direcionado aos dependentes do plano, confirmando as condições já citadas acima pois devido a hierarquia do plano de saúde dentro da empresa, logo são poucos colaboradores que podem ter como dependentes sogro e sogra, e são esses predominantes na segunda componente.

De um modo geral todas as variáveis possuem uma contribuição de correlação que varia de moderada a forte, ou seja uma maior participação e outras menos. Para uma melhor visualização das análises calculadas já comentadas acima pode ser verificada conforme o gráfico que explicita as duas primeiras componentes principais junto a correlação componentes principais versus variáveis original onde temos componente 1 versus componente 2.

GRÁFICO 2 - CORRELAÇÃO COMPONENTE 1 *VERSUS* COMPONENTE 2



FONTE: AUTOR (2010)

Visualizando o gráfico fica visível a distinção entre as duas componentes principais pois apresentam grupos distintos, no primeiro grupo a direita obtemos a primeira componente que esta mais voltado ao titular do plano, e no segundo grupo a esquerda estão outros voltados para os dependentes.

4.3 ESCORES DA COMPONENTES PRINCIPAIS

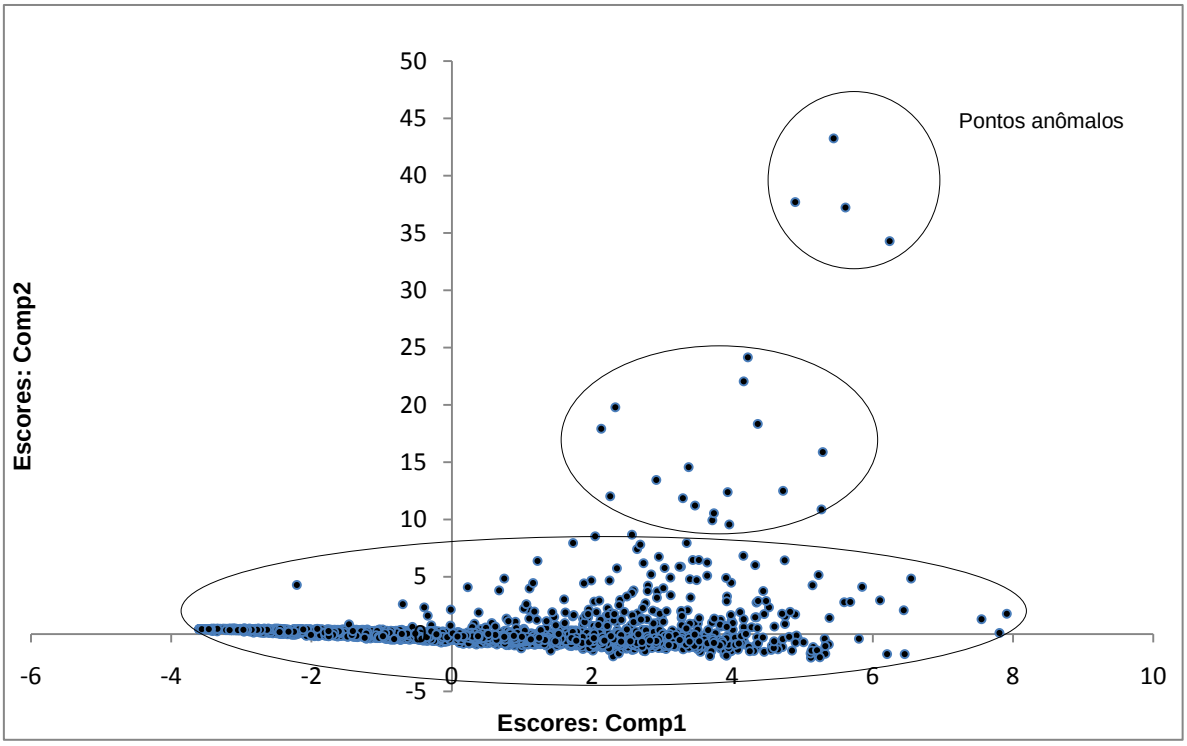
Da mesma forma que as amostras têm coordenadas definidas pelas variáveis originais, elas também possuem coordenadas relativas aos novos eixos e são denominadas escores (do inglês*: scores*). A contribuição que cada variável original exerce sobre uma determinada componente é denominada peso (do inglês*: loading*) que, matematicamente, pode ser definida como sendo o cosseno do ângulo entre o eixo da variável original e o eixo da componente principal.

Os escores são os valores dos componentes principais. Após a redução de p para k dimensões, os k componentes principais serão os novos indivíduos e toda análise desta pesquisa é feita utilizando-se dos escores desses componentes. Os gráficos (03) e (04) são a representação dos escores das componentes principais padronizados.

Esses dados são a matriz de carregamentos de cada variável nas componentes principais ao ser multiplicada pela matriz original de dados e fornecerá a matriz de contagens (escores) de cada caso em relação às componentes principais.

Esses valores estão dispostos num diagrama de dispersão, onde os eixos são as duas componentes mais importantes e assim mostrar o relacionamento entre os casos condicionados pelas variáveis medidas. Elaborado gráfico da primeira componente1 *versus* componentes 2.

GRÁFICO 3 - DISPERSÃO DOS ESCORES COMPONENTES 1 VERSUS COMPONENTE 2

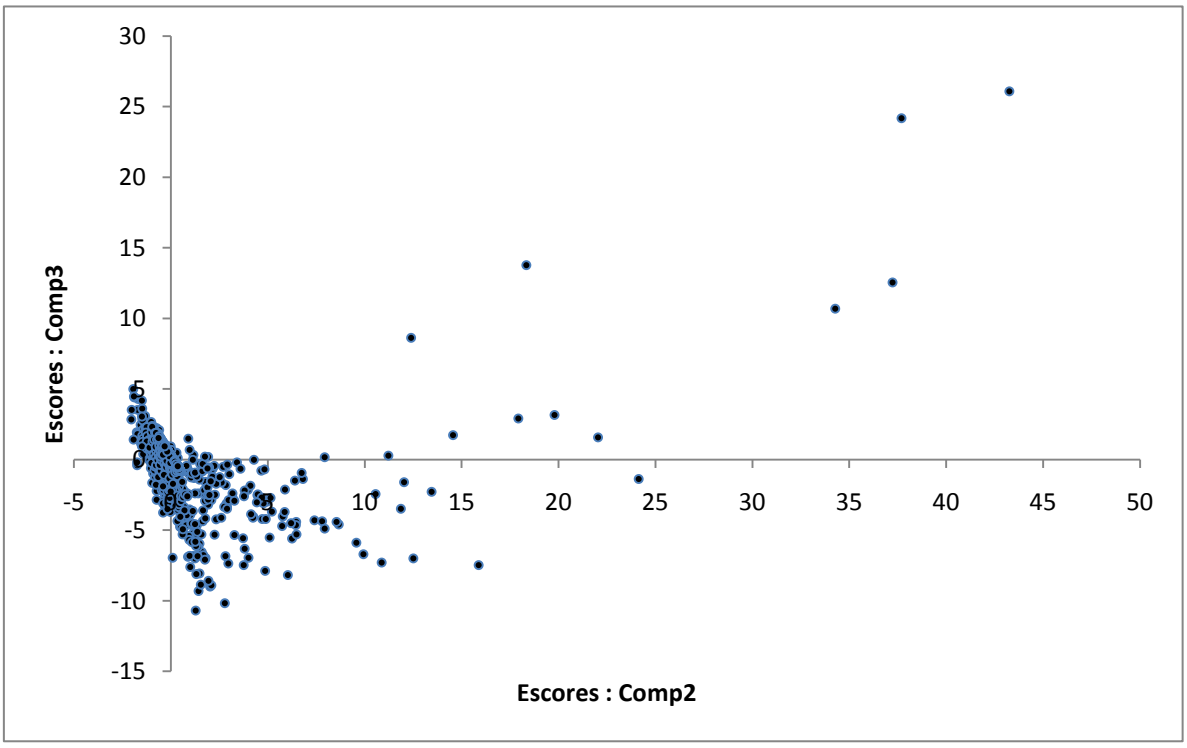


FONTE: AUTOR (2010)

Os escores representam a combinação da dispersão espacial dos dados originais em cada tempo, sendo não correlacionados entre si. Neste gráfico, alguns pontos apresentam uma dispersão muito anômala (*outliers*) dos dados da matriz dos escores.

Para uma melhor análise temos abaixo outro gráfico agora sendo os escores da componente 2 versus componente 3.

GRÁFICO 4 - DISPERSÃO DOS ESCORES FATORIAIS COMPONENTE 2 VERSUS COMPONENTE 3



FONTE: AUTOR (2010)

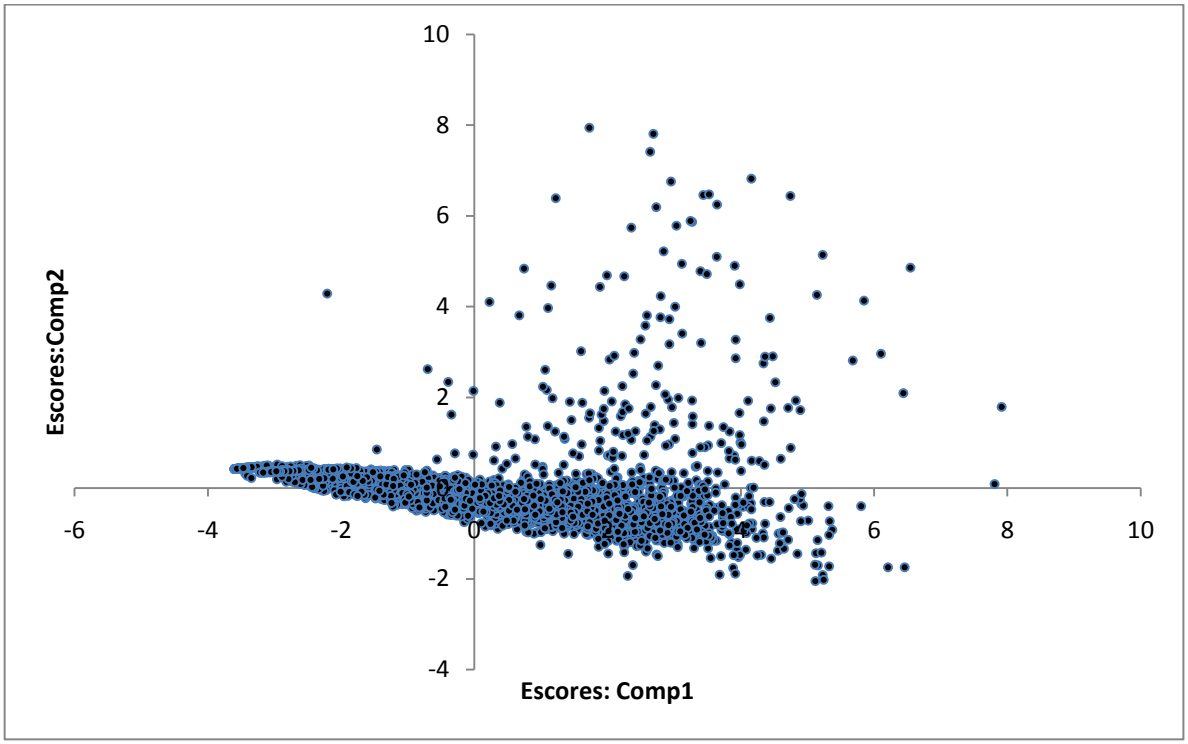
Neste segundo gráfico também temos a presença de pontos bem dispersos representando pontos anômalos.

Inicialmente os pontos anômalos são identificados pois estes tendem a prejudicar no resultado da análise de regressão, essas anomalias representam, em geral, dados irrelevantes, erros grosseiros quando comparados com a maioria dos dados.

Para o caso em estudo utilizou-se o método Z-escores em virtude da base de dados ser superior a mil variáveis, este método permite considerar como normais

valores menores de ±3, fato da grande quantidade de informações, e uma análise de observação gráfica considera-se não *outliers* quando as variâncias apresentarem uma variação, tanto positiva quanto negativa de oito, e decidiu-se que valores com variância acima de ±8 foram excluídos da análise, onde identificou-se 24 pontos na base original dos dados.

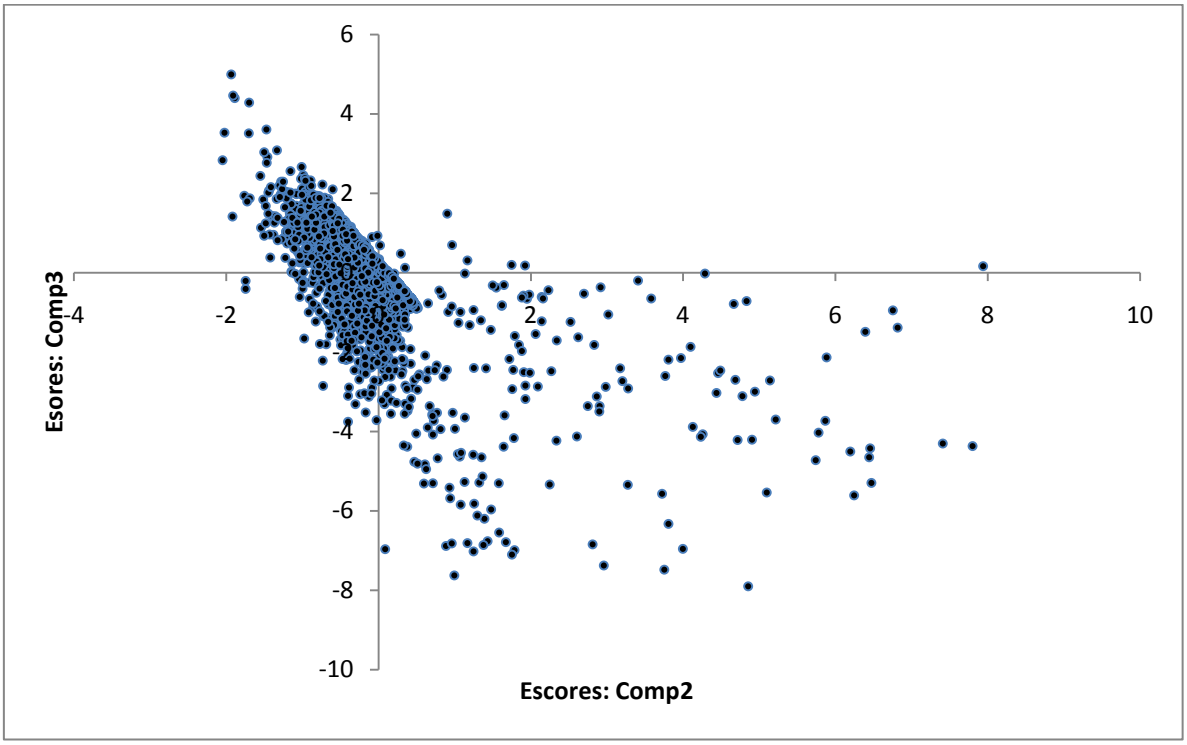
GRÁFICO 5 - DISPERSÃO DOS ESCORES FATORIAIS COMPONENTE 1VERSUS COMPONENTE2 SEM OS PONTOS ANÔMALOS



.

FONTE: AUTOR (2010)

GRÁFICO 6 - DISPERSÃO DOS ESCORES FATORIAIS COMPONENTE 2 VERSUS COMPONENTE3 SEM OS PONTOS ANÔMALOS



FONTE: AUTOR (2010)

Uma análise mais refinada desses pontos anômalos, passível de identificar quais informações fazem os mesmos obterem uma variação tão diferenciada dos outros valores. Expresso em percentis tabela (05) do número de dependentes representados pelos pontos anômalos diante dos dados originais.

TABELA 5 - PERCENTAGEM: NÚMERO DE DEPENDENTES DOS PONTOS ANÔMALOS / VARIÁVEIS ORIGINAIS

Dependentes	Nº dependentes Pontos Anômalos	Nº dependentes Variáveis Original	Representação em %
cônjuge	5	3061	0.16
filho	3	2182	0.14
tutelado	6	1	600
mãe	6	1464	0.41
pai	6	461	1.30
sogro	19	130	14.62
sogra	15	17	88.24

FONTE: AUTOR (2010)

Visto a grande diferença em porcentagem entre a base de dados originais e os pontos anômalos, destaca-se os tutelados que praticamente em todas as suas participações ocorrem erros grosseiros, quais apresentam uma média de 27 procedimentos anuais e 20,5 anos. Observado também sua grande influência no agrupamento entre os maiores valores dos escores disposto no gráfico (03).

Decorrente da situação analisada acima as variáveis que envolve dependentes tutelado não podem fazer parte deste estudo por apresentarem muita incoerência em relação as outras variáveis da base de dados.

Implicando ao pesquisador uma nova análise das Componentes Principais, envolvendo agora somente 18 variáveis, excluindo as variáveis X_{11} (Quantidade de procedimentos realizados pelo tutelado em 2008); X_{12} (Idade do paciente tutelado até 31 / 12 /2008), obtemos uma nova situação. A tabela (06) abaixo mostra os autovalores e a porcentagem variância explicada e acumulada.

TABELA 6 - AUTO VALORES E VARIÂNCIA EXPLICADA % 18 VARIÁVEIS

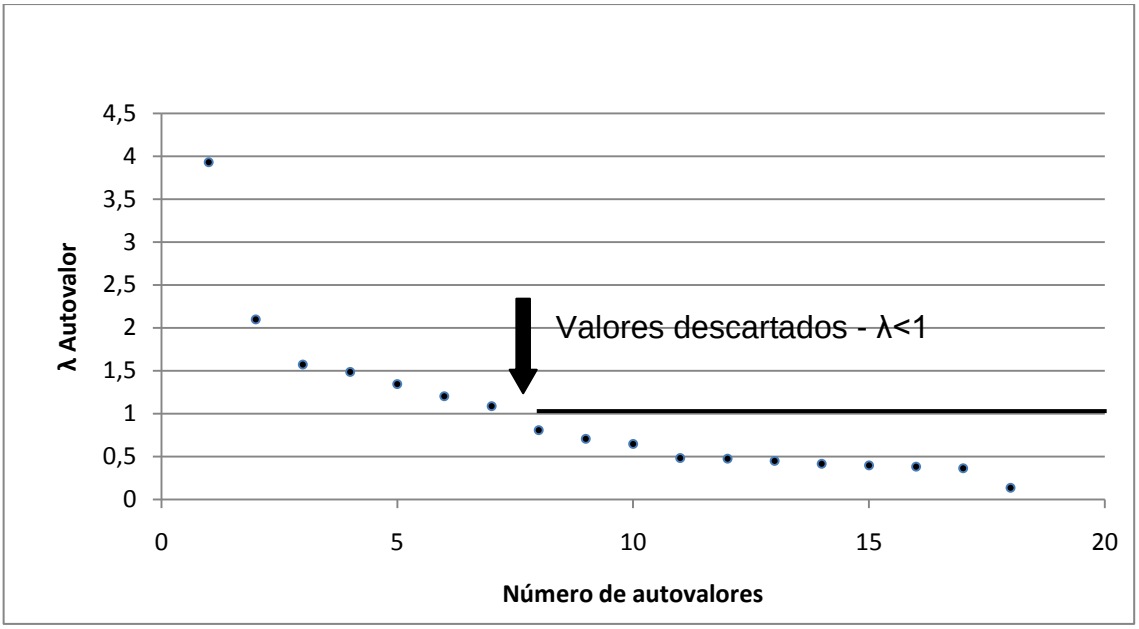
Ordem	λ Autovalores	Variância explicada (%)	Variância explicada acumulada (%)
1	3.9304	21.84	21.84
2	2.1007	11.67	33.51
3	1.5734	8.74	42.25
4	1.4859	8.25	50.5
5	1.3469	7.48	57.98
6	1.2034	6.69	64.67
7	1.088	6.04	70.71
8	0.8085	4.49	75.21
9	0.7064	3.92	79.13
10	0.6493	3.61	82.74
13	0.485	2.69	85.43
14	0.477	2.65	88.08
15	0.449	2.49	90.58
16	0.417	2.32	92.89
17	0.3965	2.2	95.1
18	0.3823	2.12	97.22
19	0.3642	2.02	99.24
20	0.136	0.76	100

FONTE: AUTOR (2010)

Neste novo cenário temos uma realidade de sete variáveis significativas, conforme critério de Kaiser resultado do λ> 1, qual representa 70,71% de todas as variáveis.

Estas variáveis identificadas estão disposta no gráfico (07).

GRÁFICO 7 - VALORES DA MATRIZ CORRELAÇÃO 18 VARIÁVEIS



FONTE: AUTOR (2010)

Destaque a primeira componente principal, com uma representação de 21.84% de todo o modelo.

QUADRO 5 - CORRELAÇÃO COMPONENTES PRINCIPAIS *versus* VARIÁVEIS ORIGINAL 18 VARIÁVEIS

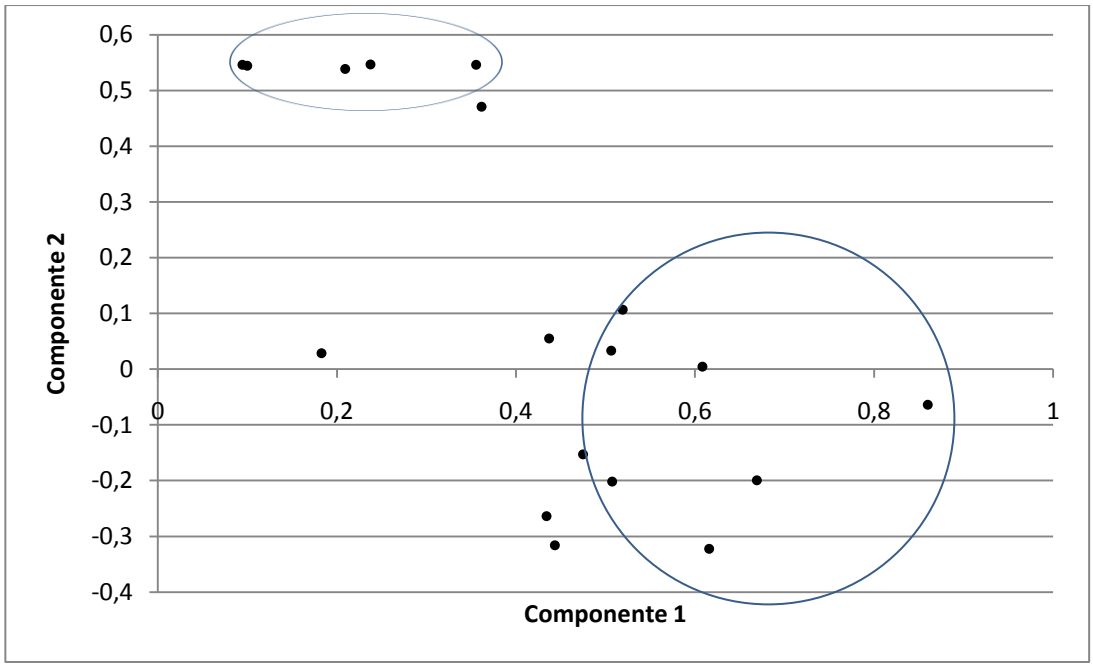
COMPONENTES PRINCIPAIS VERSUS VARIÁVEIS ORIGINAIS																		
Var	CP 1	CP 2	CP 3	CP 4	CP 5	CP 6	CP 7	CP 8	CP 9	CP 10	CP 13	CP 14	CP 15	CP 16	CP 17	CP 18	CP 19	CP 20
x1	**0.8599**	-0.0638	-0.2643	-0.1719	-0.0595	0.0727	-0.0032	0.0482	-0.0142	0.0923	0.0022	0.0299	-0.0174	-0.1436	0.003	-0.17	0.0953	0.2783
x2	**0.6083**	0.0045	0.3041	0.1976	0.0829	-0.1025	0.0425	-0.4941	0.0654	-0.048	-0.3067	-0.3373	0.0375	0.0788	0.046	0.0769	-0.0566	0.0432
x3	0.4748	-0.1532	0.405	0.0579	0.0406	-0.1452	-0.0946	0.1632	**0.6782**	-0.2033	0.0827	0.13	-0.007	-0.0017	-0.01	-0.0012	0.0093	-0.0014
x4	**0.669**	-0.1995	0.0081	-0.2358	-0.1083	0.135	-0.0089	0.0018	0.0611	**0.516**	0.1498	0.0414	0.0182	0.2168	-0.039	0.263	-0.1194	-0.021
x5	0.4371	0.0549	0.4438	0.3553	0.1905	-0.2397	-0.033	-0.2976	-0.3224	-0.0892	0.3295	0.2529	-0.1096	-0.0234	-0.03	0.0288	-0.0008	0.0254
x6	0.1828	0.0287	0.4842	0.3333	0.1979	-0.316	-0.1763	**0.5451**	-0.189	0.2652	-0.1473	-0.1373	0.0296	-0.0423	0.007	-0.0594	0.0119	0.0001
x7	0.4434	-0.316	0.3218	-0.2958	-0.1935	0.2592	0.0858	0.2149	-0.2829	-0.3614	-0.2253	0.1988	0.143	0.1391	-0.014	0.105	-0.0472	0.0114
x8	**0.6158**	-0.3224	0.3026	-0.2992	-0.2008	0.2709	0.1354	-0.0578	-0.0657	0.0557	0.0958	-0.1639	-0.1244	-0.227	0.038	-0.2211	0.0964	-0.1739
x9	0.434	-0.2639	-0.3934	-0.1666	0.083	-0.2793	-0.4248	0.1048	-0.1103	-0.2781	0.2641	-0.2958	0.1667	0.0249	0.015	0.0753	-0.0335	-0.031
x10	**0.5073**	-0.2017	-0.4443	-0.0452	0.1418	-0.3091	-0.3873	-0.1353	-0.0147	0.077	-0.305	0.2876	-0.1174	-0.0066	-0.007	-0.0902	0.0207	-0.1198
x13	**0.5063**	0.0329	-0.3713	0.2438	0.1496	-0.0472	0.486	0.2387	-0.0189	-0.1612	0.0017	-0.1104	-0.3756	0.1817	-0.108	0.0446	0.0231	-0.0269
x14	**0.5191**	0.1065	-0.3004	0.3048	0.151	-0.1044	**0.5321**	0.0223	0.0271	0.0443	0.0148	0.1041	0.412	-0.1469	0.083	-0.0073	-0.0301	-0.0759
x15	0.3616	0.4709	-0.1125	0.2803	0.0866	**0.5161**	-0.2805	0.1058	0.0069	-0.0725	-0.0034	0.0194	-0.1269	-0.1286	0.305	0.0787	-0.2338	-0.0157
x16	0.3555	**0.5458**	-0.0155	0.2263	0.1487	0.4762	-0.2708	-0.0229	0.0157	-0.0078	0.0166	-0.0192	0.1592	0.1577	-0.304	-0.1031	0.2122	-0.0464
x17	0.2373	**0.5467**	-0.0272	-0.0005	**-0.6194**	-0.2337	-0.045	0.0362	-0.0103	-0.0394	-0.0654	-0.0064	-0.0521	-0.2723	-0.159	0.2892	0.1022	-0.0205
x18	0.2093	**0.5385**	0.0164	-0.0096	**-0.6221**	-0.2845	0.014	0.0056	-0.0138	-0.0043	0.0611	0.0117	0.0333	0.2904	0.162	-0.272	-0.0872	-0.0182
x19	0.0999	**0.5443**	0.1151	**-0.5483**	0.409	-0.1257	0.0786	0.015	-0.01	-0.0015	-0.0001	0.0165	-0.0185	0.0751	0.277	0.1195	0.3048	-0.0143
x20	0.0944	**0.5459**	0.1138	**-0.5757**	0.3608	-0.1386	0.0957	0.0045	-0.0003	-0.0279	-0.0149	-0.0116	-0.0114	-0.0957	-0.241	-0.0977	-0.335	-0.0047

FONTE: AUTOR (2010)

Baseado na tabela de correlação temos uma situação de análise mais realista do que no quadro 04, todas as principais informações estão concentradas na 1ª componente com uma correlação de média para alta representando as variáveis mais influentes com um índice de 21.84%, sendo elas: X_1 (Nº de dependentes do Titular do Plano de Saúde); que se destaca com uma variância de 0,86 de índice de correlação; X_2 (Tempo de serviço na empresa do titular), com um índice de correlação de 0,61; X_4 (Sexo do titular do plano); X_8 (Idade do paciente cônjuge até 31/ 12/ 2008); X_{10} (Idade do Paciente filho(a) até 31/ 12/ 2008); X_{13} (Quantidade de procedimentos realizados pela mãe do titular em 2008); X_{14} (Idade do paciente mãe até 31/ 12/ 2008).

A segunda componente com uma representação de 11,67% especifica um grupo distinto de dependentes com idades avançadas.

GRÁFICO 8 - CORRELAÇÃO COMPONENTE 1 *VERSUS* COMPONENTE 2 18 VARIÁVEIS



FONTE:AUTOR (2010)

Dentro de uma condição mais realista a representação da primeira componente esta correlacionada com os dados pessoais dos titulares do plano de saúde,com destaque as variáveis mãe do titular, expressando a formação de um grupo próximo ao eixo das abscissas, influenciado pelo grande número de

dependentes, privilégio de colaboradores com benefícios oferecidos pela empresa para lideres, gestores, diretores.

A segunda componente com uma representação 11,7%, vem como uma complementação da primeira demonstrando um grupo seleto de dependentes, que só podem ter como titulares gestores e diretores.

A regressão linear múltipla produz uma análise mais significativa dos dados, para fins deste trabalho não será utilizada as variáveis originais mas sim a sua variância. Para a análise da regressão utiliza-se a matriz do escores das componentes principais que possuem uma maior representação no índice de correlação junto as variáveis selecionadas através da primeira componente, pois esses, como visto em capítulos anteriores tendem a possuir pouca multicolinearidade.

Da decomposição da matriz dos dados X, produzi-se uma matriz de scores , que será utilizada na determinação dos coeficientes de regressão (preditores), ou seja, construir uma equação para representar Sinistro R$ total gasto plano de saúde em 2008.

Neste estudo o sinistro Y representa a variável dependente que é mensurada em função das variáveis descritas na tabela (06) (variáveis independentes).

O interesse é o Y total gasto plano de saúde em 2008, será a nossa variável dependente Y e as outras sete variáveis como sendo a nossas variáveis independentes X, ou seja uma função :

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_{2i} +, \dots, +\beta_k X_k + \varepsilon_i$$

que representa o valor de Y do plano de saúde das variáveis independentes, ou seja os pesos baseado nas informações dos colaboradores, com que cada um representa no consumo do plano de saúde.

Diante da decomposição da matriz X, é necessário uma decomposição da variável Y, já que as unidades de medidas não são as mesmas, usualmente em estatística a padronização e feita através de média zero e variância 1.

Para isso usamos o softwer *MINITAB* para a o cálculo da análise de regressão múltipla como segue na tabela (07).

TABELA 7 - ESTATÍSTICA DE REGRESSÃO DA VARIAVEL Y PADRONIZADA (MÉDIA ZERO E VARIÂNCIA 1)

Estatística de regressão	
R-Quadrado	0,075
R-quadrado ajustado	0.073
Erro padrão	0,96
Observações	4984

FONTE: AUTOR (2010)

QUADRO 6 - ANOVA VARIÁVEL Y PADRONIZADA (MÉDIA ZERO VARIÂNCIA 1)

Fonte de variação	Graus de liberdade	Soma dos Quadrados dos erros	Quadrado médio dos erros	Estatistica - F	P- Valor
Modelo	7	371.582	53.083	57.28	0
Erro	4977	4612.418	0.927		
Total	4984	4984			

FONTE:AUTOR (2010)

Pelo p-valor todas as variáveis possuem contribuição no modelo ,não podendo ser descartadas nenhuma delas,infelizmente o modelo em geral explica somente 7,5% de todas as variáveis, revelando que a padronização com média zero e variância um não é consistente para o modelo.

Torna-se muito difícil a padronização de variáveis para que se ajustem ao modelo e apresentem uma boa representação do mesmo, para finalidade deste de obter uma variável com variância mais homogênea é utilizada a transformação logarítmica,pois quando a variável resposta é positiva, esta transformação também modifica o seu intervalo de variação, para o qual é mais coerente com a definição da distribuição normal estas estão apresentadas no quadro (07) abaixo.

QUADRO 7 - CÁLCULO TABELA ANOVA DA VARIÁVEL PADRONIZADA LN(Y)

Fonte de variação	Graus de liberdade	Soma dos Quadrados dos erros	Quadrado médio dos erros	Estatistica - F	P- Valor
Modelo	7	5108.46	729.78	1014.38	0
Erro	4977	3580.61	0.72		
Total	4984	8689.07			

FONTE: AUTOR (2010)

O p-valor do teste quadro (07), indica que não existe variável explicativa que apresenta efeito significativamente diferente de zero.

TABELA 8 - ESTATÍSTICA DE REGRESÃO DA VARIÁVEL PADRONIZADA LN(Y)

Estatística de regressão	
R-Quadrado	0,586
R-quadrado ajustado	0,587
Erro padrão	0.8482
Observações	4984

FONTE: AUTOR (2010)

QUADRO 8 - COEFICIENTES DA REGRESÃO LINEAR MÚLTIPLA

Coeficientes de Regressão múltipla				
PREDITORES	Estimativas de β_j	Estimativas Desvio-Padrão	Estatistica Teste-T	Teste-P P-Valor
Constante	6.9254	0.01201	576.47	0
CP_1	0.4377	0.006060	72.22	0
CP_2	-0.1285	0.008290	-15.50	0
CP_3	-00.472	0.009856	4.78	0
CP_4	0.5322	0.01336	39.82	0
CP_5	-0.0554	0.01491	-3.71	0
CP_6	-0.578	0.01725	-3.35	0.001
CP_7	-0.03494	0.0174	-2.01	0.045

FONTE: AUTOR (2010)

Após a análise exploratória dos dados, obtem-se a equação para as variáveis do custo de um plano de saúde:

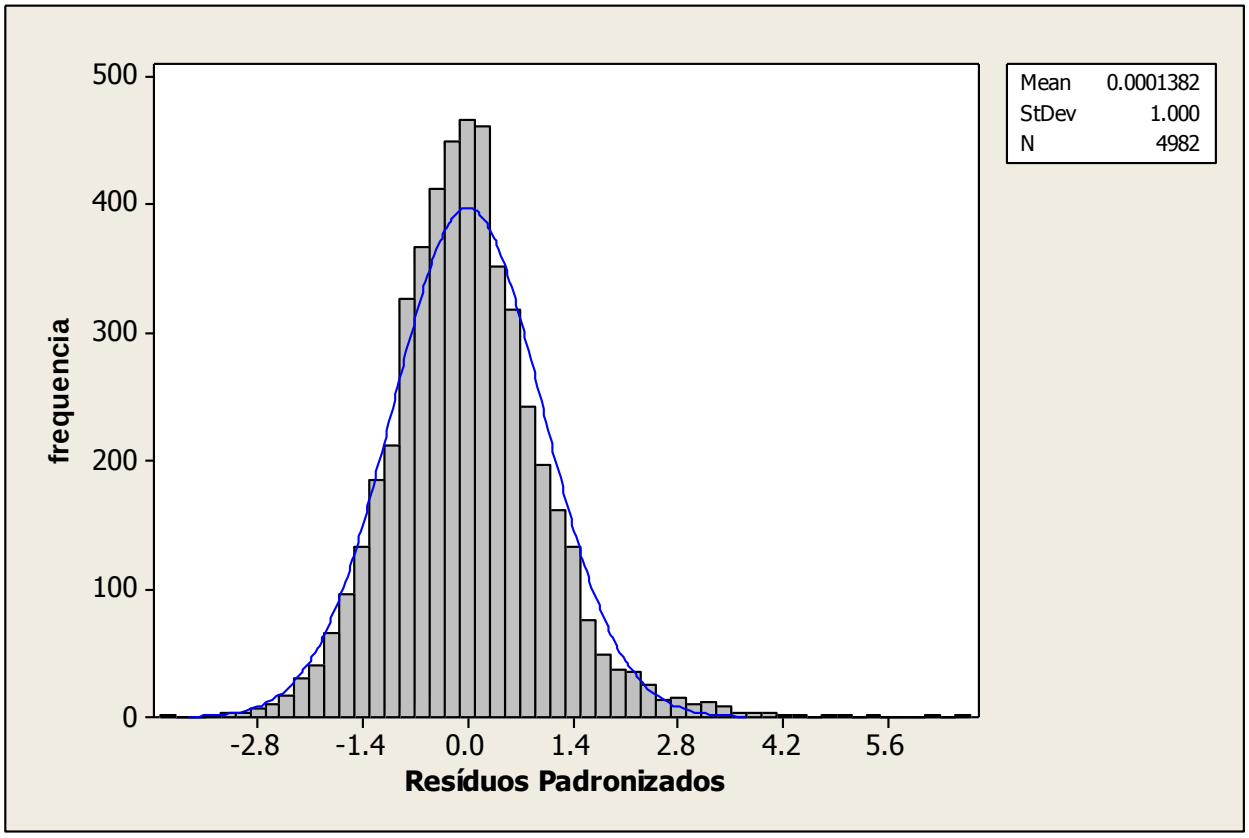
$$LN\ Y= 6.93+ 0.438CP_1 - 0.128CP_2 +0.0471CPX_3 + 0.532CP_4 – 0.0553CP_5 - 0.0578CP_6 – 0.0349CP_7$$

O coeficiente de determinação e predição encontrado foi $R^2 = 0,588$ conforme a tabela (08) obtemos um modelo ajustado pois explica 58,8% do relacionamento entre as sete componentes principais que influenciam no gasto do plano de saúde(Sinistro R$ do plano de saúde).

O valor de $F_{critico}$ (5%;7;4977)=2.012< $F_{calculado}$ =1014.38 descarta, que todas as variáveis são iguais, e possuem representação no modelo em questão. Como todas as variáveis análise teste P estão dentro da normalidade de 95%, podemos atribuir que todas as variáveis independentes tem sua contribuição no custo do plano de saúde. A partir da análise do p-valor do teste Tabela (08), verifica-se que todos os parâmetros do modelo são significativamente diferentes de zero ao nível de significância de 5%.

A seguir é avaliado a adequação do modelo ajustado, através de métodos gráficos, onde os resíduos do modelo desempenham papel fundamental. Objetivando também detectar a presença de *outliers*, utilizando-se alternativamente os chamados resíduos padronizados gráfico (09) que tem média zero e variância unitária MONTGOMERY(2003).

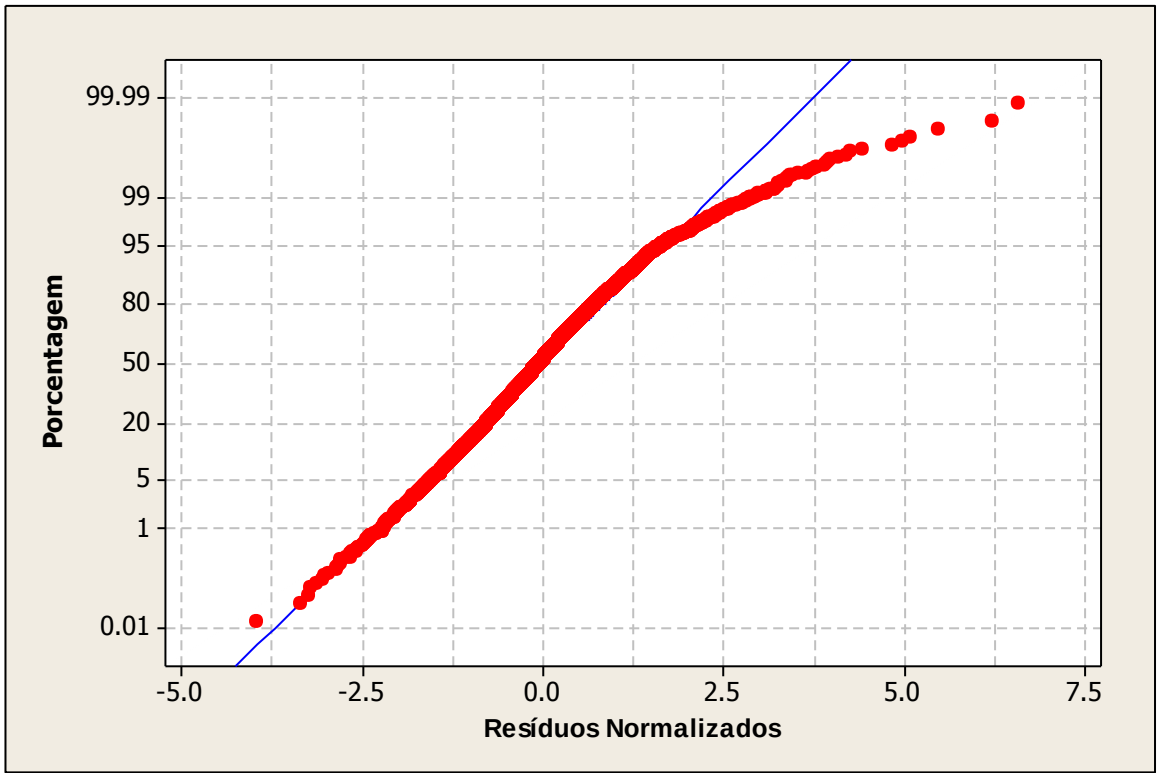
GRÁFICO 9 – HISTOGRAMA DOS RESIDUOS PADRONIZADOS



FONTE : AUTOR (2010)

A figura mostra um bom ajuste para a equação proposta que desta forma pode prever dentro das condições estudadas, com uma boa aproximação, as variáveis que influenciam no sinistro $ plano de saúde.

GRÁFICO 10 - RESÍDUOS COM DISTRIBUIÇÃO NORMAL

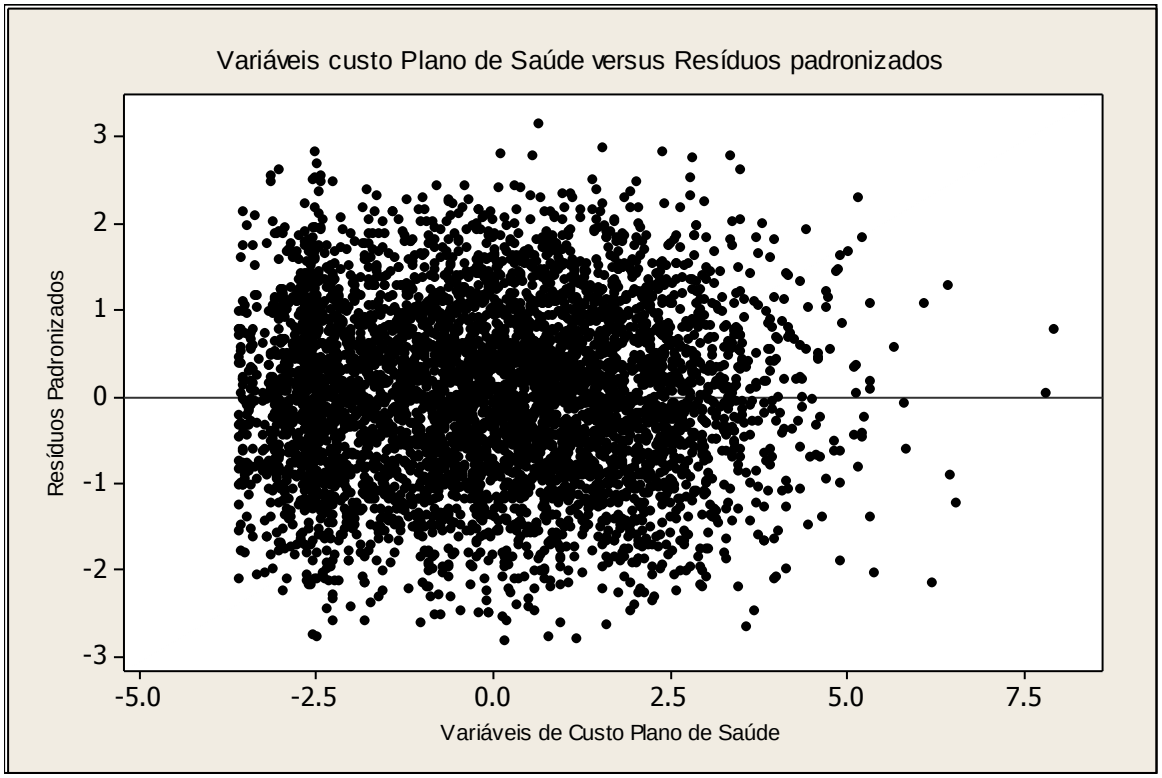


FONTE : AUTOR (2010)

O histograma e o Gráfico da distribuição normal dos resíduos gráficos (9) e (10), que mostra a proximidade dos resíduos em relação à reta esperada, indicam que os resíduos padronizados seguem uma distribuição aproximadamente normal. Portanto, a hipótese de normalidade dos erros pode ser considerada satisfeita.

Uma outra forma de verificar se o modelo acima e realmente bem ajustado,e de acordo com os pressupostos da regressão, os resíduos devem distribuir-se aleatoriamente em torno de 0 (zero), tanto no modelo global como em relação a cada variável. Caso tal não se verifique, será normalmente necessário alterar o modelo, incluindo ou retirando variáveis, ou realizando alguma transformação que adéqüe melhor o modelo aos dados, verificar análise gráfico de resíduos.

GRÁFICO 11 - VARIÁVEIS ORIGINAIS VERSUS RESÍDUOS PADRONIZADOS



FONTE : AUTOR (2010)

Através da análise do gráfico de dispersão entre os resíduos padronizados e as variáveis de custo plano de saúde gráfico (11), percebe-se que os resíduos do modelo se encontram aleatoriamente distribuídos em torno de zero, não apresentando pontos anômalos, indicando que a hipótese de independência e homocedasticidade dos erros não é violada.

5. CONSIDERAÇÕES FINAIS

5.1 CONCLUSÕES

A Análise das Componentes Principais (ACP) se mostrou bastante útil no sentido da sumarização dos dados e conseqüentemente na redução de variáveis, que podem ser utilizadas para o estudo de gastos de um plano de saúde, discriminando satisfatoriamente estes fatores em diferentes dimensões. Demonstrando mais uma vez a importância da análise de componentes principais (PCA) no uso em grande quantidade de dados.Neste trabalho uma aplicação prática de um estudo de caso, para a identificação dos fatores que elevam o custo de um plano de saúde empresarial.

Com aplicação da (ACP) obteve-se uma redução nas 18 variáveis iniciais para somente (07) sete componentes principais, sem perda significativa de informação, o resultado foi satisfatório já que as mesmas representaram mais de 70% das informações contidas na matriz de dados. Isto foi feito seguindo critério de Kaiser com escolha dos autovalores maiores que um ($\lambda>1$).

Tão importante afirmação que na análise de correlação variáveis originais versus componentes principais, destaque a primeira componente l com uma representação de 19,51%, evidenciando as variáveis com alto índice de correlação > 0.6, sendo: Nº de dependentes do titular do plano de saúde; Tempo de serviço na empresa do titular do plano; Sexo do titular do plano; Idade do paciente cônjuge até 31/12/2008; Idade do paciente filho(a) até 31/12/2008, em torno deste primeiro eixo nota-se a formação de um grupo mais generalizado, pois as informações estão mais centralizadas diretamente ao titular do plano de saúde e não a seus dependentes. Com destaque ao número de dependentes do titular do plano de saúde, devido hierarquia do RH, tem-ser um seleto grupo de titulares portadores dos planos, Q440 – Diretores e Q990 – Gestores que somente estes podem estender os direitos a todos os seus familiares.

Outra contribuição importante no emprego da análise por componentes principais (ACP), é o cálculo dos escores, estes que carregam o peso individual das informações de cada colaborador e sua participação no sinistro $ do plano de saúde, com isso evidenciou-se algumas amostras anômalas presentes no conjunto de

dados que influenciaram negativamente no primeiro cálculo das Componentes principais, que verificada através de análise gráfica dos escores da ACP, observou-se valores aberrantes nas variáveis que envolviam os tutelados grupo de dependentes do titular do plano de saúde, havendo a necessidade de se excluir essas variáveis do trabalho para melhor confiabilidade na pesquisa. Foi extremamente importante a eliminação dos 24 pontos (anômalos) e duas variáveis, sendo: Quantidade de procedimentos realizados pelo tutelado em 2008 e idade do paciente tutelado até 31/12/2008, que nesse estudo representam informações pessoais dos colaboradores em relação ao sinistro $ do plano de saúde.

O ajuste do modelo de regressão linear múltipla aconteceu a partir do uso dos escores das componentes principais (ACP), que demonstra que essas (07) sete componentes principais, estão significativamente relacionadas com o auto valor anual do sinistro $ do plano de saúde, com uma representação de 59% entre todas as variáveis.

Pode a empresa estudada ter uma previsão de quanto cada colaborador custa para a empresa através de informações reduzidas e de seus dependentes, permitindo a mesma proporcionar políticas de intervenção junto a esses em relação ao alto custo do plano de saúde e seu aumento.

5.2 SUGESTÃO PARA FUTURAS PESQUISAS

Utilização destas poderosas metodologias de análise multivariada, na identificação dos principais tipos de ocorrências de saúde dos colaboradores, que elevam o custo do plano de saúde de uma empresa.

REFERÊNCIAS

ANS**, Caderno de Informação da Saúde Suplementar,** ANS, março de 2006

Bahia, L. **Mudanças e padrões das relações público-privado: seguros e planos de saúde no Brasil.** Tese de Doutorado. Rio de Janeiro: Escola Nacional de Saúde,1999 Pública. Fundação Oswaldo Cruz.

Bahia,L, **Os Planos de Saúde Empresariais no Brasil: Notas para a Regulação Governamental** Instituto de Estudos em Saúde Coletiva. Universidade Federal do Rio de Janeiro. Rio de Janeiro, RJ, Brasil,2008
www.ans.gov.br/portal/upload/forum_**saude**/forum_bibliografias/abrangenciada**regulacao**/AA7.pdf
ultimo acesso 21 /10/ 2008

Empresativa **Informações sobre programas de qualidade de vida nas empresas**,Blog, Outubro 2008.
http://empresativa.blogspot.com/2008_09_01_archive.html
Ultimo acesso 23/10/2008

FERREIRA, D, F, **ANÁLISE MULTIVARIADA_**LAVRAS, MG, 1996
Disponível em: http://www.dex.ufla.br/~danielff/dex522.pdf ultimo acesso em 20/05/2008.

FONSECA, J.S.; MARTINS, G. **Curso de estatística**. 6. ed. São Paulo, Atlas, 1999.

HAIR JR., J. F. et al. Análise Multivariada de Dados.5a edição. Porto Alegre: Bookman, 2005.

(IBGE) INSTITUTO BRASILEIRO DE PESQUISA **Pesquisa Nacional por Amostra de Domicílio**
http://www.ibge.gov.br/home/estatistica/populacao/trabalhoerendimento/pnad2003/saude/saude2003.pdf ultimo acesso em 20/05/2009.

JONHSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis.** 4. ed. New Jersey: Prentice-Hall, inc., 1998.

MARQUES, J, Mendes **Análise Multivariada aplicada à pesquisa:** Notas de aula.

Departamento de Estatística, Universidade Federal do Paraná, Curitiba,2006.

MCKINSEY, Empresa de Consultoria, **JORNAL FORLUZ ESPECIAL,**Belo Horizonte ,MG, setembro,2008.
http://www.forluz.org.br/calandra/filesmng.nsf/3E1A2D3496FD8AD3032574C6006997E6/$File/JFEspecial.pdf
ultimo acesso em 18/09/2009.

MEYER, L, Paul, PROBABILIDADE Aplicações à Estatística, Rio de Janeiro,R.J,LTC---- Livros Técnicos e Científicos Editora S.A. ano 2000.

Miranda,C,R. **Gerenciamento de Custos em Planos de Assistência à Saúde,** Projeto ANS/PNUD, Novembro 2003
http://www.ans.gov.br/portal/upload/biblioteca/TT_AS_20_ClaudioMiranda_GerenciamentodeCusto.pdf,Ultimo acesso 22/10/2008

MOITA, J.M, Neto. **Estatística multivariada Uma visão didática-metodológica,**2004_disponível em: criticanarede.com/cien_estatistica.html, ultimo acesso em 14/05/2008

MONTGOMERY, Runger C.GEORGE (2003). Estatística **APLICADA E PROBABILIDADES PARA ENGENHEIROS,**Rio de Janeiro,RJ,LTC,AS**.**2003, 2 rd ed.

PEREIRA, J.C. **Análise de dados qualitativos: estratégias metodológicas para as ciências da saúde, humanas e sociais**. 2 ed. São Paulo: Editora da Universidade de São Paulo, 1999. 156 p.

PONTES,A.C, Fonseca **ANÁLISE DE VARIÂNCIA MULTIVARIADA COM A UTILIZAÇÃO DETESTES NÃO-PARAMÉTRICOS E COMPONENTES PRINCIPAIS BASEADOS EM MATRIZES DE POSTOS.**
disponível em : http://www.fcav.unesp.br/RME/fasciculos/v19/A10_Artigo.pdf, ultimo acesso em 18/05/2008

R.Sampaio, E. Cataldo, R. Riquelme, **Introdução ao MATLAB_**, Rio de Janeiro, RJ, 1997, AEB—Agencia Espacial Brasileira

Revista digital você RH **Check-up econômico**,,São Paulo, Editora Abril S.A, 2008.
http://revistavocerh.abril.com.br/noticia/melhoresp/conteudo_281691.shtml
Ultimo acesso 22/10/2008

Rosenburg,C, , **O insustentável custo da saúde nas empresas**, Revista Digital Portal Exame,2005.

http://portalexame.abril.com.br/revista/exame/edicoes/0848/carreira/m0056959.html

Ultimo acesso 10/12/2008

Silva A.da Alves: **Relação Entre Operadoras de Planos De Saúde e Prestadores de serviços – Um Novo Relacionamento Estratégico,**Porto Alegre, Julho de 2003 http://www.ans.gov.br/portal/upload/biblioteca/TT_AR_6_AAlvesdaSilva_RelacaoOperadorasPlanos.pdf ultimo aceso 28/10/2009

_____(1998) Pesquisa Nacional por Amostra de Domicílios, **pesquisa suplementar Saúde na PNAD 1998,**
http://www.ibge.gov.br/home/estatistica/populacao/trabalhoerendimento/pnad98/saude/análise.shtm
Ultimo acesso 22/10/2008

____ (2003) Pesquisa nacional por amostra de domicílios: **acesso e utilização de Serviços de saúde 2003**. Rio de Janeiro: Fundação Instituto Brasileiro de Estatística; Departamento de População e Indicadores Sociais.
http://www.ibge.gov.br/home/estatistica/populacao/trabalhoerendimento/pnad2003/saude/saude2003.pdf
Ultimo acesso 22/10/2008

_____Towers Perrin (2004) **Pesquisa de Planos de Benefícios no Brasil,** Rio de Janeiro,RJ,2004.

http://www.towersperrin.com/hrservices/pt_BR/research/Pesquisa_de_Beneficios.pdf

Ultimo acesso 10/12/2008

ANEXOS

Devido ao tamanho da base de dados todos os anexos estão disponíveis em Compact Disk (CD), junto a este.