MAURÍCIO KOUBAY DO AMARAL

APLICAÇÃO DA ESTATÍSTICA MULTIVARIADA NA ANÁLISE DAS OBRIGAÇÕES CONDICIONAIS DO PROGRAMA BOLSA FAMÍLIA

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciências, Curso de Pós-graduação em Métodos Numéricos em Engenharia – Programação Matemática, Setores de Tecnologia e de Ciências Exatas, Universidade Federal do Paraná.

Orientador: Prof. Dr. Jair Mendes Marques

CURITIBA
2006

TERMO DE APROVAÇÃO

MAURÍCIO KOUBAY DO AMARAL

APLICAÇÃO DA ESTATÍSTICA MULTIVARIADA NA ANÁLISE DAS OBRIGAÇÕES CONDICIONAIS DO PROGRAMA BOLSA FAMÍLIA

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

Orientador: _____

PROF. D.Sc. Jair Mendes Marques
Programa de Pós-Graduação em Métodos Numéricos em Engenharia, UFPR.

_____

PROF. D.Sc. Jair Mendes Marques
Programa de Pós-Graduação em Métodos Numéricos em Engenharia, UFPR.

_____

PROF. Dr. Mário Romero Pelegrini de Souza
Centro Universitário – Bom Jesus.

Curitiba, 10 de julho de 2006.

DEDICATÓRIA

A Deus.

À minha Família.

A meus Amigos.

AGRADECIMENTOS

A Deus, pelo dom da vida.

Ao Professor Jair Mendes Marques, pela orientação, pelo exemplo, pela confiança e pela amizade.

Aos meus Pais: João Mendes do Amaral e Aracy Koubay do Amaral, pelos exemplos de força e bravura para enfrentar situações difíceis, e pela conduta e caráter sempre pautados na retidão.

Aos meus irmãos: Tatiana e Murilo, pelo apoio nos momentos importantes desta caminhada.

Ao meu sobrinho João Vitor, pelas alegrias e brincadeiras nas horas de descontração.

A minha cunhada Tatiely, pelos esclarecimentos dos Programas do Governo Federal e pelos empréstimos e dicas de livros da área.

Aos companheiros do mestrado, Leonardo, Marcos, Ricardo, Roberto e Ronaldo, um agradecimento especial por estarem sempre me incentivando.

A UTFPR, em especial a coordenação do curso de Sistema de Informação da Unidade de Ponta Grossa, pelo apoio e oportunidade.

Ao companheiro e amigo Geraldo Ranthum pelo apoio nos momentos difíceis desta caminhada.

A Secretaria da Criança e Ação Social do Município de Tibagi, Lilian Lorena Santos Scheraiber, pelo fornecimento do banco de dados.

SUMÁRIO

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E LISTA DE SIGLAS

ADLF	- Análise Discriminante Linear de Fisher
Bid	- Banco Interamericano de Desenvolvimento
BNDES	- Banco Nacional de Desenvolvimento Econômico e Social
CadÚnico	- Cadastramento Único
CEF	- Caixa Econômica Federal
CGI	- Conselho Gestor Interministerial
CMI	- Centro de Mídia Independente
CRD	- Esposo ou companheiro reside no domicílio
EAM	- Amamentando
ECA	- Estatuto da Criança e do Adolescente
FBP	- Famílias beneficiárias do Programa
FDLF	- Função Discriminante Linear de Fisher
FE	- Freqüenta escola
FNBP	- Famílias não-beneficiárias do Programa
CGPAN	- Coordenação Geral da Política de Alimentação e Nutrição
GI	- Grau de Instrução
IMG	- Se grávida, informar mês de gestação
LOAS	- Lei Orgânica da Assistência Social
MDS	- Ministério de Desenvolvimento Social e Combate à Fome
MRLM	- Modelo de Regressão Logístico Múltiplo
NPVR	- Número de Pessoas que vivem da renda desta família
PBF	- Programa Bolsa Família
PPGF	- Participa de algum programa do Governo Federal
pop1	- Base de dados das Famílias beneficiárias
pop2	- Base de dados das Famílias não-beneficiárias
pop_a_d	- Base de dados das Famílias beneficiárias após descarte de variáveis
pop_a_dn	- Base de dados das Famílias não-beneficiárias após o descarte de variáveis.
RC	- Raça/Cor
RL	- Regressão Logística
Sast	- Secretaria da Criança e Ação Social da Prefeitura Municipal de Tibagi
SE	- Série Escolar
SENARC	- Secretaria Nacional de Renda de Cidadania
SETP	- Secretaria de Estado do Trabalho, Emprego e Promoção Social
SMT	- Situação no mercado de trabalho
TDF	- Total da Despesa Familiar
TD	- Tem algum tipo de deficiência
TRF	- Total da Renda Familiar

RESUMO

No presente estudo, aplicou-se técnicas estatísticas multivariadas aos dados de famílias cadastradas no Programa Bolsa Família do município de Tibagi, a fim de elaborar um instrumento de alocação que ajudasse no controle dos cadastros das Famílias do Programa. Os resultados encontrados com a aplicação dos modelos estatísticos foram preparados através do aplicativo Excel e do *software* MATLAB®. A metodologia adotada inicia-se com a aplicação do teste T^2 de Hotelling, onde se testa a existência de diferenças estatísticas entre os grupos de famílias beneficiárias e não-beneficiárias do Programa. Passando, em seguida, ao descarte de *outlier*, que faz a verificação das famílias que possuem dados cadastrados adversos dos demais. Já o descarte de variáveis usado na seqüência utiliza a técnica de Componentes Principais. Após o estudo criterioso no conjunto de dados, encontram-se duas funções discriminantes. Uma através da análise da Função Discriminante Linear de Fisher e a outra da análise do Modelo de Regressão Logístico Múltiplo. O ajuste do modelo foi feito com base nos estimadores de máxima verossimilhança obtidos através do método de Levenberg-Marquardt. As duas funções de classificação alocam novos cadastros em um dos grupos pré-estabelecidos. Os resultados encontrados com a aplicação dos conceitos geram informações necessárias para a criação de uma política social para elaboração de programas sociais que venham somar aos esforços do Governo Federal na transferência de renda. Com o surgimento das funções de classificação, houve a necessidade de um conceito que avaliasse as mesmas, sendo este desenvolvido através da abordagem de Lachenbruch, encontrando a probabilidade de classificação correta de 78,4 e 79,6%, respectivamente.

Palavras-chave: Programa Bolsa Família, Análise Estatística Multivariada, Método de Levenberg-Marquardt, Abordagem de Lachenbruch.

ABSTRACT

In this study, multivariate statistics techniques were applied to the registered families' data from the Family Allowance Program in Tibagi City, in order to elaborate an allocation instrument to help to control the families' registers from the Allowance Program. The found results with the application of the statistical models had been prepared through the applicatory Excel and of *software* MATLAB®. The adopted methodology, is initiated with the application of the Hotelling T^2 test, where if it tests the existence of statistics differences among the groups of beneficiary families and non-beneficiary families from the program. Then it was used the *outlier* disposal, which verifies the families that have different registered data from the others. But the disposal of variables used in the sequence uses the Main Components Technique. Once the data has been carefully analyzed there were found two Discriminate Functions. One was found by the analysis of Fisher's Linear Discriminate Function and the other by the Multiple Logistic Regression Model. The model adjustment was done based on the estimators of maximum verosimilitude which were found by Levenberg-Marquardt Method. Both classification functions put new registers into the pre-established groups. The results found by the application of the concepts give us the necessary information to create a social policy to elaborate social programs that will be added to the Federal Government's efforts related to the income transference. The appearance of the classification functions there was the need of a concept which could evaluate them and that was developed by the Lachenbruch approaching. The Lachenbruch approaching found the right classification probability of 78,4 and 79,6%, respectively.

Key Words: Family Allowance Program, Analysis Multivariate Statistics, Levenberg-Marquardt Method, Lachenbruch Approaching.

CAPÍTULO I

1 INTRODUÇÃO

1.1 PROBLEMA

De acordo com a Agência Folha, “O Conselho de Acompanhamento e Promoção Social do Bolsa Família no município de Manhuaçu (MG, 287 km de Belo Horizonte) identificou 300 famílias que estavam recebendo irregularmente o benefício. Foram descobertos beneficiários que tinham carros, motos e até fazendas.”

Segundo o site da Secretaria de Estado do Trabalho, Emprego e Promoção Social (SETP),

> O Ministério do Desenvolvimento Social e Combate à Fome (MDS) estão promovendo, a partir deste segundo semestre de 2005, a descentralização da Gestão de Benefícios do Programa Bolsa Família. Atualmente, é a SENARC que realiza as atividades de bloqueio, desbloqueio e cancelamento de benefícios, a partir da solicitação do prefeito ou do gestor municipal do Bolsa Família. Esta medida permitirá aos gestores municipais do programa administrar, em sua própria cidade, a transferência de renda às famílias participantes do programa.

De acordo com o site do Grupo Banco Mundial, “17 de junho de 2004 – A Diretoria Executiva do Banco Mundial aprovou hoje um empréstimo de US$ 572,2 milhões para apoiar o programa de transferência de renda familiar do Governo brasileiro, o Bolsa Família”.

E ainda sobre o mesmo referencial, tem-se:

> O apoio do Banco ao programa será dividido em duas partes. A Fase I (do segundo semestre de 2004 até o final de 2006), apoiada pelo empréstimo de US$ 572,2 milhões aprovado hoje, abrangerá a etapa de transição, tendo como foco a consolidação dos quatro principais programas de transferência condicional de renda (Bolsa Escola, Bolsa Alimentação, Cartão Alimentação e Auxílio-Gás) e o aprimoramento da arquitetura básica do programa Bolsa Família. A Fase II (2007-08), que será apoiada por um segundo empréstimo, partirá dessas bases para consolidar ainda mais a rede de proteção social e aprofundar as melhorias técnicas.

O governo brasileiro está negociando com o Banco Interamericano de Desenvolvimento (BID), a concessão de um financiamento de US$ 200 milhões para capacitação profissional de pessoas que estão sendo beneficiadas pelo Bolsa Família.”

Como cita Carlos Galves em relação à Economia e Justiça Social:

> Os que dispõem de recursos, cuidam de si mesmos. Aos que não dispõe, – para levar um mínimo de vida humana decente, com seus familiares, – o Estado tem a obrigação de fornecer-lhes os recursos para tal, e sobre tudo para que, pela educação e pelo exercício da liberdade, se evadam da alienação, pelas vias que a democracia – sociedade aberta – lhes abre em todas as faixas da vida humana.

Trabalhos desenvolvidos nesta área serão úteis para auxiliar na implantação, administração e controle do PBF. A preocupação encontra-se no desenvolvimento do PBF, começando pela descentralização de poderes e no investimento que o Governo Federal estuda investir.

Para diminuir a preocupação com este problema, faz-se a aplicação dos conceitos das técnicas estatísticas multivariadas no modelo de funcionamento do PBF,

verificando que o estudo deste modelo de trabalho leva ao envolvimento de muitas variáveis. Aqui, tomou-se o cuidado da escolha das principais para aplicação do sistema. Mas não é apenas a escolha das variáveis que podem causar problemas. Também, tem-se a tabulação dos dados, e isto dependendo da maneira como é feita, traz resultados pouco informativos.

Com o desenvolvimento deste trabalho espera-se contribuir, junto ao PBF através da aplicação de dois métodos da Estatística Multivariada: Análise Discriminante Linear de Fisher (ADLF) e o Modelo de Regressão Logístico Múltiplo (MRLM), onde estes discriminam e alocam famílias cadastradas em grupos beneficiários e não-beneficiários do PBF, a fim de minimizar parte deste quadro de fraudes e elaborar um estudo da análise do cadastramento de novas famílias.

A aglomeração de várias variáveis na análise de dados é, muitas vezes, imprescindível em muitas áreas da pesquisa, principalmente no cruzamento da mesma no PBF. O inter-relacionamento de variáveis tomadas em uma mesma amostra ocorre naturalmente, em decorrência de sua natureza única. Em geral, as diferenças existentes entre grupos ou populações, não é dependente de apenas uma variável e sim de um conjunto delas.

Existem situações ainda em que, quando analisadas separadamente, não são detectadas diferenças significativas entre as populações (ou tratamentos ou grupos) para as variáveis em estudo. Porém, quando a análise é feita de forma global, multivariada, as diferenças ficam evidenciadas e são detectadas pelos testes estatísticos. Isso pode ocorrer tanto pelo acúmulo de diferenças das variáveis individuais como por diferenças existentes entre combinações lineares dessas variáveis.

Uma preocupação refere-se à fundamentação teórica do PBF. Existem muitas publicações de matérias na área, mas estas estão em páginas na internet, principalmente no site do MDS e da Caixa Econômica Federal (CEF). Com o

desenvolvimento do trabalho poderá visualizar a situação. Isto ocorre por se tratar da implantação de um programa “novo” de desenvolvimento social.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

O objetivo geral deste trabalho é mostrar que os resultados alcançados através da aplicação dos conceitos e de técnicas estatísticas multivariadas na base de dados do PBF podem, de forma eminente, ajudar na fiscalização da distribuição de recursos e no monitoramento das famílias cadastradas no programa, a fim de, identificarem falhas e desenvolver um trabalho de prevenção, auxiliando as famílias enquanto estas não estão com os benefícios liberados.

1.2.2 Objetivos Específicos

Buscando atingir o objetivo geral é necessário alcançar os seguintes objetivos específicos.

- Investigar se as populações multivariadas têm o mesmo vetor de médias.
- Reduzir o número de varáveis para efeito de análise e interpretação, sem perda significativa de informação.
- Analisar quais as variáveis que explicam maior parte da variabilidade total dos dados.
- Obter combinações interpretáveis das variáveis.
- Determinar funções das variáveis observadas que permitam classificar ou alocar observações no grupo mais adequado.
- Interpretar os resultados obtidos.

1.3 ESTRUTURA DA DISSERTAÇÃO

A estrutura desta dissertação apresenta no capítulo I os objetivos gerais, específicos e o verdadeiro significado da escolha do tema. Em seguida o capítulo II trata da Revisão da Literatura necessária para o conhecimento das Leis que regem o PBF e a compreensão do entendimento da aplicação da análise estatística multivariada.

No capítulo III, encontra-se o material utilizado e a metodologia desenvolvida para a realização do trabalho. O capítulo IV trás, os resultados obtidos através da utilização do PBF e a análise de todo o procedimento estatístico utilizado.

No capítulo V, têm-se as conclusões com base no estudo realizado e sugestões para trabalhos futuros.

CAPÍTULO II

2 REVISÃO DA LITERATURA

2.1 PROGRAMA BOLSA FAMÍLIA (PBF)

2.1.1 Histórico

O Decreto nº. 5.209 de 17 de setembro de 2004, regulamenta a Lei nº. 10.836, de 09 de janeiro de 2004 editada pela Medida Provisória nº.132 de 20 de outubro de 2003. Sendo necessário para sua apresentação uma orientação teórica das leis e regulamentos, assim como, dados estatísticos do mesmo.

Segundo o decreto do PBF, os objetivos são:

- Promover o acesso à rede de serviços públicos, em especial, de saúde, educação e assistência social;
- Combater a fome e promover a segurança alimentar e nutricional;
- Estimular a emancipação sustentada das famílias que vivem em situação de pobreza e extrema pobreza;
- Combater a pobreza;
- Promover a intersetorialidade, a complementaridade e a sinergia das ações sociais do Poder Público.

Segundo a teoria malthusiana, que faz uma análise profunda a respeito da explosão demográfica do Planeta, afirmando que jamais teríamos uma sociedade feliz,

devido à tendência (estatística) de que as populações sempre cresceriam mais que os meios de sua subsistência e que, enquanto o crescimento populacional tenderia a seguir um ritmo de progressão geométrica, a produção de alimentos cresceria segundo uma progressão aritmética. Assim, a população tenderia a crescer além dos limites de sua sobrevivência, e disso resultariam a fome e a miséria.

Com tudo isto, pode-se verificar que o problema relacionado com a fome e a miséria já era estudado há muitos anos por economistas clássicos. Onde a população e a atividade econômica eram olhadas na perspectiva do tempo, futuro adentro.

De acordo com o quadro 2.1.1 em janeiro de 2006, mais de 8,6 milhões de famílias já estavam sendo atendidas pelo PBF, chegando a um valor investido de 537.782.206,00 reais com um valor médio do benefício de 62,21 reais.

O PBF apresenta-se em diversos aspectos, tais como: o histórico do programa, objetivos, público-alvo, recursos, parcerias e resultados alcançados, assim como, desenvolvem um sistema que auxilia no controle do mesmo. Discriminando grupos de famílias cadastradas e alocando novos cadastros em grupos pré-determinados.

Segundo o decreto nº. 5.209, de 17 de setembro de 2004 do PBF, houve a criação de um órgão de assessoramento imediato do Presidente da República, o Conselho Gestor Interministerial (CGI) do PBF, com a finalidade de formular e integrar políticas públicas, definir diretrizes, normas e procedimentos sobre o desenvolvimento e implantação do PBF, bem como apoiar iniciativas para a instituição de políticas públicas sociais, visando promover a emancipação das famílias beneficiadas pelo programa nas esferas: Federal, Estadual, do Distrito Federal e Municipal, tendo as competências, composição e funcionamentos estabelecidos em ato do Poder Executivo.

O CGI é composto pelos titulares dos seguintes Órgãos e Entidades:

- Ministro do Desenvolvimento Social e Combate à Fome, que o presidirá;

- Ministro do Planejamento, Orçamento e Gestão;
- Ministro da Fazenda;
- Ministro – Chefe da Casa Civil;

De acordo com a Lei nº. 10.836, de 09 de janeiro de 2004 do PBF, o CGI tem o apoio de uma Secretaria-Executiva, com a finalidade de coordenar, supervisionar, controlar e avaliar a operacionalização do programa, compreendendo o Cadastramento Único[1] (CadÚnico), a supervisão do cumprimento das condicionalidades, o estabelecimento de sistema de monitoramento, avaliação, gestão orçamentária e financeira dentre outras funções.

A Secretaria – Executiva é composta pelos seguintes Órgãos:

- Ministério da Educação;
- Ministério da Saúde;
- Ministério de Desenvolvimento Social e Combate à Fome
    - Secretaria Nacional de Segurança Alimentar e Nutricional;
    - Secretaria Nacional de Assistência Social.

2.1.2 O Programa Bolsa Família e a Transferência de Renda

O Programa Bolsa Família é um programa de transferência de renda destinado às famílias em situação de pobreza, com renda *per capita* de até R$ 120 mensais, que

[1] Instituído pelo Decreto nº. 3.877, de 24 de julho de 2001, o CadÚnico é um instrumento para identificação das famílias em situação de pobreza de todos os municípios brasileiros. Esse banco de dados auxilia no planejamento e avaliação das ações sociais, proporcionando ao Governo Federal, Estadual e Municipal uma visão abrangente da população de baixa renda do Brasil, na medida em que possui os dados sócio-econômicos das famílias com renda mensal per capita de até meio salário mínimo. Programas que utilizam o Cadastro Único**:** PBF, Programa de Erradicação do Trabalho Infantil e Projeto Agente Jovem de Desenvolvimento Social e Humano.

associa à transferência do benefício financeiro o acesso aos direitos sociais básicos – saúde, alimentação, educação e assistência social.

Para Vieira, “As políticas de proteção garantem a cobertura de vulnerabilidades a redução de riscos sociais e defendem um padrão básico de vida”.

Segundo Carlos Galves, para Keynes e seus discípulos, estes deram um passo avante, e atinaram com que o desenvolvimento só se efetiva, se sustentado por uma demanda agregada[2], que absorva a produção a mais, e a mantenha sempre solicitada: O desenvolvimento econômico é algo que empenha tanto o desenvolvimento da oferta, como o desenvolvimento da demanda.

O Art. 2º § 1º da Lei nº. 10.836, de 09 de janeiro de 2004, descreve o conceito de família utilizada pelo Programa como sendo: “Família, a unidade nuclear, eventualmente ampliada por outros indivíduos que com ela possuam laços de parentesco ou de afinidade, que forme um grupo doméstico, vivendo sob o mesmo teto e que se mantém pela contribuição de seus membros”.

Conforme descreve a Lei nº. 10.836, de 09 de janeiro de 2004 do PBF, “renda familiar mensal corresponde à soma dos rendimentos brutos auferidos mensalmente pela totalidade dos membros da família, excluindo-se os rendimentos concedidos por programas oficiais de transferência de renda, nos termos do regulamento”.

Também, o Programa busca promover a inclusão social, contribuindo para a emancipação das famílias beneficiárias, construindo meios e condições para que elas possam sair da situação de vulnerabilidade em que se encontram.

[2] Demanda Agregada: o conjunto de todos os tipos de gasto de dinheiro que se fazem no país.

2.1.3 Público – Alvo e Recursos

A Lei nº. 10.836, de 09 de janeiro de 2004 do PBF unifica todos os benefícios sociais do governo federal num único programa. A medida proporcionou maior agilidade na liberação do dinheiro a quem precisa, reduziu burocracias e criou mais facilidade no controle dos recursos, dando assim, mais transparência ao programa.

O PBF unificou os seguintes programas:

I. Programa Nacional de Renda Mínima à educação – "Bolsa Escola", instituído pela Lei nº. 10.219, de 11 de abril de 2001;
II. Programa Nacional de renda Mínima vinculado à saúde – "Bolsa Alimentação", instituído pela Medida Provisória nº. 2.206-1, de 6 de setembro de 2001;
III. Programa Auxílio-Gás, instituído pelo Decreto nº. 4.102, de 24 de janeiro de 2002;
IV. Programa Nacional de Acesso à alimentação – PNAA – "Cartão Alimentação", criado pela Lei nº. 10.689, de 13 de junho de 2003.

As famílias elegíveis são compostas por dois grupos:

1. Famílias em situação de extrema pobreza, com renda mensal per capita até R$ 60,00;
2. Famílias pobres e extremamente pobres com crianças e jovens entre zero e 16 anos incompletos (Grupo 1 e 2), com renda mensal até de R$ 120,00 per capita. Inicialmente, serão atendidas pelo programa as famílias que já estão no Cadastro Único.

Ainda, segundo a Lei acima, o PBF oferecerá às famílias dois tipos de benefícios: o básico (fixo) e o variável.

O benefício básico será concedido às famílias em situação de extrema pobreza. O valor deste benefício será de R$ 50,00 mensais, independentemente da composição e do número de membros do grupo familiar.

O benefício variável, no valor mínimo de R$ 15,00, será concedido às famílias pobres e extremamente pobres que tenham, sob sua responsabilidade, crianças e adolescentes na faixa de 0 a 16 anos incompletos, até o teto de 3 (três) benefícios por família, ou seja, R$ 45,00.

As famílias em situação de extrema pobreza poderão acumular o benefício básico e o variável, chegando ao máximo de R$ 95,00 mensais (R$ 50,00 do benefício básico mais R$ 45,00 do benefício variável). As famílias em situação de pobreza com renda entre R$ 61 e R$ 120,00 podem receber até R$ 45,00.

Nenhuma família que passe a integrar o novo Programa sofrerá redução ou cancelamento do benefício.

De acordo com a Lei nº. 10.836, de 09 de janeiro de 2004 do PBF, as suas despesas correrão à conta das dotações alocadas nos programas federais de transferência de renda e no CadÚnico, bem como de outras dotações do Orçamento da Seguridade Social da União que vierem a ser consignadas ao programa. O Poder Executivo deverá compatibilizar a quantidade de beneficiários do PBF com as dotações orçamentárias existentes.

Os valores dos benefícios e os valores referenciais para caracterização de situação de pobreza ou extrema pobreza poderão ser majorados pelo Poder Executivo, em razão da dinâmica sócio-econômica do País e de estudos técnicos sobre o tema Art. 2º § 6º da Lei do PBF.

Os benefícios serão pagos mensalmente por meio de cartão magnético bancário, fornecido pela Caixa Econômica Federal (CEF), com a respectiva identificação do

responsável mediante o Número de Identificação Social (NIS) de uso do Governo Federal.

Conforme o Art.24, do Decreto nº.5.209 de 17 de setembro de 2004, os valores postos à disposição do titular do benefício, não sacados ou não recebidos por noventa dias, serão restituídos ao PBF, conforme disposto em contrato com o Agente Operador.

2.1.4 Manual de Preenchimento dos Formulários do Cadastramento Único

O Manual de preenchimento dos formulários do Cadastramento único encontra-se disponível no site da Caixa Econômica Federal (Disponível em: <http://www1.caixa.gov.br/cidade/asp/personaliza/iPaginaRedesenho.asp?pagina=4560000456> Acesso em 08 mar. 2006.).

Segundo o Manual, “Ele tem por objetivo orientar o entrevistador sobre o correto preenchimento dos formulários, no processo de coleta de dados, viabilizando a formação de um banco de dados único. Este banco de dados único é compartilhado pelos gestores dos programas sociais do Governo Federal”.

O comentário do manual, neste caso, serve para informar que os dados utilizados para processamento no programa seguiram os valores informados nos campos das respectivas variáveis no momento da coleta de dados.

2.1.5 Parcerias e Associações

O PBF é realizado com a participação do Governo Federal, Estados e Municípios. Em relação à gestão do Programa, o Art. 8º da Lei nº. 10.836, de 09 de janeiro de 2004 do PBF estabelece o seguinte: “A execução e a gestão do PBF são públicas e governamentais e dar-se-ão de forma descentralizada, por meio da conjugação de esforços entre os entes federados, observada a intersetorialidade, a participação comunitária e o controle social”.

Assim, observa-se que a gestão do PBF envolve os três níveis de governo, sendo a União, os Estados e os Municípios parceiros na execução do programa.

Em âmbito local, o controle e a participação social do PBF serão realizados por um conselho ou por um comitê instalado pelo Poder Público municipal. Os entes federados devem oferecer serviços educacionais e de saúde; os Municípios são responsáveis ainda pela inscrição das famílias pobres no CadÚnico.

A Prefeitura Municipal é a responsável pela realização do cadastramento, bem como pela atualização dos dados do cadastro.

São atribuições do Município:

- Planejar e executar o cadastramento;
- Analisar os dados do cadastro em âmbito municipal;
- Estimular o uso deste cadastro pelas diversas Secretarias Municipais;
- Zelar pela qualidade das informações coletadas;
- Digitar, transmitir e acompanhar o retorno dos dados enviados à Caixa;
- Manter atualizada a base de dados do CadÚnico;
- Prestar apoio e informações às famílias de baixa renda sobre o CadÚnico;

- Arquivar os formulários em local adequado por 5 anos.

O MDS esclarece que a aliança com Estados e Municípios permite aumentar o valor dos benefícios, ampliando a cobertura da população assistida, bem como pretende facilitar o acesso das famílias integrantes do programa aos micro-créditos, qualificação profissional e alfabetização.

2.1.6 Alguns Resultados Alcançados pelo Programa Bolsa Família

Com o objetivo de facilitar a superação da situação de pobreza, o PBF estabelece um conjunto de condicionalidades (ações/contrapartidas sociais) que devem ser cumpridas pelo grupo familiar para que possam permanecer no programa. O PBF também prevê ações complementares, que não têm o mesmo caráter compulsório, denominadas "fortes recomendações". Tanto as condicionalidades quanto às recomendações envolvem a concretização de direitos sociais e constitucionais: saúde, educação, alimentação e assistência.

A primeira condicionalidade estabelecida pelo programa é o acompanhamento de saúde e do estado nutricional das famílias. Todos os membros da família beneficiária devem participar do acompanhamento de saúde.

A outra condicionalidade estabelecida é a freqüência à escola. Todas as crianças em idade escolar devem estar matriculadas e freqüentando o ensino fundamental.

E a última condicionalidade é a educação alimentar, onde todas as famílias beneficiárias devem participar de ações de educação alimentar oferecida pelo Governo Federal, estadual ou municipal, quando oferecidas.

As condicionalidades visam certificar o compromisso e a responsabilidade das famílias atendidas. Representam o acesso a direitos que, a médio e longo prazo,

aumentam a autonomia das famílias, na perspectiva da inclusão social. Elas também ampliam as condições para o aumento nas oportunidades de geração de renda das famílias. Nesse sentido, as condicionalidades representam resultados alcançados e que ainda serão, na medida em que estabelecem ações para a melhoria da qualidade de vida das famílias assistidas pelo programa.

Segundo a apresentação do PBF no evento realizado em Brasília pela Coordenação Geral da Política de Alimentação e Nutrição (CGPAN) nos dias 16 a 18 de maio de 2005, a meta para a quantidade de famílias atendidas pelo PBF é:

- Dezembro 2003: 3,6 milhões;
- Dezembro 2004: 6,5 milhões;
- Dezembro 2005: 8,7 milhões;
- Dezembro 2006: 11,2 milhões.

De acordo com os dados disponibilizados pelo quadro 2.1.1, em janeiro de 2006, destacavam-se os Estados da Bahia, Minas Gerais, São Paulo, Ceará e Pernambuco, respectivamente, como os maiores beneficiários do PBF.

Já os Estados que mais possuem municípios atendidos pelo PBF são: Minas Gerais, São Paulo, Rio Grande do Sul e Bahia, respectivamente.

Em relação às famílias atendidas por Estado, constatam-se, novamente, Bahia, Minas Gerais, São Paulo, Ceará e Pernambuco, respectivamente como os maiores beneficiários. Isso explica o fato desses Estados receberem os montantes mais elevados do PBF.

QUADRO 2.1.1 – DEMONSTRATIVO DO PROGRAMA BOLSA FAMÍLIA POR UNIDADE DE FEDERAÇÃO

Demonstrativo - Resumo Bolsa Família por UF

Ref.: Janeiro/2006

FOME ZERO

Ministério do Desenvolvimento Social e Combate à Fome

BRASIL UM PAÍS DE TODOS

UF	Total de municípios na UF	Municípios recebendo Bolsa Família	% Participação	Famílias atendidas	Valor Investido (R$)	Valor médio do benefício (R$)
Acre	22	0	0,00%	39.325	2.626.373,00	66,79
Alagoas	102	0	0,00%	254.517	16.708.495,00	65,65
Amazonas	62	0	0,00%	152.941	10.589.250,00	69,24
Amapá	16	0	0,00%	11.355	781.830,00	68,85
Bahia	417	0	0,00%	1.070.103	70.441.776,00	65,83
Ceará	184	0	0,00%	733.662	48.089.850,00	65,55
Distrito Federal	1	0	0,00%	52.543	2.891.459,00	55,03
Espírito Santo	78	0	0,00%	158.308	9.150.879,00	57,80
Goiás	246	0	0,00%	184.918	10.000.859,00	54,08
Maranhão	217	0	0,00%	533.818	37.201.000,00	69,69
Minas Gerais	853	0	0,00%	978.562	57.314.720,00	58,57
Mato Grosso	139	0	0,00%	116.985	6.561.112,00	56,09
Mato Grosso do Sul	77	0	0,00%	85.002	4.715.296,00	55,47
Pará	143	0	0,00%	340.507	23.494.274,00	69,00
Paraíba	223	0	0,00%	335.072	21.964.081,00	65,55
Pernambuco	186	0	0,00%	635.812	40.785.602,00	64,15
Piauí	222	0	0,00%	282.471	18.963.400,00	67,13
Paraná	399	0	0,00%	435.662	23.380.595,00	53,67
Rio de Janeiro	91	0	0,00%	307.802	18.280.921,00	59,39
Rio Grande do Norte	167	0	0,00%	237.454	14.926.006,00	62,86
Rio Grande do Sul	497	0	0,00%	390.146	22.535.280,00	57,76
Roraima	15	0	0,00%	17.803	1.231.410,00	69,17
Rondônia	52	0	0,00%	71.539	4.433.757,00	61,98
Santa Catarina	293	0	0,00%	138.247	7.728.358,00	55,90
São Paulo	645	0	0,00%	849.563	48.299.504,00	56,85
Sergipe	75	0	0,00%	153.751	9.968.687,00	64,84
Tocantins	139	0	0,00%	76.334	4.717.432,00	61,80
Total Brasil	5.561	0	0,00%	8.644.202	537.782.206,00	62,21

Fonte: Ministério do Desenvolvimento Social (MDS), 2006.

O quadro 2.1.2 mostra a relação das Unidades da Federação e os Programas Sociais de Combate à Fome. Pode-se visualizar que os maiores gastos são dos seguintes programas: Bolsa Família, Auxilio Gás, Bolsa Escola, Cartão Alimentação e o Bolsa Alimentação, respectivamente. Também, pode-se notar que o número de famílias por tipos de programas segue a mesma ordem citada nesse parágrafo.

QUADRO 2.1.2 – DEMONSTRATIVO DA TRANSFERÊNCIA DE RENDA ÀS FAMÍLIAS DO PROGRAMA BOLSA FAMÍLIA POR UNIDADE DE FEDERAÇÃO

Demonstrativo - Programas de Transferência de Renda por UF
Ref.: Janeiro/2006

Ministério do Desenvolvimento Social e Combate à Fome

UF	Programas de Transferência de Renda									
	Bolsa Família		Bolsa Escola		Bolsa Alimentação		Cartão Alimentação		Auxílio Gás	
	Famílias	Total (R$)	Famílias	Total (R$)	Famílias	Total (R$)	Famílias	Total (R$)	Famílias	Total (R$)
Acre	39.325	2.626.373,00	9.906	194.250,00	413	7.170,00	309	15.450,00	19.698	293.257,50
Alagoas	254.517	16.708.495,00	49.768	830.985,00	1.010	17.880,00	2.314	115.700,00	79.387	1.184.790,00
Amazonas	152.941	10.589.250,00	32.771	620.925,00	502	8.685,00	0	-	45.058	674.617,50
Amapá	11.355	781.830,00	11.276	236.580,00	107	1.800,00	0	-	15.009	224.775,00
Bahia	1.070.103	70.441.776,00	258.746	4.332.765,00	3.516	61.605,00	17.989	899.450,00	478.760	7.154.475,00
Ceará	733.662	48.089.850,00	145.176	2.350.875,00	1.246	20.700,00	11.298	564.900,00	302.407	4.404.697,50
Distrito Federal	52.543	2.891.459,00	14.036	248.985,00	0	-	6	300,00	26.930	402.300,00
Espírito Santo	158.308	9.150.879,00	25.680	458.400,00	295	4.785,00	183	9.150,00	54.312	810.832,50
Goiás	184.918	10.000.859,00	46.024	849.855,00	512	8.445,00	70	3.500,00	90.591	1.349.467,50
Maranhão	533.818	37.201.000,00	109.607	1.838.235,00	755	13.035,00	2.354	117.700,00	178.228	2.654.325,00
Minas Gerais	978.562	57.314.720,00	149.715	2.535.615,00	1.201	20.730,00	8.449	422.450,00	334.146	4.978.275,00
Mato Grosso	116.985	6.561.112,00	16.730	298.770,00	203	3.435,00	33	1.650,00	35.549	527.857,50
Mato Grosso do Sul	85.002	4.715.296,00	14.597	271.695,00	517	8.745,00	0	-	39.263	583.995,00
Pará	340.507	23.494.274,00	122.962	2.342.415,00	1.955	32.505,00	0	-	162.327	2.427.622,50
Paraíba	335.072	21.964.081,00	43.917	671.235,00	641	10.860,00	5.863	293.150,00	100.163	1.489.350,00
Pernambuco	635.812	40.785.602,00	100.806	1.649.265,00	2.057	36.060,00	15.740	787.000,00	232.384	3.466.972,50
Piauí	282.471	18.963.400,00	50.267	784.440,00	641	10.800,00	5.336	266.800,00	102.471	1.524.322,50
Paraná	435.662	23.380.595,00	75.746	1.332.390,00	1.973	32.670,00	269	13.450,00	180.686	2.680.702,50
Rio de Janeiro	307.802	18.280.921,00	113.867	2.147.625,00	269	4.770,00	3	150,00	187.581	2.805.675,00
Rio Grande do Norte	237.454	14.926.006,00	44.785	774.030,00	463	8.010,00	6.417	320.850,00	101.731	1.509.802,50
Rio Grande do Sul	390.146	22.535.280,00	63.044	1.044.255,00	127	2.100,00	564	28.200,00	139.119	2.066.272,50
Roraima	17.803	1.231.410,00	3.256	64.380,00	336	5.730,00	0	-	5.440	81.300,00
Rondônia	71.539	4.433.757,00	13.762	262.800,00	178	2.970,00	1	50,00	20.374	304.162,50
Santa Catarina	138.247	7.728.358,00	24.894	424.230,00	614	10.380,00	92	4.600,00	61.488	907.102,50
São Paulo	849.563	48.299.504,00	112.037	1.970.100,00	1.257	22.410,00	22	1.100,00	237.638	3.536.662,50
Sergipe	153.751	9.968.687,00	18.171	295.050,00	1.971	33.255,00	1.936	96.800,00	47.476	707.670,00
Tocantins	76.394	4.717.432,00	21.957	409.425,00	565	9.225,00	0	-	37.673	562.237,50
Total Brasil	8.644.202	537.782.206,00	1.693.503	29.239.575,00	23.324	398.760,00	79.248	3.962.400,00	3.315.889	49.313.520,00

Fonte: Ministério do Desenvolvimento Social (MDS), 2006.

Cerca de 5 milhões de famílias já faziam parte do cadastro do PBF em outubro de 2004. Desse total, cerca de 20% participavam pela primeira vez e os outros 80% já fazia parte de outros programas. Ainda, de acordo com a mesma reportagem, até o final de 2006 o governo federal pretende transferir todas as famílias que estão cadastradas nos outros programas – Bolsa Escola, Cartão Alimentação e Vale Gás – e incluí-las no PBF. (Nunes, 2004)

O problema da pobreza é um dos principais desafios que devem ser superados atualmente. A solução desses problemas passa por uma descentralização do planejamento e gestão de políticas públicas. Nesse contexto, destaca-se a necessidade da aplicação de diversas teorias inter-relacionadas buscando resultados concretos para prevenção e controle dos recursos.

Nesse sentido, ressalta-se que, a partir do ano de 2000, o Governo Federal adotou uma série de medidas visando dar transparência das ações dos governantes. Cita-se, como exemplo, a Lei de Responsabilidade Fiscal[3] (Lei Complementar nº. 101, de 4/5/2000) que exige, dentre outras disposições, a divulgação de relatórios contábeis pela internet. Com essa iniciativa, o Governo Federal está auxiliando a criação de uma cultura no setor público voltada para a transparência das ações na esfera pública.

Em relação ao PBF, além das informações básicas sobre o programa, estão disponibilizadas na Internet, dados como: municípios atendidos, quantidade de famílias atendidas, valor dos benefícios por Unidades da Federação bem como os valores totais transferidos as famílias brasileiras beneficiadas pelo programa.

[3] A Lei de Responsabilidade Fiscal - LRF (Lei Complementar nº. 101, de 04 de maio de 2000) estabelece normas de finanças públicas voltadas para a responsabilidade na gestão fiscal, mediante ações em que se previnam riscos e corrijam os desvios capazes de afetar o equilíbrio das contas públicas, destacando-se o planejamento, o controle, a transparência e a responsabilização, como premissas básicas.

2.2 ANÁLISE ESTATÍSTICA MULTIVARIADA

2.2.1 Introdução

Quando se trabalha com uma grande quantidade de variáveis, leva-se em consideração: a importância do banco de dados e a agilidade no processo para a obtenção dos resultados. Considerando isto, deu-se interesse no estudo das ciências que buscam transformar dados em conhecimento.

Para os órgãos governamentais responsáveis pela formulação de políticas e supervisão do sistema financeiro, é de suma importância que se crie “sistemas” baseados em variáveis que deixe o governo em uma melhor posição para prever a ocorrência de gastos e superação da situação de pobreza.

Segundo Moita Neto (2004), “A denominação “Análise Multivariada” corresponde a um grande número de métodos e técnicas que utilizam simultaneamente todas as variáveis na interpretação teórica do conjunto de dados obtidos.”

De acordo com JOHNSON & WICHERN (1998), em problemas que envolvem p variáveis $(p>1)$, tomando-se n observações de cada vetor aleatório $\underline{X}$ de dimensão p, tem-se que as medidas observadas X_{jk}, com $j = 1,2,\text{K}\ ,n$ e $\text{k} = 1,2,\text{K}\ ,p$, podem ser arranjadas em uma matriz de dados genérica $n \times p$, conforme abaixo:

$$ {}_{n}X_{p} = \begin{bmatrix} X_{11} & X_{12} & \Lambda & X_{1p} \\ X_{21} & X_{22} & \Lambda & X_{2p} \\ \text{M} & \text{M} & \text{O} & \text{M} \\ X_{n1} & X_{n2} & \Lambda & X_{np} \end{bmatrix} $$

A representação da matriz de dados correspondente a n observações do vetor $\underline{X}' = \left[X_1, X_2, \text{K}\ , X_p \right]$ de dimensão p, composto por p variáveis aleatórias, pode ser

${}_nX_p = (X_{jk})$. No entanto, essa matriz correspondente a uma amostra aleatória de tamanho n do vetor $p-$dimensional $\underline{X}$, ou seja, $X_1, X_2, \text{K}\ X_p$.

2.2.2 Estatísticas Descritivas

Segundo Ferreira (1996), quando se tem um banco de dados grande existe um sério obstáculo para qualquer tentativa de extração de informações visuais pertinentes ao mesmo. Muitas das informações contidas nesse banco de dados podem ser obtidas por cálculo de certos números, conhecidos como estatísticas descritivas. Por exemplo, a média aritmética ou média amostral, é uma estatística descritiva que fornece informação de posição, isto é, representa um valor central para o conjunto de dados. Como um outro exemplo, a média das distâncias ao quadrado de cada dado em relação à média, fornece uma medida de dispersão, ou variabilidade.

As estatísticas descritivas que mensuram posição, variação e associação linear são enfatizadas. As descrições formais destas medidas estão apresentadas a seguir. A média amostral, simbolizada por $\overline{\underline{X}}$, que estima o vetor médio populacional $\underline{\mu}$ é dado por:

$$\overline{\underline{X}}_k = \frac{1}{n}\sum_{j=1}^{n} \underline{X}_{jk} \qquad k = 1,2,\text{K}\ , p \tag{2.1}$$

A matriz de covariância populacional Σ é definida por:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \Lambda & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \Lambda & \sigma_{2p} \\ \text{M} & \text{M} & \text{O} & \text{M} \\ \sigma_{p1} & \sigma_{p2} & \Lambda & \sigma_p^2 \end{bmatrix}$$

onde σ^2 é a variância da variável aleatória X_j e σ_{jk} é a covariância entre as variáveis X_j e X_k.

Para estimar a matriz de covariância populacional, Σ, utiliza-se a matriz de covariância amostral, S, que é dada por:

$$S = \begin{bmatrix} s_{11} & s_{12} & \Lambda & s_{1p} \\ s_{21} & s_{22} & \Lambda & s_{2p} \\ \text{M} & \text{M} & \text{O} & \text{M} \\ s_{p1} & s_{p2} & \Lambda & s_{pp} \end{bmatrix}$$

Uma medida de variação é fornecida pela variância amostral da variável aleatória X_j.

$$S_k^2 = S_{kk} = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_{jk} - \overline{X}_k\right)^2 \quad k = 1,2,\text{K}\ , p \tag{2.2}$$

A raiz quadrada da variância amostral, $\sqrt{S_{kk}}$, é conhecida como desvio padrão amostral. Esta medida de variação está na mesma unidade de medida das observações.

Uma medida de associação entre as observações de duas variáveis, variáveis k e k', é dada pela covariância amostral:

$$S_{kk'} = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_{jk} - \overline{X}_k\right)\left(X_{jk'} - \overline{X}_{k'}\right) \quad k, k' = 1,2,\text{K}\ , p \tag{2.3}$$

A matriz de correlação amostral que estima o parâmetro de correlação populacional é:

$$R = \begin{bmatrix} 1 & r_{12} & \Lambda & r_{1p} \\ r_{21} & 1 & \Lambda & r_{2p} \\ \text{M} & \text{M} & \text{O} & \text{M} \\ r_{p1} & r_{p2} & \Lambda & 1 \end{bmatrix}$$

Para a correlação amostral, a medida de associação linear entre duas variáveis não depende da unidade de mensuração. O coeficiente de correlação amostral para $k-ésima$ e $k'-ésima$ variável, é definido por:

$$r_{kk'} = \frac{S_{kk'}}{\sqrt{S_{kk}}\sqrt{S_{k'k'}}} = \frac{\sum_{j=1}^{n}\left(X_{jk} - \overline{X}_k\right)\left(X_{jk'} - \overline{X}_{k'}\right)}{\sqrt{\sum_{j=1}^{n}\left(X_{jk} - \overline{X}_k\right)^2}\sqrt{\sum_{j=1}^{n}\left(X_{jk'} - \overline{X}_{k'}\right)^2}} \tag{2.4}$$

Verifica-se que $r_{kk'} = r_{k'k}$ para todo k e k'. O coeficiente de correlação amostral é a versão estandardizada da covariância amostral, onde o produto das raízes das variâncias das amostras fornece a estandardização. O coeficiente de correlação amostral pode ser considerado como uma covariância amostral. Suponha que os valores X_{jk} e $X_{jk'}$ sejam substituídos pelos valores padronizados,

$$\left(X_{jk} - \overline{X}_k\right)\Big/\sqrt{S_{kk}} \quad \text{e} \quad \left(X_{jk'} - \overline{X}_{k'}\right)\Big/\sqrt{S_{k'k'}} \ .$$

Esses valores padronizados são expressos sem escalas de medidas (adimensionais), pois são centrados em zero e expressos em unidades de desvio padrão. O coeficiente de correlação amostral é justamente a covariância amostral das observações estandardizadas.

2.2.3 T^2 DE HOTELLING

2.2.3.1 Introdução

Conforme JONHSON & WICHERN (1998), o teste T^2 de Hotelling avalia se dois vetores de médias são iguais e compara a resposta média da população π_1 com a

da população π_2 com tamanhos das amostras n_1 e n_2. Nestas amostras, calculam-se estatísticas que estimam parâmetros populacionais $\underline{\mu}_i$ e Σ_i.

O teste T^2 de Hotelling verifica se existem diferenças significativas entre os grupos formados pelas famílias beneficiárias ou não do PBF.

Os pressupostos para aplicar o teste são:

- As amostras aleatórias das diferentes populações são independentes;
- As populações têm a mesma matriz de covariância Σ, ou seja, $\Sigma_i = \Sigma, \forall_i$;
- Cada população é normal multivariada.

Para testar a hipótese de que os vetores médios são iguais, usa-se o teste baseado na distância quadrática com:

$$H_0 : \mu - \mu_0 = \delta_0 \quad \text{vs} \quad H_1 : \mu - \mu_1 \neq \delta_0$$

onde, H_0 é a hipótese nula e H_1 é a hipótese (bilateral) alternativa.

Considera-se que:

$$E\left(\underline{\overline{X}}_1 - \underline{\overline{X}}_2\right) = E\left(\underline{\overline{X}}_1\right) - E\left(\underline{\overline{X}}_2\right) = \underline{\mu}_1 - \underline{\mu}_2 = 0 \tag{2.5}$$

e

$$V\left(\underline{\overline{X}}_1 - \underline{\overline{X}}_2\right) = V\left(\underline{\overline{X}}_1\right) + V\left(\underline{\overline{X}}_2\right) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_p \tag{2.6}$$

onde, S_p é a matriz de covariância amostral conjunta, dada por:

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \tag{2.7}$$

que estima a matriz de covariância populacional Σ. Sendo o nível de significância adotado (α), e a estatística do teste:

$$T^2 = \left[\underline{\overline{X}}_1 - \underline{\overline{X}}_2 - \delta_0\right]^t \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_p\right]^{-1} \left[\underline{\overline{X}}_1 - \underline{\overline{X}}_2 - \delta_0\right] \sim \left(\frac{n_1 + n_2 - 2}{n_1 + n_2 - p - 1}\right) p F_{p, n_1+n_2-p-1}(\alpha)$$

Assim, a regra de decisão para o teste, em um nível de significância α, tem a seguinte forma:

$$T^2 \frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - 2)p} > F_{p, n_1+n_2-p-1}(\alpha) \qquad (2.8)$$

Considerando:

$$A = T^2 \frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - 2)p} \quad \text{e} \quad B = F_{p, n_1+n_2-p-1}(\alpha)$$

Rejeita-se H_0 se $A > B$, caso contrário, aceita-se H_0.

2.2.4 ANÁLISE DE COMPONENTES PRINCIPAIS

2.2.4.1 Introdução

Segundo BARROSO (2003), a “Análise de Componentes Principais é uma técnica estatística que transforma um conjunto de p variáveis em um conjunto com um número menor (k) de variáveis aleatórias não-correlacionadas, que explica uma parcela substancial das informações do conjunto original”.

O método das Componentes Principais procura explicar a estrutura da variância e covariância de um vetor aleatório através de poucas combinações lineares das variáveis originais.

Segundo CHAVES NETO (1997), os principais objetivos da Análise de Componentes Principais são:

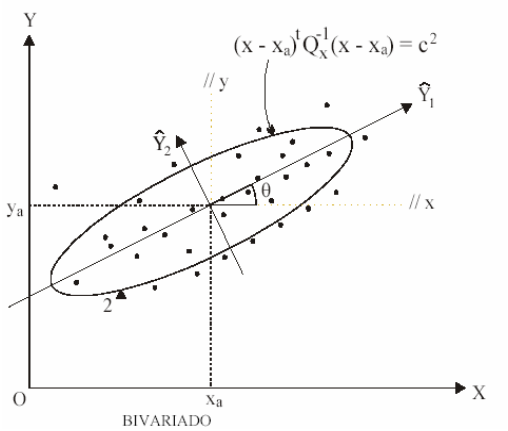
"Reduzir o número de variáveis e analisar quais as variáveis ou, quais os conjuntos de variáveis explica a maior parte da variabilidade total, revelando que tipo de relacionamento existe entre eles".

Conforme Moita Neto (2004) "A análise de componentes principais é uma técnica estatística poderosa que pode ser utilizada para redução do número de variáveis e para fornecer uma visão estatisticamente privilegiada do conjunto de dados. A análise de componentes principais fornece as ferramentas adequadas para identificar as variáveis mais importantes no espaço das componentes principais".

2.2.4.2 Componentes Principais Populacionais

Segundo JOHNSON & WICHERN, (1998), algebricamente as componentes principais são combinações lineares das p variáveis originais $X_1, X_2, \text{K}\ X_p$ que compõe o vetor aleatório $\underline{X}$. Geometricamente, as combinações lineares representam à seleção de um novo sistema de coordenadas, obtido por rotação do sistema original, sendo que os novos eixos representam as direções com variabilidade máxima. Como exemplo, tem-se a representação da estrutura de componentes principais para $p = 2$.

FIGURA 2.5.1 – REPRESENTAÇÃO GEOMÉTRICA DAS COMPONENTES PRINCIPAIS



Fonte: Curso de Pós-Graduação em Ciências Geodésicas da Universidade Federal do Paraná.

As componentes Principais são obtidas a partir da matriz de covariância Σ ou da matriz de correlação ρ, que resumem a estrutura de relacionamento das p variáveis originais que compõe o vetor $\underline{X}$. Então, da matriz de covariância Σ ou da matriz de correlação ρ, obtém-se os autovalores $\lambda_1 \geq \lambda_2 \geq,\mathrm{K}\ ,\geq \lambda_p$ e os respectivos autovetores $\underline{e}_1,\underline{e}_2,\mathrm{K}\ ,\underline{e}_p$. E, com estes entes algébricos se constrói as combinações lineares que definem as componentes principais, ou seja, $Y_i = \underline{e}'_i\,\underline{X} \quad i = 1,2,\mathrm{K}\ ,p$.

As componentes principais são combinações lineares, $Y_i \quad i = 1,2,\mathrm{K}\ ,p$, não correlacionadas, uma vez que a matriz dos autovetores P, abaixo, é ortogonal,

$$P = \begin{bmatrix} e_{11} & e_{12} & \Lambda & e_{1p} \\ e_{21} & e_{22} & \Lambda & e_{2p} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ e_{p1} & e_{p2} & \Lambda & e_{pp} \end{bmatrix}$$

A variância da Componente Principal $Y_i = \underline{e}'_i\,\underline{X} \quad i = 1,2,\mathrm{K}\ ,p$ é dada por:

$$V(Y_i) = V(\underline{e}'_i\,X) = \underline{e}_i{}'V(X)\underline{e}_i{}' = \underline{e}_i{}'\,\Sigma\,\underline{e}_i{}' \tag{2.9}$$

e a covariância entre as componentes Y_j e Y_k é nula, ou seja, $cov(Y_j,Y_k) = 0$.

Portanto, define-se:

- A primeira componente principal como a combinação linear $Y_1 = \underline{e}'_1 \underline{X}$ que maximiza a variância de Y_1, sob a restrição $\underline{e}_1{}' \underline{e}_1 = 1$;
- A segunda componente principal como a combinação linear $Y_2 = \underline{e}'_2 \underline{X}$ que maximiza $V(\underline{e}'_2 \underline{X})$, sob a restrição $\underline{e}_2{}' \underline{e}_2 = 1$ e $\text{cov}(\underline{e}'_1 \underline{X} \underline{e}'_2 \underline{X}) = 0$;
- A i-ésima componente principal como a combinação linear $Y_i = \underline{e}'_i \underline{X}$ que maximiza $V(\underline{e}'_i \underline{X})$, sob a restrição $\underline{e}_i{}' \underline{e}_i = 1$ e $cov(\underline{e}'_i \underline{X} \underline{e}'_k \underline{X}) = 0 \; \forall \, k \neq i$.

2.2.4.3 Componentes Principais da Amostra

Geralmente os parâmetros da estrutura de covariância, Σ ou ρ, são desconhecidos, então a obtenção das componentes principais é feita a partir de seus estimadores, que são a matriz de covariância amostral definida por 2.7 ou a matriz de correlação amostral R, esta definida abaixo:

$$R = D^{-1} S D^{-1} \tag{2.10}$$

onde D é a matriz desvio padrão amostral e $\overline{\underline{X}}$ é o vetor médio amostral, dados respectivamente por:

$$D = \begin{bmatrix} s_1 & 0 & \Lambda & 0 \\ 0 & s_2 & \Lambda & 0 \\ \text{M} & \text{M} & \text{O} & \text{M} \\ 0 & 0 & \Lambda & s_p \end{bmatrix} \qquad \overline{\underline{X}} = \begin{bmatrix} \underline{X}_1 \\ \underline{X}_2 \\ \text{M} \\ \underline{X}_p \end{bmatrix}$$

Então, obtêm-se as estimativas dos elementos da estrutura de covariância do vetor aleatório $\underline{X}$, ou seja, os autovalores $\hat{\lambda}_i \;\; i = 1,2,\text{K}\,,p$ e os correspondentes autovetores $\underline{\hat{e}}_i$ e se constroem as componentes principais amostrais

$\hat{Y}_{\hat{i}} = \underline{\hat{e}}_i \underline{X} \quad i = 1,2,\text{K}\ , p$. As propriedades das componentes principais se mantêm e são obtidas com base em estimadores.

A obtenção das componentes principais com base nas informações da matriz de correlação é preferida, devido ao fato de se conseguir eliminar o efeito de escala nos valores das componentes do vetor de variáveis originais $\underline{X}$. Como é bem conhecida, a matriz de correlação é uma matriz de covariância, mas de variáveis padronizadas. Assim, consegue-se eliminar a influência da escala na magnitude das variâncias.

Os autovalores e os autovetores da matriz de correlação são a essências do método das componentes principais. Os autovetores definem as direções de máxima variabilidade e os autovalores especificam as variâncias. Quando os primeiros autovalores são muito maiores que os demais, a maior parte da variância total pode ser explicada por um número menor do que as p dimensões do vetor $\underline{X}$.

2.2.4.4 Critérios para definição do número de Componentes Principais Extraídas. (MARDIA et al), (BARROSO 2003), (JOHNSON & WICHERN 1998)

A resposta definitiva para a questão de quantos componentes principais deverá ser retida, não existe, o que se pode considerar é a quantidade de variância total explicada.

Na literatura existem vários critérios que auxiliam nessa tomada de decisão, mas estes podem levar a resultados diferentes. Resumindo os principais critérios:

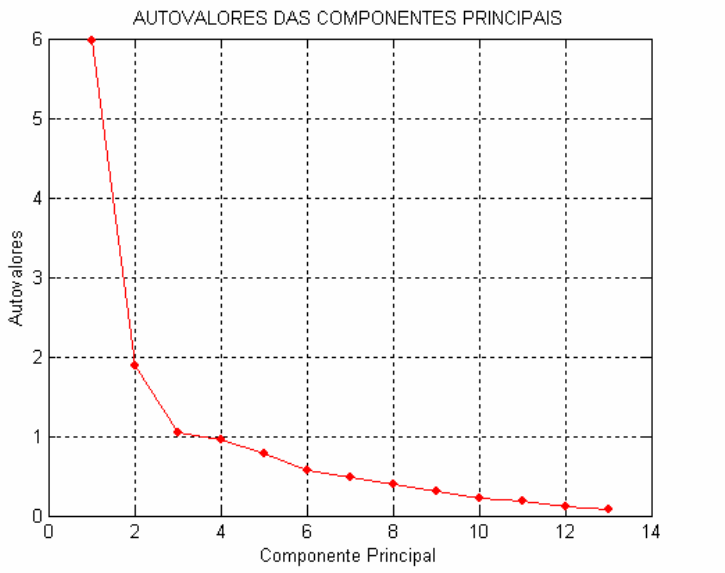
- Critério de Kaiser (1958). Esse critério sugere manter na análise as componentes principais correspondentes aos autovalores maiores do que a média dos autovalores, se a análise é baseada na matriz de covariância, ou as componentes principais correspondentes aos autovalores maiores

que um, se a matriz de correlação é usada. Seguindo esse critério, podem se descartar as componentes principais com contribuições importantes;

- Reter o número de componentes principais que acumulem pelo menos certa porcentagem da variabilidade total dos dados, por exemplo, 70%;
- Reter as componentes principais que acumulem pelo menos uma certa porcentagem da variabilidade de cada uma das variáveis originais, por exemplo, 50%.

Uma ferramenta que pode auxiliar na escolha do número de componentes principais a reter é o "*scree plot*". Nesse gráfico, representam-se os autovalores. Comumente, a diferença entre os primeiros autovalores é grande e diminui para os últimos. A sugestão é fazer o corte quando a variação passa a ser pequena.

FIGURA 2.5.2 – REPERSENTAÇÃO GRÁFICA DO "SCREE PLOT"



A definição e descrição para uma possível maneira de descartarem variáveis (variáveis redundantes) usando a técnica de Componentes Principais. Eliminando-se variáveis dessa categoria, obtém-se uma nova matriz de dados com menor ordem.

O procedimento para descarte é o seguinte:

1º. Determine os autovalores $\lambda_j \quad j = 1,2,K\ ,p$ e autovetores $\underline{e}_j \quad j = 1,2,K\ ,p$ da matriz de correlação das variáveis independentes, ou seja, das covariáveis.

2º. Considere o autovetor $(\underline{e}_j)$ correspondente ao menor autovalor $(\lambda_j < \lambda_l\ ,\ \forall\ j \neq l \quad \text{para}\ l = 1,2,K\ p)$. Descarta-se então a variável cujo coeficiente no autovetor for o maior (valor absoluto). É claro que o autovetor com menor autovalor é o menos importante e uma variável importante nele será a menos importante no conjunto.

3º. O número de variáveis descartadas pode ser igual ao número de autovalores menores ou iguais a 0,70.

2.3 ANÁLISE DISCRIMINANTE

2.3.1 Introdução

Segundo JOHNSON & WICHERN (1988), a análise discriminante é uma técnica multivariada que tem por objetivo tratar dos problemas relacionados com separar conjuntos distintos de objetos (itens ou observações) e alocar novos objetos em conjuntos previamente definidos. Quando empregada como procedimento de classificação não é uma técnica exploratória, uma vez que ela conduz a regras bem definidas, as quais podem ser utilizadas para classificação de outros objetos.

Segundo CHAVES NETO (1997), os objetivos primordiais da técnica da análise discriminante são:

"Determinar qual variável (ou quais variáveis) discriminam (diferenciam, separam) esses grupos (através da ANOVA/MANOVA) e determinadas as variáveis que melhor discriminam os grupos, utilizá-las para criar funções discriminantes que serão utilizadas para alocar novos indivíduos, objetos ou observações no grupo mais adequado (a função discriminante otimiza a alocação)."

Uma função que separa pode servir para alocar, e da mesma forma uma regra alocadora pode sugerir um procedimento discriminatório. Na prática, o primeiro e o segundo objetivo, frequentemente, sobrepõem-se e a distinção entre separação e alocação torna-se confusa.

2.3.2 Separação e Classificação para Duas Populações

Segundo JOHNSON & WICHERN (1988), "discriminar" e "classificar" foi introduzida por R. A. Fisher no primeiro tratamento moderno dos problemas de separação.

2.3.2.1 Método de Fisher para Duas Populações

Conforme JOHNSON & WICHERN (1998), o método de Fisher consiste basicamente em separar duas classes de objetos, ou fixar um novo objeto em uma das classes. A idéia de Fisher foi transformar as observações multivariadas $\underline{X}$ nas observações univariadas Y tal que os $Y's$ nas populações π_1 e π_2 fossem separadas

tanto quanto possível. Utilizou-se para isso combinações lineares dos $\underline{X}$ criando os $Y's$.

Seja μ_{1y} a média dos $Y's$ obtidos dos $\underline{X}'s$ pertencentes a π_1 e μ_{2y} a média dos $Y's$ obtidos dos $\underline{X}'s$ pertencentes a π_2, então Fisher selecionou a combinação linear que maximiza a distância quadrática entre μ_{1y} e μ_{2y} relativamente à variabilidade dos $Y's$. Assim, seja:

$\underline{\mu}_1 = E(\underline{X}/\pi_1)$: valor esperado de uma observação multivariada de π_1

$\underline{\mu}_2 = E(\underline{X}/\pi_2)$: valor esperado de uma observação multivariada de π_2

e supondo a matriz de covariância

$$\Sigma = \left[\left(\underline{X} - \underline{\mu}_i \right) \left(\underline{X} - \underline{\mu}_i \right)' \right]; \quad i = 1,2$$

como sendo a mesma para ambas as populações, então considerando a combinação linear

$$Y = \underline{C}' \underline{X} \tag{2.11}$$

tem-se

$$\mu_{1y} = E(Y/\pi_1) = E\left(\underline{C}' \underline{X}/\pi_1\right) = \underline{C}' E(\underline{X}/\pi_1) = \underline{C}' \underline{\mu}_1 \tag{2.12}$$

da mesma forma

$$\mu_{2y} = E(Y/\pi_2) = E\left(\underline{C}' \underline{X}/\pi_2\right) = \underline{C}' E(\underline{X}/\pi_2) = \underline{C}' \underline{\mu}_2 \tag{2.13}$$

e

$$\sigma^2{}_Y = V(\underline{Y}) = V\left(\underline{C}' \underline{X}\right) = \underline{C}' V(X) \underline{C} = \underline{C}' \Sigma \underline{C} \tag{2.14}$$

Seguindo-se ainda a idéia de Fisher, a melhor combinação linear é obtida da razão entre o quadrado da distância entre as médias e a variância de Y. Tem-se:

$$\frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma^2{}_Y} = \frac{\left(\underline{C}^{'}\underline{\mu}_1 - \underline{C}^{'}\underline{\mu}_2\right)^2}{\underline{C}^{'}\sum\underline{C}} = \frac{\underline{C}^{'}\left(\underline{\mu}_1 - \underline{\mu}_2\right)\left(\underline{\mu}_1 - \underline{\mu}_2\right)^{'}\underline{C}}{\underline{C}^{'}\sum\underline{C}} = \frac{\left(\underline{C}^{'}\underline{\delta}\right)^2}{\underline{C}^{'}\sum\underline{C}} \tag{2.15}$$

onde: $\underline{\delta} = \underline{\mu}_1 - \underline{\mu}_2$ e $V(\underline{X}) = \Sigma$

Dado que $\underline{C} = k\Sigma^{-1}\underline{\delta},\ \forall k \neq 0$, definição encontra-se em LIMA (2002), e fazendo $k = 1$, encontramos:

$$\underline{C} = \Sigma^{-1}\left(\underline{\mu}_1 - \underline{\mu}_2\right) \tag{2.16}$$

Portanto, substituindo o valor de $\underline{C}$, na equação 2.11, tem-se:

$$Y = \left(\underline{\mu}_1 - \underline{\mu}_2\right)\Sigma^{-1}\underline{X} \tag{2.17}$$

que é a conhecida como FDLF.

Assim, tomando-se:

$$Y_0 = \left(\underline{\mu}_1 - \underline{\mu}_2\right)^{'}\Sigma^{-1}\underline{X}_0 \tag{2.18}$$

como valor da FDLF para uma nova observação $\underline{X}_0$ e considerando o ponto médio entre as duas populações univariadas

$$m = \frac{1}{2}(\mu_{1Y} + \mu_{2Y}) = \frac{1}{2}\left(\underline{C}^{'}\underline{\mu}_1 + \underline{C}^{'}\underline{\mu}_2\right) = \frac{1}{2}\underline{C}^{'}\left(\underline{\mu}_1 + \underline{\mu}_2\right) = \frac{1}{2}\left[\left(\underline{\mu}_1 - \underline{\mu}_2\right)^{'}\Sigma^{-1}\left(\underline{\mu}_1 + \underline{\mu}_2\right)\right] \tag{2.19}$$

ou ainda:

$$E(Y_0/\pi_1) - m \geq 0 \tag{2.20}$$

$$E(Y_0/\pi_2) - m < 0 \tag{2.21}$$

ou seja, se $X_0 \in \pi_1$, é esperado que $Y_0 \geq m$. Por outro lado se $\underline{X}_0 \in \pi_2$ é esperado que $Y_0 < m$.

Desta forma tem-se a Regra de Classificação:

Alocar $\underline{X}_0$ em π_1, se $Y_0 \geq m$

Alocar $\underline{X}_0$ em π_2, se $Y_0 < m$

Como na maioria das vezes os parâmetros $\underline{\mu}_1, \underline{\mu}_2$ e Σ não são conhecidos, devem-se usar seus estimadores. Então, supondo que se tem n_1 observações da variável aleatória multivariada $\underline{X}_1$ de dimensão p, que corresponde a uma amostra aleatória da população π_1 e n_2 observações da variável aleatória multivariada $\underline{X}_2$ de dimensão p, que corresponde a uma amostra aleatória da população π_2, os resultados amostrais correspondentes são:

$$\underline{\overline{X}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\underline{X}_{i1} \tag{2.22}$$

$$S_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(\underline{X}_{i1} - \underline{\overline{X}}_1\right)\left(\underline{X}_{i1} - \underline{\overline{X}}_1\right)' \tag{2.23}$$

$$\underline{\overline{X}}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}\underline{X}_{i2} \tag{2.24}$$

$$S_2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(\underline{X}_{i2} - \underline{\overline{X}}_2\right)\left(\underline{X}_{i2} - \underline{\overline{X}}_2\right)' \tag{2.25}$$

Assim, $\underline{\overline{X}}_1$ estima μ_1, $\underline{\overline{X}}_2$ estima μ_2 e a matriz de covariância conjunta (estimada) é dado pela equação 2.7.

A FDLF estimada será dada por

$$\hat{Y} = \underline{\hat{C}}\,\underline{X} = \left(\underline{\overline{X}}_1 - \underline{\overline{X}}_2\right)' S_p^{-1}\,\underline{X} \tag{2.26}$$

e, a estimativa do ponto médio entre as médias amostrais univariadas é:

$$\hat{m} = \frac{1}{2}\left[\left(\underline{\overline{X}}_1 - \underline{\overline{X}}_2\right)' S_p^{-1}\left(\underline{\overline{X}}_1 + \underline{\overline{X}}_2\right)\right] \tag{2.27}$$

e, finalmente a regra de classificação:

$$\text{Alocar } \underline{X}_0 \text{ em } \pi_1, \text{ se } Y_0 \geq \hat{m}$$

$$\text{Alocar } \underline{X}_0 \text{ em } \pi_2, \text{ se } Y_0 < \hat{m}$$

A combinação linear particular da expressão 2.26 maximiza a razão:

$$\frac{\left(\overline{Y}_1 - \overline{Y}_2\right)^2}{S_y^2} = \frac{\left(\hat{\underline{C}}_1 \underline{\overline{X}}_1 - \hat{\underline{C}}_2 \underline{\overline{X}}_2\right)^2}{\hat{\underline{C}}' S_p \hat{\underline{C}}} = \frac{\left(\hat{\underline{C}}' \underline{d}\right)^2}{\hat{\underline{C}}' S_p \hat{\underline{C}}} \tag{2.28}$$

onde:

$$\underline{d} = \underline{\overline{X}}_1 - \underline{\overline{X}}_2 \tag{2.29}$$

e

$$S_y^2 = \frac{\sum_{i=1}^{n_1}\left(Y_{i1} - \overline{Y}_1\right)^2 \sum_{i=1}^{n_2}\left(Y_{i2} - \overline{Y}_2\right)^2}{n_1 + n_2 - 2} \tag{2.30}$$

2.4 REGRESSÃO LOGÍSTICA

2.4.1 Introdução

Segundo LIMA (2002), a função logística surgiu em 1845 ligada ao problema do crescimento demográfico. Neste caso, a função também desempenha papel importante. A partir da década de 30, esta metodologia passou a ser aplicada no âmbito da biologia. Em relação a problemas econômicos e sociais, somente nos anos 60 é que os modelos logísticos começaram a ser utilizados.

Este modelo é aplicável quando a variável dependente é dicotômica, ou seja, possui duas possibilidades de resultado, sendo uma oposta à outra, como é o caso ora aplicado, onde a variável dependente assume dois valores possíveis: 0 ou 1.

Para VICENTE (2001), a Regressão Logística (RL), ou LOGIT, é útil para situações nas quais se deseja predizer a presença ou ausência de uma característica, ou resultado, baseado em valores de um conjunto de variáveis independentes. A LOGIT pode estimar a probabilidade máxima depois de transformar a variável dependente em variável de base logarítmica. Deste modo a LOGIT calcula a probabilidade de “um” certo evento acontecer. Pode assim calcular mudanças nas inter-relações dos logs da variável dependente, e não mudanças na própria variável.

Para CASTRO JUNIOR (2003), assim como a ADLF, a RL é apropriada quando a variável dependente é do tipo não-métrico. Devido aos pressupostos rígidos da ADLF, a RL tornou-se preferida em estudos desta natureza. A RL se parece muito com uma regressão múltipla e por isso também seu uso é bastante apreciado entre os pesquisadores.

2.4.2 Modelo de Regressão Linear Múltiplo (HOFFMANN, 1942) (LIMA, 2002)

2.4.2.1 Introdução

Tem-se uma regressão linear múltipla quando se admite que o valor da variável dependente (resposta), seja função linear de duas ou mais variáveis independentes. O modelo estatístico de uma regressão linear múltipla com p variáveis independentes é:

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \mathrm{K} + \beta_{p-1} X_{p-1\,j} + \varepsilon_j \qquad j = 1,\mathrm{K}\ ,n \tag{2.31}$$

ou

$$Y_j = \beta_0 + \sum_{i=1}^{p-1} \beta_i X_{ij} + \varepsilon_j \tag{2.32}$$

onde: n é o número de observações e $p-1$ o número de variáveis.

Utilizando notação matricial o modelo fica

$$\underline{Y} = X\,\underline{\beta} + \underline{\varepsilon} \tag{2.33}$$

onde:

$\underline{Y}$: Variável resposta
X : Matriz do modelo
$\underline{\beta}$: Vetor de parâmetros a ser estimado
$\underline{\varepsilon}$: Vetor de erros aleatórios

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \mathrm{M} \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{11} & X_{21} & \mathrm{K} & X_{p1} \\ 1 & X_{12} & X_{22} & \mathrm{K} & X_{p2} \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & & \mathrm{M} \\ 1 & X_{1n} & X_{2n} & \mathrm{K} & X_{pn} \end{bmatrix}, \quad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \mathrm{M} \\ \beta_{p-1} \end{bmatrix}, \quad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \mathrm{M} \\ \varepsilon_n \end{bmatrix}$$

Com a finalidade de determinar o estimador de mínimos quadrados ordinários do vetor $\underline{\beta}$, deve-se minimizar a soma dos quadrados dos erros e supõe-se que estas variáveis aleatórias sejam independentes e identicamente distribuídas, ou seja:

$$\varepsilon_i \sim .(0,\sigma^2) \text{ e } \varepsilon_i \sim (0,\sigma^2 I_n) \tag{2.34}$$

Dado o modelo da expressão 2.33**,** admitem-se sobre o mesmo as seguintes suposições:

- A variável dependente (Y_j) é função linear das variáveis independentes $(X_{ij},\ i=1,...,n)$;
- Os valores das variáveis independentes são fixos;
- $E(\varepsilon_j)=0$, ou seja, $E(\underline{\varepsilon})=\underline{0}$, onde $\underline{0}$ representa um vetor de zeros;
- Os erros são independentes, isto é, $E(\varepsilon_j,\varepsilon_h)=0$ para $j\neq h$;
- Os erros são homocedásticos, isto é, $E(\varepsilon_j^2)=\sigma^2,\ \forall j$;
- Os erros têm distribuição normal, $\varepsilon_i \sim N(0,\sigma^2)$.

As três primeiras pressuposições são necessárias para demonstrar que os estimadores de mínimos quadrados são não-tendenciosos e as cinco primeiras pressuposições permitem demonstrar que tais estimadores são estimadores lineares não tendenciosos de variância mínima (Teorema de Gauss-Markov).

2.4.2.2 Estimativas dos Parâmetros de Acordo com o Método dos Mínimos Quadrados

Sejam $\underline{\hat{\beta}}$ e $\underline{\hat{\varepsilon}}$ os vetores das estimativas dos parâmetros e dos erros, respectivamente, isto é,

$$\underline{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \text{M} \\ \hat{\beta}_{p-1} \end{bmatrix} \quad \text{e} \quad \underline{\hat{\varepsilon}} = \begin{bmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \text{M} \\ \hat{\varepsilon}_n \end{bmatrix}$$

Então, o modelo estimado é:

$$\underline{\hat{Y}} = X\,\underline{\hat{\beta}} \tag{2.35}$$

pois, tem-se que:

$$\underline{\varepsilon} = \underline{Y} - X\,\underline{\hat{\beta}} = \underline{Y} - \underline{\hat{Y}} \tag{2.36}$$

onde:

$$\hat{Y} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \text{M} \\ \hat{Y}_n \end{bmatrix}$$

A soma dos quadrados dos desvios é dada por

$$\underline{Z} = \underline{\varepsilon}'\,\underline{\varepsilon} = \left(\underline{Y}' - \underline{\hat{\beta}}' X'\right)\left(\underline{Y} - X\,\underline{\hat{\beta}}\right) = \underline{Y}'\underline{Y} - \underline{Y}' X\,\underline{\hat{\beta}} - \underline{\hat{\beta}}' X'\underline{Y} + \underline{\hat{\beta}}' X' X\,\underline{\hat{\beta}} \tag{2.37}$$

As matrizes $\underline{Y}' X\,\underline{\hat{\beta}}$ e $\underline{\hat{\beta}}' X'\underline{Y}$ são iguais, pois uma é a transposta da outra e cada uma tem apenas um elemento. Então

$$\underline{Z} = \underline{Y}'\underline{Y} - 2\underline{\hat{\beta}}' X'\underline{Y} + \underline{\hat{\beta}}' X' X\,\underline{\hat{\beta}} \tag{2.38}$$

A função $\underline{Z}$ apresenta ponto de mínimo para os valores de $\underline{\hat{\beta}}$ que tornem sua diferencial identicamente nula, isto é:

$$dZ = -2\left(d\hat{\beta}'\right)X'\underline{Y} + \left(d\hat{\beta}'\right)X' X\,\underline{\hat{\beta}} + \underline{\hat{\beta}}' X' X\,(d\hat{\beta}) \equiv 0 \tag{2.39}$$

Como $\left(d\hat{\beta}'\right)X' X\,\underline{\hat{\beta}} = \underline{\hat{\beta}}' X' X\,(d\hat{\beta})$, por serem matrizes com apenas um elemento e uma ser a transposta da outra, segue-se que:

$$-2\left(d\hat{\beta}^{'}\right)X^{'}\underline{Y}+2\left(d\hat{\beta}^{'}\right)X^{'}X\,\underline{\hat{\beta}}\equiv 0 \tag{2.40}$$

ou

$$\left(d\hat{\beta}^{'}\right)\left(X^{'}X\,\underline{\hat{\beta}}-X^{'}\underline{Y}\right)\equiv 0 \tag{2.41}$$

Portanto, a diferencial de $\underline{Z}$ será identicamente nula para:

$$X^{'}X\,\underline{\hat{\beta}}=X^{'}\underline{Y} \tag{2.42}$$

que é o sistema de equações normais.

Se $X^{'}X$ é não singular, existe a matriz inversa $\left(X^{'}X\right)^{-1}$. Pré-multiplicando os dois membros da equação acima por $\left(X^{'}X\right)^{-1}$, obtém-se:

$$\underline{\hat{\beta}}=\left(X^{'}X\right)^{-1}X^{'}\underline{Y} \tag{2.43}$$

O estimador não viciado uniformemente de mínima variância (UMVU) da variância do erro ε_i, $V(\varepsilon_i)=\sigma^2$, é o quadrado médio dos resíduos, isto é:

$$\hat{V}(\varepsilon_i)=\hat{\sigma}^2=S^2=\frac{\underline{Z}}{n-p}=\frac{\sum_{i=1}^{n}\varepsilon_i^{\,2}}{n-p}=\frac{\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}{n-p} \tag{2.44}$$

Considerando ainda o estimador para $\underline{\hat{\beta}}$ determinado anteriormente, tem-se:

$$\hat{V}(\underline{\hat{\beta}})=\hat{V}[\left(X'X\right)^{-1}X'\underline{Y}] \tag{2.45}$$

Mas, usando a propriedade de variância $\left(V\left(kX\right)\right)=kV(X)k',\ k=\text{constante}$):

$$\hat{V}(\underline{\hat{\beta}})=\left(X'X\right)^{-1}X'\hat{V}(\underline{Y})X\left(X'X\right)^{-1} \tag{2.46}$$

Ainda, por hipótese $V(\underline{Y})=\hat{\sigma}^2 I_n$, logo:

$$\hat{V}(\underline{\hat{\beta}})=S^2\left(X'X\right)^{-1} \tag{2.47}$$

Não há dificuldades computacionais para trabalhar com o procedimento exposto anteriormente, quando se tem variável politômica. Porém quando a variável resposta é dicotômica, aparecem limitações, pois para o problema analisado, tem-se:

$$Y_i = \begin{cases} 1 & \text{com } P(Y_i = 1) = \theta \\ 0 & \text{com } P(Y_i = 0) = 1 - \theta \end{cases}$$

Usando o modelo $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$ em observações binárias e considerando as mesmas como quantitativas, tem-se:

$$E(Y_i) = \sum y_i . P(Y_i = y_i) = 1.P(Y_i = 1) + 0.P(Y_i = 0) = 1.\theta_i + 0.(1 - \theta_i) = \theta_i \quad (2.48)$$

$$E(Y_i^2) = \sum y_i^2 . P(Y_i = y_i^2) = 1^2 . P(Y_i = 1) + 0^2 . P(Y_i = 0) = 1.\theta_i + 0.(1 - \theta_i) = \theta_i \quad (2.49)$$

Assim,

$$E(Y_i) = P(Y_i = 1) = \beta_0 + \sum_{i=1}^{n} \beta_i X_{ij} = \theta_i \quad (2.50)$$

Por definição,

$$V(Y_i) = E[Y_i^2 - E(Y_i)]^2 \quad (2.51)$$

Ou, ainda de outra forma:

$$V(Y_i) = E(Y_i^2) - [E(Y_i)]^2 = \theta_i - \theta_i^2 = \theta_i(1 - \theta_i) \quad (2.52)$$

Portanto,

$$V(Y_i) = \theta_i(1 - \theta_i) \quad (2.53)$$

Assim a condição de variância constante para os resíduos não se verifica (pois, depende de θ_i). Tem-se ainda outra dificuldade, já citada anteriormente, que é o fato do modelo oferecer valores estimados fora do intervalo $0 \leq \theta_i \leq 1$.

Segundo CASTRO JUNIOR (2003), ao contrário da ADLF, a RL não baseia suas predições em escores discriminantes. A RL aborda os mesmos tipos de problemas que a ADLF, e de uma forma mais parecida com a regressão múltipla. A diferença é que a RL prediz diretamente a probabilidade de um evento ocorrer, que pode ser qualquer valor entre zero e um. Os valores preditos devem estar limitados ao intervalo de zero a um, e para definir essa relação, a RL utiliza uma relação entre a variável dependente e as variáveis independentes que se assemelha a uma curva em forma de um S, conforme pode ser visto na figura 2.5.3. Para valores muito baixos da variável independente, a probabilidade se aproxima de zero. À medida que o valor da variável independente aumenta, a probabilidade aumenta rapidamente, mas devido à característica da curva, passa a aumentar lentamente e tende assintoticamente para o valor um, mas nunca o ultrapassa.

2.4.3 A Transformação de Logit

O modelo logit é um modelo de resposta qualitativa, pois é utilizado com o propósito de modelar o comportamento de um tomador de decisão que deve escolher entre um conjunto finito de alternativas. Estes modelos são aplicáveis a um conjunto mais extenso de situações de pesquisa que a análise discriminante.

Conforme, visto no item anterior, o modelo de regressão linear múltiplo mostrou-se inadequado, pois Y é uma variável aleatória dicotômica (binária, Bernoulli).

Assim $Y \sim b(1,\theta) \Rightarrow P(Y = y) = \theta^{y}(1-\theta)^{1-y}$ com $0 \leq \theta_i \leq 1$ e $y = 0,1$

A transformação logit é definida por:

$$\text{logit } P(X) = ln\left[\frac{P(X)}{1-P(X)}\right] \tag{2.54}$$

onde:

$$P(X) = \frac{e^{z}}{1+e^{z}}, \tag{2.55}$$

com

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \text{K} + \beta_p X_p \tag{2.56}$$

Assim,

$$\text{logit } P(X) = ln\left[\frac{\frac{e^{z}}{1+e^{z}}}{1-\frac{e^{z}}{1+e^{z}}}\right] \tag{2.57}$$

$$\text{logit } P(X) = ln\left[\frac{\frac{e^{z}}{1+e^{z}}}{\frac{1+e^{z}-e^{z}}{1+e^{z}}}\right] \tag{2.58}$$

$$\text{logit } P(X) = ln\left[e^{z}\right] = Z \tag{2.59}$$

ou ainda:

$$\text{logit } P(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \text{K} + \beta_p X_p \tag{2.60}$$

Tendo-se apenas uma variável independente, $\underline{X}$, e um conjunto de pares de observações $(X_1,Y_1),(X_2,Y_2),\text{K}\,(X_n,Y_n)$, a expressão 2.60, fica:

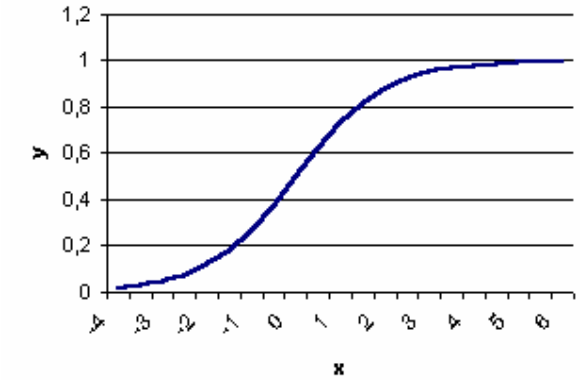
$$\text{logit } P(X) = \beta_0 + \beta_1 X \tag{2.61}$$

Observa-se que esta função é linear e é derivada da função matemática 2.55, que varia monotonamente no intervalo [0,1]. Assim se a função possui um valor crítico

(máximo ou mínimo) usando o logaritmo de $f(X)$ o valor determinado será o mesmo. Ainda, tem-se esta função é simétrica em torno de $Z = 1/2$.

Para a variável dependente Y, relacionada com uma única variável independente X a função é chamada de sigmóide. Assim, quando $X \to -\infty \Rightarrow Y \to 0$, quando $X \to +\infty \Rightarrow Y \to 1$, e quando $X \to 0 \Rightarrow Y \to 1/2$, o que facilmente se observa pela figura 2.5.3.

FIGURA 2.5.3 – GRÁFICO DA FUNÇÃO SIGMÓIDE ASSIMÉTRICA



Fonte: ANNeF - Artificial Neural Networks Framework

2.4.4 Modelo de Regressão Logística

O modelo de RL é, por definição, apropriado para estudos em que a variável de resposta assume valores 0 ou 1 e formula uma equação de relação não-linear entre as variáveis explicativas e a variável de resposta, devido à forma funcional do método, que apresenta funções exponenciais relacionando as variáveis explicativas com a variável de resposta.

O método de RL foi escolhido pelo fato do problema em questão ter uma variável dependente binária, o que torna o método mais apropriado que os demais,

além do fato de ser computacionalmente simples. No entanto, em geral não se pode afirmar qual é o melhor método. Isso depende do problema estudado, da estrutura de dados, das variáveis explicativas disponíveis (inclusive a quantidade de variáveis) e o objetivo da classificação.

Dado a variável aleatória resposta Y assumindo apenas dois resultados possíveis zero ou um. E o vetor $\underline{X}' = [\, X_1, X_2, \mathrm{K}\, , X_p \,]$, um vetor de dimensão p, composto de variáveis aleatórias independentes e ainda tomando-se n observações independentes, podem-se escrever o modelo de RL, na forma:

$$P(X) = \frac{e^{\underline{\beta}'\underline{X}}}{1 + e^{\underline{\beta}'\underline{X}}} \tag{2.62}$$

Onde:

$$\underline{\beta}' = [\beta_0, \beta_1, \mathrm{K}\, , \beta_p] \text{ e } \underline{X} = [1, X_1, X_2, \mathrm{K}\, , X_p]$$

Para SOUZA (2000)**,** RL é uma técnica comumente usada para a análise de dados com resposta binária ou politômica. Normalmente esta análise é realizada usando-se aproximações assintóticas. Quando o tamanho da amostra é pequeno ou os dados são esparsos, a solução assintótica pode não existir, sendo recomendado o uso do método exato. A idéia principal deste método é gerar distribuições de permutações exatas da estatística suficiente dos parâmetros de interesse do modelo de regressão logística, condicionada à estatística suficiente dos parâmetros remanescentes.

2.4.4.1 Modelo de Regressão Logística Simples

Seja a amostra aleatória composta de n pares de observações (X_i, Y_i) com $i = 1,2,\mathrm{K}\, , n$ onde os $\underline{Y}'s$ representam os valores observados de uma variável

dicotômica, e os $\underline{X}'s$ representam os valores observados de uma única variável independente.

Assim a equação 2.62**,** quando $\underline{Y}$ é uma variável dicotômica, e tem-se apenas uma variável independente tornando-se:

$$P(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{2.63}$$

Esta expressão nos fornece a probabilidade condicional de que Y é igual a 1, dado o valor de X, ou seja, $P(Y=1 \mid X)$.

Faz-se necessário estimar os valores para os parâmetros $\beta's$ e então se determina o modelo de RL.

Então a probabilidade condicional de que Y é igual à zero, fica:

$$P(Y=0 \mid X) = 1 - P(X)$$

Portanto, para as n observações têm-se:

$$P(Y_i=1 \mid X_i) = P(X_i) \qquad \text{e} \qquad P(Y_i=0 \mid X_i) = 1 - P(X_i) \tag{2.64}$$

Para fornecer estimativas para os parâmetros que maximizam a probabilidade de obter o conjunto observado de dados, utiliza-se o método da Máxima Verossimilhança.

Uma forma conveniente para expressar a contribuição da função de verossimilhança para os pares (X_i, Y_i) é através da fórmula:

$$F(X_i) = P(X_i)^{Y_i}[1 - P(X_i)^{1-Y_i}] \tag{2.65}$$

Desde que assumido que as observações são independentes, a função de verossimilhança é obtida como o produto dos termos dados na equação 2.66, como segue:

$$l(\beta) = \prod_{i=1}^{n} F(X_i) = \prod_{i=1}^{n} [P(X_i)]^{Y_i} [1 - P(X_i)]^{1-Y_i} \tag{2.66}$$

Segundo LIMA (2002), os estimadores de Máxima Verossimilhança destes parâmetros são escolhidos de forma a maximizar essa função, ou seja, deseja-se determinar o estimador para β que maximize a expressão 2.66. Analisando essa função do ponto de vista matemático é mais fácil trabalhar com o logaritmo da mesma. Assim, tem-se a função de log-verossimilhança dada por:

$$L(\beta) = \ln(l(\beta)) = \ln\left[\prod_{i=1}^{n} F(X_i)\right] = \ln\left[\prod_{i=1}^{n} [P(X_i)]^{Y_i} [1 - P(X_i)]^{1-Y_i}\right] \tag{2.67}$$

assim,

$$L(\beta) = \sum_{i=1}^{n} [Y_i \ln(P(X_i))] + (1 - Y_i) \ln[1 - P(X_i)] \tag{2.68}$$

então,

$$L(\beta) = \sum_{i=1}^{n} \left[Y_i \ln \frac{e^{\beta' X}}{1 + e^{\beta' X}} + (1 - Y_i) \ln\left(1 - \frac{e^{\beta' X}}{1 + e^{\beta' X}}\right)\right] \tag{2.69}$$

logo,

$$L(\beta) = \sum_{i=1}^{n} \left[Y_i \ln \frac{e^{\beta' X}}{1 + e^{\beta' X}} + (1 - Y_i) \ln\left(\frac{1}{1 + e^{\beta' X}}\right)\right] \tag{2.70}$$

mas, $\ln\left(\frac{e^{\beta' X}}{1 + e^{\beta' X}}\right) = \ln e^{\beta' X} - \ln\left(1 + e^{\beta' X}\right) = \beta' X - \ln\left(1 + e^{\beta' X}\right)$ e

$$\ln\left(\frac{1}{1 + e^{\beta' X}}\right) = \ln 1 - \ln\left(1 + e^{\beta' X}\right) = -\ln\left(1 + e^{\beta' X}\right)$$

que substituindo em 2.70 fica

$$L(\beta) = \sum_{i=1}^{n} \left[Y_i \left(\beta' X - \ln\left(1 + e^{\beta' X}\right) \right) + (1 - Y_i)\left(- \ln\left(1 + e^{\beta' X}\right) \right) \right] \tag{2.71}$$

assim,

$$L(\beta) = \sum_{i=1}^{n} \left[Y_i \beta' X - Y_i \, ln\left(1 + e^{\beta' X}\right) - ln\left(1 + e^{\beta' X}\right) + Y_i \, ln\left(1 + e^{\beta' X}\right) \right] \tag{2.72}$$

portanto:

$$L(\beta) = \sum_{i=1}^{n} \left[Y_i \beta' X - ln\left(1 + e^{\beta' X}\right) \right] \tag{2.73}$$

Para determinar o valor de β que maximiza $L(\beta)$ deve-se derivar $L(\beta)$ em relação à β_0 e β_1 igualando o conjunto de resultados à zero. Desta forma, tem-se a seguir as chamadas equações de verossimilhança, que são derivadas da expressão 2.73 em relação à β.

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} \left(Y_i - \frac{e^{\beta' X}}{1 + e^{\beta' X}} \right) = \sum_{i=1}^{n} \left[Y_i - P(X_i) \right] = 0 \tag{2.74}$$

e

$$\frac{\partial L(\beta)}{\partial \beta_1} = \sum_{i=1}^{n} \left(X_i Y_i - X_i \frac{e^{\beta' X}}{1 + e^{\beta' X}} \right) = \sum_{i=1}^{n} X_i \left[Y_i - P(X_i) \right] = 0 \tag{2.75}$$

Verifica-se mediante a aplicação do modelo que a soma dos valores observados de Y é igual à soma dos valores esperados (preditos).

O valor de β dado pela solução das equações 2.74 e 2.75 é chamado estimador de máxima verossimilhança e denotado por $\hat{\beta}$.

No modelo de RL, onde as equações são não-lineares, utiliza-se para estimar os parâmetros métodos iterativos, o que exige a utilização de *softwares* específicos,

STATGRAPHICS, SAS, por exemplo, ou a construção de programas computacionais DELPHI, MATLAB.

2.4.4.2 Modelo de Regressão Logística Múltiplo

Segundo LIMA (2002), o método de estimação para determinar o ajuste do modelo, ou seja, estimar $\beta' = (\beta_0, \beta_1, \mathrm{K}, \beta_p)$ usado no caso multivariado será o mesmo do caso univariado. E a função de verossimilhança é aproximadamente idêntica à dada na equação 2.63 com uma mudança, sendo que $P(X)$ é definido por:

$$P(X) = \frac{e^{\beta' X}}{1 + e^{\beta' X}} \tag{2.76}$$

Para a estimação dos parâmetros oriundos das soluções das equações de verossimilhança é utilizado método de Levenberg-Marquardt descrito abaixo:

Algoritmo:

Dada uma solução atual X_k para o conjunto de parâmetros $\beta's$ do Modelo de Regressão Logístico Múltiplo, tem-se:

Passo 1: Calcular $f(X_k)$;
Passo 2: Escolher um valor modesto para λ $(\lambda = 0.0001 \text{ por exemplo})$;
Passo 3: Resolver o sistema de equações $\sum_{j=1}^{p} \alpha'_{lj} . \Delta X_j = \mathrm{A}_l$ para determinar ΔX e X_{k+1} e avaliar $f(X_{k+1})$;

i) Se $f(X_{k+1}) \geq f(X_k)$ aumentar λ por um fator, 10 vezes (ou um outro valor substancial) e voltar ao passo 3;
ii) Se $f(X_{k+1}) < f(X_k)$ diminuir λ por um fator, 10 vezes (ou um outro valor substancial) e atualizar a solução atual, ou seja, X_{k+1} recebe o valor de X_k, voltar ao passo 3.

O critério de parada foi estabelecido a partir de duas considerações: a primeira se relaciona ao fato que analisando as derivadas parciais da função log-verossimilhança, tem-se que a soma dos observados deve ser igual a soma dos preditos e a segunda está baseada no fato de uma nova solução não trazer alteração no valor da função objetivo, sendo assim o processo é suspenso apenas quando as duas condições descritas anteriormente forem satisfeitas.

2.5 AVALIAÇÃO DA FUNÇÃO DE CLASSIFICAÇÃO

2.5.1 Critério *TPM* (*Total Probability of Misclassification*)

Uma forma de avaliar o desempenho de um procedimento de classificação consiste no cálculo da Taxa de Erro de Reconhecimento (*total probability of misclassifications*) (TPM) sendo dada por:

$$TPM = p_1 \int_{R_2} f_1(x)dx + p_2 \int_{R_1} f_2(x)dx \tag{2.77}$$

onde, p_1 e p_2 são as probabilidades de uma observação pertencer a π_1 e π_2, respectivamente.

O valor mínimo para a quantidade acima, chamado Taxa ótima de Erro (*optimum error rate*) (OER),

$$OER = p_1 \int_{R_2} f_1(x)dx - p_2 \int_{R_1} f_2(x)dx \tag{2.78}$$

é obtido pela escolha adequada das regiões $(R_1 \text{ e } R_2)$, onde as regiões são determinadas por:

$$R_1 : \frac{f_2(x)}{f_1(x)} \geq \frac{p_2}{p_1} \qquad \text{e} \qquad R_2 : \frac{f_2(x)}{f_1(x)} < \frac{p_2}{p_1} \tag{2.79}$$

Uma medida da performance que não depende da forma da distribuição e que pode ser calculada para qualquer procedimento de classificação é a Taxa aparente do erro (que é definida como fração das observações no treinamento amostral). Ela é calculada da matriz de confusão que mostra a situação real das observações nos grupos versus o reconhecimento. Para n_1 observações de π_1 e n_2 observações de π_2, a matriz de confusão tem a forma:

TABELA 2.5.1 – MATRIZ DE CONFUSÃO

Classificação Real	Classificação Prevista	
	π_1	π_2
π_1	$n_{1/1}$	$n_{1/2}$
π_2	$n_{2/1}$	$n_{2/2}$

onde: $n_{i/j}$ = número de observações de π_i classificadas como de π_j e considerada correta se $i = j$ ou incorreta se $i \neq j$.

$n_{1/1}$: número de itens de π_1 corretamente reconhecido como de π_1;
$n_{2/1}$: número de itens π_2 misturados com de π_1;
$n_{2/2}$: número de itens π_2 corretamente reconhecido como de π_2;
$n_{1/2}$: número de itens π_1 misturados com de π_2.

A taxa aparente de erro $(APER)$ é dada por:

$$APER = \frac{n_{1/2} + n_{2/1}}{n_1 + n_2} \tag{2.80}$$

é interpretada como a proporção de itens ou observações no conjunto de treinamento que são reconhecidos incorretamente.

2.5.2 Abordagem de Lachenbruch

É uma técnica para avaliar a eficiência da regra de classificação, e segue os passos:

1º. Comece com o grupo da população π_1.Omita uma observação deste grupo e construa uma função baseada nas $n_1 - 1$ e n_2 observações.
2º. Reconheça (classifique), usando a função, a observação não incorporada.

3º. Repita os passos 1 e 2 até que todas as n_1 observações de π_1 sejam classificadas. Seja $n_{1/2}$ o número de observações reconhecidas erroneamente neste grupo.

4º. Repita os passos 1a 3 para as n_2 observações de π_2. Seja $n_{2/1}$ o número de observações reconhecidas erroneamente neste grupo.

então,

$$\hat{P}(2|1) = \frac{n_{2/1}}{n_1} \tag{2.81}$$

e

$$\hat{P}(1|2) = \frac{n_{1/2}}{n_2} \tag{2.82}$$

e, a proporção total esperada de erro é:

$$\hat{E}(AER) = \frac{n_{1/2} + n_{2/1}}{n_1 + n_2} \tag{2.83}$$

Assim, obtém-se uma regra de reconhecimento e classificação construída com as n observações amostrais e testadas com todas referidas observações.

CAPÍTULO III

3 MATERIAL E MÉTODOS

3.1 CARACTERIZAÇÃO DA AMOSTRA E DAS VARIÁVEIS

Os dados utilizados neste trabalho são provenientes do site da Caixa Econômica Federal (CEF)[4], onde este divulga quais são as famílias beneficiárias do PBF nos últimos quatro meses, neste caso, foram “baixados” e gravados os dados referentes ao mês de agosto de 2005; da Secretaria da Criança e Ação Social da Prefeitura Municipal de Tibagi (Sast) que forneceu o banco de dados do cadastro do PBF. Como estes dados não são abertos, o trabalho virá a somar esforços no controle e a participação social no PBF.

No site da CEF, obtiveram-se os dados referentes às 755 famílias com o benefício liberado. Sendo estes compactados, no formato “zip”. Já a Sast disponibilizaram as informações referentes ao banco de dados do cadastramento das 1449 famílias no formato “cxa” e “zip”.

O formato “cxa” é uma extensão utilizada no aplicativo CadÚnico da CEF e pelos Órgãos responsáveis pelo PBF. O Software pode ser “baixado” junto ao site da CEF.

Com a obtenção dos dados do cadastramento, deu-se início no cruzamento dos mesmos, ou seja, as famílias que estavam recebendo o benefício com aquelas

[4] BRASIL, Caixa Econômica Federal, **Sistema de Benefício Por Município.** Disponível em: <https://webp.caixa.gov.br/sibec/consulta/beneficio/04.01.00-00_00.asp> Acesso em: 08 mar. 2006.

cadastradas no PBF disponibilizados pela Sast. Com a ocorrência disto, teve-se a formação dos grupos: beneficiários e não-beneficiários do PBF. Aqui, utilizou-se uma ferramenta do pacote Microsoft Office chamada de Excel.

Verificam-se, logo após, que as famílias passaram de 1449 para 1372, ou seja, ocorreram grupos de famílias cadastradas mais de uma vez, estes apenas cadastrados não participando do recebimento do benefício.

Considerou-se, para o cruzamento dos dados, o valor da renda mensal per capita familiar, como sendo, de R$ 50,00 para famílias em condições de extrema pobreza e de R$ 100,00 para famílias pobres e extremamente pobres com crianças e jovens entre zero e 16 anos incompletos, ou seja, estes eram os valores impostos pela legislação no período.

O aplicativo Excel é uma ferramenta que permite a edição de diversas extensões como, por exemplo: “xls”, “txt”, etc. Para este caso utilizou-se a abertura de arquivos “txt”, precisando apenas em seguida, ajustar o assistente de importação de texto conforme os dados cadastrados no CadÚnico.

Em seguida, dividiram-se os dados em dois grupos.

- Grupo 1 – Base de dados das Famílias beneficiárias (pop1);
- Grupo 2 – Base de dados das Famílias não-beneficiárias (pop2).

Após a separação dos grupos, deu-se início a definição das variáveis que seriam utilizadas no trabalho, como:

- Total da renda familiar (TRF);
- Total da despesa familiar (TDF);
- Esposo ou companheiro reside no domicílio (CRD);
- Tem algum tipo de deficiência (TD);
- Raça/Cor (RC);
- Freqüenta escola (FE);
- Grau de instrução (GI);
- Série Escolar (SE);

- Situação no mercado de trabalho (SMT);
- Número de pessoas que vivem da renda desta família (NPVR);
- Se grávida, informar o mês de gestação (IMG);
- Amamentando (EAM);
- Participa de algum programa do Governo Federal (PPGF).

Como se observa foram determinadas 13 (treze) variáveis iniciais para o estudo do trabalho, sendo estas determinadas através de estudos criteriosos entre o Mestrando e os Responsáveis dos Órgãos pelo controle e participação no PBF do Município de Tibagi.

Depois da base de dados pronta para ser utilizada, deu-se início a aplicação dos conceitos estatísticos verificados no capítulo II. Verificando a diferença no vetor de médias entre os dois grupos, de acordo com item 2.2.3. Sendo em seguida, aplicado o descarte de pontos (*outlier*), verificado através do método das duas primeiras componentes principais (padronizadas). Com o descarte de observações poderá se perceber se alguns dados foram digitados de forma incorreta.

Após o descarte de observações, continuou-se com o descarte de variáveis, que no início eram treze escolhidas sem critérios estatísticos. Este descarte além de reduzir o número de variáveis e diminuir o custo do levantamento dos dados, ele também evita a colinearidade, que atrapalha a regra de classificação a ser construída.

Para o desenvolvimento deste trabalho, levaram-se a utilização da aplicação de dois conceitos estatísticos (FDLF e MRLM). Estes envolvidos por dois motivos: a necessidade da pesquisa e a fundamentação teórica. Os dois métodos estatísticos aparecem com o objetivo de determinar a melhor maneira de separar os grupos de famílias em beneficiários e não-beneficiários do PBF.

O descarte de *outlier eliminou* duas famílias com dados anormais em algumas das variáveis, passando de 1372 para 1370 pessoas cadastradas. Já o descarte de variáveis, alicerçado na definição 2.2.4.4, conseguiu de forma significativa, uma diminuição do banco de dados eliminando seis variáveis, passando de 1370 x 13 para

1370 x 7. Neste último banco de dados que se deu a aplicação dos conceitos estatísticos (ADLF e MRLM).

3.2 APLICAÇÃO DOS MÉTODOS PROPOSTOS

Com a base dos dados determinada, começou a etapa da aplicação dos métodos propostos: FDLF e MRLM.

O método da FDLF aplicado sobre a matriz de ordem 1370 x 7 resultou na tabela 2.5.5, que traz os resultados da estimação dos seus coeficientes. Já a avaliação da eficiência desta função na classificação foi verificada pela abordagem de Lachenbruch como mostra a tabela 2.5.6.

A aplicação do segundo método no conjunto de dados acima foi verificada pelos resultados alcançados, como mostra a tabela 2.5.8, onde esta traz os coeficientes e os erros padrões estimados para o MRLM, tendo por base os estimadores de máxima verossimilhança obtidos através do método de Levenberg-Marquardt. Já a avaliação da eficiência desta função na classificação foi verificada pela abordagem de Lachenbruch como mostra a tabela 2.5.9.

3.3 RECURSOS UTILIZADOS

Num primeiro momento como citado no índice 3.1, utilizou-se a internet através do site da CEF para fazer *download* do banco de dados das famílias beneficiárias do PBF. Já a Sast disponibilizou o banco de dados através de um disquete de tamanho de 3,5 polegadas com capacidade de armazenamento de 1,44 Mega Bytes.

Com as duas bases de dados, deu-se início a utilização do aplicativo Excel para realizar o cruzamento dos dados. Estes dados foram processados em uma plataforma Microsoft®, utilizando Windows® XP SP2, com processador Pentium® IV, 3000 Mega Hertz, 1 Giga Bytes de Memória (RAM) e placa de vídeo 256 Mega Byte.

Após esta formatação da amostra, deu-se início a alteração na programação das funções no *software* MATLAB® elaborado por José Donizetti de Lima, LIMA (2002), optando por este, pela facilidade de entendimento, pela rapidez e bom desempenho no processamento dos dados, além da estrutura da aplicação dos métodos estatísticos. A adaptação ocorrida no programa levou a eliminação de algumas rotinas desnecessárias para esta dissertação.

Aqui, seguindo o modelo do programa desenvolvido pelo Mestre José Donizetti de Lima, LIMA (2002), optou-se também pela criação de uma rotina de controle e várias sub-rotinas.

Como o trabalho utilizava a aplicação de conceitos estatísticos semelhantes ao do autor citado acima, houve apenas adaptações e aperfeiçoamentos do programa para a aplicação na base de dados do PBF. O código fonte das funções programadas segue no apêndice I.

CAPÍTULO IV

4 RESULTADOS

4.1 RESULTADOS

4.1.1 Resultados da Análise Estatística dos Dados

Os resultados encontrados utilizando a base de dados do PBF foram os seguintes: Inicialmente utilizou-se a primeira função do programa (T^2 de Hotelling), onde esta verifica se as populações, representadas por suas amostras, são distintas em suas várias características médias.

Após selecionar a Função T^2 de Hotelling e definir o nível de significância alfa, o valor de delta e as populações, como mostra a figura 2.5.4. A escolha da análise baseada em matrizes de covariâncias iguais seguiu a teoria citada no tópico 2.2.3 deste trabalho, obtendo como resultados os valores abaixo:

FIGURA 2.5.4 – CAIXA DE DIÁLOGO PARA O TESTE T^2 DE HOTELLING

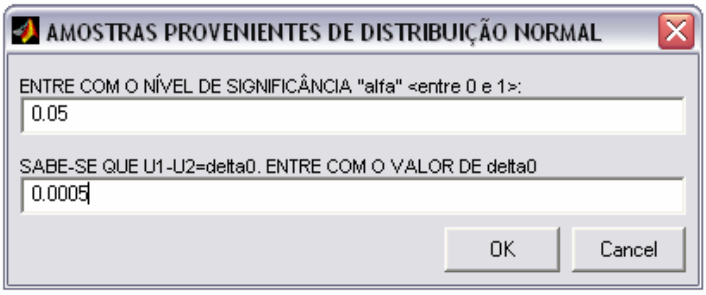


TABELA 2.5.2 – TABELA REFERENTE AOS RESULTADOS DO TESTE T^2 DE HOTELLING

CASO DE IGUALDADE DE MATRIZES DE COVARIÂNCIAS: RESULTADOS PARA ANÁLISE

Nº. de obs.	Nº. de var.	Nº. de elem. pop1	Nº. de elem. pop2.
1372	13	740	632

VALOR ESTATÍSTICO DO TESTE	T^2 AJUSTADO	T^2 TEÓRICO
1773,6347	135,2384	1,7274

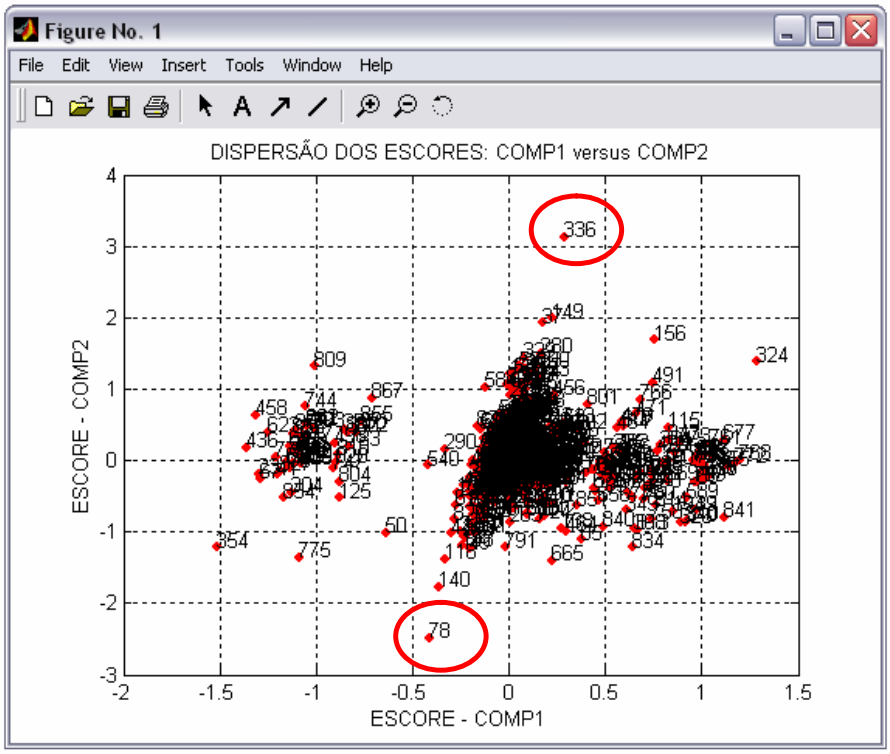
H0 REJEITADO, PARA NÍVEL DE SIGNIFICÂNCIA DE 5%, POIS
T^2 (n1+n2-p-1) / ((n1+n2-2)*p) > Fp,n1+n2-p-1 (alfa)
OU SEJA, 135.2384 > 1.7274

Comparando-se os resultados encontrados para os dois grupos de famílias, tem-se que 135,2384 é muito maior que 1,7274, indicando que eles estão centrados em médias diferentes, ao nível de significância de 5%. Assim rejeita-se H0, ou seja, realmente trata-se de grupos de famílias distintas: Beneficiários e não-beneficiárias do PBF, justificando-se o prosseguimento do trabalho.

Como o cadastramento pode ser disponibilizado em vários locais e por várias pessoas surge à preocupação com a veracidade na coleta dos dados. Dado isso, utilizam-se o descarte de *outlier*, onde este faz a verificação das famílias que possuem dados adversos dos demais. Para as duas primeiras componentes principais padronizadas, adotou-se o descarte das observações que possuíam valores fora do intervalo [-2,2].

Na figura 2.5.5, tem-se a visualização gráfica para os resultados das amostras utilizadas para o descarte de *outlier*. Verifica-se a plotagem dos escores das observações para as duas primeiras componentes principais, mostrando as observações a serem descartadas (*outlier*).

FIGURA 2.5.5 – DESCARTE DE *OUTLIER* VIA ESCORES DAS COMPONENTES PRINCIPAIS



A figura 2.5.5, mostra que as observações a serem descartadas são: 78, 336. Os cálculos foram feitos a partir da matriz de correlação, o que seria análogo ao que se determinaria ao usar variáveis padronizadas.

Em seguida, deu-se a aplicação da função descarte de variáveis, aqui se pode notar a eliminação das variáveis de acordo com a teoria citada no item 2.2.4.4, deste trabalho.

Os resultados encontrados estão na tabela 2.5.3, onde logo abaixo se darão o processo da forma com que foram eliminadas as variáveis.

TABELA 2.5.3 – RESULTADOS DO TESTE – DESCARTE DE VARIÁVEIS

****

*** AUTOVALORES E AUTOVETORES ORDENADOS DA MATRIZ CORRELAÇÃO ***

****

AUTOVETORES CORRESPONDENTES

e1	e2	e3	e4	e5	e6	e7	e8	e9	e10	e11	e12	e13	
-0.3486	-0.0510	0.0558	-0.0631	-0.1043	-0.2542	0.5184	0.2945	0.0450	-0.1916	-0.0105	-0.6174	-0.1334	TRF
-0.3639	-0.0450	0.0354	-0.0256	-0.0895	-0.2395	0.4017	0.2435	0.0093	0.0194	0.0266	**0.7472**	0.1350	TDF
-0.3434	-0.0278	0.0860	0.0173	-0.0908	0.3339	0.3746	0.2282	0.3235	**0.5772**	0.3353	-0.0638	0.0844	CRD
-0.0211	-0.0095	-0.7925	-0.5757	-0.1778	0.0327	0.0206	0.0186	-0.0432	0.0567	0.0362	-0.0036	0.0052	TD
-0.3098	-0.0873	0.0557	-0.0620	-0.0972	0.5384	-0.3712	**0.5891**	0.2025	-0.1258	-0.2173	0.0213	-0.0286	RC
-0.3825	-0.0067	0.0336	0.0093	-0.0261	-0.0166	-0.2020	-0.0331	-0.2901	0.2630	0.2379	-0.2227	**0.7417**	FE
-0.3141	-0.1856	-0.0143	-0.0435	0.1532	0.4903	0.2239	-0.4885	0.2056	-0.3574	0.3745	0.0440	-0.0067	GI
-0.1741	-0.5425	-0.1501	0.0737	0.4138	-0.0225	0.0821	-0.1628	0.1483	0.3944	**-0.5121**	-0.0517	0.0073	SE
-0.3764	-0.0313	0.0382	0.0227	-0.0405	0.0110	-0.1833	-0.0365	-0.3432	0.4346	0.3265	0.0193	-0.6364	SMT
-0.2683	0.3862	-0.0483	-0.0509	0.1164	-0.3342	-0.2723	-0.1083	**0.7259**	0.1460	0.1137	-0.0302	-0.0279	NPVR
-0.0792	0.0620	-0.5205	0.7961	-0.2766	0.0582	0.0386	0.0032	0.0503	-0.0304	0.0006	-0.0190	0.0001	IMG
-0.0260	0.4907	-0.1993	0.0775	0.7391	0.1683	0.1713	0.2441	-0.2202	-0.0125	0.0246	0.0024	-0.0025	EAM
-0.1894	0.5132	0.1368	-0.0971	-0.3139	0.3015	0.2346	-0.3401	-0.0789	0.2169	-0.5089	-0.0135	-0.0022	PPGF

AUTOVALORES

λ 1	λ2	λ3	λ4	λ5	λ6	λ7	λ8	λ9	λ10	λ11	λ12	λ13
5.9728	1.8789	1.0352	0.9647	0.7820	0.5777	0.4849	**0.4002**	**0.3071**	**0.2182**	**0.1847**	**0.1081**	**0.0857**

Como, pode-se verificar em negrito estão os autovalores e os autovetores da matriz de correlação das variáveis independentes que foram utilizados para o processo de eliminação.

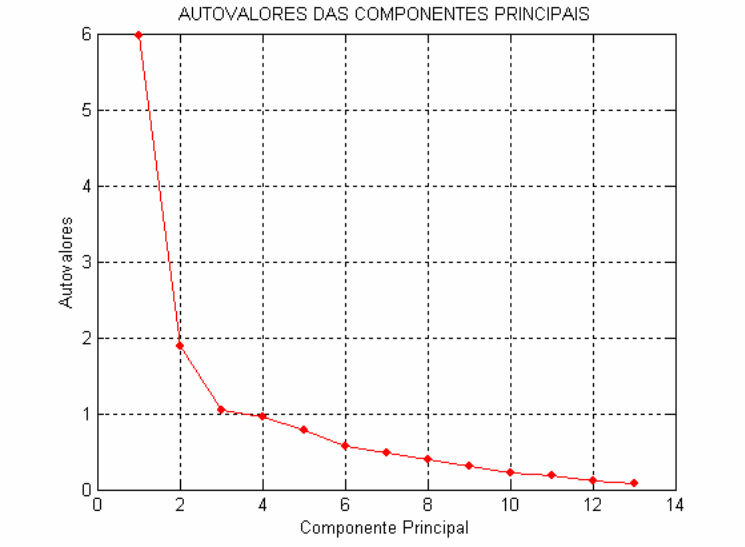
Seguindo o item 2.2.4.4, para a eliminação das variáveis iniciou-se o processo, primeiramente, pelo autovalor de menor valor, sendo este autovalor 0.0857, dirigindo-se em seguida para a matriz de autovetores para encontrar o autovetor em módulo de maior valor, neste caso, 0.7417. Com este valor definido, elimina-se essa linha referente a esta variável, neste caso a sexta da base de dados. Esta variável é: Freqüenta Escola (FE).

A segunda variável a ser eliminada refere-se ao segundo menor autovalor, neste caso 0,1081, que tem como autovetor 0,7472 referente à segunda linha que tem a variável: Total da despesa familiar (TDF);

Seguindo o mesmo raciocínio têm-se as seguintes variáveis em ordem eliminadas:

- Série Escolar (SE);
- Esposo ou companheiro reside no domicílio (CRD);
- Número de pessoas que vivem da renda desta família (NPVR);
- Raça/Cor (RC).

FIGURA 2.5.6 – AUTOVALORES DAS COMPONENTES PRINCIPAIS



Como se visualiza na tabela 2.5.3, o número de variáveis eliminadas foram seis, deixando um restante de sete para a base de dados. Ficando as seguintes variáveis para seres estudadas através da FDLF e MRLM:

- Total da renda familiar (TRF);
- Tem algum tipo de deficiência (TD);
- Situação no mercado de trabalho (SMT);
- Se grávida, informar o mês de gestação (IMG);
- Amamentando (EAM);
- Grau de Instrução (GI);
- Participa de algum programa do Governo Federal (PPGF).

Na tabela 2.5.4, pode-se verificar as porcentagens das variâncias das variáveis explicadas pelos autovalores da matriz de correlação. Verificando-se um valor alto para o primeiro autovalor.

TABELA 2.5.4 – PROPORÇÃO DE VARIÂNCIA EXPLICADA PELOS AUTOVALORES DA MATRIZ CORRELAÇÃO

ORDEM	AUTOVALORES	VAR. EXPL. (EM%)	VAR. EXPL. ACUM. (EM%)
1	5.9658	45.89	45.89
2	1.8772	14.44	60.33
3	1.0350	7.96	68.29
4	0.9649	7.42	75.72
5	0.7815	6.01	81.73
6	0.5773	4.44	86.17
7	0.4827	3.71	89.88
8	0.4020	3.09	92.97
9	0.3067	2.36	95.33
10	0.2178	1.68	97.01
11	0.1847	1.42	98.43
12	0.1185	0.91	99.34
13	0.0858	0.66	100.00

Como o objetivo principal deste trabalho é apresentar duas metodologias para o reconhecimento de padrões das famílias cadastradas no PBF, com relação à liberação do benefício fornecido pelo governo federal, o banco de dados utilizado é composto pelas variáveis acima que não foram eliminadas pelo processo de descarte de variáveis.

4.1.2 Resultados da Função Discriminante Linear de Fisher

Tendo a base de dados formatada após o descarte, dá-se início ao processo da aplicação da FDLF. Aqui, reformulou-se a base para um número menor de variáveis, onde este novo arquivo chamou-se: dados_descartes. Os resultados dos coeficientes estimados da FDLF encontram-se na tabela 2.5.5.

TABELA 2.5.5 – COEFICIENTES ESTIMADOS DA FDLF

VARIÁVEL	COEFICIENTES DA FDLF (C)
TRF	0.0013
TD	0.8669
GI	0.0005
SMT	0.3631
IMG	-0.0440
EAM	1.1011
PPGF	1.9086

A tabela 2.5.5 informa os valores dos coeficientes das variáveis do CadÚnico. Já os valores dos coeficientes multiplicados pelas suas respectivas variáveis formam a função de classificação. Esta serve para alocar o cadastro da família num dos grupos pré-estabelecidos: Beneficiários e não-beneficiários do PBF.

Pode-se visualizar na tabela 2.5.5, os três maiores coeficientes da FDLF: PPGF, EAM e TD, respectivamente.

A função de classificação construída a partir dos resultados da tabela 2.5.5, encontra-se abaixo.

$$Y = (TRF*0.0013) + (TD*0.8669) + (GI*0.0005) + (SMT*0.3631) - (IMG*0.0440) + (EAM*1.1011) + (PPGF*1.9086) \quad (4.1)$$

Para o Município de Tibagi a expressão 4.1, auxilia no controle do PBF, através da verificação das famílias cadastradas, assim como, na alocação de novos cadastros.

A expressão 4.1, informa o cadastro das famílias que deveriam estar sendo beneficiadas e estão fora do programa, assim como, famílias que precisaram algum dia do benefício, receberam, e que não pediram seu desligamento do PBF.

Depois da expressão construída, e o ponto médio estimado, nesse caso $(\hat{m} = 5.5376)$, dá-se a formulação de um escore para cada nova família cadastrada. Se esse escore for maior que o ponto médio $(\hat{m})$, classificar-se-á no grupo das famílias beneficiárias; caso contrário classificar-se-á no grupo das famílias não-beneficiárias.

O percentual corretamente classificado para cada grupo (pop_a_d e pop_a_dn) utilizando a FDLF está mostrado na tabela 2.5.6. Os testes realizados nos grupos originais, informaram resultados satisfatórios quanto a sua alocação.

A probabilidade de classificação correta para a FDLF, segundo a tabela 2.5.6, é de: 92,7%, e a proporção de observações classificadas incorretamente de: 7,3%.

TABELA 2.5.6 – RESULTADOS DE CLASSIFICAÇÃO PARA A FDLF

GRUPOS	pop_a_d certo	pop_a_d errado	pop_a_dn certo	pop_a_dn errado
Nº. de famílias	738	0	533	99
Taxa	100.0 %	0.0 %	84.34 %	15.66 %
PROBABILIDADE DE CLASSIFICAÇÃO CORRETA = 92.7 %				

Já para a tabela 2.5.7, a probabilidade de classificação correta para a FDLF é de 78,4%, avaliada pela Abordagem de Lachenbruch. Quanto maior a taxa de acerto mais eficiente será a regra de classificação, diminuindo a probabilidade de cometer os erros tipo I e II.

Na tabela 2.5.7, verifica-se as probabilidades de classificação para a FDLF, podendo-se notar a probabilidade maior para a classificação de grupos de famílias não-beneficiárias como certa em relação ao grupo de famílias beneficiárias.

TABELA 2.5.7 – RESULTADOS DE CLASSIFICAÇÃO PARA A FDLF UTILIZANDO ABORDAGEM DE LACHENBRUCH

GRUPOS	pop_a_d certo	pop_a_d errado	pop_a_dn certo	pop_a_dn errado
Nº. de famílias	544	194	530	102
Taxa	73.71 %	26.29 %	83.86 %	16.14 %
PROBABILIDADE DE CLASSIFICAÇÃO CORRETA = 78.4 %				

A tabela 2.5.7, fornece como proporção de observações classificadas incorretamente, utilizando a FDLF, a porcentagem de: 21,6%.

4.1.3 Resultados do Modelo de Regressão Logístico Múltiplo

O próximo passo refere-se à aplicação do MRLM na mesma base de dados utilizada na FDLF. Os resultados encontrados nesta parte da pesquisa estão representados pela tabela 2.5.8, referentes aos coeficientes e os erros padrões estimados para o MRLM. Verificando-se que os três maiores coeficientes do MRLM são os mesmos da FDLF, ou seja, PPGF, EAM e TD.

TABELA 2.5.8 – COEFICIENTES ESTIMADOS DO MRLM

VARIÁVEL	COEFICIENTE DO MRLM	ERRO PADRÃO
CONSTANTES	-6.4113	1.0382
TR	0.0020	0.009
TD	0.7950	0.9503
GI	-0.0583	0.0705
SMT	0.3328	0.0312
IMG	-0.1301	0.0988
EAM	1.5024	0.1360
PPGF	2.4001	0.2806

A tabela 2.5.8, informa que a função de classificação para o MRLM é:

$$Z = - 6.4113 + (0.0020*TRF) + (0.7950*TD) – (0.0583*GI) + (0.3328*SMT) – (0.1301*IMG) + (1.5024*EAM) + (2.4001*PPGF) \quad (4.2)$$

Como, pode-se notar a expressão 4.2, resultam nas mesmas informações da expressão 4.1, no caso na utilização desta para o controle do PBF. As duas expressões trazem a mesma aplicabilidade, diferenciando apenas no método escolhido.

A aplicação da expressão $P(X) = \frac{e^{z}}{1+e^{z}}$, com os novos valores dos coeficientes estimados resulta na probabilidade da nova família cadastrada pertencer ao grupo beneficiário do PBF. O valor adotado para o *cut off score*, o qual maximiza a probabilidade de acerto, foi igual a 0,5.

A probabilidade de classificação correta para o MRLM é de 79,6%, avaliada pelo método de Lachenbruch. Na tabela 2.5.9, verifica-se as probabilidades de classificação para a MRLM.

TABELA 2.5.9 – RESULTADOS DE CLASSIFICAÇÃO PARA O MRLM UTILIZANDO ABORDAGEM DE LACHENBRUCH

GRUPOS	pop_a_d certo	pop_a_d errado	pop_a_dn certo	pop_a_dn errado
Nº. de famílias	536	202	554	78
Taxa	72.63%	27.37%	87.66%	12.34%
PROBABILIDADE DE CLASSIFICAÇÃO CORRETA = 79.6%				

Também, visualiza-se na tabela 2.5.9, a probabilidade maior de classificação do grupo de famílias não-beneficiárias em relação ao grupo de famílias beneficiárias.

Na tabela 2.5.9, a proporção de observações classificadas incorretamente é de: 20,4%, menor que a classificação utilizando a FDLF.

CONCLUSÃO

A aplicação dos conceitos estatísticos, como FDLF e do MRLM, na amostra, gerou índices e pesos que refletem com segurança informações necessárias para obter conhecimento e elaborar estudos futuros.

O desenvolvimento de pesquisas envolvendo a área social é de grande valia para a sociedade. Os estudos realizados neste trabalho indicaram o aproveitamento deste, e a aplicação de conhecimentos no aprofundamento do desenvolvimento e controle de programas sociais no país.

Segundo o site OPOVO Online, “O Ministério do Desenvolvimento Social e Combate à Fome (MDS), responsável pela execução do Programa Bolsa Família, vai notificar as famílias de crianças que não cumpriram a freqüência escolar entre fevereiro e abril, alertando-as de que podem perder o benefício, caso os estudantes continuem faltando nos meses subseqüentes”.

Com a informação obtida no parágrafo anterior e os dados da tabela 2.5.3, referente aos resultados do teste de descarte de variáveis, ficou claro a falta de importância dada à questão educacional. Sendo a variável “freqüenta escola (FE)”, a primeira a ser eliminada, seguindo em terceiro lugar pela “série escolar (SE)”, significando que as variáveis descartadas através das técnicas das componentes principais comprovam que os conceitos aplicados configuram com a realidade do sistema.

Identificou-se, através da tabela 2.5.4, que a porcentagem explicada para as sete variáveis escolhidas inicialmente para a elaboração do trabalho refletiu um valor de 89,88%, ou seja, utilizou-se uma porcentagem considerada para o desenvolvimento do mesmo.

Os coeficientes encontrados pelos testes estatísticos determinaram valores grandes e positivos para as variáveis PPGF, EAM e TD. Já a variável IMG apresentou valores negativos para seus pesos, tanto para a FDLF quanto para o MRLM. Sendo importantes para a classificação ou alocação de um novo cadastro.

As probabilidades de classificação correta para a FDLF e o MRLM foram de 78,4% e 79,6%, respectivamente. Ambas avaliadas pelo método de Lachenbruch. Com isso, verificou-se a proporção de observações classificadas incorretamente para os dois métodos, FDLF e MRLM, sendo as probabilidades iguais a: 21,6% e 20,4%, respectivamente.

Como se verificou, os métodos aplicados informaram valores próximos para as probabilidades de classificação correta indicando que as técnicas estatísticas utilizadas comprovaram seus conceitos. Também, através disso, pode-se escolher o método a adotar no tratamento dos dados pelo gestor do PBF do município de Tibagi.

Os resultados encontrados com a aplicação destes conceitos ajudam no controle e desenvolvimento da região, através do conhecimento da realidade das famílias cadastradas no PBF, obtendo assim, informações sociais que possam auxiliar na criação de novos programas que venham a somar aos esforços do governo federal no combate a fome, a pobreza, as desigualdades sociais e a inclusão social.

Conforme se notou nos resultados encontrados, a aplicação de técnicas estatísticas tem importância fundamentada na explicação e orientação para o desenvolvimento de trabalhos ligados a base de dados, informando tanto preventivamente como controlando possíveis aplicações de recursos nas áreas envolvidas.

No trabalho pôde-se observar, desde a primeira aplicação, o sucesso das informações obtidas. Os resultados encontrados para os testes, T^2 de Hotelling, o descarte de *outlier* e o descarte de variáveis, mostraram condições necessárias para que a aplicação de técnicas futuras como: aplicação da FDLF e do MRLM pudesse ser

desenvolvida. Estes testes determinaram o seguinte: O primeiro teste, T^2 de Hotelling verificou a existência de diferenças significativas entre os grupos formados pelas famílias beneficiárias ou não do PBF, rejeitando $H0$. O segundo, descarte de *outlier* eliminou dois grupos que possuíam dados do cadastro diferentes dos demais e o terceiro eliminou algumas variáveis do banco de dados, passando de 1370x13 para 1370x7, diminuindo o tempo computacional.

O MRLM foi o método que apresentou melhores resultados em relação a dois itens: a probabilidade de classificação correta no acerto global (79,6%) e a probabilidade de classificar uma família do grupo não-beneficiário como pertencente ao mesmo (87,66%). Mas foi a FDLF que indicou probabilidade mais alta para classificar famílias cadastradas beneficiárias como pertencentes ao mesmo grupo (73,71%).

Com o conhecimento da situação das famílias beneficiárias, os órgãos envolvidos com programas sociais podem formular programas próprios de transferência de renda somando-se ao esforço do Governo Federal.

A informações dos resultados encontrados neste trabalho, dirige-se ao município de Tibagi, por se tratar da sua base de dados do CadÚnico. Isto não impede a aplicação deste trabalho em outras regiões.

Em trabalhos futuros, pode-se fazer um comparativo entre os métodos estatísticos aqui aplicados, e os demais métodos existentes, por exemplo: Redes Neurais e Programação Matemática, assim como aplicá-los em outras regiões do país.

REFERÊNCIAS

AGÊNCIA FOLHA, **Conselho aponta fraudes no Bolsa Família em MG.** Disponível em:<http://www1.folha.uol.com.br/folha/brasil/ult96u74834.shtml> Acesso em: 08 mar. 2006.

ALVES, V. **Avaliação de imóveis urbanos baseada em métodos estatísticos multivariados.** Curitiba. 2005.134 f. Dissertação (Mestrado em Métodos Numéricos em Engenharia) – Setores de Tecnologia e de Ciências Exatas, Universidade federal do Paraná.

BARROSO, L.P.**Análise Multivariada.** 48ª Reunião da RBRAS e 10º SEAGRO – 7 a – Lavras MG. Departamento de Ciências e Exatas. Universidade Federal de Lavras.

BERNARDO, D.C. dos R; SALAZAR. G. T. **PROGRAMA BOLSA FAMÍLIA: VALORIZANDO AS PARCERIAS E AS SINGULARIDADES REGIONAIS.** Disponível em: <http://www.achegas.net/numero/vinteedois/denise_e_german_22.htm> Acesso em: 08 mar. 2006.

BITTENCOURT, J.R. **Artificial Neural Networks Framework.** Disponível em: <http://www.inf.unisinos.br/~jrbitt/annef/html/lib.htm> Acesso em: 08 de mar. 2006.

BRASIL, Ministério do Desenvolvimento Social e Combate à Fome. **Cria o Programa Bolsa Família e dá outras providências.** Lei nº. 10.836, de 09 de janeiro de 2004. Disponível em:<http://www.mds.gov.br/bolsafamilia/Lei_Bolsa_Familia.pdf> Acesso em: 08 mar. 2006.

BRASIL, Ministério do Desenvolvimento Social e Combate à Fome. **Regulamenta a Lei nº.10.836, de 9 de janeiro de 2004, que cria o Programa Bolsa Família, e dá outras providências.** Decreto nº. 5.209, de 17 de setembro de 2004. Disponível em:<http://www.mds.gov.br/bolsafamilia/Decreto_Bolsa_Familia.pdf> Acesso em: 08 mar. 2006.

BRASIL, Ministério do Desenvolvimento Social e Combate à Fome (MDS). **Programas Sociais**. Disponível em: < http://www.mds.gov.br/ Acesso em: 08 mar. 2006.

BRASIL, Ministério da Saúde, Apresentação do PBF em Eventos, **Coordenação Geral da Política de Alimentação e Nutrição (CGPAN)**. Disponível em:<http://dtr2004.saude.gov.br/nutricao/evento/reuniao_nacional/documentos/apresentacao_mds.pdf>. Acesso em: 08 mar. 2006.

BRASIL, Ministério do Desenvolvimento Social e Combate à Fome (MDS). **QUADRO 2.1.1 – Demonstrativo do PBF por UF.** Disponível em: <http://www.mds.gov.br/ascom/bolsafamilia/bf_poruf_part.pdf>. Acesso em: 08 mar. 2006.

BRASIL, Caixa Econômica Federal, **Sistema de Benefício Por Município.** Disponível em: <https://webp.caixa.gov.br/sibec/consulta/beneficio/04.01.00-00_00.asp> Acesso em: 08 mar. 2006.

BRASIL, Caixa Econômica Federal, **Manuais Operacionais – Manual de Importação da Base Caixa – Versão 6.0.** Disponível em: <http://www1.caixa.gov.br/cidade/asp/personaliza/iPaginaRedesenho.asp?pagina=4560000456> Acesso em 08 mar. 2006.

BRASIL, Caixa Econômica Federal, **Cadastramento único – Download.** Disponível em:http://www1.caixa.gov.br/cidade/asp/personaliza/iPaginaRedesenho.asp?pagina=4560000456> Acesso em: 08 mar. 2006.

BRASIL, Ministério do Desenvolvimento Social e Combate à Fome (MDS**). QUADRO 2.1.2 – Demonstrativo da Transferência de Renda às Famílias do PBF por UF.** Disponível em: <http://www.mds.gov.br/ascom/bolsafamilia/transfrenda_uf.pdf> Acesso em: 08 mar. 2006.

BRASIL, **Lei de Responsabilidade Fiscal.** Disponível em:<http://www.tesouro.fazenda.gov.br/hp/lei_responsabilidade_fiscal.asp> Acesso em: 08 mar. 2006.

BRASIL, Secretaria de Estado do Trabalho, Emprego e Promoção Social (SETP), **Gestão de Benefícios do Programa Bolsa Família.** Disponível em: <http://www.setp.pr.gov.br/caixa/uteis/infBF7.pdf > Acesso em: 27 mar. 2006

BRAÚLIO, S. N. **Proposta de uma metodologia para a avaliação de imóveis urbanos baseado em métodos estatísticos multivariados.** Curitiba. 2005.134 f. Dissertação (Mestrado em Métodos Numéricos em Engenharia) – Setores de Tecnologia e de Ciências Exatas, Universidade federal do Paraná.

CHAVES NETO, A . **Análise Multivariada aplicada à pesquisa:** Notas de aula. Departamento de Estatística, Universidade Federal do Paraná, Curitiba, 1997.

FERREIRA, D.F. **Análise Multivariada:** Material Didático**.** Disponível em: <http://www.dex.ufla.br/danielff/dex522.pdf> Acesso em: 08 mar. 2006.
FIGURA 2.5.1 – **Representação Geométrica das Componentes Principais.** Disponível em:<http://www.cienciasgeodesicas.ufpr.br/boletim/pdf/bcg9-1/5Art91_2.pdf >Acesso em: 08 mar.2006.

GALVES, C. Manual de economia política atual. 14. ed. Rio de Janeiro:Forense Universitária, 1996.

GAZETAWEB. **Governo negocia verba do Bid para o Bolsa Família.** Disponível em: <http://gazetaweb.globo.com/Canais/Noticias/Noticias.php?n=108760> Acesso em: 27 mar. 2006.

GROBE, J. R. **Aplicações da Estatística Multivariada na Análise de Resultados em Experimentos com Solo e Animais.** Curitiba. 2005.134 f. Dissertação (Mestrado em Métodos Numéricos em Engenharia) – Setores de Tecnologia e de Ciências Exatas, Universidade federal do Paraná.

GRUPO BANCO MUNDIAL, **Apoio ao Bolsa Família.** Disponível em: <http://www.obancomundial.org/index.php/content/view_document/2004.html> Acesso em: 08 de mar. 2006.

HOFFMANN, R. **Análise de regressão: uma introdução à econometria.** São Paulo: Hucitec, 1977.

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

LIMA, J.D. de. **A Análise Econômico-Financeira de Empresas sob a Ótica da Estatística Multivariada.** Curitiba. 2002.167 f. Dissertação (Mestrado em Métodos Numéricos em Engenharia) – Setores de Tecnologia e de Ciências Exatas, Universidade federal do Paraná.

MARDÍA, K. V.; KENT, J.P.; BIBBY, J.M. **Multivariate analysis,** London: Academic Press, 1979, p.175-178.

NETO, J.M. MOITA. **Estatística Multivariada Uma visão Didática Metodológica.** Disponível em: < http://www.criticanarede.com/cien_estatistica.html> Acesso em: 08 mar. 2006.

NUNES, J.C. **Bolsa Família completa um ano com 5 milhões de famílias beneficiadas.** Radio Brás**,** Brasília, DF, 20 out. 2004. Disponível em: <http://www.radiobras.gov.br/materia_i_2004.php?materia=204242&editoria=&q=1> Acesso em: 08 mar. 2006.

OPOVOOnline, **Prazo para notificar freqüência escolar ao Bolsa Família começa dia 31.** Disponível em:<http://www.opovo.com.br/brasil/614923.html > Acesso em: 23 jul. 2006.

SOUZA, M.C.F.M. de C. e **Regressão Logística Exata para Dados de Resposta Binária.** Disponível em: <http://www.est.ufmg.br/posgrad/dissert10.html > Acesso em: 08 mar. 2006.

VIEIRA, E. Mínimos Sociais e Seguridade Social. **O Estado e a sociedade civil perante o Eca e a Loas,** nº.56, p.26-29, mar. 1998.