UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

MÉTODO DE ORIENTAÇÃO À MODELAGEM DE DADOS MENSURADOS EM PROPORÇÃO

Ângelo Márcio Oliveira Sant'Anna

Porto Alegre, 2006

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

MÉTODO DE ORIENTAÇÃO À MODELAGEM DE DADOS MENSURADOS EM PROPORÇÃO

Ângelo Márcio Oliveira Sant'Anna

Orientador: Prof.ª Carla Schwengber ten Caten, Dr.ª

Banca Examinadora:

José Luis Duarte Ribeiro, Dr.
Prof. Depto. de Engenharia de Produção e Transporte / UFRGS

Liane Werner, Dr.ª
Prof. Depto. de Estatística / UFRGS

Flávio Augusto Ziegelmann, Ph.D.
Prof. Depto. de Estatística / UFRGS

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Produção como requisito parcial à obtenção do título de
MESTRE EM ENGENHARIA DE PRODUÇÃO

Área de concentração: Engenharia da Qualidade

Porto Alegre, Março / 2006

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Engenharia de Produção e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação em Engenharia de Produção.

_____

Prof. ª Carla Schwengber ten Caten, Dr.ª
PPGEP / UFRGS
Orientadora

_____

Prof. Luis Antônio Lindau, Ph.D.
Coordenador PPGEP / UFRGS

Banca Examinadora:

José Luis Duarte Ribeiro, Dr.
Prof. Departamento de Engenharia de Produção e Transporte / UFRGS

Liane Werner, Dr.ª
Prof.ª Departamento de Estatística / UFRGS

Flávio Augusto Ziegelmann, Ph.D
Prof. Departamento de Estatística / UFRGS

Dedico este trabalho ao meu pai Walmir, pelo carinho, amor, compreensão, incentivo, estímulo e orientações. Minha gratidão!

AGRADECIMENTOS

A Deus e aos Orixás que me deram forças para vencer mais uma etapa.

À minha mãe, Marinalva, e meu padrasto, Waldomiro, pelo amor incondicional, pelos valores transmitidos, incentivo, esforço e luta desde os meus primeiros passos.

Aos meus irmãos, em especial a Valmar, pelo carinho e estímulo; a minha cunhada Cleise.

À minha dinda, Naty, e meu tiozão, Clóvis, pelo carinho e incentivo em todos os momentos.

A minha namorada Vanessa, pela compreensão, carinho, amor e companheirismo.

À minha orientadora Carla ten Caten, pelo estímulo, orientação e visão prática da Engenharia.

À professora Lia Guimarães, pela confiança, incentivo, amizade, pela oportunidade de aprender sobre a Ergonomia e pelo carinho incondicional em todos os momentos.

Aos professores do Programa de Mestrado em Engenharia de Produção – UFRGS, por suas contribuições.

A família Carballo, em especial a Mariana, pelo apoio, atenção e afeto de todos, quando dos meus primeiros momentos nesta cidade.

Ao meu grande amigo Gustavo (guguinha), pela sua amizade, afeto e apoio; amigo em todos os momentos.

Ao meu grande amigo/irmão Alex (Léo), pela sua amizade, pelo carinho e pelas sugestões e participação, mesmo que distante.

A todos os amigos e colegas do LOPP - PPGEP, pelo companheirismo, e bom convívio nestes dois anos, em especial a Cleber, Cristiane, Fabiano, Flávia, Leandro e Morgana, pela amizade e alegrias compartilhadas.

Aos amigos, Mara, Fernanda (baiana), Gustavo (gugão), Daniel e Cristóvão, pelos momentos de amizade e parcerias.

A todos que participaram direta e indiretamente da concretização deste ideal e conclusão deste trabalho.

“A única coisa que interfere com meu aprendizado,

é a minha educação”.

Albert Einstein

“Procure ser um homem de valor,

em vez de procurar ser um homem de sucesso”.

Albert Einstein

“Nenhuma conquista é importante,

quando não se tem alguém para dividi-las”.

Roger Martin

RESUMO

A implementação de técnicas estatísticas, como modelos de regressão, permite conhecer os efeitos dos fatores sobre a característica de qualidade de um produto, contribuindo na melhoria da qualidade de produtos e processos. O objetivo desta dissertação consiste em elaborar um método que oriente à modelagem de dados mensurados em proporção, levando em consideração a classificação das variáveis dependentes e independentes, com enfoque no Modelo de Regressão Beta e no Modelo de Quase-verossimilhança. O método é ilustrado com um estudo em uma empresa curtidora da região do Vale do Rio dos Sinos no Rio Grande do Sul. A modelagem realizada neste estudo referiu-se a proporção de produtos refugados no processo de produção por erro de classificação. Os Modelos de Regressão Beta e de Quase-verossimilhança apresentaram bom ajuste e mostraram-se adequados na modelagem da proporção de produtos por erros de classificação. Esses modelos podem ser estendidos a todos os processos industriais que envolvam a produção de produtos não conformes às especificações de fabricação (defeituosos). O método elaborado apresentou facilidade de entendimento e clareza dos passos para a escolha dos modelos de regressão usados na modelagem de dados mensurados em proporção.

Palavras-chave: Modelos de Regressão, Proporção, Modelo de Regressão Beta, Modelo de Quase-verossimilhança, Controle de Qualidade.

ABSTRACT

The implementation of statistical techniques, as regression models, allows to know the effects of the factors on the characteristic of quality of a products, contributing in the improvement of the quality of products and processes. The objective of this dissertation consists of elaborating a method to guide the modelling of data measured in proportion, taking into account the classification of the dependent and independent variables, with focus in Beta Regression Model and in the Quasi-likelihood Model. The method is illustrated with a study in a company of leather of the area of the valley of Rio of the Bells in Rio Grande do Sul. The modelling accomplished in this study referred the proportion of products rejected in the production process by mistake of classification. Beta Regression Model and Quasi-likelihood Model presented good adjustment and were shown appropriate in the modelling of the proportion of products for classification mistakes. These models can be extended the all of the industrial processes that involve the production of products out-of-specifications (defective). The elaborated method presented easiness and clarity of the steps for choice of the regression models used in the modelling of data measured in proportion.

Key word: Regression Models, Proportion, Beta Regression Model, Quasi-likelihood Model, Quality Control.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

O cenário mundial atual é de intensa competitividade devido ao desenvolvimento tecnológico rápido de produtos e processos, visando buscar itens que tenham características de qualidade sem defeito. Esta competitividade vem obrigando as empresas a aprimorarem-se rápida e progressivamente na implementação de técnicas e conhecimentos científicos para fazer frente ao crescimento constante da competição. Também é de conhecimento que, em processos de manufatura, a implementação de técnicas permite eliminar desperdícios, reduzir os índices de produtos refugados, diminuir a necessidade da realização de inspeção e aumentar a satisfação dos clientes.

Em um processo de manufatura, pode ser definido um conjunto de causas ou fatores que tem como objetivo produzir determinado efeito e que apresenta uma ou mais respostas observáveis, por exemplo, um produto conforme às especificações recomendadas. Muitas vezes não se conseguem controlar todas as causas de variação, pois certas causas são inerentes ao processo (Montgomery, 2001). Causas de variação que interferem num processo podem gerar a produção de itens não conformes às especificações preestabelecidas, os quais podem ser mensurados avaliando-se a sua proporção.

A estatística objetiva explicar por que, eventualmente, ocorre a produção de itens não conformes ou defeituosos e descobrir que causas poderiam estar influenciando tal produção. Deseja-se também saber em quanto cada causa afeta o resultado. Tais questionamentos conduzem ao problema de construção de um modelo de regressão em que a variável dependente, que descreve a proporção de produtos não conformes (defeituosos), é uma variável dependente contínua. Esse fato afeta a escolha de um modelo de regressão.

Segundo Montgomery e Peck (1992), modelos de regressão consistem numa técnica estatística de investigação e modelagem que relaciona a variável dependente a demais variáveis independentes. Assim, deseja-se descrever os efeitos de um conjunto de informações adicionais, chamados de variáveis explicativas ou independentes, sobre a proporção de produtos não conformes (defeituosos), e a modelagem desses efeitos pode ser uma estratégia eficiente. Conforme Hamada e Nelder (1997), um modelo de regressão que apresenta um bom ajuste usualmente permite gerar boas estimativas das probabilidades dos efeitos associados à variável dependente.

Segundo Cox (1996), a modelagem da proporção em um determinado conjunto de observações, por meio de um modelo de regressão linear normal, nem sempre é recomendada, uma vez que este modelo requer a suposição de que as proporções seguem a distribuição normal. Segundo Kieschnick e McCullogh (2003), o uso do modelo de regressão linear normal na modelagem de proporções ou frações como variável dependente, é um modelo falho, pois possibilita a previsão de valores fora do limite do intervalo [0,1].

Os Modelos Lineares Generalizados apresentam-se como uma nova forma de investigação e modelagem de dados em proporção. Conforme Myers *et al.* (2002), a teoria dos Modelos Lineares Generalizados apresenta opções para a distribuição da variável dependente, permitindo que dados provenientes de uma distribuição de probabilidade Binomial possam ser modelados usando a distribuição original dos dados.

Outra forma de relacionar a variável dependente e demais independentes, num processo de investigação e modelagem de dados, foi proposta por Ferrari e Cribari-Neto (2004), cuja estrutura de regressão baseia-se na suposição de que os dados mensurados em proporção seguem a distribuição de probabilidade Beta. Este procedimento é chamado de Modelo de Regressão Beta.

1.1 TEMA

O tema desta dissertação contempla modelos de regressão utilizados na modelagem de dados mensurados em proporção, ou seja, variável dependente contínua restrita no intervalo [0,1], mais especificamente o Modelo de Regressão Beta (MRB) e o Modelo de Quase-verossimilhança (MQV), que é pertencente à classe dos Modelos Lineares Generalizados (MLG).

1.2 OBJETIVOS

1.2.1 *Objetivo Geral*

O objetivo do trabalho consiste em elaborar um método que oriente à modelagem de dados mensurados em proporção, levando em consideração a classificação das variáveis dependente e independentes, com enfoque no Modelo de Regressão Beta e no Modelo de Quase-verossimilhança.

1.2.2 *Objetivos Específicos*

Pretende-se adicionalmente alcançar os seguintes objetivos específicos:

- Aplicar o método de orientação ao processo de produção de uma empresa curtidora de couro.
- Avaliar o ajuste e a adequabilidade dos modelos de regressão, baseando-se nas técnicas de diagnóstico.
- Comparar o Modelo de Regressão Beta e o Modelo de Quase-verossimilhança, identificando vantagens e desvantagens desses modelos.

1.3 JUSTIFICATIVA DO TEMA E OBJETIVO

A modelagem de um conjunto de informações é parte de um processo científico e uma maneira de aprender a respeito do comportamento de processos é investigar a influência de possíveis efeitos. Com isso, a abordagem de modelos de regressão vem despertando crescente interesse no meio industrial. Os modelos mais conhecidos são usados com variáveis dependentes contínuas, sem que estas apresentem restrições nos valores mensurados.

O uso de ferramentas estatísticas, como modelos de regressão, auxilia no controle e na melhoria da qualidade dos processos de manufatura (PARK, 1996), permitindo investigar possíveis efeitos na produção de produtos não conformes às especificações.

Uma observação importante sobre os modelos de regressão é que os dados apresentem validade sob certas suposições, como, por exemplo, um tamanho de amostra consideravelmente grande. No entanto, em virtude do tipo de processo que se tem interesse em investigar, como um processo destrutivo ou um processo complexo de coleta de dados, o tamanho de amostra obtido é pequeno e, conseqüentemente, as estimativas dos parâmetros e capacidade de previsão podem sofrer distorções.

Segundo Kieschnick e McCullogh (2003), um modelo de regressão linear normal utilizado na investigação de valores em proporção ou fração como variável dependente é um modelo falho, pois não satisfaz as pressuposições necessárias ao uso, produzindo: (*i*) não normalidade do termo de erro; (*ii*) heterocedasticidade, ou seja, não homogeneidade de variância dos valores e (*iii*) possibilidade de a probabilidade estimada (proporção predita) estar fora do limite do intervalo [0,1].

A partir dessas constatações, necessita-se de um modelo de regressão que possua flexibilidade de adaptação para a distribuição de probabilidade da variável dependente. Este trabalho apresenta dois modelos de regressão que contemplam a adaptação aos dados em proporção.

O primeiro é o Modelo de Quase-verossimilhança, pertencente à classe dos Modelos Lineares Generalizados. Segundo McCullagh e Nelder (1989), os Modelos Lineares Generalizados apresentam um leque de opções para a distribuição da variável dependente, permitindo a escolha de uma distribuição de probabilidade para o ajuste adequado do modelo aos dados. Assim, dados provenientes de uma distribuição de probabilidade Binomial ou de Poisson podem ser modelados usando a distribuição original dos dados. Por conseguinte, não há necessidade de pressuposição de normalidade aos dados (CORDEIRO, 1986; DOBSON, 1990 e MYERS *et al.*, 2002).

O segundo é o Modelo de Regressão Beta, que apresenta características importantes na modelagem de dados, a saber: (*i*) distribuição de probabilidade que melhor se ajusta aos dados em proporção; (*ii*) não normalidade do termo de erro; (*iii*) variabilidade dos dados não constante; (*iv*) probabilidade estimada (proporção predita) contida no intervalo [0,1] e (*v*) modelagem adequada dos dados em proporção para tamanho de amostra pequeno (TORRES, 2005), que se fazem de fundamental importância. Segundo Martínez (2004), esse modelo

permite gerar estimativas precisas e seguras dos parâmetros, sem necessidade de violar pressuposições para uso.

Os modelos de regressão Beta e de Quase-verossimilhança são indicados para resolver problemas de modelagem de dados quando mensurados em proporção. Além de serem menos conhecidos, suas descrições na literatura não é ampla: mesmo em artigos que, por vezes, apresentam os modelos, não são detalhados aspectos importantes da análise, tais como propriedades dos modelos, estatísticas e gráficos indicados para verificar a sua adequação.

O desenvolvimento de um método que oriente à modelagem de dados mensurados em proporção, conforme certas características relevantes, é importante, pois o método permite que nos experimentos realizados e que serão analisados, sejam escolhidos adequadamente quais modelos podem ser utilizados. Conforme Hair *et al.* (1998), para o uso da abordagem estatística em experimentos e análise de um processo, é necessário previamente possuir uma idéia do que será estudado, de como os dados serão coletados, da natureza dos dados (discretos ou contínuos) e um entendimento qualitativo de como serão analisados.

No estudo de aplicação deste trabalho, o processo de classificação de couro é um fator crítico na empresa curtidora, uma vez que a subjetividade na classificação das especificações gera condições para uma maior variabilidade nos seus critérios, o que conduzem a refugos por erros de classificação no produto final. Os produtos rejeitados (refugados), gerados por defeitos no processo de manufatura, constituem o principal problema qualitativo das empresas no meio industrial. Segundo Helfer (1991), empresas do ramo de curtimento de couro vêem este problema agravado pela crescente complexidade dos produtos e pelas exigências dos clientes. Desta forma, começam a ser colocadas exigências que apontam a um maior e melhor controle dos processos.

Arriba (2005) relata que os defeitos devidos a uma escolha errada da matéria-prima, se traduzem para o processo como refugos. Os produtos produzidos, a partir destes refugos, são vendidos por um preço muitas vezes até 50% menor que o produto produzido pela matéria-prima original. Assim, quanto mais precisa for a classificação inicial da matéria-prima, menor o risco de ter refugos por erros de classificação.

1.4 MÉTODO

Uma vez definidos os objetivos deste trabalho, torna-se necessário estabelecer o método pelo qual estes objetivos serão buscados.

1.4.1 *Método de Pesquisa*

Segundo Jung (2004), toda pesquisa que utilizar métodos científicos é dita científica, não importando se o propósito é de ordem teórica ou aplicada. De acordo com Silva e Menezes (2001), é importante caracterizar e classificar uma pesquisa científica, de forma a delinear as etapas para a sua realização. As formas clássicas de classificação são: do ponto de vista da sua natureza (aplicada e básica), da forma de abordagem do problema (quantitativa e qualitativa), de seus objetivos (descritiva, explicativa e exploratória) e dos procedimentos técnicos a serem adotados (bibliográfica, documental, estudo de caso, experimental, expost-facto, levantamento, participante e pesquisa-ação) conforme Figura 1.

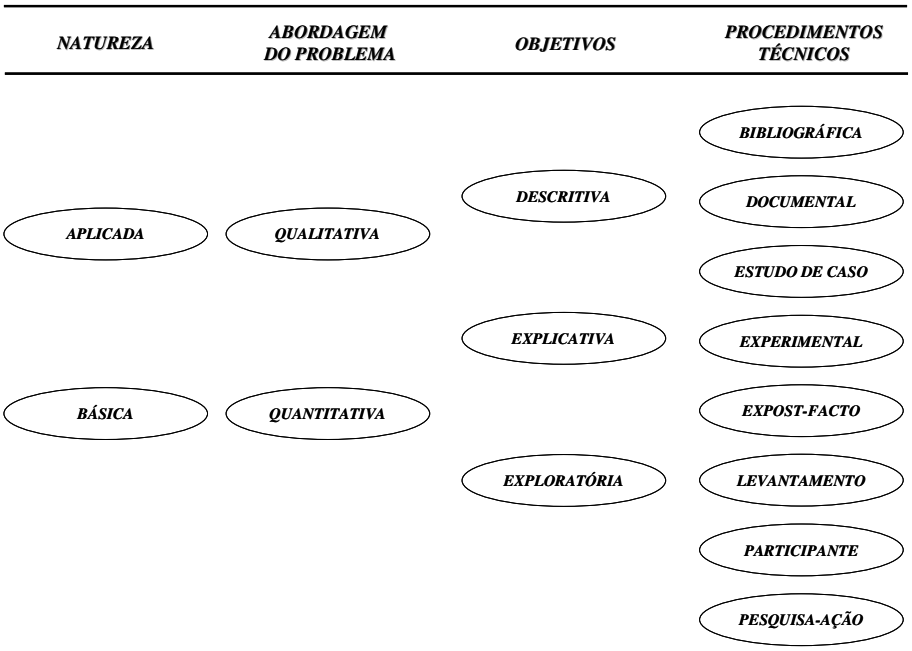


Figura 1 – Classificações das pesquisas segundo Silva e Menezes (2001)

A partir da classificação apresentada por Silva e Menezes (2001), este trabalho se caracteriza como: (*i*) uma pesquisa aplicada, uma vez que objetiva gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos, envolvendo verdades e

interesses locais. A pesquisa aplicada é difundida no tempo e no espaço, mas é limitada no contexto da aplicação, pois admite-se que os problemas podem ser entendidos e resolvidos apenas com o conhecimento; (*ii*) uma pesquisa quantitativa, pois requer o uso de técnicas estatísticas na análise de informações obtidas; (*iii*) uma pesquisa explicativa por possuir um objetivo explicativo, visando a identificar características da relação entre variáveis em estudo, contribuindo para explicar a razão de ocorrência do fenômeno. (*iv*) uma pesquisa bibliográfica, elaborada a partir de material já publicado, constituído de livros, artigos de periódicos, artigos em anais de congressos e materiais disponibilizados na internet e pesquisa experimental, por haver o interesse em observar a influência dos efeitos das variáveis no objeto de estudo.

1.4.2 *Método de Trabalho*

O método de trabalho seguiu as etapas apresentadas na Figura 2. Na seqüência são detalhadas cada uma das etapas.

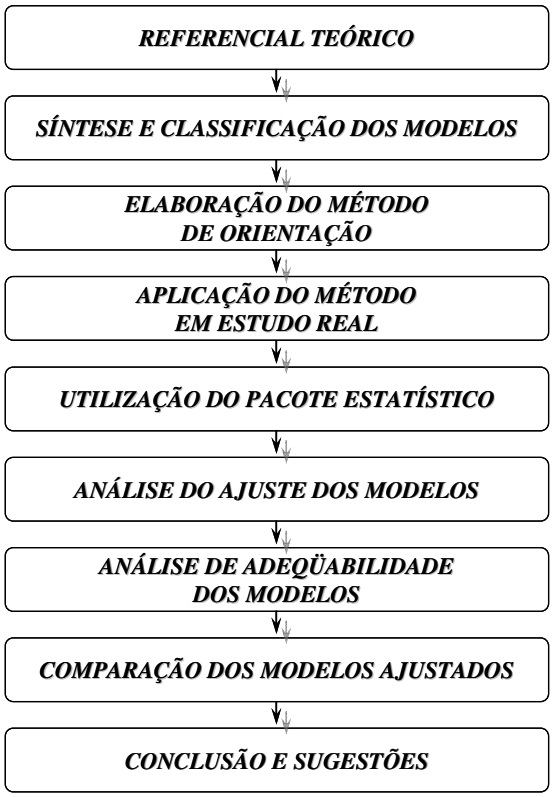


Figura 2 – Etapas de execução do método de trabalho

A presente pesquisa iniciou com uma revisão de literatura abordando os temas: (*i*) Gráfico de Representação de Sistemas, (*ii*) Modelos de Regressão e (*iii*) Medidas de Diagnóstico para a verificação da adequação dos modelos, com o objetivo de consolidar informações e conceitos relevantes e necessários ao desenvolvimento deste trabalho.

Na seqüência, realizou-se uma síntese e classificação dos modelos de regressão quanto à natureza das variáveis dependente e independentes. É importante salientar que não é aconselhável tratar as diferentes classificações de medição com os mesmos modelos de regressão. Portanto, é essencial observar quais são os "tipos" possíveis de variáveis dependentes e independentes existentes no conjunto de dados em estudo, pois a identificação e classificação destas variáveis nos levam a considerar classes de modelos de regressão diferentes.

Posteriormente, elaborou-se um método de orientação à modelagem de dados medidos em proporção, considerando a classificação da natureza das variáveis dependente e independentes, em particular os que modelam uma única variável dependente (modelagem univariada). O desenvolvimento do método foi baseado na literatura, na classificação prévia dos modelos de regressão e na experiência empírica do autor.

Aplicou-se o método na empresa Bracol Couros, que é afiliada ao Grupo Bertin Ltda, produtora de couro acabado e fornecedora para as indústrias calçadista e de artefatos de couro da região do Vale do Rio dos Sinos. Para a empresa, um dos principais interesses é conhecer quais são os fatores de maior influência na produção de refugo por erros de classificação no produto acabado causados por erros de classificação no estágio *wet blue*, e assim poder identificar os efeitos dos fatores com o intuito de definir procedimentos para reduzir a ação das fontes de variabilidade.

Na análise estatística de modelos de regressão mais sofisticados ou mais recentemente desenvolvidos, freqüentemente se encontra um obstáculo no que diz respeito aos recursos computacionais disponíveis. Os pacotes estatísticos mais comumente usados na construção dos modelos de regressão são STATGRAPHICS, SPSS e MINITAB. No entanto, esses pacotes não apresentam procedimentos para construir os modelos propostos neste trabalho: o Modelo de Quase-verossimilhança e o Modelo Beta.

A construção dos modelos de regressão propostos foi realizada no *software* ℝ 2.0.1, um programa estatístico *freeware* desenvolvido em linguagem C++, bastante simples de utilizar e que permitiu adicionar rotinas de programação na *syntax* do modelo, pois possui código aberto. Dentre outras análises, este programa pode ser utilizado para o ajuste de qualquer modelo de regressão com enfoques lineares e não-lineares. Mais detalhes podem ser vistos em R Development Core Team (2004).

Com os dados coletados pela empresa quando da verificação dos critérios de classificação da matéria-prima no estágio *wet blue*, a mensuração dos resultados obtidos na verificação foi definida como a variável dependente no estudo. Em seguida, foi feito o uso do método proposto na orientação da modelagem dos dados, possibilitando identificar quais os fatores controláveis (parâmetros do processo) e os graus de influência destes na proporção de produtos refugados por erros de classificação quando do ajuste dos modelos de regressão.

Na seqüência da modelagem dos dados, realizou-se a análise das medidas de diagnóstico para verificação da adequabilidade dos modelos ajustados, fornecendo informações relevantes na escolha dos modelos de regressão. As medidas de diagnóstico forneceram evidências quanto ao desempenho dos modelos ajustados e permitiu a realização da comparação dos modelos utilizados, identificando posteriormente vantagens e desvantagens no uso. Por último, conclusões e sugestões para trabalhos futuros foram elaboradas.

1.5 ESTRUTURA DO TRABALHO

A dissertação é composta de cinco capítulos. Neste primeiro capítulo, tem-se a visão geral, dos objetivos a serem alcançados, dos métodos e das delimitações do trabalho.

No segundo capítulo é apresentada uma revisão sobre Gráfico de Representação de Sistemas, Modelo de Quase-verossimilhança (a partir da teoria dos Modelos Lineares Generalizados) e Modelo de Regressão Beta. Ademais são apresentados aspectos básicos das medidas de diagnóstico usualmente empregadas para: (*i*) avaliar a qualidade do ajuste; (*ii*) avaliar a adequabilidade dos modelos aos dados; (*iii*) identificar observações influentes e (*iv*) capacidade de predição dos modelos em estudo.

No terceiro capítulo é apresentado um método de orientação à modelagem de dados mensurados em proporção, considerando a classificação da variável dependente e das variáveis independentes. O método apresenta um enfoque na variável dependente quantitativa contínua com restrição ao intervalo [0,1].

No quarto capítulo é discutida a aplicação do método proposto em um estudo de caso realizado numa empresa curtidora de couro, situada na região do Vale do Rio dos Sinos, em uma etapa do processo de produção, denominada de estágio *wet blue*. Foi realizada uma análise comparativa em que foram discutidas as vantagens e desvantagens do uso dos modelos de regressão sugeridos nesta dissertação.

No quinto capítulo são apresentadas as considerações finais obtidas com o desenvolvimento da dissertação e sugestões para trabalhos futuros.

1.6 DELIMITAÇÕES

A dissertação delimita-se no estudo dos modelos de regressão para variáveis dependentes contínuas com mensurações em proporção, apresentando o ajuste e a adequabilidade aos dados - os modelos de regressão Beta e de Quase-verossimilhança.

O modelo de regressão Beta não é aplicável a variável dependente quantitativa contínua que não apresente valores de mensuração compreendidos no intervalo entre zero e um (0,1). Entretanto, não há nenhuma restrição de aplicação do modelo quanto às variáveis independentes, podendo ser de natureza quantitativa e/ou qualitativa.

Os modelos de regressão apresentados no método de orientação, bem como os modelos utilizados na modelagem se restringem a uma variável dependente, não havendo restrições ao número de variáveis independentes a serem usadas.

O método elaborado delimita-se na orientação à modelagem de variáveis dependentes contínuas com mensurações em proporção.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta uma revisão de literatura sobre Gráfico de Representação de Sistemas, Modelos Lineares Generalizados (MLG), que foram propostos por Nelder e Wedderburn (1972), mais especificamente o Modelo de Quase-verossimilhança; Modelo de Regressão Beta (MRB); e Medidas de Diagnóstico. Abordando também distribuições de probabilidade, forma estrutural dos modelos, método de estimação e teste de significância. Além de uma síntese dos modelos contemplados no método proposto no Capítulo 3.

2.1 GRÁFICO DE REPRESENTAÇÃO DE SISTEMAS

Uma importante ferramenta gerencial para compreender os processos existentes ou propostos é o seu mapeamento, uma representação de forma gráfica que permita visualizar as atividades nas diversas etapas da organização, identificando oportunidades de clareza e simplificação (ARAÚJO, 2001).

A elaboração de métodos gráficos constitui a ferramenta para a compreensão dos procedimentos gerenciais do processo. Em que permite orientar com maior clareza e objetividade o fluxo de informação e sua operacionalização, possibilitando melhor resultado na análise das informações.

Segundo Oliveira (1999), um método gráfico desenvolvido para descrever o fluxo de processos e/ou procedimentos, permite ao analista o discernimento na orientação adequada quanto ao processo ou procedimento a ser utilizado. Para Araújo (2001), um método gráfico que descreve um processo existente ou proposto, usando simbologia simples, de maneira clara e objetiva, se constitui uma importante ferramenta na gestão organizacional. Harrington (1993) relata que, um método gráfico vale mais que mil procedimentos, salientando a importância da ferramenta.

Em suma, os métodos gráficos têm uma função básica: descrever um processo para que se possa evidenciar a orientação dos procedimentos, reduzir o tempo de execução das atividades e identificar as oportunidades de mudanças. Quer dizer, o essencial não é a documentação e sim a análise do processo, cujo fim é definir e implementar melhorias.

2.2 MODELO LINEAR GENERALIZADO

2.2.1 *Introdução*

Em muitas situações práticas em que se deseja realizar uma investigação entre uma variável dependente e demais variáveis independentes, cuja variável dependente apresenta restrição nos valores mensurados como proporção de algum evento de interesse, é comum usar no processo de modelagem, o modelo de regressão linear normal. Contudo, segundo Cox (1996), a modelagem da proporção utilizando um modelo de regressão linear normal nem sempre é recomendada, pois este modelo requer a suposição de normalidade aos dados. Pelo fato dos dados serem mensurados em proporção dificilmente apresentarão normalidade. Portanto, deve-se buscar uma nova forma de relacionar as variáveis independentes à variável dependente.

Uma classe de modelos conhecidos como Modelos Lineares Generalizados é apropriada para investigar o efeito de variáveis independentes sobre uma única variável dependente de comportamento não-normal. Estes modelos permitem estimar os parâmetros relacionados com cada efeito, analisar a influência e realizar previsões. Ademais, na construção destes modelos as variáveis independentes podem ser de natureza quantitativa ou qualitativa.

Segundo Hamada e Nelder (1997), a classe de Modelos Lineares Generalizados foi desenvolvida por Nelder e Wedderburn (1972) e estes modelos se baseiam em distribuições de probabilidade pertencentes à família exponencial, com um parâmetro desconhecidos, cujas médias são não-lineares num conjunto de parâmetros lineares. Conforme Lee e Nelder (1998), está classe de modelos é definida ainda por um conjunto de variáveis independentes que descreve a estrutura linear do modelo e uma função de ligação entre a média da variável dependente e a estrutura linear.

O número de produtos não conformes (m_i) em uma amostra (n_i) independente, onde $m_i < n_i$, é classificado como uma variável aleatória discreta (y_i), pois esta variável pode ser representada por um valor de grandeza no conjunto dos números reais (AGRESTI, 1996). Segundo Fahrmeir e Tutz (1994) e Paula (2004), esta variável aleatória segue a distribuição de probabilidade Binomial com os parâmetros n_i e p_i, sendo $p_i = m_i/n_i$.

Segundo Cordeiro (1986), o estudo de dados na forma de proporção é descrito formalmente como um Modelo Binomial, pois apresenta a probabilidade de sucessos de um referido evento ocorrer, em um conjunto de *n* dados investigados. McCullagh e Nelder (1989) salientam que, neste tipo de estudo, a relação entre a variável dependente e as variáveis independentes em estudo é descrita por uma função.

Conforme Crowder (1978, p.34); Prentice (1986, p.323) e Demétrios (2002, p.15), a proporção de sucessos (p_i) de um referido evento (por exemplo, produto defeituoso) segue uma distribuição de probabilidade Beta-Binomial. Onde admiti-se que a variável dependente (y_i) segue a distribuição Binomial e a proporção (p_i) em cada ocorrência ($i = 0, ..., n$) segue uma distribuição de probabilidade Beta. A combinação das distribuições de probabilidade Binomial e Beta na estrutura da modelagem produzem apenas um ajuste na função de variância da variável dependente.

2.2.2 *Família Exponencial*

Em linhas gerais, supondo uma variável aleatória *y* cuja função densidade de probabilidade depende do parâmetro θ. A distribuição de probabilidade pertence à família exponencial se pode ser escrita na forma:

$$f(y;\theta) = \exp\left[a(y)b(\theta) + c(\theta) + d(y)\right] \quad (1)$$

onde $b(\theta)$ é definido como parâmetro da distribuição de probabilidade e $a(y)$ é chamada de forma canônica.

Desta forma, muitas distribuições de probabilidade pertencem à família exponencial, como por exemplo, a distribuição Normal, Binomial e Poisson, pois podem ser escrita na forma canônica e apresentar um parâmetro θ, conforme Tabela 1.

Tabela 1– Características das principais distribuições de probabilidade da família exponencial.

Distribuição de Probabilidade	**Parâmetro** (θ)	**Forma canônica** $a(y)$	$c(\theta)$	$d(y)$
Normal	μ	y	$-\frac{\mu^2}{2\sigma^2} - \frac{\log(2\pi\sigma^2)}{2}$	$-\frac{y^2}{2\sigma^2}$
Binomial	$\log\left(\frac{\mu}{1-\mu}\right)$	y	$n\log(1-\mu)$	$\log\binom{n}{y}$
Poisson	$\log\mu$	y	$-\mu$	$-\log y!$

Adaptado de DOBSON (1990) e PAULA (2004).

Uma vez que a variável aleatória *y* segue alguma distribuição de probabilidade, implicitamente são definidas para esta variável: a esperança matemática (média), a variância, a função de variância, dentre outros parâmetros (ver Tabela 2).

Tabela 2– Média e Variância das principais distribuições de probabilidade da família exponencial

Distribuição de Probabilidade	**Esperança (média)** ($E(y)$)	**Variância** ($Var(y)$)
Normal	μ	σ^2
Normal Inversa	μ	$\mu^3\sigma^2$
Binomial	μ	$\mu(1-\mu)$
Binomial Negativa	μ	$\mu+\mu^2/\alpha$
Poisson	μ	μ
Gamma	μ	μ^2/α

Contudo, observa-se que uma variância de uma variável aleatória *y* é um produto de dois componentes, e apresenta a forma da equação (2)

$$\text{Var}(y) = \phi \cdot V(\mu) \qquad (2)$$

onde ϕ é o parâmetro de dispersão, que é a parte da variância que não depende da média e é constante para as distribuições pertencentes a família exponencial, e $V(\mu)$ é a função de variância, que depende da média. Na Tabela 3 tem-se a forma de algumas distribuições membros da família exponencial.

Tabela 3– Forma dos componentes da Variância das principais distribuições da família exponencial

Distribuição de Probabilidade	**Parâmetro Dispersão** (ϕ)	**Função de Variância** $V(\mu)$
Normal	σ^2	1
Normal Inversa	σ^2	μ^3
Binomial	1	$\mu(1-\mu)$
Binomial Negativa	1	$\mu + \mu^2/\alpha$
Poisson	1	μ
Gamma	$1/\alpha$	μ^2

2.2.3 ***Componentes do Modelo***

A formulação de um MLG compreende-se por possuir três componentes: a *componente aleatória*, que identifica a distribuição de probabilidade da variável dependente; *a componente sistemática*, que especifica a estrutura linear das variáveis independentes quantitativas e/ou qualitativas, que é utilizada como preditor linear; e a *função de ligação*, que descreve a relação funcional entre a componente sistemática e o valor esperado da componente aleatória (CORDEIRO, 1986; McCULLAGH ; NELDER, 1989; DOBSON, 1990; FAHRMEIR ; TUTZ, 1994; PAULA, 2004).

2.2.3.1 Componente Aleatória

A componente aleatória especifica uma variável aleatória y com n observações independentes e identicamente distribuídas, um vetor de médias $\mu = (\mu_1, \ldots, \mu_n)^T$ e uma distribuição pertencente à família exponencial (McCULLAGH ; NELDER, 1989; DOBSON, 1990).

Conforme Agresti (1996), em muitas aplicações, os resultados potenciais para cada observação de y são binários, como sucesso ou fracasso, ou mais geralmente, cada y_i poderia ser definido como o número de sucessos de um certo número fixo de tentativas. Desta forma, assumimos uma distribuição binomial para a componente aleatória. Para Montgomery e Peck (1992) em alguma outra aplicação, se cada observação y_i é contínua, como o peso de um lote de peças em um estudo no processo de manufatura, pode-se assumir uma componente aleatória normal.

2.2.3.2 Componente Sistemática

A componente sistemática especifica a estrutura linear das variáveis independentes quantitativas e/ou qualitativas, que é utilizada como preditor linear (McCULLAGH ; NELDER, 1989). Para Agresti (1996) a componente especifica as variáveis independentes que entram linearmente à direita da equação do modelo como preditores, conforme a equação (3)

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k \qquad (3)$$

Assim, a combinação linear das variáveis independentes é chamada de preditor linear. Segundo Paula (2004), algumas variáveis independentes (x_j) podem ser baseadas em outro formato que permita avaliar o efeito em y, por exemplo, seja $x_3 = x_1 x_2$, que permite interação entre x_1 e x_2 ou $x_3 = x_1^2$, que permite um efeito quadrático de x_1.

De acordo com Cordeiro (1986), a estrutura linear de um MLG pode ser escrita como, a equação (4)

$$\eta_i = \sum_{j=1}^{k} x_j \beta_j \tag{4}$$

onde a função linear η_i dos parâmetros desconhecidos $\beta = (\beta_1, \ldots, \beta_k)$ é denominada de preditor linear, x_j representa os valores de k $(k < n)$ variáveis independentes que são assumidas fixas e conhecidas.

2.2.3.3 Função de Ligação

A terceira componente do um MLG é a função de ligação, que descreve a relação funcional entre a componente sistemática e o valor esperado da componente aleatória (a média da variável dependente). A estrutura da função de ligação na equação do modelo pode ser representada conforme a fórmula (5)

$$g(\mu_i) = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k \tag{5}$$

podendo ser reescrito como

$$g(\mu_i) = \eta_i \tag{6}$$

em que a função $g(\mu_i)$ segundo muitos autores, é uma função estritamente monótona e duplamente diferenciável e que possibilite modelar diretamente a média da variável dependente, conforme pode ser ilustrado,

$$\mu_i = g^{-1}(\eta_i) \;\; ; \quad i = 1, \ldots, n \tag{7}$$

Esta dissertação aborda as principais funções de ligação $g(\cdot)$ que são utilizadas na modelagem dos Modelos Lineares Generalizados, quando a distribuição de probabilidade da variável dependente é Binomial, sendo:

- a função Probit: $g(\mu) = \Phi^{-1}(\mu)$, onde $\Phi(\mu)$ é a função de distribuição normal acumulada e $\Phi^{-1}(\mu)$ é a função inversa, monótona e diferenciável,

- a função Logit: $g(\mu) = log(\mu/(1-\mu))$

- a função Complemento Log-log: $g(\mu) = log\{-log(1-\mu)\}$

Estas funções são definidas conforme a distribuição de probabilidade da variável dependente, sendo estas funções de ligação contínuas e estritamente crescentes no intervalo unitário [0,1]. Em muitos casos, pode ser viável utilizar a função de ligação que melhor ajuste a relação da estrutura linear (preditor linear) e a média da distribuição da variável dependente (CORDEIRO, 1986; McCULLAGH ; NELDER, 1989; DOBSON, 1990 e PAULA, 2004).

Segundo Sant'Anna e Caten (2005), dentre as funções de ligação usadas na modelagem de dados que seguem a distribuição de probabilidade Binomial, a função de ligação Logit apresenta melhor ajuste, além de permitir facilidade de interpretação.

2.2.4 *Método de Estimação*

Esta seção apresenta a estimação dos parâmetros para o Modelo Linear Generalizado através do método clássico de máxima verossimilhança, em que os estimadores β e ϕ são obtidos a partir da maximização do logaritmo da função de verossimilhança, utilizando um algoritmo de otimização não-linear, tal como o algoritmo de Newton (Newton-Rapson, Fisher's *scoring*, etc.) descrito em detalhes por Cordeiro (1992) ou o algoritmo quasi-Newton (BFGS) descrito por Ferrari e Cribari-Neto (2004). Conforme Cordeiro e Cribari-Neto (1998), este método de estimação pode ser utilizado considerando qualquer distribuição de probabilidade para variável dependente.

O método de estimação por máxima verossimilhança dos parâmetros dos modelos de regressão pertencentes à classe dos MLG's, considera a função de log-verossimilhança baseada na amostra de *n* observações independentes, de forma geral,

$$l(\beta,\phi)=\sum_{i=1}^{n} l_i(\mu_i,\phi), \tag{8}$$

com μ_i definida de tal forma que satisfaz a equação $\mu_i = g^{-1}(\eta_i)$, que é função de β.

A função escore é obtida pela diferenciação da função de log-verossimilhança em relação aos parâmetros desconhecidos. A isto se segue que, para $j = 1, \ldots, k$, a derivada da função de log-verossimilhança apresenta a forma

$$\frac{\partial l(\beta,\phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i(\mu_i,\phi)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}. \tag{9}$$

Note que $\partial\mu_i / \partial\eta_i = 1/g'(\mu_i)$ e também,

$$\frac{\partial l_i(\mu_i,\phi)}{\partial \mu_i} = \phi\left[\log\frac{y_i}{1-y_i} - \{\delta(\mu_i\phi) - \delta(1-\mu_i)\phi\}\right], \tag{10}$$

onde $\delta(\cdot)$ é uma função digamma, isto é, $\delta(z) = \partial \log\Gamma(z)/\partial z$, $z > 0$. Desta forma y_i^* = log (y_i / (1– y_i)) e μ_i^* = {δ($\mu_i\phi$) – δ(1– μ_i)ϕ }. Conforme Martínez (2004) o valor esperado da derivada em (10) iguala-se a zero, de forma que o valor esperado da variável aleatória transformada y_i^* iguala-se a μ_i^* . Ou seja, μ_i^* = E(y_i^*), por conseguinte,

$$\frac{\partial l(\beta,\phi)}{\partial \beta_j} = \phi\sum_{i=1}^{n}\left(y_i^* - \mu_i^*\right)\frac{1}{g'(\mu_i)}x_j . \tag{11}$$

A função escore para β pode ser definida de forma matricial como:

$$U_\beta(\beta, \phi) = X^T T(y^* - \mu^*), \tag{12}$$

onde X é uma matriz (n x k) cuja t-ésima linha é x_t^T, $T = diag\{1/g'(\mu_1), ..., 1/g'(\mu_n)\}$, $y^* = (y_1^*, ..., y_n^*)^T$ e $\mu^* = (\mu_1^*, ..., \mu_n^*)^T$. De forma semelhante, tem-se que para o parâmetro de dispersão (ϕ) a derivada do logaritmo da função de verossimilhança e dado por

$$\frac{\partial l(\beta,\phi)}{\partial \phi} = \sum_{i=1}^{n}\frac{\partial l_i(\mu_i,\phi)}{\partial \phi}. \tag{13}$$

e a função escore pode ser escrita como

$$U_\phi(\beta,\phi) = \sum_{i=1}^{n}\left\{\mu_i\left(y^* - \mu_i^*\right) + \log(1-y_i) - \delta((1-\mu_i)\phi) + \delta(\phi)\right\}. \tag{14}$$

sendo $U_\phi(\beta, \phi)$ um escalar.

O próximo passo é obter a matriz de informação de Fisher para (β, ϕ) fazendo as derivadas de 2ª ordem da equação (8) em relação a β_i e ϕ. A partir disto, obteremos W = $diag\{w_1, ..., w_n\}$, com

$$w_i = \phi\{\delta'(\mu_i\,\phi) + \delta'((1-\mu_i)\phi)\}\frac{1}{\{\delta'(\mu_i)\}^2},$$

$c = (c_1, ..., c_n)$, com $c_i = \phi\{\delta'(\mu_i\,\phi)\mu_i + \delta'((1-\mu_i)\phi)(1-\mu_i)\}$, onde $\delta'(\cdot)$ é a função trigamma. Admite-se que $D = diag\{d_1, ..., d_n\}$, com $d_i = \delta'(\mu_i\,\phi)\mu_i^2 + \delta'((1-\mu_i)\phi)(1-\mu_i)^2 - \delta'(\phi)$. Pode-se provar que a matriz de informação de Fisher é dada por

$$K = K(\beta,\phi) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{pmatrix}, \tag{15}$$

onde $K_{\beta\beta} = \phi X^T W X$, $K_{\beta\phi} = K_{\phi\beta}^T = X^T T c$ e $K_{\phi\phi} = \mathrm{tr}(D)$ (FAHRMEIR e TUTZ, 1994). Observe que $K_{\beta\phi} = K_{\phi\beta}^T \neq 0$, o que indica que os parâmetros β e ϕ não são ortogonais, diferentemente do que é verificado na classe dos modelos lineares generalizados (MYERS *et al.*, 2002).

Sob condições de regularidade usuais para estimação de máxima verossimilhança (ver SEN e SINGER, 1993), quando o tamanho da amostra é grande, tem-se que

$$\begin{pmatrix} \hat{\beta} \\ \hat{\phi} \end{pmatrix} \approx N_{k+1}\left(\begin{pmatrix} \beta \\ \phi \end{pmatrix}, K^{-1}\right), \tag{16}$$

onde $\hat{\beta}$ e $\hat{\phi}$ são estimadores de máxima verossimilhança de β e ϕ, respectivamente, e N_{k+1} uma distribuição normal $(k+1)$-variada. Por esta razão é útil obter uma expressão para K^{-1}, a qual pode ser usada para obtenção dos erros padrões assintóticos das estimativas de máxima verossimilhança. Utilizando a expressão padrão para a inversa de matrizes particionadas (ver, por exemplo, MARTÍNEZ, 2004), obtem-se a inversa da matriz de informação de Fisher (15) como segue

$$K^{-1} = K^{-1}(\beta,\phi) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{pmatrix}, \tag{17}$$

onde

$$K_{\beta\beta} = \frac{1}{\phi}\left(X^T W X\right)^{-1}\left\{I_k + \frac{X^T T_{cc}^T T^T X\left(X^T W X\right)^{-1}}{\gamma\phi}\right\}, \tag{18}$$

com $\gamma = \text{tr}(D) - \phi^{-1}c^T X(X^T WX)^{-1}X^T Tc$, $K_{\beta\phi} = (K_{\phi\beta}{}^T) = 1/\gamma\phi\ (X(X^T WX)^{-1}X^T Tc)$, e $K_{\phi\phi} = \gamma^{-1}$. Sendo I_k a matriz identidade de ordem k x k.

2.2.4.1 Algoritmo de Newton-Raphson

Entre os métodos mais poderosos para solucionar sistemas de equações não-lineares, está o algoritmo de Newton-Raphson, o qual é o mais utilizado, embora apresente algumas desvantagens, como o cálculo da matriz $U(\theta^{(k)})$, bem como sua inversa, que em algumas situações é de difícil obtenção, posto que a matriz $-U'(\theta)$ pode não ser positiva definida.

Seja $\theta = (\beta^T, \phi^T)$, o vetor de parâmetro e $U(\theta) = (U_\beta(\beta, \phi)^T\ ,\ U_\phi(\beta, \phi))^T$, o vetor das funções escore de dimensão $(k+1)$ x 1. O processo iterativo de Newton-Raphson para a obtenção da estimativa da máxima verossimilhança do vetor θ é definido expandindo-se em série de Taylor até primeira ordem, a função escore $U(\theta)$ em torno de uma valor inicial $\theta^{(0)}$, tal que

$$U(\theta) \cong U(\theta^{(0)}) + U'(\theta^{(0)}).(\theta - \theta^{(0)}), \tag{19}$$

em que $U'(\theta^{(0)})$ denota a derivada de primeira ordem de $U(\theta)$ com respeito a θ^T. Fazendo $U(\theta) = 0$, então

$$\theta^{(k)} - \theta = \{-U'(\theta)\}^{-1}\ U(\theta), \tag{20}$$

logo, repetindo o procedimento acima, chega-se ao processo iterativo

$$\theta^{(k+1)} = \theta^{(k)} + \{-U'(\theta^{(k)})\}^{-1}U(\theta^{(k)}), \qquad k = 0,1,2, \ldots \tag{21}$$

Assim, o processo anterior é repetido até que a distância entre $\theta^{(k+1)} = \theta^{(k)}$ seja menor que uma tolerância especificada (NOCEDAL e WRIGHT,1999).

Kieschnick e McCullough (2003) utilizaram o método de máxima verossimilhança, através do algoritmo de Newton-Raphson, na estimação dos parâmetros de quatro modelos de regressão construídos (modelo linear censurado normal, logístico beta, logístico simplex e quase-verossimilhança) e o método de mínimos quadrados em três modelos (modelo linear normal, logístico transformado e logístico linear) realizando posteriormente uma comparação analítica entre os valores estimados pelos métodos e constatou que o método de máxima verossimilhança é melhor, pois apresenta maior consistência e precisão.

2.2.5 *Teste de Significância dos Parâmetros*

Nesta seção será apresentado o teste de significância das estimativas dos parâmetros do modelo de regressão. Ou seja, o teste de hipótese para os parâmetros desconhecidos (β'*s*) dos modelos de regressão. Para a previsão de futuras observações da variável *y* deve-se usar modelos contendo apenas parâmetros significativos (modelos parcimoniosos), obtidos a partir da execução de testes que determinem a significância de cada parâmetro.

Para testar hipóteses que lidam com modelos de regressão não linear, pode-se utilizar o teste da Razão de Verossimilhança e o teste de Wald. Ambos tendem à distribuição de probabilidade qui-quadrado com graus de liberdade dependendo dos níveis de cada variável. Segundo Agresti (1996), o teste da Razão de Verossimilhança é mais confiável para qualquer tamanho de amostra do que o teste de Wald. Por isto será abordado apenas o Teste da Razão de Verossimilhança.

É possível realizar testes assintóticos para fazer inferência sob o vetor dos parâmetros desconhecidos. Este teste verifica se há relação linear entre y_i, as variáveis independentes x_1, x_2, ..., x_k, e $\beta_j = (\beta_1, \ldots, \beta_k)^T$ onde $i = 1, \ldots, n$. Considere o teste de hipótese

$$H_0: \beta_j = \beta^{(0)}{}_j \text{ versus } H_1: \beta_j \neq \beta^{(0)}{}_j$$

Para o teste da razão de verossimilhança, a estatística de teste é dada por

$$\varpi = 2\left\{l\left(\hat{\beta},\hat{\phi}\right) - l\left(\tilde{\beta},\tilde{\phi}\right)\right\} \tag{22}$$

em que $l(\beta,\phi)$ é logaritmo natural da função de máxima verossimilhança e $l_i\left(\hat{\beta}^T,\hat{\phi}\right)^T$ é o estimador de máxima verossimilhança restrito de $l(\beta^T,\phi)^T$ obtido pela imposição hipótese nula. Sob condições gerais de regularidade e sob H_0, $\varpi \rightarrow \chi^2_k$. Ou seja, sob a hipótese nula, ϖ tende a distribuição qui-quadrado com *k* graus de liberdade.

2.2.6 *Modelo de Quase-verossimilhança*

De um modo geral, para alguns modelos de regressão realizar a modelagem de um conjunto de observações, primeiro deve-se assumir que os dados seguem uma distribuição de probabilidade conhecida e que esta pertença à família exponencial, em alguns casos não é

adequado escolher uma distribuição de probabilidade a *priori* para os dados, pois os dados podem não seguir tais distribuições de probabilidade. Nestes casos, Weddeburn (1974) propôs os modelos de quase-verossimilhança (MQV's) pertencentes à classe dos MLG's, pois estes modelos apresentam uma componente sistemática (estrutura linear das variáveis independentes) e função de ligação que relaciona a média (μ_i) da variável dependente à estrutura linear das variáveis independentes (x_j).

A característica destes modelos de regressão, é que não há a necessidade de assumir a princípio alguma distribuição de probabilidade para a variável dependente. Por conseguinte, a esperança matemática e a variância da variável aleatória não são conhecidas a *priori*.

Seja y_i uma variável aleatória qualquer de interesse, que assume a $E[y_i] = \mu_i$ e uma variância definida por $Var[y_i] = \phi * V(\mu_i)$, onde a função de variância $V(\mu_i)$ é uma função conhecida da média μ_i e ϕ é o parâmetro de dispersão constante. A função de quase-verossimilhança para um modelo de regressão é definida pela equação

$$Q(y_i ; \mu_i) = \int_y^{\mu} \frac{y_i - t}{\phi \cdot V(t)} dt \quad , \tag{23}$$

Segundo Cox (1996), quando se modela um conjunto de dados usando os MQV's, a variância é modelada como uma função da média da variável dependente, multiplicada ainda por um parâmetro de dispersão constante. Desta forma, a distribuição da variável dependente ficará determinada quando a função de variância escolhida coincidir com a função de variância de alguma distribuição de probabilidade pertencente à família exponencial.

O Modelo de Quase-verossimilhança utilizado na modelagem de um conjunto de dados mensurados em proporção é descrito a partir de uma variável aleatória (y_i) que assume a esperança matemática e a variância da forma $E[y_i] = \mu_i$ e $Var[y_i] = \phi * V(\mu_i)$, respectivamente. Onde a função de variância é definida por $V(\mu_i) = \mu(1-\mu)$. Assim, a função para a variável aleatória acima descrita, apresenta a forma da equação (24)

$$Q(y_i ; \mu_i) = \frac{1}{\phi} \int_y^{\mu} \frac{y_i - \mu_i}{\mu_i (1 - \mu_i)} d\mu \tag{24}$$

e o logaritmo da função de quase-verossimilhança fica nesse caso dado por

$$Q(y_i;\mu_i) = y_i \ln\left(\frac{\mu_i}{1-\mu_i}\right) + \ln(1-\mu_i) \tag{25}$$

que conforme McCullagh e Nelder (1989), a função acima corresponde: a função de variância $V(\mu) = \mu(1-\mu)$ e a função de log-verossimilhança da distribuição de probabilidade Binomial é dada por

$$L(y_i;\mu_i) = y_i \ln\left(\frac{\mu_i}{1-\mu_i}\right) + n_i \ln(1-\mu_i) \tag{26}$$

Nota-se portanto que a principal diferença entre como formam-se as equações (25) e (26) está em que, quando se usa a função de quase-verossimilhança para estimar os coeficientes (parâmetros desconhecidos) do modelo de regressão, apenas se define a relação da variância da variável dependente com a sua média, não sendo necessário definir anteriormente uma distribuição de probabilidade.

De acordo com Cox (1996), uma vantagem da flexibilidade de uso dos MQV's na modelagem de uma variável dependente de conjunto de dados, é que poderíamos utilizar uma função de variância que melhor se ajuste aos dados, sem assumir a *priori* uma distribuição de probabilidade para esta variável dependente. Além disso, esta função de variância pode não pertencer a nenhuma distribuição de probabilidade da família exponencial. Por exemplo, uma função de variância do tipo $\mathrm{Var}[y_i] = \mu^2(1-\mu)^2$, apresenta o logaritmo da função de quase-verossimilhança da forma dada na equação (27)

$$Q(y_i;\mu_i) = \int_y^{\mu} \frac{y_i - t}{\phi \cdot V(t)} = \int_y^{\mu} \frac{y_i - \mu_i}{\phi \cdot \mu_i^{\,2}(1-\mu_i)^2}\, d\mu_i, \tag{27}$$

que pode ser reescrita como

$$Q(y_i;\mu_i) = \frac{1}{\phi}\left[(2y-1)\log\left(\frac{\mu_i}{1-\mu_i}\right) - \frac{y_i}{\mu_i} - \left(\frac{1-y_i}{1-\mu_i}\right)\right], \tag{28}$$

para $0 < \mu < 1$ e $0 \leq y \leq 1$. Portanto, a função acima demonstrada não corresponde a função de verossimilhança de nenhuma distribuição de probabilidade pertencente a família exponencial (PAULA, 2004).

2.2.6.1 Estimativas dos Coeficientes

A estimação dos parâmetros β e ϕ dos Modelos de Quase-verossimilhança é realizada pela maximização da função de quase-verossimilhança, produzindo as mesmas estimativas dos coeficientes dos parâmetros dos modelos que utilizam a função de lo-verossimilhança, portanto, pode-se usar o mesmo algoritmo de estimação dos parâmetros visto na seção 2.2.4.1.

McCullagh e Nelder (1989) descrevem um algoritmo iterativo, similar ao algoritmo visto na seção 2.2.4.1 quando as formas da função de variância não são iguais aos da família exponencial. Cox (1996) demonstra que, para funções de variância que não pertence à família exponencial, o algoritmo apresentado na seção 2.2.4.1 pode ser utilizado, pois fornece estimativas consistentes e precisas.

2.2.6.2 Teste de Significância

Para testar a significância dos coeficientes do modelo de regressão pelo teste da razão de verossimilhança tem-se a estatística de quase-*deviance*. Pode-se dizer que a quase-*deviance* está para a modelagem pela função de quase-verossimilhança como a *deviance* está para a função de verossimilhança. Por analogia, a quase-*deviance* de um modelo qualquer é definida como o desvio deste modelo em relação ao modelo nulo, sendo:

$$D_i(y_i, \hat{\mu}_i) = -2\phi[Q_i(y_i\,;\hat{\mu}_i) - Q_i(y_i\,;y_i)] = -2\phi[Q_i(y_i\,;\hat{\mu}_i)] = 2\int_{y_i}^{\mu_i} \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)}, \tag{29}$$

em que $Q_i(y_i, \hat{\mu}_i)$ é a função de máxima verossimilhança do modelo sob pesquisa e $Q_i(y_i\,;y_i)$ é a função de máxima verossimilhança do modelo nulo.

Para o Modelo de Quase-verossimilhança definido pela equação (24), a estatística de quase-*deviance* é expressa da forma

$$D_i(y_i, \hat{\mu}_i) = -2\phi[Q_i(y_i\,;\hat{\mu}_i)] = 2\int_{y_i}^{\mu_i} \frac{y_i - \hat{\mu}_i}{\phi \cdot [\hat{\mu}_i(1-\hat{\mu}_i)]}, \tag{30}$$

2.3 MODELO DE REGRESSÃO BETA

2.3.1 *Introdução*

A proporção de defeitos em um determinado conjunto de observações, ou lotes de itens produzidos, é uma importante informação a respeito do comportamento do processo de produção de uma empresa, pois a produção de itens não conformes em um processo acarreta à empresa, custos diretos e indiretos. Segundo Park (1996), é fundamental que as empresas busquem desenvolver um processo robusto, de produção de bens e fornecimento de serviços.

Sabe-se que, apenas o desenvolvimento de um processo robusto não garanta a empresa o fim da produção de produtos não conformes (defeituosos), em virtude da variabilidade inerente ao processo. Assim, se faz importante a realização de etapas complementares, como por exemplo, a modelagem das informações produzidas pelo processo. Neste contexto, Montgomery e Peck (1992) relatam que, a construção de um bom modelo de investigação permite gerar estimativas de possíveis efeitos que influenciam no processo.

Em muitos casos, o interesse em um modelo de regressão está em analisar a possível relação entre uma única variável dependente e duas ou mais variáveis independentes. Entretanto, quando a variável dependente é mensurada em proporção, apresentando valores no intervalo unitário ($0 \leq Y \leq 1$), a relação entre as variáveis dependente e independentes apresenta restrição no domínio da função ($0 < E(Y) < 1$) (PAULA, 2004).

Conforme Cordeiro (1986) e McCullagh e Nelder (1989), a variável dependente mensurada em proporção assume que os dados seguem a distribuição de probabilidade Binomial. Em virtude disso, torna-se necessário uma transformação na variável dependente de tal forma que esta variável possa assumir valores reais ($-\infty < E(Y^*) < \infty$) e então modelar a média da variável dependente transformada em relação as demais variáveis independentes.

Para outros autores, tais como, Wiley *et al.* (1989, p.99); Johnson *et al.* (1995, p.217); McDonald e Xu (1995, p.144); Kieschnick e McCullough (2003, p.194) e Ferrari e Cribari-Neto (2004, p.15), a proporção de sucessos de um referido evento é uma variável aleatória contínua, com mensurações positivas e restritas ao intervalo [0,1], a qual segue uma distribuição de probabilidade Beta, indexada pelos parâmetros (p , q), onde $p > 0$ e $q > 0$.

Partindo deste princípio, Ferrari e Cribari-Neto (2004) propuseram um procedimento alternativo na modelagem de dados mensurados em proporção cuja estrutura do modelo de regressão permite modelar as relações, lineares e não-lineares, entre as variáveis independentes e a variável dependente. O passo inicial deste procedimento é assumindo que as proporções apresentam distribuição de probabilidade Beta.

O modelo proposto apresenta uma estrutura de regressão baseada na suposição de que a variável dependente tem distribuição de probabilidade Beta, que as variáveis independentes formam uma estrutura linear nos parâmetros desconhecidos e que permite modelar a média da variável dependente em relação às demais variáveis independentes através de uma função. Conforme Martínez (2004), este modelo de regressão permite gerar estimativas precisas e seguras dos parâmetros, mesmo que o conjunto de dados coletado para a investigação seja consideravelmente pequeno. Segundo Oliveira (2004), o Modelo Beta apresenta estimativas precisas e confiáveis em amostras de dados mensurados próximos de zero e próximos de um.

2.3.2 *Família Beta*

Segundo Crowder (1978), a variável aleatória mensurada em proporção segue uma distribuição Beta-binomial, pois esta distribuição permite leve flexibilidade no ajuste aos dados. Conforme Prentice (1986), a distribuição Beta-binomial apresenta flexibilidade no ajuste de seus parâmetros, pois sua função densidade de probabilidade pode assumir diferentes formas dependendo dos valores que indexam esta distribuição. Martínez (2004) e Prentice (1986) demonstram que a distribuição Beta-binomial, bem como a distribuição Uniforme, dentre outras, pertencem à família de distribuições Beta.

Johnson *et al.* (1995) relatam que, a distribuição Beta é versátil e uma variedade de incertezas podem ser modeladas por ela, sendo que sua flexibilidade encoraja seu uso nas aplicações. Wiley *et al.* (1989) desenvolvem um Modelo Beta para estimar a probabilidade de transmissão de HIV durante o contato sexual entre um indivíduo infectado e outro indivíduo sadio.

Supondo uma variável aleatória y cuja função densidade de probabilidade depende dos parâmetros p e q. A família de distribuições Beta é composta de todas as distribuições com função densidade de probabilidade escrita na forma,

$$f(y;p,q)=\begin{cases}\dfrac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\,y^{p-1}(1-y)^{q-1} & ,\quad 0<y<1,\ p>0\ ,\ q>0,\\ \\ 0, & ,\ \text{caso contrário}\end{cases} \tag{31}$$

onde $\Gamma(p)$ é uma função gama avaliada no ponto p, ou seja, $\Gamma(p)=\int_0^{\infty} y^{p-1}e^{-y}dy$ com $p > 0$. A função $f(y) = f(y; p, q)$ é efetivamente uma função densidade. Nota-se ainda, que a função $f(y; p, q)$ assume valores estritamente positivos, pois para qualquer valor de y pertencente ao intervalo (0,1), a função densidade descrita é crescente, ou seja, $f(y) \geq 0$.

Ferrari e Cribari-Neto (2004), se baseiam na suposição de que as proporções seguem uma distribuição de probabilidade Beta, para realizar uma parametrização da esperança matemática (média) e da variância utilizando os parâmetros (p e q) da distribuição de probabilidade Beta.

A média e a variância da variável aleatória y são, respectivamente,

$$E(y)=\frac{p}{p+q} \tag{32}$$

e

$$Var(y)=\frac{pq}{(p+q)^2(p+q+1)}. \tag{33}$$

Pode-se observar que os parâmetros p e q são parâmetros de ajuste da distribuição, pois através da escolha de valores para p e q podem ser obtidas diferentes distribuições de probabilidade, por exemplo, quando $p = q = 1$; a distribuição Beta se reduz a distribuição Uniforme, porém se $p = q = ½$; a distribuição Beta será a distribuição Arco seno, que é utilizada para análise de passeios aleatórios. Destacando que, quando $p = q$, as densidades terão formas simétricas, do contrário serão assimétricas. Assim, a distribuição Beta na realidade se apresenta uma família de distribuições. Mais detalhes da família de distribuições Beta, podem ser vista em Prentice (1986), McDonald e Xu (1995) e Krysicki (1999).

Conforme Ferrari e Cribari-Neto (2004) e Kieschnick e McCullough (2003) a distribuição Beta é uma função densidade de probabilidade que não pertence à família exponencial, pois a sua distribuição não pode ser escrita na forma canônica e apresentar um parâmetro θ.

2.3.3 *Componentes do Modelo*

Semelhante a estrutura do Modelo Linear Generalizado, o Modelo Beta é caracterizado por três componentes: *componente aleatória*, que identifica a distribuição de probabilidade da variável dependente; *componente sistemática*, que especifica a estrutura linear das variáveis independentes, denominada de preditor linear; e a função que relaciona a média da variável dependente à estrutura linear das variáveis independentes, sendo uma função estritamente monótona e duplamente diferenciável que transforma valores do intervalo (0,1) em valores reais, denominada *função de ligação*.

2.3.3.1 Componente Sistemática

Sejam $y_1, \ldots, y_n$ variáveis aleatórias identicamente distribuídas, em que cada y_i, $i = 1, \ldots, n$, tem a densidade em (31), com média $\mu_i = (\mu_1, \ldots, \mu_n)$ e parâmetro de precisão desconhecido ϕ. O modelo de regressão Beta é definido pela distribuição de probabilidade Beta e por uma estrutura linear,

$$\eta_i = \sum_{j=1}^{k} x_j \beta_j \tag{34}$$

onde a função linear η_i dos parâmetros desconhecidos $\beta = (\beta_1, \ldots, \beta_k)$ é denominada de preditor linear. x_j representa os valores de k ($k < n$) variáveis independentes que são assumidas fixas e conhecidas.

2.3.3.2 Função de Ligação

Semelhante a estrutura do Modelo Linear Generalizado, o Modelo Beta apresenta a função que relaciona a variável dependente e a estrutura linear das variáveis independentes, sendo uma função estritamente monótona e duplamente diferenciável que transforma valores do intervalo (0,1) em valores reais, denominada função de ligação.

Segundo Dobson (1990) e McCullagh e Nelder (1989), a estrutura da função de ligação na equação do modelo, pode ser representada como,

$$g(\mu_i) = \alpha + \beta_1 x_1 + ... + \beta_k x_k, \tag{35}$$

podendo ser reescrito como

$$g(\mu_i) = \eta_i\,, \tag{36}$$

em que a função $g(\mu_i)$ é uma função estritamente monótona e duplamente diferenciável e que possibilite modelar diretamente a média da variável dependente, conforme pode ser ilustrado

$$\mu_i = g^{-1}(\eta_i); \quad i = 1, ..., n \tag{37}$$

Similar aos Modelos Lineares Generalizados, as funções de ligação ($g(\cdot)$) utilizadas no Modelo de Regressão Beta podem ser: a função Logit ($g(\mu) = log(\mu\ /\ 1-\mu)$), a função Probit ($g(\mu) = \Phi^{-1}(\mu)$), a função Complemento Log-log ($g(\mu) = log\{-log(1-\mu)\}$) e a função Log-log ($g(\mu) = -log\{-log(\mu)\}$). Essas quatro funções de ligações apresentadas são contínuas e estritamente crescentes no intervalo unitário [0,1].

Na prática é comum realizar a escolha da função de ligação que melhor ajuste a relação da estrutura linear (predito linear) e a média da distribuição da variável dependente. Ferrari e Cribari-Neto (2004) afirmam que, no processo de modelagem de algum conjunto de dados medidos em proporção, tais funções de ligação podem apresentar similaridade no ajuste destes dados.

Dentre as funções de ligação acima citadas, uma particularmente usada é a ligação Logit, que é descrita da seguinte forma,

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = x_j\beta_j\,, \tag{38}$$

onde x_j, $i = 1, ..., n$, e $j = 1, ..., k$, podendo ser reescrito como

$$\frac{\mu_i}{1-\mu_i} = e^{x_j\beta_j}\,, \tag{39}$$

ou escrito como

$$\mu_i = \frac{e^{x_j\beta}}{1+e^{x_j\beta}}\,, \tag{40}$$

A equação (40) é a função inversa de $g(\mu_i)$. Nesta forma, o parâmetro de regressão β tem uma importante interpretação que é a razão de chaces (*odds ratio*), definida por e^{β_j}.

Considere a aplicação apresentada no Capítulo 4 , como uma ilustração do cálculo da razão de chances, para a proporção por erro de classificação das peças de couro segundo os níveis ("sim" e "não") da variável independente "rebaixamento". Seja a estimativa do parâmetro do modelo de regressão Beta $\beta_7 = -\,0{,}8183$, o cálculo da razão de chances deste coeficiente fica: $\exp[\beta_7] = \exp[-\,0{,}8183] = 0{,}44$, indicando que o fato do estado de textura da superfície da matéria-prima estar rebaixado (nível "sim" da variável rebaixamento), implica em que as chances de produção da proporção por erro de classificação diminuem em 44% em relação ao nível "não" da variável rebaixamento.

2.3.4 *Método de Estimação*

Os estimadores de máxima verossimilhança dos parâmetros β e ϕ, são obtidos através das equações $U_\beta(\beta,\phi) = 0$ e $U_\phi(\beta,\phi) = 0$ (descritas na seção 2.2.4), não apresentando solução analítica em forma fechada. Sendo assim necessário obter os estimadores pela maximização numérica do logaritmo da função de verossimilhança, utilizando um processo iterativo.

A função de log-verossimilhança baseada em n observações independentes para o Modelo de Regressão Beta é descrito como,

$$l(\beta,\phi) = \sum_{i=1}^{n} l_i(\mu_i,\phi), \tag{41}$$

onde $l_i(\mu_i,\phi) = \log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log\Gamma((1-\mu_i)\phi) + (\mu_i\phi - 1)\log y_i + \{(1-\mu_i)\phi\}\log(1-y_i)$

com μ_i definida de tal forma que satisfaz a equação $\mu_i = g^{-1}(\eta_i)$, que é função de β. A função escore é obtida pela diferenciação da função de log-verossimilhança em relação aos parâmetros desconhecidos (β, ϕ), isto é,

$$\frac{\partial l(\beta,\phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i(\mu_i,\phi)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}. \tag{42}$$

e

$$\frac{\partial l(\beta,\phi)}{\partial \phi} = \sum_{i=1}^{n} \frac{\partial l_i(\mu_i,\phi)}{\partial \phi}. \tag{43}$$

Assim, podemos assumir que o Modelo de Regressão Beta como função da equação (31) é um modelo regular, e seus estimadores de máxima verossimilhança são únicos. Desta forma, o método de estimação por máxima verossimilhança dos parâmetros do Modelo Beta utilizado é o algoritmo quasi-Newton, também chamado de algoritmo BFGS.

2.3.4.1 Algoritmo BFGS

Este algoritmo foi desenvolvido por Broyden, Fletcher, Goldfart e Shanno. Ele utiliza o princípio do algoritmo de Newton-Raphson, se diferenciando apenas na forma de como é calculado o passo de iteração, já que neste caso o passo não depende da *k*-ésima iteração e da matriz hessiana, pois ao invés de trabalhar com $U'(\theta^{(k)})^{-1}$, é utilizado uma seqüência de matrizes simétricas e positivas definidas $Q^{(k)}$ tal que,

$$\lim_{k\to\infty} Q^{(k)} = U'\left(\theta^{(k)}\right)^{-1} \tag{44}$$

Comumente, toma-se como matriz inicial, $Q^{(0)}$, a matriz identidade de mesma ordem, pois ela também é positiva definida e simétrica, assim conduzindo a aproximações $Q^{(k)}$ positivas definidas e simétricas. A forma recursiva para obter tais matrizes é dada pela expressão

$$Q^{(k+1)} = Q^{(k)} - \frac{Q^{(k)} s^{(k)} \left(s^{(k)}\right)^T Q^{(k)}}{\left(s^{(k)}\right)^T Q^{(k)} s^{(k)}} + \frac{y^{(k)} \left(y^{(k)}\right)^T}{\left(y^{(k)}\right)^T s^{(k)}}, \qquad k = 0,1,2, \ldots \tag{45}$$

onde $s^{(k)} = \theta^{(k+1)} - \theta^{(k)}$ e $y^{(k)} = U(\theta^{(k+1)}) - U(\theta^{(k)})$. Assim, mesmo que θ esteja distante do ponto de máximo, as matrizes $Q^{(k)}$ garantem que as iterações se darão da direção crescente. De forma semelhante ao método de Newton-Raphson, o máximo é obtido pela recorrência

$$\theta^{(k+1)} = \theta^{(k)} - \alpha^{(k)} Q^{(k)} U(\theta^{(k)}), \qquad k = 0,1,2, \ldots \tag{46}$$

onde $\alpha^{(k)}$ é um escalar determinado por algum procedimento de busca linear a partir de $\theta^{(k)}$ na direção $-Q^{(k)}\mathrm{U}(\theta^{(k)})$ de forma que $f(y;\theta^{(k)})$ cresça nessa direção, provocando a convergência do algoritmo. Maiores detalhes podem ser vistos em Nocedal e Wright (1999).

Kieschnick e McCullough (2003) utilizaram o algoritmo de Newton-Raphson na estimação dos parâmetros desconhecidos, em quatro dos sete modelos de regressão construídos. Ferrari e Cribari-Neto (2004) propuseram o uso do algoritmo de quasi-Newton (BFGS) na estimação dos parâmetros do Modelo de Regressão Beta.

No uso do procedimento iterativo é sugerido uma estimativa para o ponto inicial de β, a estimativa de mínimos quadrados ordinários do vetor de parâmetros, obtida a partir de uma regressão linear da resposta transformada $g(y_1)$, $g(y_2)$, ... , $g(y_n)$ em X, isto é, $(X^T X)^{-1}X^T z$, em que $z = (g(y_1), g(y_2), \ldots , g(y_n))^{\mathrm{T}}$. Quanto ao parâmetro de precisão ϕ, também é necessário uma valor inicial, neste caso, baseia-se no fato de que $Var\,(y_i) = \mu_i\,(1-\mu_i)\,/(1+\phi)$ e pode ser facilmente escrita da forma $\phi = \{\mu_i\,(1-\mu_i)\,/\,Var\,(y_i)\}-1$. Note que ao expandir até a primeira ordem a função $g(y_i)$ em série de Taylor em torno do ponto μ_i e tomando a variância, deduz-se que

$$Var\,(g(y_i)) \approx Var\,\{g(y_i) + (y_i - \mu_i)\,g'(\mu_i)\} = Var\,(y_i)\{g'(\mu_i)\}^2, \tag{47}$$

isto é,

$$Var\,(y_i) \approx Var\,(g(y_i))\,/\{g'(\mu_i)\}^2, \tag{48}$$

Então, a sugestão dada foi considerar como estimativa inicial para ϕ

$$\frac{1}{n} = \sum_{i=1}^{n} \frac{\breve{\mu}_i\left(1-\breve{\mu}_i\right)}{\breve{\sigma}_i^2} - 1, \tag{49}$$

onde $\breve{\mu}_i$ é obtido aplicando $g^{-1}(\cdot)$ ao i-ésimo valor ajustado da regressão linear de $g(y_1)$, $g(y_2)$, ... , $g(y_n)$ em X, isto é, $\breve{\mu}_i = g^{-1}\,(x_i^{\mathrm{T}}(X^T X)^{-1}X^T z)$, e $\breve{\sigma}_i^2 = \breve{e}^{\mathrm{T}}\breve{e}\,/\,[(n-k)\{g'(\breve{\mu}_i)\}^2]$, sendo $\breve{\mathrm{e}} = z - (X^T X)^{-1}X^T z$ o vetor de resíduos de mínimos quadrados ordinários da regressão linear sob a variável transformada. Oliveira (2004) relata que, estes valores iniciais proporcionaram resultados satisfatórios em seus experimentos.

2.3.5 *Teste de Significância dos Parâmetros*

Esta seção apresenta o teste de significância dos parâmetros para o Modelo de Regressão Beta. Ou seja, o teste de hipótese para os parâmetros desconhecidos (β'*s*) dos modelos. No processo de modelagem, os modelos devem apresentar apenas parâmetros significativos, pois permite realizar boas previsões dos valores da variável *y*. Segundo Oliveira (2004), o teste Escore apresenta melhor desempenho nas estimações dos parâmetros desconhecidos do Modelo Beta, para qualquer tamanho de amostra, considerando as distribuições de probabilidade Normal, t-Student, Exponencial e Qui-quadrado.

Para descrever o teste Escore, o qual verifica a inferência sob o vetor dos parâmetros desconhecidos. Sendo o teste de hipótese

$$H_0: \beta_j = \beta^{(0)}_j \text{ versus } H_1: \beta_j \neq \beta^{(0)}_j$$

considere U_β um vetor coluna *k* dimensional contendo os primeiros *k* elementos da função escore de β e $K_{\beta\beta}$ a matriz *k* x *k* formada das *k* primeiras linhas e as *k* primeiras colunas da matriz de K^{-1}. Pode-se demonstrar, usando a expressão (12) que a estatística do teste Escore podes ser escrita como

$$\upsilon = \tilde{U}_\beta^T K_{\beta\beta} \tilde{U}_\beta, \tag{50}$$

em que o til expressa nas estruturas, indica que as quantidades estão sendo avaliadas do estimador de máxima verossimilhança restrita. Sob condições gerais de regularidade e sob H_0, $\upsilon \rightarrow \chi^2_k$. Ou seja, sob a hipótese nula, υ tende a distribuição qui-quadrado com *k* graus de liberdade.

2.3.6 *Modelo Beta*

O objetivo desta seção é apresentar um modelo que contempla uma variável dependente contínua, restrita a mensuração em proporção, em relação as variáveis independentes. Esta variável assume uma distribuição de probabilidades Beta, a qual possui uma densidade indexada pelos parâmetros *p* e *q*. Para obter uma estrutura de regressão que permita modelar a média de uma variável dependente em função de um conjunto de variáveis independentes e que contenha um parâmetro de precisão (ou dispersão), Ferrari e Cribari-Neto

(2004) apresentam uma parametrização diferente da distribuição de densidade Beta com finalidade de conseguir uma estrutura de regressão associada a um parâmetro de precisão.

Seja uma variável aleatória y de distribuição Beta, ou seja, $y \sim Beta\ (p\ ;\ q)$. Desta forma, os parâmetros de locação e precisão da variável aleatória y são: $\mu = p\ /\ (p+ q)$ e $\phi = p+ q$, respectivamente, isto é, $p = \mu\phi$ e que $q = (1 - \mu)\phi$. Então, a equação (31), será escrita da seguinte forma,

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}\, y^{\mu\phi-1}(1-y)^{[(1-\mu)\phi]-1}, \quad 0 < y < 1,\ \mu > 0,\ \phi > 0 \qquad (51)$$

Por conseguinte, as equações (32) e (33) torna-se-ão

$$E(y) = \mu \qquad (52)$$

e

$$Var\,(y) = \frac{V(\mu)}{1+\phi}, \qquad (53)$$

em que $V(\mu) = \mu(1- \mu)$ de forma que, μ é a média da variável dependente y e ϕ pode ser interpretado como um parâmetro de precisão no sentido que, para μ fixado, quanto maior o valor de ϕ, menor a variância de y. Oliveira (2004) demonstra que a densidade Beta pode apresentar diferentes formas dependendo dos valores dos parâmetros (μ,ϕ). Vale a pena salientar que, quando $\mu = ½$ a distribuição é simétrica e que quando $\mu \neq ½$ a mesma é assimétrica. Em particular para $\mu = ½$ e $\phi = 2$ a equação (31) se reduz a densidade da distribuição Uniforme. Mais detalhe veja Ferrari e Cribari-Neto (2004) e Kieschnick e McCullough (2003).

Nesta dissertação é assumido que a variável dependente está restrita ao intervalo unitário [0,1], ou seja, mensurada em proporção. No entanto, o modelo proposto é utilizado em situações mais gerais, em que a variável dependente é restrita ao intervalo $(a;\ b)$, onde a e b são constantes conhecidas, com $a < b$. Neste caso, em vez de modelar y diretamente, utiliza-se $(y - a)\ /\ (b - a)$ que ficará então definido no intervalo [0,1].

2.4 MEDIDAS DE DIAGNÓSTICO

2.4.1 *Introdução*

Em análises estatísticas, em particular na análise de regressão, surge sempre uma pergunta importante: qual é o melhor modelo? Para responder a esta pergunta é necessário verificar se o modelo ajustado é adequado para descrever os dados em estudo. Rao e Wu (2005) sugerem que se escolha o modelo que mais se aproxima do modelo verdadeiro a partir de um conjunto de modelos candidatos, pois o modelo verdadeiro é definido como o modelo adequado.

Uma etapa essencial na análise de um ajuste de regressão é a verificação de possível violação de qualquer uma das suposições feitas para o modelo, especialmente para a parte aleatória (y_i) e pelo componente sistemático (η_i), bem como a existência de observações extremas com alguma interferência desproporcional nos resultados do ajuste. Tal etapa, conhecida como análise de diagnóstico, se inicia com a análise de resíduos para detectar a presença de pontos extremos (*outliers*) e avaliar a adequação da distribuição proposta para a variável dependente.

A adequação de um modelo é avaliada pela sua capacidade preditiva e definida a partir dos próprios dados utilizados na determinação do modelo. Modelos com bom desempenho estatístico apresentam pequena discrepância entre os dados reais e seus respectivos valores preditos. Ademais, segundo Cordeiro e Lima Neto (2004), na adequação do modelo aos dados é fundamental a análise de ferramentas gráficas como, avaliação do gráfico de resíduos, a observação de pontos influentes (valores que influenciam na estimativa da média da variável dependente), alavanca generalizada proposta por Wei *et al.* (1998) e a distância de Cook.

Em uma modelagem, variáveis independentes e/ou a interação destas variáveis só devem ser acrescentadas ao modelo se apresentarem sobre o comportamento da variável dependente, um nível explanatório significativo (*p-value*). Para garantir a inclusão somente de variáveis significativas no modelo, procedem-se testes de significância estatística.

2.4.2 *Tipos de Medidas de Diagnóstico*

2.4.2.1 Coeficiente de Determinação

Em muitas aplicações, o coeficiente de determinação (R^2) é uma medida global da qualidade do ajuste, utilizado como indicador numérico que permite comparar o desempenho de diferentes modelos, contudo, não é uma boa estratégia, pois o mesmo sempre aumenta com a inclusão de novas variáveis independentes. Para contornar este problema foi criado um coeficiente de determinação ajustado, denominado "pseudo" R^2 (R_p^2) que é definido como o quadrado do coeficiente de correlação amostral entre $g(y)$ e $\hat{\eta}$. Observe que $0 \leq R_p^2 \leq 1$ e, quando $R_p^2 = 1$ existe uma concordância perfeita entre $\hat{\mu}$ e y, consequentemente, melhor será o ajuste.

Segundo Rao e Wu (2005), embora o desempenho do R_p^2 não seja muito bom sob certas circunstâncias, o mesmo é uma ferramenta suporte em muitos estudos de modelagem de dados. Ademais, dentre os possíveis modelos propostos, o melhor modelo é aquele que maximiza o pseudo R^2.

2.4.2.2 Desvio (*Deviance*)

É uma outra medida utilizada para verificar o grau da qualidade de ajuste do modelo, e em muitos casos, serve como valor padrão para a comparação entre diversos modelos que serão ajustados. Assim, quanto menor o seu valor, melhor é a qualidade do ajuste do modelo. Esta medida foi originalmente proposta por Nelder e Wedderburn (1972) no contexto dos Modelos Lineares Generalizados.

A análise de *deviance* é feita através da comparação dos valores da medida *deviance* dos modelos ajustados. Segundo Atkinson e Riani (2000), a análise da *deviance* é verificada como duas vezes a diferença entre o máximo do logaritmo da verossimilhança do modelo nulo e do modelo sob pesquisa. Esta medida embora denominada de *deviance*, também conhecida como desvio (CORDEIRO, 1986).

Ferrari e Cribari-Neto (2004) propuseram utilizar a *deviance* como critério de análise da qualidade do ajuste para os Modelos de Regressão Beta, sendo,

$$D(y;\mu,\phi) = \sum_{i=1}^{n} 2\{l_i(\tilde{\mu},\phi) - l_i(\mu_i,\phi)\} \qquad (54)$$

em que $\tilde{\mu}$ é solução de $\partial l_i / \partial \mu_i = 0$, isto é, $\phi\,(y_i^* - \mu_i^*) = 0$, $l_i(\tilde{\mu},\phi)$ é a função de máxima verossimilhança do modelo sob pesquisa e $l_i(\mu,\phi)$ é a função de máxima verossimilhança do modelo saturado.

Segundo Myers e Montgomery (1997), quando ϕ é grande, $\mu_i^* \approx log\{\mu_i/(1-\mu_i)\}$ e desta forma $\tilde{\mu}_i = y_i$. Para ϕ conhecido, podemos definir uma medida de discrepância como $D(y;\hat{\mu},\hat{\phi})$, em que $\overline{\mu}$ é o estimador de máxima verossimilhança de μ sob o modelo em pesquisa. E quando ϕ é desconhecido, uma aproximação para essa quantidade é $D(y;\hat{\mu},\hat{\phi})$, denominada usualmente de desvio do modelo sob pesquisa.

A diferença entre as medidas possui distribuição de probabilidade qui-quadrado (χ^2), sendo que o número de graus de liberdade corresponde à diferença no número de parâmetros que o modelo possuir. Diversos modelos, começando com um simples, que contém apenas o intercepto (modelo nulo) e indo até aquele com todas as variáveis independentes sob investigação, são então dispostos sucessivamente para análise das medidas de *deviance*. Desta forma, o modelo que apresentar o menor valor, representa o modelo de melhor ajuste.

O desvio (*deviance*) é sempre maior ou igual à zero, e decresce quando ocorre a inclusão de variáveis independentes na componente sistemática, até atingir o valor zero para o modelo saturado ou completo. Note que a utilização de um grande número de variáveis independentes visando a redução da *deviance* resulta em um modelo com alto grau de complexidade de interpretação, portanto, o mais indicado é escolher modelos simples com desvio moderado, localizados entre os modelos mais complicados e os que se ajustam bem aos dados. Conforme Myers e Montgomery (1997), usualmente costuma-se proceder a análise de *deviance* utilizando o ponto crítico $\chi^2_{(n-p)}(\alpha)$ da distribuição qui-quadrado ao nível de significância igual a α, sendo n o número de observações e p o número de parâmetros do modelo. Portanto, se

$$D(y;\mu,\phi) \leq \chi^2_{(n-p)}(\alpha), \qquad (55)$$

pode-se considerar que há evidências que o modelo proposto esteja bem ajustado aos dados, a um nível de φ % de significância, caso contrário deve-se descartar o modelo pois o mesmo pode ser considerado inadequado.

2.4.2.3 Resíduo Componente do Desvio (Resíduo *Deviance*)

Conforme Lee e Nelder (1998), o resíduo *deviance* é o que mais se aproxima da distribuição normal e recomendam a análise gráfica destes resíduos na verificação da adequação do modelo de regressão. Para cada observação (*i*) da variável dependente *y*, pode-se definir o desvio $r_i^d = D_i(y_i;\hat{\mu}_i)$, de tal modo que

$$D(y;\hat{\mu},\hat{\phi}) = \sum_{i=1}^{n}\left(r_i^d\right)^2 , \tag{56}$$

em que

$$r_i^d = \text{sinal}(y_i - \hat{\mu}_i)\left\{2\left[l_i(\tilde{\mu},\hat{\phi}) - l_i(\hat{\mu}_i,\hat{\phi})\right]\right\}^{1/2}. \tag{57}$$

sendo que a *i*-ésima observação contribui com a quantidade $\left(r_i^d\right)^2$ para o desvio e uma observação com um valor absoluto grande de r_i^d, pode ser vista como discrepante.

O resíduo ordinário padronizado é definido como

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\overline{Var(y_i)}}} \tag{58}$$

onde $\hat{\mu}_i = g^{-1}\left(x_i^T\hat{\beta}\right)$ e $\overline{Var(y_i)} = \{\hat{\mu}_i(1-\hat{\mu}_i)\}/\left(1+\hat{\phi}\right)$. Um gráfico desses resíduos contra a ordem das observações (*i*) não deveria mostrar nenhuma tendência e sim uma aleatoriedade. Através desse gráfico é possível verificar se existem pontos suspeitos de serem atípicos, ou seja, discrepantes em relação ao outros pontos.

Um outro gráfico importante é o de r_i contra $\hat{\eta}_i$ (valores ajustados), que é utilizado para verificar se a função de ligação está bem especificada, no caso deste não apresentar nenhuma tendência, ou seja, distribuído aleatoriamente.

2.4.2.4 Critério de Informação de Akaike (AIC)

Akaike (1974) *apud* Rao e Wu (2005) propõe o critério AIC (Akaike Information Criterion), que foi desenvolvido através dos estimadores de máxima verossimilhança (EMV), para decidir qual o modelo mais adequado quando se utiliza muitos modelos com quantidades diferentes de parâmetros, isto é, selecionar um modelo que esteja bem ajustado com um número reduzido de parâmetros. O AIC foi o primeiro critério baseado na informação de Kullback-Leibler (K-L) e assintoticamente não viesado para K-L. O critério AIC supõe que o modelo verdadeiro pertence ao conjunto de modelos candidatos. Sendo

$$AIC = -2l\left(\hat{\mu},\hat{\phi}\right) + 2(k+1) \tag{59}$$

onde $l\left(\hat{\mu},\hat{\phi}\right)$ é a função de máxima verossimilhança do modelo ajustado e k o número de parâmetro do modelo.

Na modelagem se busca o melhor modelo, o qual represente satisfatoriamente o mecanismo que gerou os dados. Portanto, o que se precisa é de uma medida da distância entre um modelo verdadeiro e vários modelos ajustados. Segundo Davison (2001), o AIC não propicia uma seleção consistente de modelos, e que em certas aplicações, indicam modelos mais complexos do que poderiam ser. Conforme Huvich e Tsai (1989), o AIC não é adequado para modelagem com pequenas amostras. Sendo um critério assintoticamente eficiente, mas não é assintoticamente consistente.

Huvich e Tsai (1989) desenvolveram o critério de seleção de modelos, o AICc, derivando a discrepância esperada da informação de Kullback-Leibler diretamente dos modelos de regressão. E demonstram que o AICc, tem melhor desempenho que o AIC em pequenas amostras, que também adotaram a suposição de que o modelo verdadeiro pertence ao conjunto de modelos ajustados. Como no caso do AIC, os parâmetros associados ao modelo candidato são estimados por máxima verossimilhança, sendo

$$AIC = -2l\left(\hat{\mu},\hat{\phi}\right) + 2(k+1)\left(\frac{n}{n-k-2}\right) \tag{60}$$

Conforme Davison (2001), o critério AICc pode aumentar apreciavelmente a probabilidade de se escolher o modelo verdadeiro, particularmente na seleção de modelos de regressão e de séries temporais. Rao e Wu (2005) propuseram a inclusão de uma validação

cruzada ao critério AIC para seleção de modelos, no sentido de que com probabilidade igual a 1, para todo n grande, o critério escolhe o modelo verdadeiro a partir de um conjunto de modelos candidatos com dimensão finita.

Kuha (2004) analisa os critérios AIC e AICc por meio de simulação e com um conjunto de dados reais, para demonstrar que se pode obter informações úteis para a seleção do modelo verdadeiro. Ambos os critérios apresentaram boas aproximações e consistências, isto significa que eles identificam bons modelos candidatos, mas o autor defende o uso de ambos os critérios juntos, pois quando os critérios concordam no melhor modelo, isto fornece confiança na escolha. Segundo Torres (2005), no estudo de simulação com vários tamanhos de amostra, o critério AICc apresentou melhor desempenho do que o critério AIC. Embora, nos resultados da aplicação, foi observado desempenho similar por parte dos dois critérios.

2.4.2.5 Gráfico de Probabilidade meio-normal com Envelope

Como a distribuição dos resíduos não é conhecida, gráficos de probabilidade meio-normal com envelopes simulados são ferramentas de diagnóstico muito úteis (ATKINSON, 1985, p.34). A proposta é acrescentar ao gráfico meio-normal usual um envelope simulado que pode ser usado para decidir se as observações são consistentes com o modelo ajustado.

Este gráfico é construído a partir da simulação de k valores (estatísticas de ordem) para cada valor estimado pelo modelo saturado, em seguida calcula suas médias, valores mínimos e máximos de cada valor estimado. Esses valores mínimos e máximos das k estatísticas de ordem produzem o envelope. Assim, o gráfico apresentará um intervalo para cada valor estimado ordenadamente contra os escores meio-normais

$$\Phi^{-1}\{(i + n - 1/8) / (2n + ½)\}, \qquad (61)$$

onde $\Phi(\cdot)$ é a função de distribuição acumulada da distribuição normal padrão e n é o número de observações.

Os correspondentes valores dos resíduos absolutos que se encontram fora dos limites fornecidos pelo envelope simulado ou proximidade dos pontos aos limites do envelope merecem uma pesquisa adicional. Caso ocorram tendências não aleatórias dos resíduos dentro

do envelope gerado, podem indicar escolha incorreta da distribuição de probabilidade para os dados, ou da função de ligação.

2.4.2.6 Alavanca Generalizada

A alavanca generalizada é um componente importante na análise de influência em modelos de regressão. Usualmente, é medido pelos elementos h_{ij} da matriz H que é conhecida como matriz de projeção ou "matriz chapéu" ($H = X(X'X)^{-1}X'$) e é usado para avaliar a importância individual de cada observação no próprio valor ajustado. Segundo Cordeiro (1986), no modelo linear normal, por exemplo, é muito razoável utilizar h_{ii} como uma medida da influência da i-ésima observação sobre o próprio valor ajustado.

Supondo que todos os pontos exerçam a mesma influência sobre os valores ajustado, pode-se esperar que os elementos h_{ii} da diagonal da matriz H sejam definidos por k / n, onde k é o somatório dos elementos h_{ii} e n é o número de observações definido pelas variáveis independentes. Uma sugestão é examinar aqueles pontos tais que $h_{ii} \geq 2k/n$, que são conhecidos como grandes pontos de alavanca. Ou seja, o valor de h_{ii} associado a i-ésima observação y_i é duas vezes maior que a média de todos os h_{ii} da diagonal da matriz H.

Recentemente, Wei *et al.* (1998) generalizaram a definição de pontos de alavanca para modelos bastante gerais onde a variável dependente seja contínua. Nessa generalização além dos modelos de regressão utilizados incluíram os métodos de estimação por máxima verossimilhança e Bayesiano. Contudo neste trabalho, não abordaremos o enfoque Bayesiano.

A alavanca generalizada proposta por Wei *et al.* (1998) é definida como

$$GL\left(\tilde{\theta}\right) = \frac{\partial \tilde{y}}{\partial y^{T}} = \left(\frac{\partial \tilde{y} i}{\partial y_{s}}\right)_{n \times n}, \qquad (62)$$

onde θ é um vetor m-dimensional tal que $E(y) = \mu(\theta)$ e $\tilde{\theta}$ é um estimador de θ, com $\tilde{y} = \mu\left(\tilde{\theta}\right)$. A partir desta definição, se pode observar que a alavanca generalizada para a i-ésima observação é a razão súbita de mudança do i-ésima valor predito em relação ao s-ésimo valor da variável dependente. Assim, essa medida de influência das observações é definida a partir do modelo ajustado. Desta forma, as observações com $\partial \tilde{y}_i / \partial y_i = GL_{ii}$ grandes são consideradas pontos de alavanca.

Considerando $\hat{\theta}$ um estimador de máxima verossimilhança de θ, assumindo que exista e seja único, e que o logaritmo da função de verossimilhança tem derivadas contínuas de segunda ordem em relação à θ e a *y*. Wei *et al.* (1998) mostraram que a alavanca generalizada é expressa por

$$GL(\theta) = \frac{\partial \mu}{\partial \theta^T}\left(-\frac{\partial^2 l}{\partial \theta \partial \theta^T}\right)^{-1}\left(\frac{\partial^2 l}{\partial \theta \partial y^T}\right), \tag{63}$$

2.4.2.7 Distância de Cook's

Outra medida importante na análise de diagnóstico é detecção de observações influentes, ou seja, identificação de pontos que exercem um peso desproporcional nas estimativas dos parâmetros do modelo ajustado. Desta forma se está interessado em conhecer o grau de dependência entre o modelo ajustado e o vetor de observações *y*, posto que, é preocupante se pequenas modificações nestas observações produzem mudanças bruscas nas estimativas dos parâmetros do modelo. Entretanto, se tais observações não alterarem os principais resultados do ajustamento do modelo, pode-se confiar mais no modelo encontrado.

Segundo Paula (2004), a distância de Cook é bastante utilizada para detectar a influência de cada observação nas estimativas dos parâmetros de regressão. Assim, esta medida identifica a influência da retirada da *i*-ésima observação sobre as estimativas dos parâmetros do modelo, sendo definido por

$$D_i = \frac{\left(\hat{\beta}_{(i)} - \hat{\beta}\right)^T X^T X \left(\hat{\beta}_{(i)} - \hat{\beta}\right)}{k\hat{\sigma}^2}, \tag{64}$$

onde D_i representa uma soma ponderada dos desvios entre as estimativas baseadas em $\hat{\beta}$ e $\hat{\beta}_{(i)}$. Assim, essa quantidade obtida pela soma mede a distância quadrática entre $\hat{\beta}$ e $\hat{\beta}_{(i)}$. Para evitar ajustar o modelo (*n* + 1) vezes, utiliza-se a aproximação usual para a distância de Cook, dada por

$$D_i = \frac{h_{ii}\left(r_i^*\right)^2}{k\left(1-h_{ii}\right)^2}, \tag{65}$$

para a *i*-ésima observação, a distância de Cook combina o resíduo padronizado r_i^* com a medida de alavanca h_{ii} e k o posto da matriz H, sendo, portanto, uma medida global de quão atípica esta *i*-ésima observação se apresenta no ajuste do modelo. Por conseguinte, D_i será grande quando a *i*-ésima observação fornecer r_i^* grande ou quando h_{ii} for próximo de um. Segundo Cook e Weisberg (1982), as observações serão consideradas influentes quando

$$D_i \geq \mathrm{F}_{\mathrm{k,\ n-k}}\,(0{,}50). \tag{66}$$

No entanto, McCullagh e Nelder (1989) propõem que seja utilizada a distância de Cook modificada, definida por

$$T_i = \left(\frac{n-k}{k}\frac{h_{ii}}{\left(1-h_{ii}\right)}\right)^{1/2} \cdot \left|\left(r_i^d\right)^2\right|, \tag{67}$$

em que $\left(r_i^d\right)^2$ é o desvio ou *deviance* residual, h_{ii} a medida de alavanca, k o posto da matriz H e n o número de observações em estudo. As observações serão consideradas influentes quando

$$T_i > 2\sqrt{k/n}\,. \tag{68}$$

2.5 SÍNTESE DOS MODELOS COMTEMPLADOS NO MÉTODO

Esta seção apresenta uma síntese dos modelos de regressão contemplados no método de orientação proposto no Capítulo 3 , com intuito de apresentar a estrutura e particularidade dos modelos.

2.5.1 *Modelo Linear Normal*

O modelo de regressão linear Normal permite prever a relação, entre uma variável dependente quantitativa contínua e uma ou mais variáveis independentes, podendo ser quantitativa ou qualitativa, caso possua apenas dois níveis, caso contrário, utilizam-se

variáveis *dummy*. Segundo Cordeiro e Lima Neto (2004), o modelo de regressão linear normal é um modelo que ajusta, a um conjunto de dados, uma equação que representa a relação entre as variáveis dependentes e independentes de forma linear, podendo ser simples ou múltipla. De acordo com Fahrmeir e Tutz (1994), a relação linear é definida pelos parâmetros do modelo de regressão quando estão elevados somente à primeira potência e não porque a variável dependente *y* é função linear das variáveis independentes.

Sendo a relação entre duas ou mais variáveis de forma linear nos parâmetros, o modelo de regressão é denominado de modelo múltiplo e apresenta a seguinte forma:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \varepsilon_i, \tag{69}$$

em que *y* é a variável dependente; α e β_i são coeficientes de regressão; ε é o erro aleatório; o termo de erro deve apresentar $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ e $Cov(\varepsilon_i,\varepsilon_j) = 0$; para todo $i \neq j$. Sendo que, o termo de erro deve seguir a distribuição normal com média zero e variância constante.

A expressão da equação (69) pode ser expressa de forma matricial, assim a relação entre duas ou mais variáveis fica:

$$y = \beta X + \varepsilon, \tag{70}$$

ou seja,

$$y = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}; \quad X = \begin{bmatrix} 1 & (x_{11} - \bar{x}_1) & \cdots & (x_{k1} - \bar{x}_k) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 1 & (x_{1n} - \bar{x}_1) & \cdots & (x_{kn} - \bar{x}_k) \end{bmatrix}; \quad \beta' = \begin{bmatrix} \alpha \\ \cdot \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix}; \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix} \tag{71}$$

em que *y* é o vetor de ordem (*n x 1*) de observações da variável dependente; β é o vetor de ordem (*k+1 x 1*) dos coeficientes de regressão, incluindo o intercepto α; *X* é a matriz de ordem (*n x k+1*) dada pelos valores das variáveis independentes, sendo que a primeira coluna é de valores um, buscando obter o valor do intercepto e, por fim, ε é o vetor de ordem (*n x 1*) dos erros aleatórios.

De acordo com Montgomery (1997), o modelo de regressão linear simples ou múltipla pode, também, ser utilizado para analisar dados que provenham de experimentos planejados ou não. Segundo Werkema e Aguiar (1996), a variável dependente utilizada na

modelagem pelo modelo Normal linear deve seguir a distribuição de probabilidade Normal e o termo de erro apresentará aleatoriedade.

2.5.2 *Modelo Logístico Linear*

O modelo de regressão Logístico linear permite prever a relação, entre uma variável dependente qualitativa ou categórica e uma ou mais variáveis independentes, podendo ser qualitativa ou quantitativa. Segundo Agresti (1996), denota-se a resposta de uma variável dependente qualitativa nominal como 0 ou 1, de forma usual como fracasso ou sucesso, a respeito de algum evento de interesse. Conforme Hosmer e Lemeshow (1989) pode-se prever as probabilidades de um evento possuir alguma característica ($y = 1$), contra a possibilidade da ausência de característica ($y = 0$). A distribuição que especifica a probabilidade de ocorrer sucesso em um referido evento é descrito como $P(y = 1) = \pi$ e $P(y = 0) = (1 - \pi)$ de ocorrer fracasso.

Assim, considerando um conjunto de k variáveis independentes, denotado pelo vetor X_i (x_1, x_2, ..., x_k), onde a relação entre a variável dependente e demais independentes é descrita como a probabilidade condicional de $P(y = 1)$ dado os valores das variáveis independentes X_i, definida por $P(y = 1/ X) = \pi(x)$.

Segundo Dobson (1990), a equação mais adequada para a modelagem dessa probabilidade é dada pelo modelo logístico:

$$g(x)= \alpha + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon_i, \tag{72}$$

em que $g(x)$ é uma função que modela a média da variável dependente; α e β_i são coeficientes de regressão; ε é o erro aleatório.

O modelo descreve como a proporção de sucessos ($y = 1$) é influenciada pelas variáveis independentes. A proporção esperada de sucessos ($y = 1$) é denotada por $\pi = E(y)$.

Assim, a função de regressão logística é definida como:

$$\pi(x)=\left(\frac{e^{g(x)}}{1+e^{g(x)}} \right) \tag{73}$$

Segundo Adimari e Ventura (2001), por razões de simetria o modelo de regressão logística expressa os coeficientes como logits. Os logits são medidas da força do relacionamento entre as variáveis e, como são simétricos podem ser comparados.

A equação de regressão linear fornece o Logit ou o log de chances:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = g(x) \tag{74}$$

isto é, o modelo de regressão logística é o logaritmo natural da probabilidade de ocorrência num grupo, dividido pela probabilidade de ocorrência no outro grupo. A expressão (74) apresenta o quociente $\pi / 1-\pi$ é denominado *odds*, que representa quantas vezes o sucesso é mais provável que o fracasso (AGRESTI, 1996).

2.5.3 ***Modelo Probit***

O modelo de regressão Probit permite prever a relação, entre uma variável dependente qualitativa ou categórica e uma ou mais variáveis independentes quantitativas. Segundo Dobson (1990), na análise do Modelo Probit é assumindo que a variável dependente y_i é qualitativa nominal, ou seja, a ocorrência de um evento assumindo apenas dois valores, como fracasso e sucesso. E todas as variáveis independentes são quantitativas. Uma variável dependente qualitativa nominal pode ser usada para prever as probabilidades de um evento possuir alguma característica (y = 1), contra a possibilidade da ausência de característica (y = 0) (HOSMER e LEMESHOW, 1989).

Segundo Rocha e Dantas (2003), o modelo acima apresenta a função de distribuição acumulada, denotada por $F(x) = P(X \leq x)$, $-\infty \leq x \leq +\infty$. Desta forma, a função de distribuição acumulada da distribuição Normal padrão, é descrita como

$$P_i = P(y_t = 1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{g(x)} e^{-t^2/2} \partial t \tag{75}$$

em que t é uma variável normal padronizada, $t \sim N(0,1)$; P_i representa a probabilidade de ocorrer um evento de interesse, e é medida pela área da curva normal padrão de $-\infty \leq g(x)$. Podendo ser reescrita como

$$P(y_t = 1) = \int_{-\infty}^{g(x)} \Phi(t)\partial t = \Phi(\beta' x) \tag{76}$$

em que a função $\Phi(\cdot)$ é a notação usual para a distribuição Normal padrão acumulada; e β é o vetor de parâmetro das variáveis independentes consideradas.

2.5.4 *Modelo Logit*

O modelo de regressão Logit permite prever a relação, entre uma variável dependente qualitativa ou categórica e uma ou mais variáveis independentes qualitativas. Segundo Agresti (1996), o modelo Logit se diferencia do modelo Log-linear pelo fato de não assumir associação entre as variáveis independentes com a variável dependente, ou seja, assume que a variável dependente *y* é associada com cada variáveis independentes x_1, x_2, ..., x_n, mas os efeitos de cada variável independente em *y* são os mesmos para cada combinação de níveis das outras variáveis. Ou seja, o modelo Logit não assume a associação (yx_1x_2) nem associações de maior ordem ($yx_1x_2...x_n$).

Segundo Paula (2004), o modelo Logit tem resultados idênticos aos do modelo de regressão Logístico linear. Ademais, o modelo é originado da função distribuição Logística acumulada e é representado por

$$L(\pi) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_i X_i \tag{77}$$

em que π_i é a probabilidade de sucesso de um referido evento, $(1 - \pi_i)$ é a probabilidade de fracasso de um referido evento, β_i*'s* são coeficientes desconhecidos do modelo de regressão, X_i é uma matriz de variáveis independentes, e $L(\pi_i)$ é o log da razão de chances (probabilidades).

Observa-se que:

$$1 - \pi_i = \frac{1}{1 + \exp(\beta_0 + \beta_i X_i)} \tag{78}$$

Assim, a expressão (78) pode ser reescrita como:

$$\frac{\pi_i}{1-\pi_i} = \frac{1+\exp(\beta_0+\beta_i X_i)}{1+\exp(-\beta_0+\beta_i X_i)} \tag{79}$$

Assim, π_i / (1 – π_i) é a razão de probabilidade de um evento ocorrer. Segundo Haad e McConnel (2005), o modelo logit supõe que o log da razão de chances se relaciona linearmente com X_i's e com os parâmetros. Desta forma, π_i varia de 0 a 1, o logit L(π_i) varia de – ∞ a + ∞. Desta forma, as probabilidades se situam entre 0 e 1 e os logits não se restringem a esses limites.

2.5.5 ***Modelo Log-linear***

O modelo de regressão Log-linear permite prever a relação, entre uma variável dependente quantitativa discreta e uma ou mais variáveis independentes, podendo ser qualitativas e/ou quantitativas. Segundo Demétrios (2002), o modelo Log-linear corresponde ao caso onde $y \sim Poi(\mu)$, com parâmetro natural da distribuição de Poisson sendo igual a logμ, em estudos de tabelas de contingência com qualquer número de variáveis.

Considere uma tabela de contingência de classificação cruzada de tamanho I x J, com n observações. As probabilidades (π_{ij}) das células da tabela de contingência são determinadas pelos totais marginais das linhas e colunas, expressa como

$$\pi_{ij} = \pi_{i*} + \pi_{*j}\,, \quad i = 0, 1, 2, ...,I \qquad j = 0, 1, 2, ..., J \tag{80}$$

em que as freqüências esperadas ($\mu_{ij} = n\pi_{ij}$) é = $\mu_{ij} = \pi_{i*}\ \pi_{*j}$, para todo i e j. Segundo Agresti (1996), as probabilidades referente as células (π_{ij}) podem ser descritas como parâmetros das distribuições de probabilidade Binomial e Multinomial. No entanto, os modelos Log-lineares são os mais recomendados, pois utiliza o parâmetro μ_{ij} ao invés de π_{ij}, como valores esperados que seguem a distribuição de Poisson.

Segundo Cordeiro e Lima Neto (2004), os modelos Log-lineares são recomendados para a análise de dados de contagem, mesmo quando o tempo de observação não é o mesmo para cada unidade amostral. Em particular, se tem um conjunto de k tabelas 2 x 2, uma modelagem possível para a taxa média por unidade de tempo em cada célula, é descrita como:

$$\log\mu_{11} = \alpha,$$

$$\log\mu_{21} = \alpha + \beta,$$

$$\log\mu_{1i} = \alpha + \gamma_i,$$

$$\log\mu_{2i} = \alpha + \beta + \gamma_i + \delta_i, \text{ para } i = 2, ..., k. \qquad (81)$$

assim, tem-se a reparametrização $(\mu_{11}, \mu_{21}, ..., \mu_{1k}, \mu_{2k}) \rightarrow (\alpha, \beta, \gamma_2, \delta_2, ..., \gamma_k, \delta_k)$. A razão de taxas na *i*-ésima tabela fica definida por

$$\upsilon_i = \mu_{2i} / \mu_{1i} = \exp(\beta + \delta_i), \text{ com } \delta_1 = 0. \qquad (82)$$

portanto, testar $H_0 = \upsilon_i = ... = \upsilon_k$ é o mesmo que testar na nova parametrização $H_0 = \delta_2 = ... = \delta_k = 0$, o que significa não haver interação entre as tabelas. É importante salientar que δ_i é o efeito da *i*-ésima tabela com relação à primeira tabela.

2.5.6 ***Modelo Poisson***

O modelo de regressão Poisson é utilizado na modelagem, quando a variável dependente é quantitativa discreta, ou seja, uma contagem ou uma taxa, e as variáveis independentes podem ser quantitativas e/ou qualitativas. Segundo McCullagh e Nelder (1989), a distribuição de Poisson é definida como limite da distribuição Binomial quando $np = \mu$ fixo e $n \rightarrow \infty$. Desta forma, a distribuição supõe que a variável de interesse assume valores inteiros não-negativos e, em particular, não existe um limite superior. A função de probabilidade de Poisson é expressa por

$$P(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}, \quad \mu > 0, \quad y = 0, 1, 2, ... \qquad (83)$$

em que *y* é a variável aleatória, μ é a média da distribuição de probabilidade e $y \sim Poi(\mu_i)$.

Conforme Dean (1992), quando se supõe distribuição de Poisson para os dados, espera-se que a média e variância dos dados sejam iguais, ou seja, $E(y) = \mu$ e $Var(y) = \mu$. Cordeiro e Lima Neto (2004) relatam que o modelo Poisson, ao contrário do modelo de regressão linear Normal, supõe que a variância seja dependente da média e pode ser aplicado para modelar o número de ocorrência de um dado evento de interesse.

Segundo Paula (2004), quando a média da distribuição é grande ($\mu \to \infty$), y segue aproximadamente uma distribuição Normal de média μ e desvio padrão $\sqrt{\mu}$. Assim, se desejar aplicar um modelo de regressão linear Normal para explicar μ, ocorrerá o inconveniente de o desvio padrão depender da média, o que inviabiliza o uso de um modelo de regressão linear Normal homocedástico, quer dizer, um modelo de regressão linear Normal com variâncias constante. Segundo Fahrmeir e Tutz (1994), uma forma de contornar o problema de pressuposição de variância constante, ou seja, não dependente da média, é através da aplicação de uma transformação na variável dependente y de modo que seja alcançado a normalidade dos dados e variância constante, mesmo que aproximadamente.

Segundo Dean (1992), numa aplicação prática, uma variável dependente que segue a distribuição de probabilidade Poisson, dificilmente tenderá a normalidade. Segundo Sant'Anna e Caten (2004), há situações em que os dados seguem a distribuição de probabilidade Poisson, mas o Modelo de Poisson não apresenta bom ajuste, em função de um variabilidade excedente (superdispersão). Nestes casos, se faz necessário uma adequação na estrutura do modelo e no método de estimação dos parâmetros, pois a validade das inferências depende de quão bem o modelo de regressão descreve os dados observados.

2.5.7 *Modelo Binomial Negativa*

O modelo de regressão Binomial Negativa também é utilizado na modelagem, quando a variável dependente quantitativa é discreta, ou seja, uma contagem ou uma taxa. As variáveis independentes podem ser quantitativas e/ou qualitativas. Segundo Paula (2004), em certas situações práticas, o Modelo Binomial Negativa apresenta superioridade em relação ao Modelo de Poisson, em termos do desempenho dos estimadores e controle da variabilidade dos dados, pois não necessita que a variância seja igual a média.

Assim, supondo que a variável dependente assume valores inteiros não-negativos e, em particular, não existe um limite superior. A função de probabilidade Binomial Negativa é expressa por

$$P(Y=y)=\binom{y+k-1}{y}\cdot p^{k}\cdot(1-p)^{y}, \quad y=0,\,1,\,2,\,....;\;\; 0<p<1; \tag{84}$$

em que k é o número de sucesso, conforme um determinado tempo de espera de ocorrência; p é a probabilidade do evento ocorrer e $y \sim BN(k, p)$ (MURTEIRA, 1990).

Segundo Paula (2004), considera-se um conjunto de k variáveis independentes, denotado pelo vetor X_i (x_1, x_2, ..., x_k), onde a relação entre a variável dependente y e demais independentes é descrita da forma

$$g(x) = \alpha + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon_i, \quad (85)$$

em que $g(x)$ é uma função que modela a média da variável dependente; α e β_i são coeficientes de regressão; ε é o erro aleatório.

Segundo Cordeiro e Lima Neto (2004), este modelo pertence à classe dos Modelos Lineares Generalizados, e admite que, como a variável dependente segue a distribuição Binomial Negativa, ou seja, $y \sim BN(\mu_i, \phi)$, a esperança matemática (média) é representada por μ_i e variância seja $\mu_i + \mu^2_i / \phi$.

2.5.8 *Modelo de Quase-verossimilhança*

O modelo de regressão de Quase-verossimilhança permite prever a relação, entre uma variável dependente quantitativa discreta ou contínua e uma ou mais variáveis independentes, podendo ser qualitativa ou quantitativa. Este modelo de regressão é apresentado na seção 2.2, e descrito detalhadamente na seção 2.2.6.

2.5.9 *Modelo Beta*

O modelo de regressão Beta permite prever a relação, entre uma variável dependente quantitativa contínua, com restrição no valor mensurado e uma ou mais variáveis independentes, podendo ser qualitativa ou quantitativa. Este modelo de regressão é apresentado na seção 2.3, descrevendo características de uso, estrutura do modelo, distribuição de probabilidade, método de estimação e teste de significância.

3 MÉTODO PROPOSTO

3.1 INTRODUÇÃO

Este capítulo apresenta o método de orientação à modelagem de dados mensurados em proporção, considerando a classificação da natureza das variáveis dependente e independentes. Ademais, o método proposto apresenta uma estrutura que permite o discernimento quanto à escolha inicial de outros modelos de regressão contemplados.

Segundo Jung (2004), a estatística, enquanto área do conhecimento, não apresenta uma teoria restrita, mas uma estrutura teórica geral que poderia ser usada em todas as situações e realidades. Além disso, esta característica de apresentar uma estrutura teórica geral é a marca da natureza dos seus saberes. Conforme Larson e Farber (2003), antes de escolher a análise estatística apropriada é necessário que seja realizada a classificação da variável de interesse, pois a adequação da técnica está diretamente relacionada ao tipo de variável em questão.

O método aqui proposto leva em consideração a classificação das variáveis independentes e da variável dependente no estudo, pois a caracterização da natureza destas variáveis em qualitativa ou quantitativa é parte importante na desenvoltura do processo de construção do mapeamento, o qual conduz ao encontro dos modelos de regressão mais adequados.

O aspecto mais simples de um tratamento estatístico de dados passa pela necessidade de se observar que "tipo" de dados devem ser coletados para o experimento. As informações contidas nos dados são importantes e compõem a base do critério científico na tomada de decisão sobre um processo de produção. Vale a pena lembrar que um bom planejamento da coleta de dados e da análise de investigação do experimento poupa muitas horas de trabalho no tratamento dos dados. Pode-se também poupar a necessidade de retomada de novos dados. Dependendo do objeto a ser estudado, suas características e sua natureza podem ser conseguidas através de um processo de mensuração específico.

3.2 CLASSIFICAÇÃO DE VARIÁVEIS

Conforme Larson e Farber (2003), a qualidade da análise estatística está relacionada com a classificação da variável em estudo, em virtude da adequabilidade da técnica estatística depender da natureza que a variável em estudo apresenta. Segundo Montgomery e Peck (1992), na realização de uma abordagem estatística, ao projetar e analisar um processo, é necessário previamente possuir uma idéia do que será estudado, de como os dados serão coletados, da natureza dos dados e um entendimento qualitativo de como serão analisados.

Define-se variável dependente como a característica de qualidade de um produto ou processo que pode ser mensurada, por exemplo, o tempo de produção de um processo de manufatura, o peso dos produtos fabricados, o número de produtos refugados, etc. Define-se variável independente como fatores que podem ou não exercer influência sobre a característica de qualidade de um produto ou processo, por exemplo, o número de máquinas no processo, a temperatura do processo de manufatura, etc.

De acordo com a estrutura numérica, as variáveis podem ser classificadas em: (*i*) Quantitativas - se os resultados das observações forem expressos sempre através de números, que representam contagens ou medidas, pertencentes a um conjunto dos números reais; (*ii*) Qualitativas - se os resultados das observações serão expressos através de categorias ou níveis que se distinguem por alguma característica não-numérica, apresentando ou não ordenamento. A Tabela 4 apresenta um resumo das variáveis por tipo de mensuração.

Tabela 4– Classificação das variáveis por tipo de mensuração

Variável	Níveis	Sub-níveis	Exemplo
Qualitativa/Categórica	Nominal	Binária/Dicotômica Polinomial/Politômica	Defeituoso/Não defeituoso; Sim/Não, etc. Fornecedor; Tonalidade; etc.
Qualitativa/Categórica	Ordinal		Grau de Satisfação; Escolaridade; etc.
Quantitativa	Discreta		Nº de peças; Nº de operadores; etc.
Quantitativa	Contínua		Proporção de defeitos; Tempo de vida útil, etc.

Fonte: Magalhães e Lima (2002).

Segundo muitos autores, o termo variável quantitativa é geralmente empregado para designar variáveis quantitativas discretas e contínuas: as discretas apresentam valores expressos como números inteiros, além da possibilidade desses números serem utilizados em forma de níveis ou distanciamento de uma escala e as contínuas, como valores expressos dentro de um intervalo de variação.

Segundo Magalhães e Lima (2002), as variáveis qualitativas são também chamadas de categóricas, e os níveis dessas variáveis são definidos como nominais e ordinais. As variáveis qualitativas nominais apresentam subníveis através da denominação, por exemplo, um produto defeituoso ou não defeituoso. Essas variáveis podem ser classificadas em binárias ou dicotômicas (quando apresentam duas categorias) e polinomiais ou politômicas (quando apresentam mais de duas categorias). As variáveis qualitativas ordinais apresentam ordenação das várias categorias, possibilitando verificar graus de intensidade entre elas. Conforme Montgomery (2001) e Hair *et al.* (1998), em estudos experimentais, é comum chamar as variáveis independentes de fatores, e os seus níveis como níveis dos fatores.

É importante salientar que não é adequado tratar as diferentes classificações de medição com os mesmos testes ou métodos estatísticos. Desta forma, é essencial observar qual é a classificação das variáveis dependentes e independentes existentes no conjunto de dados em estudo, pois a correta identificação e classificação dessas variáveis nos levam a considerar, por exemplo, classes de modelos de regressão diferentes.

Assim, a proporção de produtos não conformes (p_i) de um conjunto de informações, denominada de amostra (n_i) é classificada como uma variável aleatória contínua (y_i), podendo ser representada por um valor de grandeza no conjunto dos números reais.

Utiliza-se a análise univariada quando se estuda a distribuição de apenas uma variável dependente e a análise multivariada, para casos de duas ou mais variáveis dependentes. Indistintamente, estas análises são utilizadas com uma ou mais variáveis independentes.

Para Hair *et al.* (1998), a análise simultânea de duas ou mais variáveis de cada indivíduo ou objeto sob investigação consiste na investigação da relação de dependência entre elas com o objetivo de predizer o valor da variável dependente com base nas observações de múltiplas variáveis independentes. Conforme Park (1996), a análise das variáveis na modelagem dos dados deve ser de forma simultânea, de maneira que seus efeitos não possam ser interpretados de forma isolada. Deve-se entender que há metodologias específicas e que o tratamento matemático de qualquer conjunto de dados sempre pode ser processado ou investigado por um modelo de regressão, mas, se este não tiver sentido de validade ou relação causal, não deve ser considerado, pois o resultado não terá relação com o objetivo de conhecimento.

Um dos principais intuitos de diferenciar a aplicação dos modelos de regressão é obter sucesso na análise estatística, pois isso depende da disponibilidade de o modelo de regressão ajustar-se "satisfatoriamente" aos dados. No caso de um ou mais modelos ajustarem-se aos dados, surge o problema da escolha do "melhor modelo". Esta escolha será feita fundamentada em estudos teóricos de adequação do modelo, conforme visto na seção 2.4.

3.3 CLASSIFICAÇÃO DOS MODELOS COMTEMPLADOS NO MÉTODO

Esta seção apresenta a classificação dos modelos de regressão contemplados no método, apresentando para cada um dos modelos as restrições de uso quanto à natureza das variáveis dependente e independentes do estudo.

3.3.1 *Modelo de Regressão Linear Normal*

O modelo de regressão linear Normal permite prever a relação entre uma variável dependente quantitativa contínua e uma ou mais variáveis independentes, que podem ser quantitativas ou qualitativas, caso possua apenas dois níveis; do contrário, utilizam-se variáveis *dummy*.

3.3.2 *Modelo Logístico Linear*

O modelo de regressão Logístico linear permite prever a relação entre uma variável dependente qualitativa ou categórica e uma ou mais variáveis independentes, podendo ser qualitativas ou quantitativas.

3.3.3 *Modelo Probit*

O modelo de regressão Probit permite prever a relação entre uma variável dependente qualitativa ou categórica e uma ou mais variáveis independentes quantitativas.

3.3.4 *Modelo Logit*

O modelo de regressão Logit permite prever a relação entre uma variável dependente qualitativa ou categórica e uma ou mais variáveis independentes qualitativas.

3.3.5 *Modelo Log-linear*

O modelo de regressão Log-linear permite prever a relação entre uma variável dependente quantitativa discreta e uma ou mais variáveis independentes, que podem ser qualitativas e/ou quantitativas.

3.3.6 *Modelo Poisson*

O modelo de regressão Poisson é utilizado na modelagem quando a variável dependente é quantitativa discreta, ou seja, uma contagem ou uma taxa, e as variáveis independentes podem ser quantitativas e/ou qualitativas.

3.3.7 *Modelo Binomial Negativa*

O modelo de regressão Binomial Negativa também é utilizado na modelagem quando a variável dependente quantitativa é discreta, ou seja, uma contagem ou uma taxa. As variáveis independentes podem ser quantitativas e/ou qualitativas.

3.3.8 *Modelo de Quase-verossimilhança*

O modelo de regressão de Quase-verossimilhança permite prever a relação entre uma variável dependente quantitativa discreta ou contínua e uma ou mais variáveis independentes, que podem ser qualitativas ou quantitativas.

3.3.9 *Modelo Beta*

O modelo de regressão Beta permite prever a relação entre uma variável dependente quantitativa contínua, com restrição no valor mensurado, e uma ou mais variáveis independentes, que podem ser qualitativas ou quantitativas.

3.4 ESTRUTURA DO MÉTODO

Segundo Oliveira (1999), o método gráfico deve apresentar uma estrutura clara e concisa do fluxo de processos e/ou procedimentos, permitindo que o analista tenha o discernimento de proceder à escolha adequada quanto às etapas do processo ou procedimento a serem seguidos. Conforme Luporini e Pinto (1998), a compreensão do método, após a etapa do levantamento de dados, é essencial para a análise do processo em estudo, pois a técnica de investigação empregada poderá não servir adequadamente.

A elaboração do método de orientação apresentado na Figura 3 foi fundamentada através do estudo de cada modelo, associado com a característica das variáveis dependente e independentes que constituem o processo da análise estatística. Esta caracterização, em qualitativa ou quantitativa foi importante na elaboração do método, conduzindo à apresentação dos modelos de regressão mais adequados.

A estrutura do método foi constituída a partir do estudo e classificação de cada modelo de regressão iniciando com a classificação da variável dependente e, posteriormente, com a classificação das variáveis independentes. Como foi mostrado na seção 3.2, é importante classificar as variáveis do estudo em quantitativa ou qualitativa, em virtude da forma em que os valores são expressos. A classificação da natureza da variável dependente do estudo é o fator mais importante da análise de dados, pois permite conhecer mais sobre os dados da pesquisa e os métodos estatísticos a serem utilizados na análise.

Observa-se que qualquer elemento de um produto ou processo de produção que se tenha interesse em investigar, apresenta uma classificação correspondente em virtude da natureza de mensuração ou da forma como é medida (ver Tabela 4). Desse modo, buscou-se desenhar o fluxo das informações sobre as variáveis dependente e independentes de maneira a distinguir os passos de orientação do método.

A estrutura método proposto delimita-se no escopo dos modelos de regressão utilizados na modelagem da proporção, que é classificada como uma variável dependente quantitativa contínua. Embora uma variável quantitativa contínua apresente seus valores expressos no conjunto dos números reais, com variação de $-\infty$ a $+\infty$, o objeto deste estudo, "proporção", é classificado como uma variável quantitativa contínua e seus valores mensurados apresentam restrições, tendo variação entre 0 e 1.

Assim, buscou-se delinear de forma detalhada outras características relevantes desta variável e também a natureza das variáveis independentes que constituem o conjunto de dados, permitindo uma distinção na orientação do fluxo de informações. A estrutura dos modelos de regressão sugeridos por este trabalho permite que as variáveis independentes apresentem classificações de qualquer natureza, tanto quantitativas quanto qualitativas.

Uma informação importante é saber qual o tamanho da amostra que constitui o conjunto de dados, pois, de acordo com o estudo dos modelos de regressão, esta informação auxilia na tomada de decisão da escolha do modelo de regressão mais adequado. Uma forma de decidir sobre o tamanho da amostra é observar se o valor encontrado através do cálculo do número de observações (*n*) vezes a proporção média ($\overline{p}$) é superior ao valor nove, tal como $n\overline{p} > 9$.

Como exemplo, tem-se os passos que permitem orientar a utilização do método. Suponha um estudo onde a variável dependente é "Proporção de produtos não conformes às especificações (defeituosos)", num lote de *n = 18* produtos. Esta variável dependente apresenta $\overline{p}$ = 0,368, ou seja, que a proporção média dos produtos avaliados é de 36,8%. As variáveis independentes são: fornecedor da matéria-prima e tipo de máquina.

Iniciando os passos, a partir da classificação da variável dependente:

1. Qual a classificação da variável dependente?
   A "Proporção de produtos não conformes" é uma **variável Quantitativa**.
2. Qual a classificação da variável dependente quantitativa?
   A "Proporção de produtos não conformes" é **Quantitativa Contínua**.
3. A variável dependente quantitativa contínua é restrita ao intervalo 0 e 1?
   A "Proporção de produtos não conformes" é **Restrita ao intervalo [0,1]**.
4. Qual a classificação das variáveis independentes?
   As variáveis independentes investigadas são **Variáveis Qualitativas**.
5. O tamanho da amostra apresenta $n\overline{p} > 9$?
   O tamanho da amostra apresenta $n\overline{p}$ = 6,6, então é considerado como **Pequeno**.
6. Qual o modelo proposto?
   O modelo escolhido para este caso, foi **Modelo de Regressão Beta**.

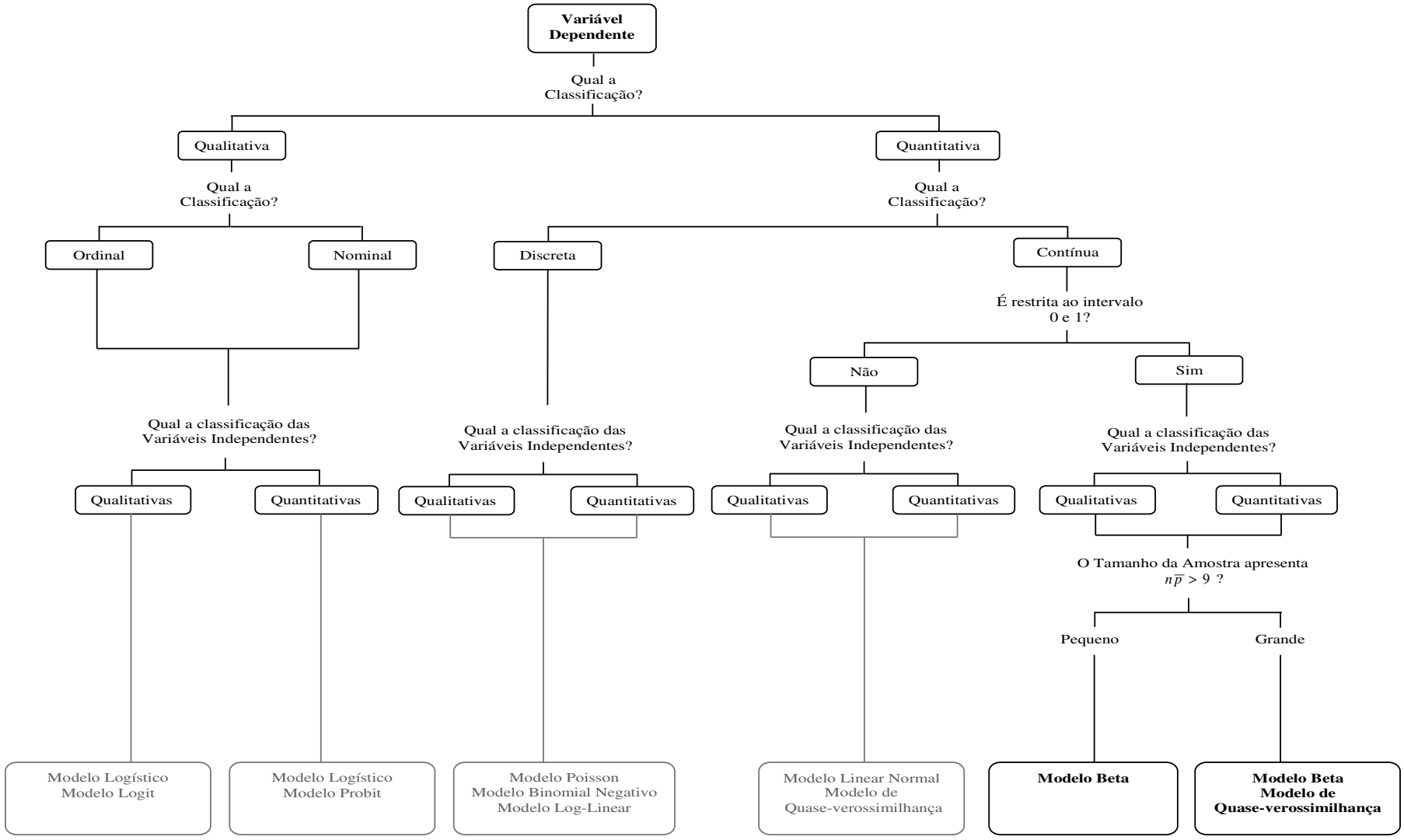


Figura 3– Método proposto para orientação à modelagem de dados mensurados em proporção

4 APLICAÇÃO DO MÉTODO

4.1 INTRODUÇÃO

Este capítulo apresenta uma aplicação do método apresentado na Figura 3 para orientação à modelagem de um conjunto de dados mensurados em proporção, utilizando os passos propostos no final da seção 3.4.

O estudo foi realizado numa empresa curtidora, a empresa Bracol Couros que é afiliada ao Grupo Bertin Ltda, localizada na região do Vale do Rio dos Sinos, produtora de couro acabado e fornecedora para as indústrias de calçados e de artefatos de couro. Das etapas do processo de manufatura da Empresa, a investigação consistirá na etapa de classificação da matéria-prima (couro) no estágio *wet blue*.

Para se conseguir resultados satisfatórios na utilização de matéria-prima e produto final, livre de erros de classificação, é necessário um processo de classificação das peles aperfeiçoado, que evite defeitos no produto final. Este processo de classificação é tanto mais crítico quanto mais nobre são os produtos fabricados (ARRIBA, 2005). Segundo Cot *et al.* (1992), o processo de classificação é a parte mais subjetiva da empresa curtidora e o qual causa as maiores perdas na cadeia produtiva. Os refugos produzidos por erros de classificação, geram um aumento do custo do produto fabricado, pelo menor aproveitamento da matéria-prima.

O erro de classificação da seleção pode ocorrer de duas maneiras diferentes. A primeira, e mais crítica, quando o classificador não reconhece um couro não apto para o produto a fabricar e classifica-o como apto. Por exemplo, um couro de classificação regular, classificado como excelente. Este erro gera custo no final do processo muito importante, tanto em desperdício do próprio material como em mão-de-obra, produtos químicos aplicados e o tempo utilizado na fabricação. A segunda maneira de ocorrer um erro de classificação é classificar como não apto para a fabricação de um produto, um couro apto. Por exemplo, um couro excelente classificado como regular. Aqui o erro se transforma em desperdício por não aproveitar um couro apto e direcionar sua aplicação para produtos que não precisariam desse tipo de seleção.

Na etapa do processo de manufatura da empresa em estudo, o estágio *wet blue*, se identifica a matéria-prima através de inspeção por classificadores, em lotes de diferentes tamanhos, quanto ao tipo de matéria-prima (couro). A inspeção consiste em avaliar se a classificação da matéria-prima corresponde à seleção inicial, por métodos cognitivos (utilizando alguns dos sentidos visão, olfato e tato) e o resultado é colocado na planilha de coleta de dados (ver APÊNDICE A).

A variável dependente para o estudo da modelagem se constituiu a proporção de produtos não conformes às especificações, ou seja, a proporção por erro de classificação da matéria-prima. Os fatores definidos pela empresa para serem investigados, pois possivelmente poderiam estar influenciando no processo de classificação da matéria-prima no estágio *wet blue*, são apresentados na Figura 4 e descritos a seguir:

- Seleção: a matéria-prima em seus diferentes tipos é classificada em grupos, conforme sua qualidade, cada um dos quais recebe o nome de seleção.

- Classificador: operador que realiza a classificação do couro nas diferentes seleções, no estágio *wet blue*.

- Procedência: a origem da matéria-prima adquirida pela empresa, apresentando características das regiões de criação de gado no país.

- Estágio da matéria-prima: constitui o estado de rebaixamento da superfície da matéria-prima. O estado rebaixado depende da etapa anterior do processo, pois o couro antes de chegar ao estágio *wet blue* (etapa de objeto deste estudo) passa por uma etapa de decapagem.

Assim os fatores em estudo, se constituíram como variáveis independentes qualitativas, pois os resultados são expressos através de categorias ou níveis sendo: (*i*) a seleção, em cinco diferentes tipos; (*ii*) o classificador que realiza a inspeção; (*iii*) a procedência da matéria-prima; e finalmente, (*iv*) o rebaixamento da matéria-prima.

Por o processo não apresentar variação das condições de trabalho, a coleta e investigação dos dados foi realizada durante um período de seis meses de acompanhamento do processo de produção da empresa, fornecendo um tamanho de amostra consideravelmente grande, 754 observações.

Seleção original que está sendo classificada (5 possíveis)

Classificador da matéria-prima (3 possíveis)

Procedência da matéria-prima (5 possíveis)

Estágio da matéria-prima classificada (2 possíveis)

BERTIN LTDA. **Unidade EV (RS)**
RESULTADO DA CLASIFICAÇÃO DE COUROS *WET BLUE* Tamanho do Lote
Período: **02 à 06.06.2004**

Itens	Seleção	Classificador	Procedência	Rebaixamento	Volume	Não Conforme	Conforme
1	A	GERALDO	LINS	Sim	105	0,76	0,24
2	A	ADELMIR	LINS	Sim	88	0,65	0,35
3	A	ADELMIR	SLMB	Não	100	0,43	0,57
4	B	VALDECIR	LINS	Não	200	0,08	0,92
5	A	GERALDO	LINS	Não	49	0,19	0,81
6	C	GERALDO	CACOAL	Sim	255	0,72	0,28
7	C	VALDECIR	CACOAL	Sim	174	0,05	0,95
8	D	GERALDO	LINS	Não	112	0,13	0,87
9	E	VALDECIR	LINS	Sim	520	0,16	0,84
10	A	ADELMIR	SLMB	Não	162	0,09	0,91
11	B	ADELMIR	LINS	Não	38	0,26	0,74
12	A	ADELMIR	REDENÇÃO	Não	400	0,11	0,89
13	E	GERALDO	REDENÇÃO	Sim	206	0,02	0,98
14	C	ADELMIR	RBTE	Não	1038	0,05	0,95
15	D	VALDECIR	REDENÇÃO	Sim	75	0,14	0,86
16	D	VALDECIR	RBTE	Não	87	0,97	0,03
17	A	ADELMIR	RBTE	Sim	24	0,67	0,33
18	E	GERALDO	LINS	Não	49	0,39	0,61
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
754	E	VALDECIR	CACOAL	Não	260	0,89	0,11

Resultado da classificação (proporção)

Figura 4– Planilha dos dados de classificação do couro no estágio wet blue

4.2 UTILIZAÇÃO DO MÉTODO

Esta seção apresenta os passos para a utilização do método elaborado no Capítulo 3 , a partir do conhecimento da variáveis dependente e independentes identificadas no processo de manufatura da empresa curtidora de couro.

Iniciando os passos propostos no método de orientação à modelagem dos dados, conforme Figura 3, tem-se:

1. Qual a classificação da variável dependente?

A "Proporção por erro de classificação" apresenta a natureza de um valor numérico, então é uma **Variável Quantitativa**.

2. Qual a classificação da variável dependente quantitativa?

A "Proporção por erros de classificação" apresenta a natureza de um valor numérico decimal, então é **Quantitativa Contínua**.

3. A variável dependente quantitativa contínua é restrita ao intervalo 0 e 1?

A "Proporção por erro de classificação" apresenta a natureza de um valor numérico **Restrito ao intervalo [0,1]**.

4. Qual a classificação das variáveis independentes?

As variáveis independentes investigadas apresentam a natureza de **Variáveis Qualitativas.**

5.O tamanho da amostra apresenta $n\overline{p} > 9$?

O tamanho da amostra do estudo apresenta $n\overline{p} =$ 139,9, então pode ser considerado como **Grande**.

6. Qual o modelo proposto?

A orientação do método apresenta dois modelos de regressão, **Modelo de Regressão Beta e Modelo de Quase-verossimilhança**.

4.2.1 *Análise dos Modelos Sugeridos*

O interesse encontra-se na modelagem da variável dependente (y) proporção por erro de classificação, pelas variáveis independentes: seleção (x_1), procedência (x_2), classificador (x_3) e rebaixamento (x_4). As variáveis independentes qualitativas foram tratadas com o auxílio de variáveis *dummy*. O uso deste tipo de variável permite uma análise de associação entre os níveis das variáveis independentes em relação a variável dependente.

A análise inicial deste conjunto de dados foi realizada por Arriba (2005), consistindo num ajuste de um modelo de regressão linear normal, com as quatro variáveis independentes e sem utilizar transformação na variável dependente. O ajuste deste modelo apresentou a variável independente "procedência" sem significância estatística, sendo retirada do modelo ajustado e um coeficiente de determinação de 0,24 (R^2 = 0,24). Técnicas gráficas não foram usadas, como também não foi examinada a diferença residual entre os valores observados e ajustados. Ademais, seu ajuste foi feito de uma forma geral, sem avaliar a influência dos níveis de cada variável independente.

Inicialmente analisou-se o conjunto de dados com intuito de verificar a disposição dos valores observados para que, em seguida fosse realizada a modelagem dos dados com base na orientação do método elaborado. A Figura 5 ilustra o gráfico da variável dependente "proporções de produtos não conformes" versus índices das observações. Pode-se observar que há uma aleatoriedade dos dados, mesmo havendo observações com valor zero e próximos de zero.

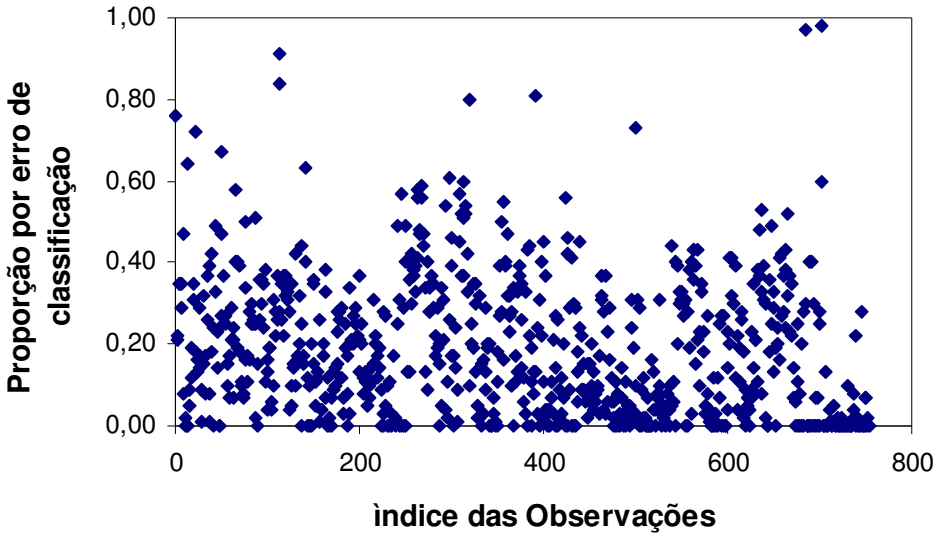


Figura 5 – Gráfico da proporção por erro de classificação versus o índice das observações

Lewis *et al.* (2001) relatam que, como a variável dependente é necessariamente não negativa, é possível que alguma transformação da variável seja conveniente para uma análise de regressão linear. A proposta de modelagem dos dados nesta dissertação foi realizada de duas maneiras, conforme os passos do método elaborado no Capítulo 3 . A primeira maneira é não assumindo *a priori* qualquer distribuição de probabilidade para a variável dependente, usando os dados originais sem transformação, apenas definindo a função de variância e a função de ligação, ajustando então o Modelo de Quase-verossimilhança descrito na seção 2.2.6. A segunda maneira é assumindo a distribuição de probabilidade Beta para a variável dependente, usando os dados originais sem transformação, ajustando um Modelo de Regressão Beta descrito na seção 2.3.6.

Conforme Tabela 5, a caracterização dos níveis das variáveis independentes foi assim definida: “seleção” (x_1) escala de 1 = baixo a 5 = alto, segundo qualidade e preço; “procedência” (x_2) escala de 1 a 5, representando os locais de origem da matéria-prima; “classificador” (x_3) escala de 1 a 3, representando os profissionais da empresa; e “rebaixamento” (x_4) escala: 1 = não e 2 = sim, segundo o estado de rebaixamento da matéria-prima.

Tabela 5– Caracterização dos níveis dos Fatores Controláveis

Fatores Controláveis	**Número de Níveis**	**Níveis Codificados**				
Seleção de Couro	5	1	2	3	4	5
Procedência	5	1	2	3	4	5
Classificador	3	1	2	3	---	---
Rebaixamento	2	1	2	---	---	---

Realizando uma análise preliminar dos efeitos das variáveis independentes e a variável dependente “proporção por erro de classificação”, nota-se na Figura 6(a) que a proporção por erro de classificação para a variável “seleção” apresenta uma tendência crescente da proporção à medida que aumenta o nível da variável “seleção”, ou seja, quanto mais nobre o couro, maior a proporção por erro de classificação. Na Figura 6(b) a variável “procedência” apresenta a origem 1 como menor produtora da proporção por erros de classificação e entre as outras origens não há uma relação de tendência com a proporção por erro de classificação, pois as proporções destas origens apresentam similaridade.

Observa-se que o avaliador 3 da variável “classificador” apresentou maior proporção por erro de classificação que os avaliadores 1 e 2. O nível rebaixado da variável “rebaixamento” apresenta menor proporção por erro de classificação, ou seja, o estado rebaixado da superfície da matéria-prima apresenta menor proporção por erro de classificação, conforme Figura 6(c) e (d).

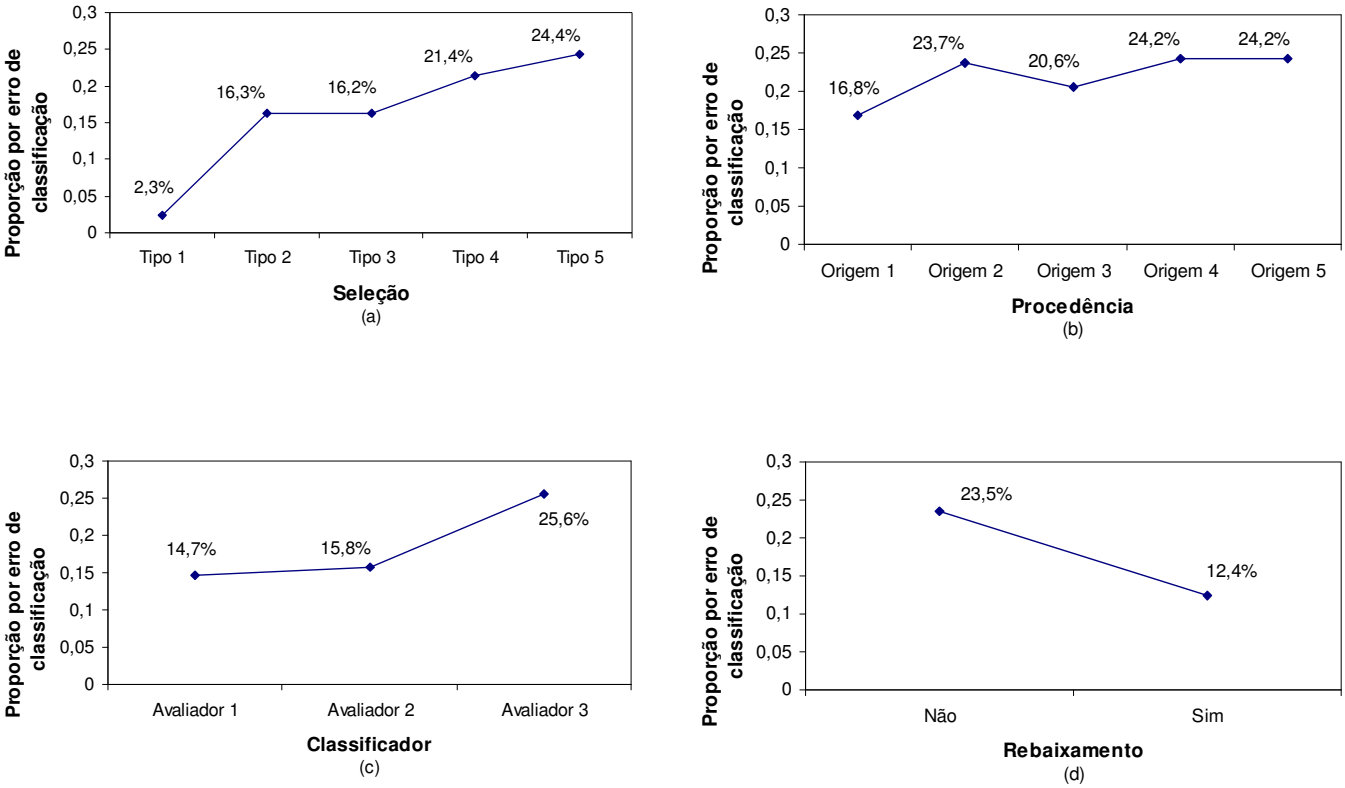


Figura 6 – Gráficos das proporções por erro de classificação em função das variáveis independentes: seleção, procedência, classificador e rebaixamento.

4.2.2 *Estrutura dos Modelos Ajustados*

Está subseção apresenta a estrutura dos modelos de regressão Beta e de Quase-verossimilhança utilizada na modelagem da proporção por erro de classificação no estágio *wet blue* da empresa curtidora de couro. No processo de modelagem as variáveis independentes: seleção (x_1), procedência (x_2), classificador (x_3) e rebaixamento (x_4) foram substituídas pelas variáveis *dummy*, constituindo a estrutura dos modelos. As variáveis independentes são então definidas como seleção tipo 2 (x_1), seleção tipo 3 (x_2), seleção tipo 4 (x_3), seleção tipo 5 (x_4), procedência 2 (x_5), procedência 3 (x_6), procedência 4 (x_7), procedência 5 (x_8), classificador 2 (x_9), classificador 3 (x_{10}) e rebaixado (x_{11}).

A construção dos modelos de regressão Beta e de Quase-verossimilhança sugeridos pelo método foi realizada no *software* ℝ 2.0.1, bem como, a verificação das medidas de diagnóstico, conforme pode ser visto no APÊNDICE B.

4.2.2.1 *Modelo de Quase-verossimilhança*

No ajuste do modelo pertencente à classe dos modelos lineares generalizados, o Modelo de Quase-verossimilhança, foi considerado inicialmente o modelo conforme a equação reescrita a seguir

$$g(\mu_i) = \beta_0 + \sum \beta_j x_{ij} + \varepsilon, \text{ com } i = (1, ..., 754) \; ; \; j = (1, ..., 11) \tag{86}$$

onde g representa a função de ligação "logit", e não é assumido a *priori* que a variável dependente (proporção por erro de classificação) possui uma distribuição de probabilidade, apenas foi escolhida uma função de variância, do tipo

$$V(\mu) = \mu(1-\mu), \tag{87}$$

e as variáveis independentes (x_1, x_2, x_3, ..., x_{10} e x_{11}) como estrutura linear do modelo.

A estimação dos parâmetros foi realizada pelo método da maximização da função de quase-verossimilhança apresentada pelo modelo, utilizando o algoritmo iterativo de Newton-Raphson, conforme discutido na seção 2.2.6.1.

4.2.2.2 *Modelo Beta*

Na abordagem do modelo de regressão Beta, assumiu-se que a variável dependente (y_i) (proporção de produtos não conformes) segue uma distribuição de probabilidade Beta com média (μ_i) e utilizando as variáveis independentes (x_1, x_2, x_3,..., x_{10} e x_{11}) como estrutura linear. E assim, foi considerado o modelo com a estrutura a seguir

$$g(\mu_i) = \beta_0 + \sum \beta_j x_{ij} + \varepsilon, \text{ com } i = (1, ..., 754) \; ; \; j = (1, ..., 11) \tag{88}$$

onde g representa a função de ligação "logit".

Para o modelo Beta, a estimação dos parâmetros foi realizada também através da maximização da função de verossimilhança apresentada pelo modelo, utilizando o algoritmo iterativo BFGS, conforme apresentado e discutido na seção 2.3.4.1.

4.2.3 *Análise do Ajuste dos Modelos*

Esta subseção apresenta a análise do ajuste dos modelos de regressão orientados pelo método, Modelos de Quase-verossimilhança e Beta aos dados do estudo. Em uma modelagem, variáveis independentes e/ou a interação destas só devem ser acrescentadas ao modelo se apresentarem sobre o comportamento da variável dependente, um nível explanatório significativo. Para garantir a inclusão somente de variáveis significativas no modelo, procedeu-se testes de significância estatística, ao nível de 5% de significância.

A Tabela 6 apresenta as estimativas dos parâmetros com respectivos erros padrões do ajuste dos dois modelos em análise e indicadores da qualidade da adequação. Verifica-se que as variáveis independentes "seleção", "classificador" e "rebaixamento" são estatisticamente significantes para explicar a variável dependente "proporção por erro de classificação", ao nível explanatório de 1%, baseado no "teste estatístico t-student". Entretanto, a variável independente "procedência" não apresenta significância estatística ao nível de 5% nos dois modelos de regressão ajustados. O coeficiente de determinação "pseudo" R_p^2 do Modelo de Quase-verossimilhança foi 0,356 e do Modelo Beta foi de 0,429.

Na Tabela 7, apresenta-se as estimativas dos parâmetros e respectivos erros padrões do ajuste dos dois modelos ajustados, nota-se que apenas as variáveis independentes "seleção", "classificador" e "rebaixamento" se mantiveram estatisticamente significantes para explicar a variável dependente "proporção por erro de classificação", ao nível explanatório de 1%.

Observa-se que a variável independente "procedência" foi retirada no ajuste dos modelos de regressão (ver Tabela 7), confirmando com os resultados apresentados por Arriba (2005), quanto a esta variável independente.

Tabela 6– Estimativas dos parâmetros e Erros padrões dos Modelos de Regressão propostos

	Modelo de Quase-verossimilhança		**Modelo Beta**	
Parâmetro	**Estimativa**	**Erro padrão**	**Estimativa**	**Erro padrão**
Intercepto	-3,7702*	0,3862	-2,4225*	0,2224
Seleção de Couro				
Tipo 1	-	-	-	-
Tipo 2	2,0507*	0,4233	1,5756*	0,2762
Tipo 3	2,2832*	0,3899	1,1068*	0,2270
Tipo 4	2,5236*	0,3892	1,2829*	0,2259
Tipo 5	2,7547*	0,3928	1,4241*	0,2307
Procedência				
Origem 1	-	-	-	-
Origem 2	0,0246	0,1360	0,0134	0,1146
Origem 3	0,0562	0,1069	0,0117	0,0889
Origem 4	0,0312	0,1598	0,0716	0,1352
Origem 5	0,1363	0,2185	0,1881	0,1886
Classificador				
Avaliador 1	-	-	-	-
Avaliador 2	0,2693*	0,1100	0,2921*	0,0900
Avaliador 3	0,4511*	0,0892	0,3756*	0,0744
Rebaixamento				
Não	-	-	-	-
Sim	-0,8317*	0,0961	-0,6671*	0,0777
Dispersão (ϕ)	0,1559*	0,5010	0,1159*	0,3958
R_p^2	0,3611		0,4295	
Deviance	115,51 (742 gl)		120,93 (742 gl)	

*Nível de significância "Teste t-student" ($p < 0,01$)
gl (graus de liberdade)

Tabela 7– Estimativas e Erros padrões dos parâmetros significativos dos Modelos de Regressão propostos

	Modelo de Quase-verossimilhança		**Modelo Beta**	
Parâmetro	**Estimativa**	**Erro padrão**	**Estimativa**	**Erro padrão**
Intercepto	-3,7692*	0,3844	-3,3144*	0,2217
Seleção de Couro				
Tipo 1	-	-	-	-
Tipo 2	2,0475*	0,4219	1,6806*	0,2761
Tipo 3	2,2857*	0,3886	1,9123*	0,2267
Tipo 4	2,5226*	0,3878	2,1206*	0,2256
Tipo 5	2,7583*	0,3913	2,6724*	0,2304
Classificador				
Avaliador 1	-	-	-	-
Avaliador 2	0,2672*	0,1095	0,1813*	0,0899
Avaliador 3	0,4483*	0,0877	0,3953*	0,0734
Rebaixamento				
Não	-	-	-	-
Sim	-0,8357*	0,0917	-0,8183*	0,0749
Dispersão (ϕ)	0,15512*	0,5015	0,12483*	0,3942
R_p^2	0,467		0,581	
Deviance	115,65 (746 gl)		121,48 (746 gl)	
AIC	– 213,3		– 224,96	

*Nível de significância "Teste t-student" ($p < 0,01$)
gl (graus de liberdade)

O coeficiente de determinação "pseudo" R_p^2 do Modelo de Quase-verossimilhança foi 0,467, e do Modelo Beta foi de 0,581. Estes coeficientes de determinação gerados são superiores ao encontrado pelo modelo de regressão linear normal. Observa-se que a medida de qualidade do ajuste para os modelos ajustados estão relativamente próximos, porém a medida do Modelo Beta indica um melhor ajuste, por estar mais próximo do valor 1.

A forma de regressão para o Modelo de Quase-verossimilhança é apresentada na equação (89)

$\hat{y}$ = -3,7692 + 2,0475(seleção 2) + 2,2857(seleção 3) + 2,5226(seleção 4) + 2,7583(seleção 5) + 0,2672(avaliador 2) + 0,4483(avaliador 3) – 0,8357(rebaixado). (89)

Para o Modelo Beta a forma de regressão é apresentada na equação (90)

$\hat{y}$ = -3,3144 + 1,6806(seleção 2) + 1,9123(seleção 3) + 2,1206(seleção 4) + 2,6724(seleção 5) + 0,1813(avaliador 2) + 0,3953(avaliador 3) – 0,8183(rebaixado). (90)

Avaliando as estimativas encontradas pelos modelos, observa-se que para o Modelo de Quase-verossimilhança tem-se exp[β_1] = exp [2,0475] = 7,75 e para o Modelo Beta exp[β_1] = exp[1,6806] = 5,37 o que significa estimar que, para o Modelo de Quase-verossimilhança a seleção tipo 2 apresenta 7,75 vezes mais chances de produzir proporção por erro de classificação que a seleção tipo 1, já para Modelo Beta a seleção tipo 2 apresenta 5,37 vezes mais chances de produzir proporção por erro de classificação que a seleção tipo 1. Bem como para o Modelo de Quase-verossimilhança, a seleção tipo 5, apresenta 15,77 vezes mais chances (exp[β_4] = exp [2,7583] = 15,77) que a seleção tipo 1 e 14,47 vezes para o Modelo Beta (exp[β_4] = exp[2,6724] = 14,47).

A variável independente "classificador", no Modelo de Quase-verossimilhança, os avaliadores 2 e 3 aumentam as chances de produzir produtos por erro de classificação em 31% (exp[β_5] = exp [0,2672] = 1,31) e 56% (exp[β_6] = exp [0,4483] = 1,56) respectivamente, em relação ao avaliador 1. Enquanto que, para o Modelo Beta, as chances são de produzir produtos por erro de classificação dos avaliadores 2 e 3 são de 20% (exp[β_5] = exp [0,1813] = 1,20) e 48% (exp[β_6] = exp [0,3953] = 1,48) respectivamente.

Para a variável independente “rebaixamento”, a estimativa do parâmetro é negativa (β_7 = -0,8357), indicando que o fato do estado de textura da matéria-prima estar rebaixado implica em que as chances de produzir proporção por erro de classificação diminuem em 43% (exp[β_7] = exp[-0,8357] = 0,43) para o Modelo de Quase-verossimilhança. Enquanto que para o Modelo Beta as chances são de 44% (exp[β_7] = exp[-0,8183] = 0,44).

4.2.4 *Análise de Adeqüabilidade dos Modelos*

O desempenho de um modelo é avaliada pela sua capacidade preditiva e definida a partir dos próprios dados utilizados na determinação do modelo. Modelos com boa adeqüabilidade apresentam pequena discrepância entre os dados reais e seus respectivos valores preditos.

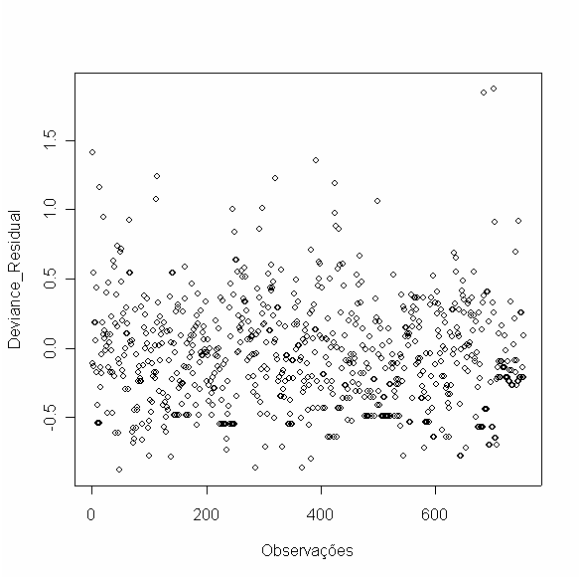
Para avaliar a adequação do modelo, apresenta-se alguns gráficos de diagnóstico como resíduos *Deviance* versus índices das observações, resíduos padronizados versus índices das observações, resíduos padronizados versus valores preditos e distância de Cook’s versus índices das observações.

Fahrmeir e Tutz (1994) recomendam que na análise dos resíduos convêm observar a aleatoriedade dos resíduos e investigar quando há valores discrepantes, pois auxiliam na verificação da adequação das funções de ligação e de variância, utilizadas nos modelos ajustados. Em relação à distância de Cook’s, McCullagh e Nelder (1989) sugerem que é conveniente analisar os casos em que $d_i > 0,5$.

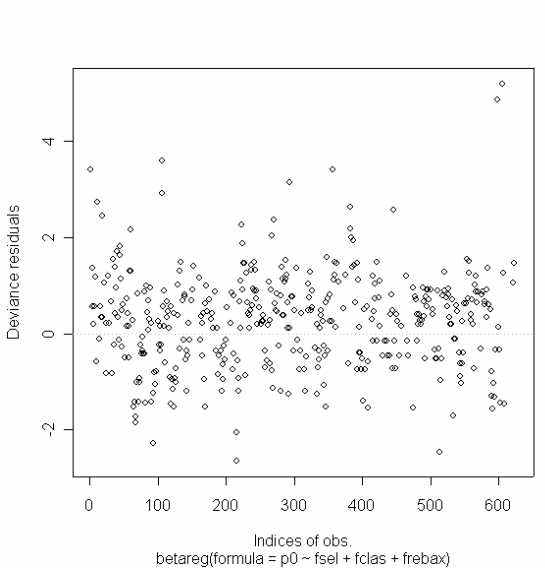
As Figura 7(a) à (d) revelam que há dois pontos com maior valor residual, tanto resíduos padronizados, quanto *deviance* correspondendo às observações 685 e 702. Isto é verificado no diagnóstico dos modelos de regressão Beta e Quase-verossimilhança. Verifica-se nas Figura 8(a) e (b) que os pontos não apresentam nenhuma tendência, ou seja, os pontos estão dispostos de forma aleatória. Indicando que a função de ligação utilizada é adequada.

Modelo de Quase-verossimilhança

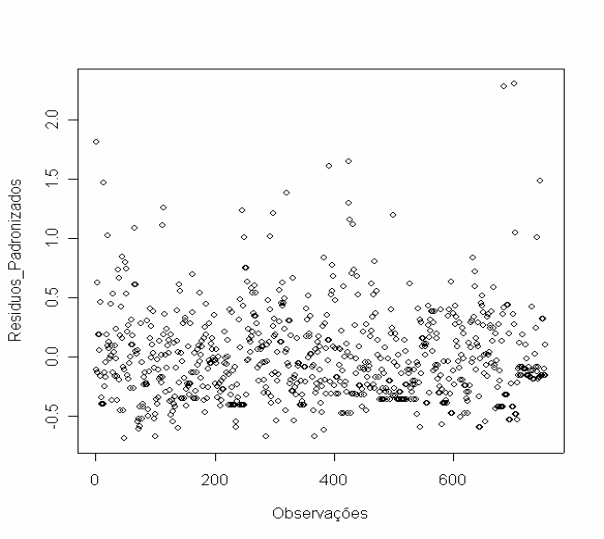
Modelo Beta



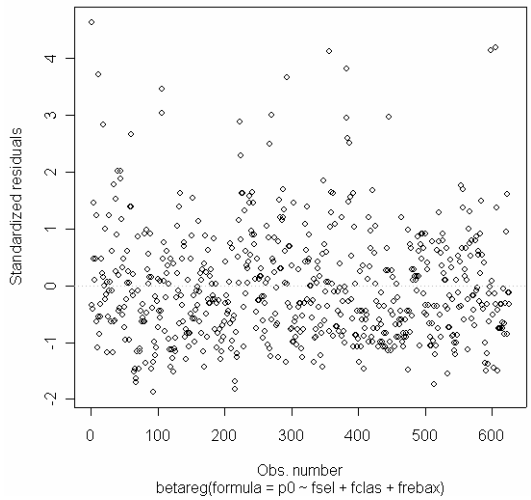
(a)



(b)



(c)

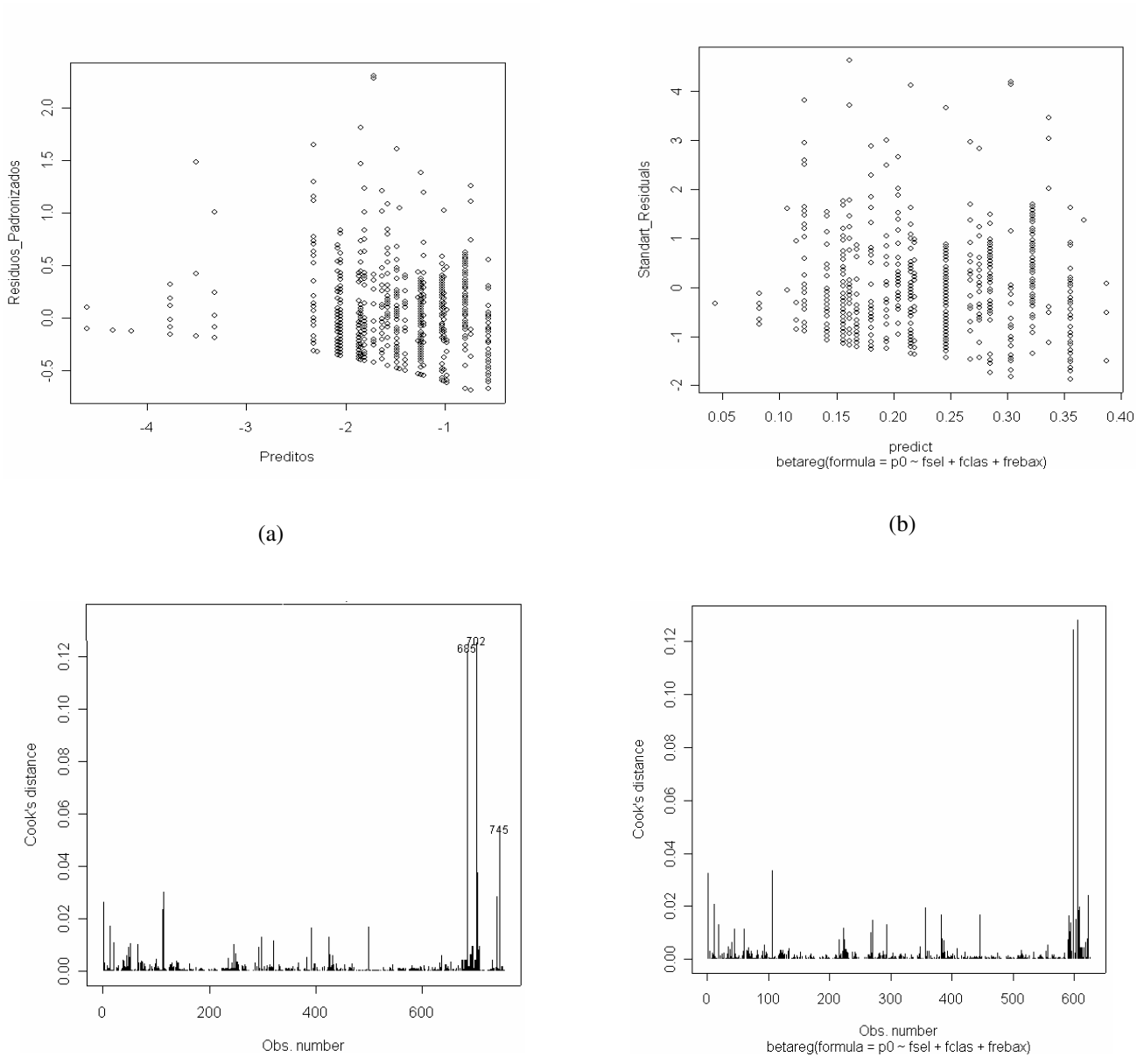


(d)

Figura 7 – Gráficos de diagnóstico, resíduo *deviance* e resíduo padronizado, para os dados com o ajuste dos Modelos de Quase-verossimilhança e Modelo Beta

Modelo de Quase-verossimilhança

Modelo Beta



(a)

(b)

(c)

(d)

Figura 8 – Gráficos de diagnóstico, resíduo padronizado e distância de Cook, para os dados com o ajuste dos Modelos de Quase-verossimilhança e Modelo Beta

Nota-se nas Figura 8(c) e (d) que as observações 685, 702 e 745 aparecem em destaque, no diagnóstico do Modelo de Quase-verossimilhança, indicando que elas são observações influentes. As observações 685 e 702 se apresentam altamente influentes, enquanto que a observação 745 apresenta influência moderada. Para o Modelo Beta, apenas as observações 685 e 702 apresentam-se como influentes. Segundo Cordeiro e Lima Neto (2004), as observações influentes não devem ser retiradas do modelo, pois sua exclusão pode implicar mudanças substanciais nas estatísticas do modelo. No entanto, se estas observações se constituírem como observações discrepantes (*outliers*) deve-se verificar a possível retirada, em virtude da possível influência nas estimativas dos parâmetros do modelo.

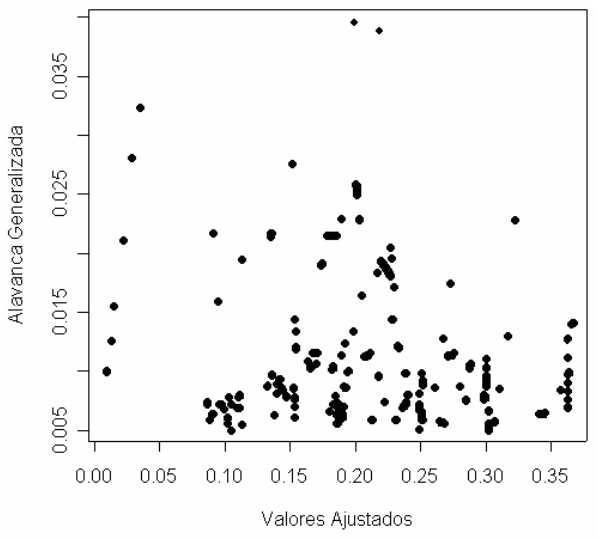
Contudo, após um estudo realizado com as observações 685 e 702 se observou que embora estas aparecerem como pontos discrepantes nos gráficos de resíduos padronizados e *deviance* dos dois modelos ajustados, não forma retiradas da modelagem.

A partir da estatística de corte ($h_{ii} \geq 2k/n$) apresentada na seção 2.4.2.6, para a verificação dos pontos de alavanca, sendo $h_{ii} = 2(8)/754 = 0{,}021$. Observa-se na Figura 9(b) que há 31 pontos que se configuram como pontos de alavanca, para o Modelo Beta. Para o Modelo de Quase-verossimilhança nota-se a detecção de apenas 5 pontos de alavanca, conforme Figura 9(a).

Tem-se na Figura 9(d) o gráfico de probabilidade meio-normal com envelope para o ajuste do Modelo Beta nota-se que a maioria dos resíduos se encontra contidos do envelope, embora haja alguns resíduos fora do envelope, e dois em destaque. Porém pode-se considerar que o modelo apresenta um ajuste satisfatório.

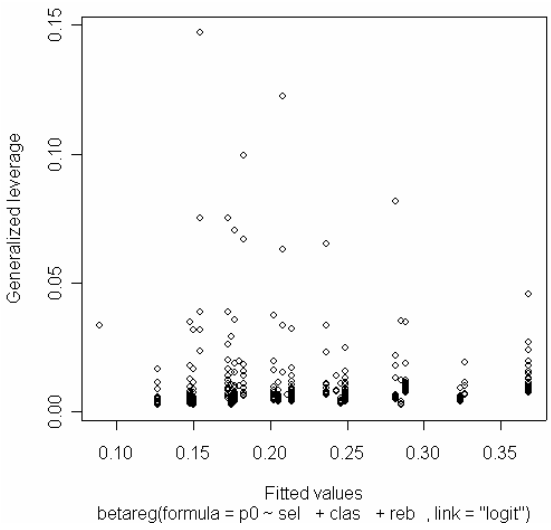
Conforme Figura 9(c), o Modelo de Quase-verossimilhança o gráfico de probabilidade meio-normal com envelope observa-se resíduos fora do envelope, considerando a necessidade de um ajuste mais adequado. Uma opção para se buscar um melhor ajuste é testar outras funções de variância que melhor se adapte a este modelo ou investigar a possibilidade de haver outros fatores (variáveis independentes) que estejam influenciando na etapa do processo de produção da empresa (estágio *wet blue*), gerando a proporção por erro de classificação.

Modelo de Quase-verossimilhança

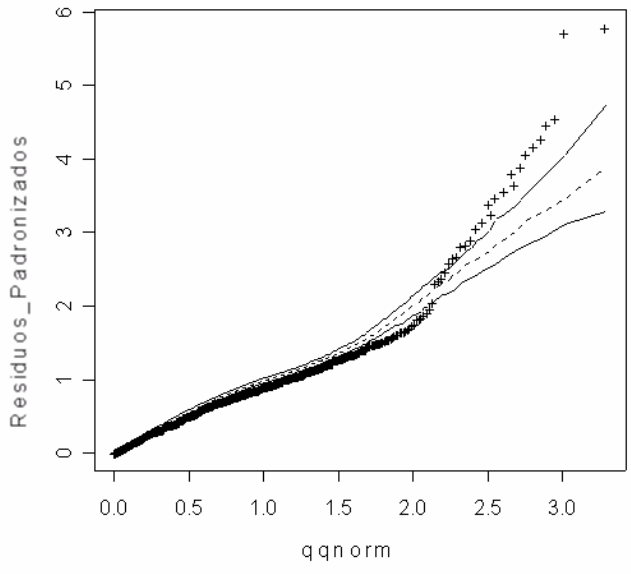


(a)

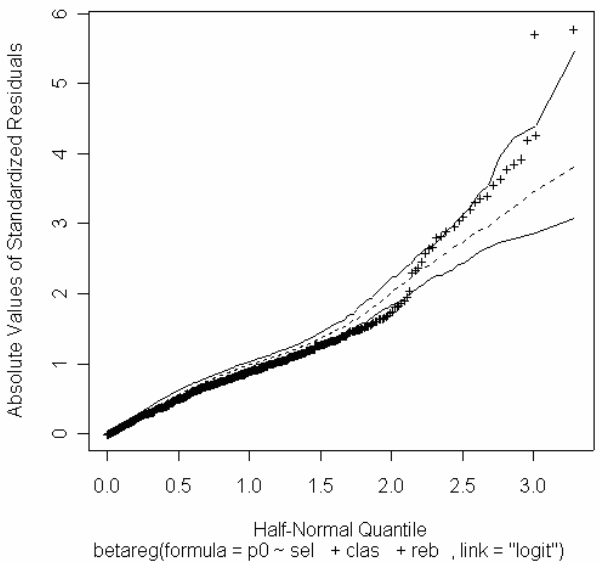
Modelo Beta



(b)



(c)



(d)

Figura 9 – Gráficos de diagnóstico, alavanca generalizada e envelope simulado, para os dados com o ajuste dos Modelos de Quase-verossimilhança e Modelo Beta

4.3 COMPARAÇÃO SOBRE OS MODELOS DE REGRESSÃO

O objetivo desta subseção é comparar os resultados estimados pelo ajuste dos modelos de regressão, partindo da observação das estimativas dos parâmetros, análises dos gráficos e dos indicadores de qualidade da adequação, como coeficiente de determinação (R_p^2), valor da *Deviance* e critério de Akaike (AIC), conforme as Tabela 6 e Tabela 7 e as Figura 7, Figura 8 e Figura 9.

Em geral pode-se afirmar que o ajuste foi satisfatório em todas as medidas de diagnóstico utilizadas na verificação dos modelos de regressão. Identificando adequadamente os fatores e níveis dos fatores quanto a sua influência e significância.

O Modelo Beta apresenta o valor das estimativas dos parâmetros e respectivos erros-padrão menores que o estimado pelo Modelo de Quase-verossimilhança. O coeficiente de determinação mostra que o Modelo Beta conseguiu explicar melhor que o Modelo de Quase-verossimilhança. Os indicadores de qualidade do ajuste, valor da *deviance* e critério de Akaike, o Modelo Beta apresenta-se como o melhor que o Modelo de Quase-verossimilhança.

Os gráficos de diagnóstico resíduos *Deviance* versus índices das observações, resíduos padronizados versus índices das observações, resíduos padronizados versus valores preditos e distância de Cook's versus índices das observações mostram similaridade no diagnóstico feito pelos modelos de regressão Beta e Quase-verossmilhança.

A análise deste estudo corrobora com os achados por Kieschnick e McCullough (2003), em que relatam a proximidade das estimativas entre o Modelo Beta e o Modelo de Quase-verossimilhança para estudos com tamanho de amostra grande. Embora os indicadores de qualidade de adequação para o Modelo Beta o indiquem como o melhor modelo.

Baseado na análise de comparação dos modelos de regressão aplicados e na literatura foi possível identificar vantagens e desvantagens de uso dos modelos de regressão Beta e Quase-verossimilhança, conforme Figura 10.

Modelo de Quase-verossimilhança	**Modelo Beta**
Vantagem	**Desvantagem**
Bibliografia Satisfatória	Bibliografia Restrita
Muitas publicações com aplicações a dados reais	Poucas publicações com aplicações a dados reais
Modelagem feita em vários *softwares*	Modelagem apenas no *software* $\mathbb{R}$
Ampla diversidade de aplicação em dados reais	Aplicação restrita a proporções
Flexibilidade de escolha da função de variância	Função de variância fixa
Diversidade de distribuições de probabilidade	Única distribuição de probabilidade
Modela dados de qualquer natureza da variável dependente	Restrição da natureza da variável dependente
Desvantagem	**Vantagem**
Incerteza no valor na precisão das estimativas dos parâmetros	Estimativas dos parâmetros mais precisas
Sugere tamanho de amostra grande	Modela qualquer tamanho de amostra
Parâmetro de regressão (β) e precisão (ϕ) ortogonais	Parâmetro de regressão (β) e precisão (ϕ) não ortogonais
Dependência entre parâmetros em função da ortogonalidade	Dependência entre parâmetros em função da ortogonalidade
Instável da presença de variâncias desiguais (Heterocedasticidade)	Boa precisão na presença de variâncias desiguais (Heterocedasticidade)
Não há rotinas computacionais prontas, deve-se realizar adaptações para o ajuste do modelo	Facilidade computacional por apresentar rotinas

Figura 10 – Vantagens e Desvantagens no uso dos Modelos de Quase-verossimilhança e Modelo Beta

5 CONSIDERAÇÕES FINAIS

A importância de conhecer e utilizar modelos de regressão vêm da necessidade de conhecer o efeito das variáveis independentes, chamadas de fatores, sobre a variável dependente, chamada de característica de qualidade. O modelo de regressão linear normal, em muitos casos, não é adequado à análise no âmbito industrial, pois as características de qualidade seguem diferentes distribuições de probabilidade, necessitando a aplicação de modelos de regressão adequados.

O tema desta dissertação contemplou modelos de regressão usados na modelagem de dados mensurados em proporção, mais especificamente o Modelo de Regressão Beta (MRB) e o Modelo de Quase-verossimilhança (MQV).

Esta dissertação apresentou uma revisão de literatura sobre Modelo de Regressão Beta e Modelo de Quase-verossimilhança, enfatizando a estrutura de regressão destes modelos e suas características, mais especificamente, propriedades dos modelos no que diz respeito a sua definição, função de ligação, métodos de estimação dos parâmetros. A revisão sobre as Medidas de Diagnóstico enfocou, além dos conceitos fundamentais, embasamento para análise de adequação e escolha do melhor modelo de regressão usado no processo de modelagem dos dados utilizados no estudo.

O objetivo principal deste trabalho foi elaborar um método que oriente à modelagem de dados mensurados em proporção, levando em consideração a classificação das variáveis dependente e independentes. A aplicação do método, a análise do ajuste e adequabilidade dos modelos e a comparação dos modelos de regressão propostos foram objetivos específicos da dissertação.

O estudo das características de cada modelo de regressão associado à classificação das variáveis dependente e independentes permitiu realizar a elaboração e estruturação do método. O método de orientação foi aplicado no processo de produção de uma empresa curtidora de couro. Foi possível observar que a “proporção por erro de classificação” é classificada como variável dependente contínua, restrita ao intervalo entre zero e um, e que o método orienta a utilização dos Modelos de Quase-verossimilhança e Beta.

Em relação ao objetivo específico, ajuste e adequabilidade dos modelos de regressão utilizados, os modelos de regressão Beta e de Quase-verossimilhança apresentaram bom ajuste e capacidade preditiva, com estimativas precisas e confiáveis dos seus parâmetros, bem como a identificação dos níveis das variáveis independentes que influenciam na proporção por erro de classificação produzido na linha de produção da empresa curtidora de couro.

Os Modelos de Quase-verossimilhança e Beta usados mostraram-se adequados na modelagem da proporção por erros de classificação, gerando os valores do coeficiente de determinação (R_p^2) superiores ao encontrado pelo modelo de regressão linear normal, sendo que o Modelo Beta apresenta um valor de coeficiente superior ao do Modelo de Quase-verossimilhança. Nos gráficos de diagnóstico, os dois modelos apresentaram similaridade nos resultados.

Em relação ao objetivo específico: comparar os Modelos de Beta e Quase-verossimilhança e identificar vantagens e desvantagens, o Modelo de Quase-verossimilhança apresentou vantagens na modelagem dos dados em proporção, por permitir flexibilidade de escolha da função de variância que melhor se ajustou ao conjunto de dados e distribuição de probabilidade mais adequada, enquanto o Modelo Beta apresentou função de variância e distribuição de probabilidade dependente da distribuição Beta. O modelo de regressão Beta apresenta vantagens em relação ao Modelo de Quase-verossimilhança na modelagem de dados em proporção, quanto à precisão das estimativas em amostras de tamanho diversos.

Foi possível observar que o Modelo de Quase-verossimilhança e o Modelo Beta surgem como proposta alternativa para o ajuste de dados sempre que estes apresentem mensurações restritas ao intervalo (0,1). Ainda se evidenciou que não há dificuldades computacionais para utilizar esses modelos de regressão, bem como as medidas de diagnóstico.

Os modelos de regressão Beta e de Quase-verossimilhança podem ser estendidos a todos os processos de manufatura que envolva produção de produtos não conformes às especificações de fabricação (defeituosos) em lotes, onde as mensurações em proporções são obtidas, posto que estes modelos se ajustam adequadamente aos dados em proporção, destacando principalmente os processos que permitem pouca coleta de dados.

Por fim, o método elaborado apresentou facilidade de entendimento e clareza dos passos para escolha dos modelos de regressão usados na modelagem de dados mensurados em proporção.

5.1 SUGESTÕES PARA TRABALHOS FUTUROS

Os modelos de regressão utilizados neste trabalho são aplicados em estudos de modelagem de dados em que se relaciona uma ou mais variáveis independentes a apenas uma variável dependente. A modelagem multivariada para as proporções de classificação da qualidade dos produtos, como por exemplo: ruim, regular e bom, poderia ser considerada em um trabalho futuro, pois a inter-relação entre as proporções das classificações não são levadas em consideração nos modelos de regressão propostos.

Na modelagem realizada, as variáveis independentes que foram investigadas no estudo da empresa curtidora de couro possuíam classificação qualitativa, em virtude do interesse da empresa. Porém outras variáveis de classificação quantitativa poderiam ser investigadas, no intuito de verificar como se comportariam o ajuste e a adequação dos modelos propostos.

A criação de um método orientativo à escolha dos modelos de regressão a serem utilizados em qualquer situação real, levando em consideração a classificação das variáveis dependente e independentes, apresenta-se como uma importante estratégia no procedimento de modelagem. A generalização desse método possibilitaria ainda melhorar os procedimentos de modelagem de dados e as análises estatísticas dos processos de produção.

REFERÊNCIAS BIBLIOGRÁFICAS

ADIMARI, G. & VENTURA, L. Robust inference for generalized linear model with application to logistic regression, **Statistics & Probability Letters**, 55, 413–419, 2001.

AGRESTI, A. **An Introduction to Categorical Data Analysis**, New York: John Wiley, 1996.

ARAÚJO, L.C.G. **Organização, Sistemas e Métodos e as modernas ferramentas de gestão organizacional**, São Paulo: Atlas, 2001.

ARRIBA, G.D. **Otimização de um processo de classificação de couros wet blue: Um caso em uma indústria curtidora**, Dissertação de Mestrado. Universidade Federal do Rio Grande do Sul – UFRGS. Escola de Engenharia – EE/PPGEP/UFRGS. Porto Alegre, RS, Brasil, 2005.

ATKINSON, A.C. **Plots, Transformation and Regression: An introduction to graphical methods of diagnostic regression analysis**, New York: Oxford University Press, 1985.

ATKINSON, A.C. & RIANI, M. **Robust Diagnostic Regression Analysis**, New York: Springer-Verlag, 2000.

COOK, R.D. & WEISBERG, S. **Residuals and Influence in Regression**, New York: Chapman and Hall, 1982.

CORDEIRO, G.M. & LIMA NETO, E.A. **Modelos Paramétricos**, In: XVI Simpósio Nacional de Probabilidade e Estatística, Águas de Lindóia, São Paulo, 246 p., 2004.

CORDEIRO, G.M. & CRIBARI-NETO, F. On bias reduction in exponential and non-exponential family regression models, **Communications in Statistics, Simulation and Computation**, 27, 485–500, 1998.

CORDEIRO, G.M. **Introdução a Teoria da Verossimilhança**, In: X Simpósio Nacional de Probabilidade e Estatística, UFRJ, Rio de Janeiro, 211 p., 1992.

CORDEIRO, G.M. **Modelos Lineares Generalizados**, In: VII Simpósio Nacional de Probabilidade e Estatística, Campinas, São Paulo, 286 p., 1986.

COT, J; MANICH, A & ARAMÓN, C. Procedimentos e Instalação para o Tratamento Integral de subprodutos da Indústria Curtidora, **Revista do Couro**. Estância Velha: ABQTIC, v.19, 1992.

COX, C. Nonlinear quasi-likehood models: applications to continuous proportions, **Computacional Statistical & Data Analysis**, 21, 449–461, 1996.

CROWDER, M.J. Beta-Binomial Anova for Proportions, **Journal of Applied Statistics**, 27, 34–37, 1978.

DAVISON, A.C. Biometrika Centenary: Theory and General Methodology, **Biometrica**, 88, 15–52, 2001.

DEAN, C. B. Testing for Overdispersion in Poisson and Binomial regression models, **Journal of the American Statistical Association**, 87, 451-457, 1992.

DEMÉTRIOS, C.G.B. **Modelos Lineares Generalizados em experimentação agronômica**, Piracicaba: ESALQ/USP, Disponível em: <http://www.lce.esalq.usp.br/ciagri/> Acesso em: 25 abr., 2002.

DOBSON, A.J. **An Introduction to Generalized Linear Models**, London: Chapman & Hall, 1990.

FAHRMEIR, L. & TUTZ, G. **Multivariate Statistical modeling based on Generalized Linear Models**, New York: Springer, 1994.

FERRARI, S.L.P & CRIBARI-NETO, F. Beta regression for modeling rates and proportions, **Journal of Applied Statistics**, 31, 799–816, 2004.

HAAD, T.C. & McCONNELL, K.E. **A simple method for bounding willingness to pay using a probit or logit model**, Greenville: ECU/East Carolina University, Disponível em: <http://www.ecu.edu/econ/wp/97/ecu9713.pdf> Acesso em: 10 nov. 2005.

HAIR, J.F.Jr.; ANDERSON, R.E.; TATHAM, R.L. & BLACK, W.C. **Multivariate Data Analysis**, 5ª ed., New Jersey: Prentice-Hall Inc, 1998.

HAMADA, M. & NELDER, J.A. Generalized linear models for quality-improvement experiments, **Journal of Quality Technology**, 29, 292–304, 1997.

HARRINGTON, H. **Aperfeiçoando processos empresariais**, São Paulo: Makron Books, 1993.

HELFER, P. J. Rebaixamento de Couros: Seus Problemas e Soluções, **Revista do Couro**, Estância Velha: ABQTIC, v.17, n.77, 1991.

HOSMER, D.W. & LEMESHOW, S. **Applied Logistic Regression**, New York: John Wiley, 1989.

HUVICH, C.M. & TSAI, C-L. Regression and Time Series Model selection in small samples, **Biometrics**, 76, 297–307, 1989.

JOHNSON, N.; KUTZ, S. & BALAKRISHNAN, N. **Continuous Univariate Distributions**, 2ª ed., New York: John Wiley, 1995.

JUNG, C.F. **Metodologia para Pesquisa e Desenvolvimento: aplicada a novas tecnologias, produtos e processos**, Rio de Janeiro: Axcel Books do Brasil, 2004.

KIESCHNICK, R. & McCULLOUGH, B.D. Regression analysis of variates observed on (0,1): percentages, proportions and fractions, **Statistical Modelling**, 3, 193–213, 2003.

KRYSICKI, W. On some new properties of the beta distribution, **Statistics & Probability Letters**, 42, 131–137, 1999.

KUHA, J. AIC and BIC – Comparisons of assumptions and performance, **Sociological Methods & Research**, 33, 188–229, 2004.

LARSON, R. & FARBER, B. **Elementary Statistics**, 2ª ed., New Jersey: Prentice-Hall Inc, 2003.

LEE, Y. & NELDER, J.A. Generalized linear models for the analysis of quality improvement experiments, **The Canadian Journal of Statistics**, 26, 95–105, 1998.

LEWIS, S.L.; MONTGOMERY, D.C. & MYERS, R.H. Examples of designed experiments with nonnormal responses, **Journal of Quality Technology**, 33, 265–278, 2001.

LUPORINI, C.E.M. & PINTO, N.M. **Sistemas administrativos: Uma abordagem moderna de O&M**, 4ª ed., São Paulo: Atlas, 1998.

MAGALHÃES, M.N. & LIMA, A.C.P. **Noções de Probabilidade e Estatística**, 4ª ed., São Paulo: Edusp, 2002.

MARTÍNEZ, R.O. **Estimação Pontual e Intervalar em um Modelo de Regressão Beta**, Dissertação de Mestrado. Universidade Federal de Pernambuco – UFPE. Instituto de Matemática – IM/UFPE. Recife, PE, Brasil, 2004.

McCULLAGH, P. & NELDER, J.A. **Generalized Linear Models**, 2ª ed., London: Chapman & Hall, 1989.

McDONALD, J.B. & XU, Y.J. A generalized of the beta distribution with applications, **Journal of Econometrics**, 66, 133–152, 1995.

MONTGOMERY, D.C. **Introduction Statistical Quality Control**, 4ª ed., New York: John Wiley, 2001.

MONTGOMERY, D.C. **Design and Analysis of Experiments**, 4ª ed., New York: John Wiley, 1997.

MONTGOMERY, D.C. & PECK, E.A. **Introduction to Linear Regression Analysis**, 2ª ed., New York: John Wiley, 1992.

MURTEIRA, B.J.F. **Probabilidade e Estatística**, 2ª ed., Lisboa: McGraw-Hill, 1990.

MYERS, R.H. & MONTGOMERY, D.C. A tutorial on generalized linear models, **Journal of Quality Technology**, 29, 274–291, 1997.

MYERS, R.H.; MONTGOMERY, D.C. & VINING, G.H. **Generalized Linear Models with applications in Engineering and the Sciences**, New York: John Wiley, 2002.

NELDER, J.A. & WEDDERBURN, R.W.M. Generalized Linear Models, **Journal of the Royal Statistical Society A**, 135, 370–384, 1972.

NOCEDAL, J. & WRIGHT, S.J. **Numerical Optimization**, New York: Springer, 1999.

OLIVEIRA, M.S. **Um Modelo de Regressão Beta: Teoria e Aplicação**, Dissertação de Mestrado. Universidade de São Paulo – USP. Instituto de Matemática e Estatística – IME/USP. São Paulo, SP, Brasil, 2004.

OLIVEIRA, D.P.R. **Sistemas, Organizações e Métodos: Uma abordagem gerencial**, 2ª ed., São Paulo: Atlas, 1999.

PARK, S.H. **Robust Design and Analysis for Quality Engineering**, London: Chapman & Hall, 1996.

PAULA, G.A. **Modelos de Regressão com apoio computacional**, São Paulo: IME/USP, Disponível em: <http://www.ime.usp.br/giapaula/livro.pdf> Acesso em: 12 mai. 2004.

PRENTICE, R.L. Binary Regression using an extended Beta-Binomial distribution, with discussion of correlation induced by covariate measurement errors, **Journal of the American Statistical Association**, 81, 321–327, 1986.

R Development Core Team. **R: A language and environment for statistical computing**, R Foundation for Statistical Computing, ISBN 3-900051-07-0, 2004.

RAO, C.R. & WU, Y. Linear model selection by cross-validation, **Journal Statistical Planning and Inference**, 128, 231–240, 2005.

ROCHA, F.J.S. & DANTAS, R.A. Avaliação de vendas de imóveis usando Modelo Probit, In XXIII Encontro Nacional de Engenharia de Produção, out/2003, Ouro Preto/MG, **CD Rom**.

SANT'ANNA, A.M.O. & CATEN, C.S. Modelagem da proporção de produtos defeituosos usando Modelo de Quase-verossimilhança, In: XXV Encontro Nacional de Engenharia de Produção, out/2005, Porto Alegre/RS, **CD Rom**.

SANT'ANNA, A.M.O. & CATEN, C.S. Comparação de modelos lineares generalizados para atributos, In: IV Semana de Engenharia de Produção e Transporte, dez/2004, Porto Alegre/RS, **CD Rom**.

SEN, P.K. & SINGER, J.M. **Large Sample Methods in Statistics: An introduction with applications**, London: Chapman & Hall, 1993.

SILVA, E.L. & MENEZES, E.M. **Metodologia de pesquisa e elaboração de dissertação**, Florianópolis: Laboratório de ensino da Universidade Federal de Santa Catarina, 2001.

TORRES, S.T.F. **Avaliação de Critérios de Seleção de Modelos para o Modelos de Regressão Beta**, Dissertação de Mestrado. Universidade Federal de Pernambuco – UFPE. Instituto de Matemática – IM/UFPE. Recife, PE, Brasil, 2005.

WEDDERBURN, R.W.M. Quasi-likelihood functions, generalized linear models and the Gauss-Newton method, **Biometrika**, 61, 439–447, 1974.

WEI, B-C.; HU, Y-Q. & FUNG, W-K. Generalized leverage and its applications, **Scandinavian Journal of Statistical**, 25, 25–37, 1998.

WERKEMA, M.C.C. & AGUIAR, S. **Análise de Regressão: Como entender o relacionamento entre as variáveis de um processo**, v. 7, Belo Horizonte: Fundação Christiano Ottoni, Escola de Engenharia da UFMG, 1996.

WILEY, J.A.; HERSCHOKORU, S.J. & PADIAU, N.S. Heterogeneity in the probability of HIV transmission per sexual contact: the case of male-to-female transmission in penile-vaginal intercourse, **Statistics in Medicine**, 8, 93–102, 1989.

APÊNDICE A

Planilha de coleta de dados de classificação em estágio *Wet Blue*

BERTIN LTDA. **Unidade EV (RS)** **CLASSIFICADOR: GERALDO**

RESULTADO DA CLASIFICAÇÃO DE COUROS *WET BLUE*

Período: 02 à 06.06.2004

					R E S U L T A D O (%)					
Seleção	**Classificador**	**Procedência**	**Rebax.**	**Meios**	**A**	**B**	**C**	**D**	**E**	**R**
A	460049228702	LINS	X	105	23,81%	57,14%	19,05%	0,00%	0,00%	0,00%
				0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
				0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
				0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
				0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
				0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
				0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
				0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	TOTAL			**105**	**23,81%**	**57,14%**	**19,05%**	**0,00%**	**0,00%**	**0,00%**
B	WALTER / ANÉSIO	LINS	X	40	0,00%	87,50%	12,50%	0,00%	0,00%	0,00%
	460054695988	LINS	X	59	0,00%	96,61%	3,39%	0,00%	0,00%	0,00%
	460056906259	SLMB	X	162	0,00%	93,83%	6,17%	0,00%	0,00%	0,00%
	MARCOS P.	LINS		38	0,00%	84,21%	15,79%	0,00%	0,00%	0,00%
				0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
				0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
				0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
				0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	TOTAL			**299**	**0,00%**	**92,31%**	**7,69%**	**0,00%**	**0,00%**	**0,00%**
C	CIDO	RBTE	X	87	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%
	PAULO	RBTE		24	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%
	460045649334	LINS		49	0,00%	0,00%	14,29%	85,71%	0,00%	0,00%
	JOÃO B.	REDENÇÃO		206	0,00%	0,00%	81,55%	18,45%	0,00%	0,00%
	ALEX / WALTER	LINS	X	200	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%
	460057844642	CACOAL	X	255	0,00%	0,00%	93,33%	6,67%	0,00%	0,00%
				260	0,00%	0,00%	82,69%	17,31%	0,00%	0,00%
				0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	TOTAL			**1.081**	**0,00%**	**0,00%**	**86,86%**	**13,14%**	**0,00%**	**0,00%**
D	ANÉSIO / ALEX	LINS		580	0,00%	0,00%	12,59%	86,21%	1,21%	0,00%
	460043761625	LINS		200	0,00%	0,00%	18,50%	81,50%	0,00%	0,00%
	460043704943	CACOAL	X	210	0,00%	0,00%	4,76%	94,29%	0,95%	0,00%
	460057526661	LINS	X	200	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%
	180057595345	SLMB		106	0,00%	0,00%	22,64%	75,47%	1,89%	0,00%
				0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
				0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
				0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	TOTAL			**1.296**	**0,00%**	**0,00%**	**11,11%**	**88,04%**	**0,85%**	**0,00%**
E	FLÁVIO / WALTER	LINS		217	0,00%	0,00%	0,00%	97,30%	2,70%	0,00%
					0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
					0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
					0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
					0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
					0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
					0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
					0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	TOTAL			**217**	**0,00%**	**0,00%**	**0,00%**	**97,30%**	**2,70%**	**0,00%**

ESPESSURA	**Meios**	**FINO**	**OK**	**GROSSO**
REBAX. LINS	1688	0%	100%	0%
REBAX. RBTE	111	0%	100%	0%
REBAX. SLMB	268	0%	100%	0%

Figura 11 – Planilha de coleta de dados de classificação dos couros no estágio *wet blue*

APÊNDICE B

Programas, Pacotes e *Syntaxes* do *software* $\mathbb{R}$ versão 2.0.1, utilizados na modelagem.

A análise dos modelos de regressão utilizados na modelagem dos dados foi feita no *software* ℝ versão 2.0.1. Os pacotes e as *syntaxes* para ajuste e adequabilidade dos modelos são apresentados a seguir. Disponível em http://www.r-project.org/.

1. Modelo de Regressão Beta

```
/*Instalação do Pacote Betareg no software ℝ */
> install.packages(choose.files('',filters=Filters[c('zip','All'),]), .libPaths()[1], CRAN = NULL)

/*Acesso a library do software ℝ para uso do Pacote Betareg*/
> local({pkg <- select.list(sort(.packages(all.available = TRUE))) if(nchar(pkg)) library(pkg, character.only=TRUE)})> install.packages(choose.files('',filters=Filters[c('zip','All'),]), .libPaths()[1], CRAN = NULL)

/*Leitura do Banco de dados indicando o endereço de localização*/
> read.table("D:/Angelo/Work/Modelos/Couros/Dados.dat",header=T,sep="")->CouroK

/*Tornar as variáveis do Banco de dados visíveis ao software ℝ */
> attach(CouroK)

/*Realiza um sumário descritivo das variáveis do Banco */
> summary(CouroK)

/*Define a variável em questão como Qualitativa */
> as.factor(select)->fsel
> as.factor(proced)->fprod
> as.factor(classif)->fclas
> as.factor(rebax)->frebax

/*Função que realiza a modelagem dos dados – ajuste do modelo de regressão*/
> betareg(formula = p0 ~ fsel + fclas + fprod + frebax, link="logit")->ModelBetaK

/*Realiza um sumário dos elementos estimados pelo modelo */
> summary(ModelBetaK)
        - Este sumário apresenta a equação do modelo ajustado, estimativas dos coeficientes, erros-padrão dos coeficientes, níveis de significância, deviance, coeficiente "pseudo"R2.

/*Repetição da modelagem – ajuste do modelo de regressão*/
> betareg(formula = pnnorm ~ fsel + fclas + frebax, link="logit")->ModelBetaK2

/*Realiza um sumário dos elementos estimados pelo modelo ajustado*/
> summary(ModelBetaK2)

/*Define uma matriz de figuras múltiplas – matriz (linha, coluna) */
> par(mfrow=c(2,2))

/*Realiza a construção dos Gráficos para adeqüabilidade do modelo ajustado*/
> plot(ModelBetaK2)

/*Realize a construção de um único gráfico específico*/
> plot("Generalized leverage","Predicted values")

/*Realize a construção de um gráfico específico, delimitando seus elementos*/
> envelope.beta(model=BetaK1,sim=100,conf=.99, pch="+",font.main=1, cex.main=1.)
```

2. Modelo de Regressão de Quase-verossimilhança

/*Instalação do Pacote Betareg no *software*/
> install.packages(choose.files(",filters=Filters[c('zip','All'),]), .libPaths()[1], CRAN = NULL)

/*Acesso a *library* do *software* R para uso do Pacote Mass*/
> local({pkg <- select.list(sort(.packages(all.available = TRUE))) if(nchar(pkg)) library(pkg, character.only=TRUE)})> install.packages(choose.files(",filters=Filters[c('zip','All'),]), .libPaths()[1], CRAN = NULL)

/*Leitura do Banco de dados indicando o endereço de localização*/
> read.table("D:/Angelo/Work/Modelos/Couros/Dados.dat",header=T,sep="")->CouroK

/*Tornar as variáveis do Banco de dados visíveis ao *software* R */
> attach(CouroK)

/*Realiza um sumário descritivo das variáveis do Banco */
> summary(CouroK)

/*Define a variável em questão como Qualitativa */
> as.factor(select)->fsel
> as.factor(proced)->fprod
> as.factor(classif)->fclas
> as.factor(rebax)->frebax

/*Função que realiza a modelagem dos dados – ajuste do modelo de regressão*/
> glm(formula = p0 ~ fsel + fclas + fprod + frebax, family = quasi(link=logit,variance ="mu(1-mu)"))->ModelQuasiK

/*Realiza um sumário dos elementos estimados pelo modelo */
> summary(ModelQuasiK)

- Este sumário apresenta a equação do modelo ajustado, estimativas dos coeficientes, erros-padrão dos coeficientes, níveis de significância, *deviance*, coeficiente de decisão ("pseudo"R^2), parâmetro de dispersão, critério de Akaike.

/*Repeticão da modelagem – ajuste do modelo de regressão*/
> glm(formula = p0 ~ fsel + fclas + frebax, link="logit")->ModelQuasiK2

/*Realiza um sumário dos elementos estimados pelo modelo ajustado*/
> summary(ModelQuasiK2)

/*Realize a construção de um único gráfico específico*/
> plot(predict(ModelQuasiK2),td,xlab="Observações", ylab="Deviance_Residual", ylim=c(a-1,b+1))
> abline(2,0,lty=2)
> abline(-2,0,lty=2)
> lines(smooth.spline(predict(ModelQuasiK2), td, df=2))

/*Realize a construção de gráficos específicos*/
> plot(residuals,.values, xlab="Resíduos_Padronizados", ylab="Observações")

> plot(residuals,.preditc values, xlab="Resíduos_Padronizados", ylab="Preditos")

> plot(di,xlab="Índice", ylab="Distância de Cook")

/*Identifica três números no gráfico desejado*/
> identify(di, n=3)

```
/*Realize a construção de um único gráfico específico*/
> X <- model.matrix(ModelQuasiK2)
> n <- nrow(X)
> p <- ncol(X)
> w <- ModelQuasiK2$weights
> W <- diag(w)
> H <- solve(t(X)%*%W%*%X)
> H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
> h <- diag(H)
> ts <- resid(ModelQuasiK2,type="pearson")/sqrt(1-h)
> td <- resid(ModelQuasiK2,type="deviance")/sqrt(1-h)
> di <- (h/(1-h))*(ts^2)
> par(mfrow=c(1,1))
> a <- min(td)
> b <- max(td)
> plot(fitted(ModelQuasiK2), h,xlab="Valores Ajustados", ylab="Alavanca Generalizada")
> dentify(fitted(ModelQuasiK2), h, n=5)

/*Realize a construção de um único gráfico específico*/
X < − model.matrix(ModelQuasiK2)
n < − nrow(X)
p < − ncol(X)
w < − ModelQuasiK2$weights
W < − diag(w)
H < − solve(t(X)%*%W%*%X)
H < − sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
h < − diag(H)
td < − resid(ModelQuasiK2,type="deviance")/sqrt(1-h)
e < − matrix(0,n,100)
for(i in 1:100){
dif < − runif(n) - fitted(ModelQuasiK2)
dif[ dif >=0 ] < 0
dif[dif < − 0] < − 1
nresp < − dif
fit < − glm(nresp _ X, family=quase(link=logit, variance=”mu(1-mu)”))
w < − fit$weights
W < − diag(w)
H < − solve(t(X)%*%W%*%X)
H < − sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
h < − diag(H)
e[,i] < − sort(resid(fit, type=‘‘deviance")/sqrt(1-h)) }
e1 < − numeric(n)
e2 < − numeric(n)
for (i in 1:n) {eo < − sort(e[i,]) e1[i] < − eo[5] e2[i] < − eo[95] }
med < − apply(e,1,mean)
faixa < − range(td,e1,e2)
par(pty=‘‘s ")
qqnorm(td, xlab=‘‘Percentis da N(0,1)", ylab=‘‘Residuos _Padronizados", ylim=faixa)
par(new=T)
qqnorm(e1,axes=F,xlab=‘‘", ylab=‘‘", type=‘‘l ", ylim=faixa, lty=1)
par(new=T)
qqnorm(e2,axes=F,xlab=‘‘", ylab=‘‘", type=‘‘l ", ylim=faixa, lty=1)
par(new=T)

qqnorm(med,axes=F,xlab=‘‘", ylab=‘‘", type=‘‘l ", ylim=faixa, lty=2)
```