versão impressa ISSN 0101-7438 / versão online ISSN 1678-5142

MODELAGEM DA FRAÇÃO DE NÃO-CONFORMES EM PROCESSOS INDUSTRIAIS

Ângelo Márcio Oliveira Sant'Anna*
Carla Schwengber ten Caten
Programa de Pós-graduação em Engenharia de Produção
Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS
angelo@producao.ufrgs.br
tencaten@producao.ufrgs.br

* *Corresponding author* / autor para quem as correspondências devem ser encaminhadas

Recebido em 07/2007; aceito em 06/2009 após 1 revisão
Received July 2007; accepted June 2009 after one revision

Resumo

Em qualquer processo industrial, pode ser definido um conjunto de causas ou fatores que produzem determinado efeito sobre uma ou mais características de qualidade de um produto que pode ou não satisfazer às especificações do cliente, gerando a produção de produtos não-conformes. A modelagem da proporção ou fração de produtos não-conformes utilizando-se o modelo de regressão linear não é adequada por pelo menos duas razões: (*i*) pressupõe que as proporções seguem a distribuição normal, que não é correto; e (*ii*) possibilita a previsão de valores fora do intervalo [0,1]. Alternativas à modelagem da proporção de não-conformes são os Modelos Lineares Generalizados e os Modelos de Regressão Beta. O objetivo deste artigo é modelar a fração de não-conformes às especificações de uma indústria curtidora com enfoque nos Modelos de Regressão Beta e Modelo Linear Generalizado. Esses modelos podem ser estendidos a processos industriais que envolvam a produção de produtos não-conformes às especificações de manufatura.

Palavras-chave: modelo linear generalizado; modelo de regressão beta; fração de não-conformes.

Abstract

In any industrial process, one can enumerate causes or factors that act on one or more quality characteristics of the resulting product such that they fail to meet customers' specifications, generating items deemed as nonconforming. Modeling the fraction or proportion of nonconforming items using linear regression models is not adequate for at least two reasons: (*i*) proportions are assumed to follow a Normal distribution, which is not correct, and (*ii*) predicted responses will not necessarily be confined in the [0,1]-interval. Alternative approaches to the modeling of nonconforming proportions are based on Generalized Linear Models and Beta Regression Models. In this paper we present a case study where the objective is to model the nonconforming fraction of items emerging from a tanning process; our analysis uses Generalized Linear Models and Beta Regression Models. The developments presented in the paper may be extended to other industrial process where the proportion of nonconforming items is easily accessible.

Keywords: generalized linear model; beta regression model; fraction nonconforming.

1. Introdução

O cenário mundial atual é de intensa competitividade devido ao desenvolvimento tecnológico rápido de produtos e processos, visando buscar produtos que tenham características de qualidade sem defeito. Em processos de manufatura, a implementação de técnicas estatísticas 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 industrial, pode ser definido um conjunto de causas ou fatores que produzem determinado efeito sobre uma ou mais características de qualidade de produtos que podem ou não satisfazer às especificações preestabelecidas pelo cliente. Muitas vezes não se consegue controlar todas as causas de variação, pois certas causas são inerentes ao processo (Montgomery, 2001). As causas de variação que podem ser controladas, também chamadas de fatores controláveis, podem interferir em um processo podendo gerar a produção de produtos com características de qualidade não-conformes às especificações preestabelecidas, os quais podem ser mensurados através da proporção ou fração de produtos não-conformes.

Segundo Montgomery *et al.* (2006), modelos de regressão consistem numa técnica estatística de modelagem e investigação que relaciona uma ou mais características de qualidade do produto, chamadas de variável dependente, com os fatores controláveis que podem afetá-las, chamados de variáveis independentes. Conforme Hamada & Nelder (1997), um modelo de regressão que apresenta um bom ajuste usualmente permite gerar boas estimativas dos efeitos dos fatores, consistindo numa estratégia eficiente de otimização, pois é possível prever a variável dependente fração de não-conformes em função do ajuste das variáveis independentes ou fatores controláveis.

Segundo Cox (1996), a modelagem da proporção em um determinado conjunto de observações, por meio de um modelo de regressão linear, nem sempre é recomendada, uma vez que este modelo requer a suposição de que as proporções seguem a distribuição normal. Segundo Kieschnick & McCullogh (2003), o uso do modelo de regressão linear 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]. Alternativas para a modelagem da fração de produtos não-conformes são os Modelos Lineares Generalizados (MLG) e os Modelos de Regressão Beta (MRB).

Conforme Myers *et al*. (2002), a teoria dos MLG 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, sem a necessidade de realizar transformações nos dados.

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

O uso destes modelos de regressão permite aumentar a precisão da estimativa dos efeitos dos fatores controláveis e da previsão dos efeitos. Desta forma, permite-se identificar com mais precisão as condições ótimas de operação do processo de manufatura.

O objetivo deste artigo é modelar a fração de não-conformes às especificações de uma indústria curtidora com enfoque no Modelo de Regressão Beta e no Modelo Linear Generalizado.

Na seção 2, é apresentada uma breve revisão sobre os conceitos básicos de MLG e MRB. Na seção 3, são apresentadas as medidas de diagnóstico para análise de adequação dos modelos de regressão. Na seção 4, um estudo de aplicação prática. Finalmente, na seção 5, são feitas considerações finais sobre a modelagem desenvolvida nesse artigo.

2. Modelos de Regressão

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 a fração ou proporção de algum evento de interesse, é comum usar no processo de modelagem, o modelo de regressão linear. Contudo, segundo Cox (1996), a modelagem da proporção utilizando um modelo de regressão linear nem sempre é recomendada, pois este modelo requer a suposição de normalidade aos dados e homocedasticidade (variâncias iguais). Assim, 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.

Os modelos de regressão abordados como alternativas à modelagem da fração de não-conformes são: os modelos lineares generalizados, que foram propostos por Nelder & Wedderburn (1972), mais especificamente o modelo de Quase-verossimilhança e o modelo de regressão Beta, que foi proposto por Ferrari & Cribari-Neto (2004).

2.1 Modelos Lineares Generalizados

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

Segundo Hamada & Nelder (1997), a classe dos MLG foi desenvolvida por Nelder & Wedderburn (1972) e estes modelos se baseiam em distribuições de probabilidade, com um parâmetro de localização desconhecido (θ), admitindo que a mesma pertença à família exponencial. Esta família contempla as distribuições Normal, Binomial, Poisson, Gamma, Exponencial, Binomial Negativa. Ademais, estes modelos admitem variância não constante e dependente da média. Conforme Lee & Nelder (1998), esta 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.

Para se admitir que uma distribuição de probabilidade pertença à família de distribuições exponencial é necessário que as funções de distribuições de probabilidade apresentem a forma

$$f(y; \theta) = \exp \{a(y)b(\theta) + c(\theta) + d(\theta)\}, \qquad (1)$$

onde $a(y)$, $b(\theta)$ e $c(\theta)$ são funções específicas, sendo θ o parâmetro natural da distribuição de probabilidade e y a variável dependente da distribuição (Dobson, 2002).

Segundo Cordeiro (1986), uma variável dependente (y) que apresenta valores na forma de proporções ou frações segue a distribuição de probabilidade Binomial com parâmetros n e p.

McCullagh & Nelder (1989) salientam que, em estudos com dados mensurados em proporção, a relação entre a variável dependente e as variáveis independentes é descrita formalmente através de um modelo de regressão Binomial.

Conforme Prentice (1986), a proporção de sucessos (p_i) de um referido evento (por exemplo, produto defeituoso) segue uma distribuição de probabilidade Beta-Binomial. Onde admite-se que a variável dependente (y_i) segue a distribuição Binomial e a proporção (p_i) obtida a partir da variável (y_i) em cada ocorrência ($i = 1, ..., n$) 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.

A forma da função da distribuição de probabilidade Binomial pertencente à família exponencial é expressa por

$$f(y;\mu) = \exp\left\{ y \log\left(\frac{\mu}{1-\mu}\right) + n\log(1-\mu) + \log\binom{n}{y} \right\} \qquad (2)$$

Observa-se que a forma da função de distribuição de probabilidade Binomial esta descrita como a forma da equação (1).

Os modelos de regressão pertencentes à classe dos MLG apresentam uma estrutura com três componentes: a componente aleatória, a componente sistemática e a função de ligação. Sendo: (*i*) Componente aleatória – identifica a distribuição de probabilidade da variável dependente *y* pertencente à família exponencial; (*ii*) Componente sistemática – especifica a estrutura das variáveis independentes, que é utilizada como preditor linear η; (*iii*) Função de ligação – descreve a relação funcional entre a componente sistemática e o valor esperado da componente aleatória (a média μ da variável dependente *y*).

A componente sistemática que compõem o modelo de regressão que consiste na estrutura das variáveis independentes é uma soma linear, definida como

$$\eta = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k , \qquad (3)$$

onde a função linear η é denominada de preditor linear, β's coeficientes desconhecidos e *k* números de variáveis independentes. Além disso, outra característica da componente sistemática de um modelo linear é que a média μ da variável dependente *y* pode ser expressa por uma função conhecida *g*(.), denominada de função de ligação (Myers *et al*., 2002), a qual descreve a relação funcional (ligação) entre a média μ e o preditor linear η, descrevendo a forma

$$g(\mu) = \eta = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k . \qquad (4)$$

Conforme Dobson (2002), a função de ligação *g*(.) é responsável pela transformação da média da variável dependente, e não de cada observação. Sant'Anna & Caten (2005) descrevem que a escolha da função de ligação depende do problema de modelagem em particular e, que em muitas vezes os modelos de regressão podem apresentar uma função de ligação diferente conforme o conjunto de observações.

A estimação dos coeficientes para o MLG é realizada através do método clássico de máxima verossimilhança, em que os coeficientes β e ϕ (parâmetro de dispersão) são obtidos a partir da maximização do logaritmo da função de verossimilhança. O procedimento de maximizar a função de verossimilhança é realizado por um algoritmo de otimização não-linear, tal como o algoritmo de Newton-Raphson descrito por Cordeiro (1986) ou o algoritmo quasi-Newton

(BFGS) descrito por Ferrari & Cribari-Neto (2004). Conforme McCullagh & Nelder (1989) o método clássico de estimação pode ser utilizado considerando qualquer função de distribuição de probabilidade para variável dependente.

Considerando uma amostra de n observações independentes, o logaritmo da função de verossimilhança (log-verossimilhança) dos modelos de regressão apresenta a forma geral,

$$l(\beta,\phi) = \sum_{i=1}^{n} l_i(\mu_i, \phi), \tag{5}$$

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

A diferenciação da função de log-verossimilhança em relação aos coeficientes desconhecidos (β) do modelo ajustado é definida pela derivada da função de log-verossimilhança, para $j = 1, ..., k$.

$$\frac{\partial l(\beta,\phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i(\mu_i,\phi)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}. \tag{6}$$

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

$$\frac{\partial l_i(\mu_i,\phi)}{\partial \mu_i} = \phi\left[\log\frac{y_i}{1-y_i} - \{\delta(\mu_i\phi) - \delta(1-\mu_i)\phi\}\right], \tag{7}$$

onde $\delta(\cdot)$ é uma função *diferencial Gamma*, isto é, $\delta(z) = \partial \log\Gamma(z)/\partial z$, $z > 0$. Desta forma $y_i^* = \log(y_i/(1-y_i))$ e $\mu_i^* = \{\delta(\mu_i\phi) - \delta(1-\mu_i)\phi\}$. Conforme Hoffmann (2003), o valor esperado da derivada em (7) iguala-se a zero, de forma que o valor esperado da variável dependente transformada y_i^* iguala-se a μ_i^*. Ou seja, $\mu_i^* = \mathrm{E}(y_i^*)$.

Na abordagem de MLG, a estimativa dos coeficientes dos modelos usualmente é obtida pela maximização da função de log-verossimilhança (EMV). No entanto, conforme algumas situações experimentais a estimação dos coeficientes pelo método tradicional de EMV fica comprometida, em virtude de não cumprimento de certos pressupostos, assim, faz-se necessário o uso do método de estimação por quase-verossimilhança (conforme pode ser visto em Sant'Anna, 2006).

Modelo de Quase-verossimilhança

Weddeburn (1974) propôs os modelos de Quase-verossimilhança (MQV) com base nos modelos pertencentes à classe dos MLG, em que os modelos apresentam duas componentes: a componente sistemática (estrutura linear das variáveis independentes) e a 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 dependente não são conhecidas a *priori*.

Seja y_i uma variável dependente qualquer de interesse, que assume a $\mathrm{E}[y_i] = \mu_i$ e uma variância definida por $\mathrm{Var}[y_i] = \phi^* \mathrm{V}(\mu_i)$, onde a função de variância $\mathrm{V}(\mu_i)$ 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

$$Q(y_i\ ;\ \mu_i) = \int_y^{\mu} \frac{y_i - t}{\phi.V(t)}\ dt \tag{8}$$

O MQV utilizado na modelagem de um conjunto de dados mensurados em proporção ou fração de não-conformes é descrito a partir de uma variável dependente (y_i) que pode assumir qualquer função de variância. Por exemplo, seja uma função $V(\mu_i) = \mu(1-\mu)$ para a variável dependente descrita acima a forma se apresenta

$$Q(y_i\ ;\ \mu_i) = \frac{1}{\phi}\int_y^{\mu} \frac{y_i - \mu_i}{\mu_i(1-\mu_i)}\ d\mu \tag{9}$$

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

$$\log\left(Q(y_i\ ;\ \mu_i)\right) = y_i \ln\left(\frac{\mu_i}{1-\mu_i}\right) + \ln(1-\mu_i) \tag{10}$$

que conforme McCullagh & Nelder (1989), a função acima corresponde a função de log-verossimilhança da distribuição de probabilidade Binomial com função de variância $V(\mu) = \mu(1-\mu)$, a qual possui a seguinte forma

$$L(y_i\ ;\ \mu_i) = y_i \ln\left(\frac{\mu_i}{1-\mu_i}\right) + n\ln(1-\mu_i) \tag{11}$$

Nota-se que a principal diferença entre como se formam as equações (10) e (11) está em que, quando se usa a função de quase-verossimilhança para estimar os coeficientes 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.

Para testar a significância dos coeficientes do modelo de regressão pelo teste da razão de quase-verossimilhança tem-se a estatística de quase-*deviance*. Pode-se dizer que a quase-*deviance* está para a modelagem pela função de quase-verossimilhança assim 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:

$$D_i(y_i, \hat{\mu}_i) = -2\phi\left[Q_i(y_i;\hat{\mu}_i) - Q_i(y_i;y_i)\right] = -2\phi\left[Q_i(y_i;\hat{\mu}_i)\right] = 2\int_{y_i}^{\mu_i} \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)}, \tag{12}$$

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 (10), a estatística de quase-*deviance* é expressa da forma

$$D_i(y_i, \hat{\mu}_i) = -2\phi\left[Q_i(y_i;\hat{\mu}_i)\right] = 2\int_{y_i}^{\mu_i} \frac{y_i - \hat{\mu}_i}{\phi\cdot\left[\hat{\mu}_i(1-\hat{\mu}_i)\right]}, \tag{13}$$

2.2 Modelo de Regressão Beta

Quando a variável dependente é mensurada em proporção ou fração, apresentando valores restritamente no intervalo unitário ($0 \leq y \leq 1$), a relação entre as variáveis dependente e independentes apresenta restrição no domínio da função ($0 < E(y) < 1$) (Cox, 1996).

Partindo deste princípio, Ferrari & Cribari-Neto (2004) propuseram o MRB, que é 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. Ademais, na construção destes modelos as variáveis independentes ou fatores controláveis podem ser de natureza quantitativa ou qualitativa.

A família de distribuições Beta é composta de todas as distribuições de probabilidade que apresente uma variável dependente y cuja função densidade de probabilidade depende dos parâmetros p e q, e que sua função densidade pode ser escrita na forma,

$$f(y;p,q) = \left\{ \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} \quad , \quad 0 < y < 1 \; , \quad p > 0 \; , \quad q > 0 \; , \right\} \qquad (14)$$

onde y é a variável aleatória, p e q são parâmetros da função densidade de probabilidade e $\Gamma(p)$ uma função gama avaliada no ponto p, ou seja, com $\Gamma(p) = \int_0^\infty y^{p-1} e^{-y} dy$, $p>0$.

A função $f(y) = f(y; p, q)$ é efetivamente uma função densidade de probabilidade com parâmetros p e q. 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, $f(y) \geq 0$.

Segundo Kieschnick & McCullough (2003) e Ferrari & Cribari-Neto (2004), a distribuição de probabilidade 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 de localização μ. A família de distribuições Beta contempla as distribuições Uniforme, Arco-seno e Dirichlet.

O MRB proposto apresenta uma estrutura de regressão baseada em três componentes: (*i*) Componente aleatória – supõe-se que a variável dependente (y) segue unicamente a distribuição de probabilidade Beta; (*ii*) Componente sistemática – que especifica uma soma linear dos coeficientes desconhecidos das variáveis independentes, conhecido como preditor linear (η); (*iii*) Função de ligação – que permite modelar a média da variável dependente em relação às demais variáveis independentes através de uma função, conhecida como $g(.)$.

A estrutura do modelo de regressão é expressa da forma,

$$g(\mu) = \eta = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \, . \qquad (15)$$

onde η é o preditor linear, β's coeficientes desconhecidos, k números de variáveis independentes, $g(.)$ a função de ligação e μ a média da variável dependente.

Conforme Kieschnick & McCullough (2003), o MRB permite gerar estimativas precisas e seguras dos coeficientes, mesmo que o conjunto de dados coletados para a investigação seja consideravelmente pequeno ou os dados mensurados sejam próximos de zero e próximos de um.

Ferrari & Cribari-Neto (2004) se baseiam na suposição de que as proporções seguem uma distribuição de probabilidade Beta, para realizar uma parametrização da esperança

matemática E(y) = p / (p+q) e da variância Var(y) = pq / [(p+q)^2 + (p+q+1)] de uma variável dependente y, utilizando os parâmetros (p e q) da distribuição de probabilidade Beta.

A estimação dos coeficientes para os MRB é obtida pela maximização da função de log-verossimilhança (EMV), em que os coeficientes β e ϕ são obtidos a partir da maximização do logaritmo da função de verossimilhança, conforme equação (5). O procedimento de maximizar a função de verossimilhança é realizado pelo algoritmo de otimização não-linear de *quasi*-Newton (BFGS).

3. Medidas de Diagnóstico

Uma etapa essencial na análise de ajuste dos modelos de regressão é a verificação da adequação dos modelos de regressão aos dados. Esta etapa, conhecida como medidas de diagnóstico, constitui um conjunto de critérios de adequação e ferramentas gráficas na avaliação da adequação do modelo aos dados. As medidas de diagnóstico fornecem subsídios para detectar: possível violação de alguma das suposições feitas para o modelo, especialmente para a componente aleatória (y) e a função de ligação (g(.)), aleatoriedade dos dados, presença de pontos extremos (*outliers*), adequação da distribuição de probabilidade proposta para a variável dependente e observação de pontos influentes. As medidas de diagnóstico propostas neste artigo para análise de adequação dos modelos de regressão são apresentadas a seguir.

3.1 Critérios de Adequação

- *Coeficiente de Determinação*: 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 esse problema, foi utilizado 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(.) e $\hat{\eta}$. Segundo Rao & Wu (2005), este coeficiente se restringe a $0 \leq R_p^2 \leq 1$ e, quando $R_p^2 = 1$ existe uma concordância perfeita entre $\hat{\mu}$ e y, consequentemente, melhor será o ajuste.

- *Deviance*: A análise de *deviance* é feita através da comparação dos valores da medida *deviance* dos modelos ajustados. Segundo Atkinson & Riani (2000), a *deviance* é obtida como duas vezes a diferença entre o máximo do logaritmo da verossimilhança do modelo nulo e do modelo sob pesquisa.

$$D(y;\mu,\phi) = \sum_{i=1}^{n} 2\{l_i(\tilde{\mu},\phi) - l_i(\mu,\phi)\} \qquad (16)$$

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 nulo. Conforme Lee & Nelder (1998), usualmente costuma-se proceder a análise de *deviance* utilizando o ponto crítico $\chi^2_{(n-k)}(\alpha)$ da distribuição qui-quadrado ao nível de significância igual a α, sendo n o número de observações e k o número de coeficientes do modelo. Portanto, se $D(y;\mu,\phi) \leq \chi^2_{(n-k)}(\alpha)$, pode-se considerar que há evidências que o

modelo sob pesquisa esteja bem ajustado aos dados, a um nível de α% de significância, usualmente $\alpha < 0{,}05$, caso contrário deve-se descartar o modelo, pois o mesmo pode ser considerado inadequado. Esta estatística de decisão também pode ser utilizada para a função de quase-*deviance* na adequação do modelo de Quase-verossimilhança.

- *Critério de Informação de Akaike*: O AIC (Akaike Information Criterion) 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 e é definido por

$$AIC = -2l\left(\hat{\mu},\hat{\phi}\right) + 2\left(k+1\right) \tag{17}$$

onde $l\left(\hat{\mu},\hat{\phi}\right)$ é a função de máxima verossimilhança do modelo ajustado e k o número de coeficientes do modelo. Segundo Rao & Wu (2005), o critério AIC 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 coeficientes, isto é, selecionar um modelo que esteja bem ajustado com um número reduzido de coeficientes. A decisão quanto ao melhor modelo ajustado é realizado escolhendo o menor valor de AIC.

3.2 Ferramentas Gráficas

- Resíduo *Deviance*: é o resíduo mais recomendado em análise gráfica de diagnóstico, pois estes resíduos são os que mais se aproximam da distribuição de probabilidade Normal na verificação da adequação ao papel de probabilidade e aleatoriedade dos resíduos. Para cada observação (i) da variável dependente y_i, pode-se definir o desvio $r_i^d = D_i\left(y_i;\hat{\mu}_i\right)$, de tal modo que

$$D\left(y;\hat{\mu},\hat{\phi}\right) = \sum_{i=1}^{n}\left(r_i^d\right)^2, \tag{18}$$

em que

$$r_i^d = \text{sinal}\left(y_i - \hat{\mu}_i\right)\left\{2\left[l_i\left(\tilde{\mu},\hat{\phi}\right) - l_i\left(\hat{\mu},\hat{\phi}\right)\right]\right\}^{1/2}. \tag{19}$$

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 (Lee & Nelder, 1998). Um gráfico desses resíduos contra o índice das observações (i) não deve mostrar nenhuma tendência e sim uma aleatoriedade.

- *Alavanca Generalizada*: tem-se mostrado uma ferramenta importante na análise gráfica de diagnóstico quanto a influência das observações em modelos de regressão, ou seja, avalia a importância individual de cada observação no próprio valor ajustado. A medida de alavancagem proposta por Wei *et al*. (1998), generaliza a definição de pontos de alavanca usados em modelos de regressão linear múltipla para outros modelos lineares pertencentes a classe dos MLG, sendo desenvolvida a partir dos elementos h_{ij} da matriz H que é conhecida como matriz de projeção ou "matriz chapéu" ($H = \text{X(X'X)}^{-1}\text{X'}$). Supondo que todos os pontos exerçam a mesma influência sobre os valores ajustados, pode-se esperar que os

elementos h_{ii} da diagonal da matriz H sejam definidos por w / n, onde w é o somatório dos elementos h_{ii} definido pelos coeficientes dos modelos e n é o número de observações.

A alavanca generalizada proposta por Wei *et al*. (1998) sugere que sejam examinados aqueles pontos tais que,

$$h_{ii} \geq 3w/n, \tag{20}$$

definidos como grandes pontos de alavanca. Ou seja, o valor de h_{ii} associado a i-ésima observação y_i é três vezes maior que a média de todos os h_{ii} da diagonal da matriz H.

- *Distância de Cook*: é uma ferramenta gráfica bastante utilizada para detectar a influência de cada observação nas estimativas dos coeficientes do modelo de regressão. Esta medida identifica a influência da retirada da i-ésima observação sobre as estimativas dos coeficientes do modelo, sendo definido por

$$D_i = \frac{\left(\hat{\beta}_{(i)} - \hat{\beta}\right)^T X^T X \left(\hat{\beta}_{(i)} - \hat{\beta}\right)}{w\hat{\sigma}^2}, \tag{21}$$

onde D_i representa uma soma ponderada dos desvios entre as estimativas baseadas nos coeficientes $\hat{\beta}$ e $\hat{\beta}_{(i)}$, w é o posto da matriz diagonal das variáveis independentes, $\hat{\sigma}^2$ é a variância estimada. Assim, essa quantidade obtida pela soma mede a distância quadrática entre $\hat{\beta}$ e $\hat{\beta}_{(i)}$. Segundo Cook & Weisberg (1982), as observações serão consideradas influentes quando $D_i \geq F_{(k,\ n-k)}\ (\alpha)$. Sabe-se que $F_{(k,\ n-k)}\ (\alpha)$ é o valor crítico da distribuição F de Snedecor (n, k) ao nível α % de significância. Geralmente, observações que apresentam $D_i > 0{,}05$, são consideradas influentes e devem ser investigadas.

- *Envelope simulado meio-Normal*: é uma ferramenta gráfica de diagnóstico muito útil em modelos lineares e não-lineares. A proposta desta medida é acrescentar ao gráfico de probabilidade 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 previsto pelo modelo ajustado e gerado médias, valores mínimos e máximos de cada valor previsto. Esses valores mínimos e máximos das k estatísticas de ordem produzem o envelope. Desta forma, o gráfico apresentará um intervalo para cada valor previsto ordenadamente contra os escores meio-normais

$$\Phi^{-1}\{(i + n - 1/8) / (2n + ½)\}, \tag{22}$$

onde $\Phi(\cdot)$ é a função de densidade acumulada da distribuição normal padrão e n é o número de observações. Segundo Atkinson & Riani (2000), caso ocorram tendências não aleatórias dos resíduos absolutos dentro do envelope simulado há indícios de escolha incorreta da distribuição de probabilidade para a variável dependente ou da função de ligação.

4. Estudo Aplicado

Esta seção apresenta um estudo para ilustrar a teoria descrita nas seções anteriores. A aplicação trata-se de uma pesquisa realizada numa empresa curtidora de couro, produtora de couro acabado e fornecedora para as indústrias de calçados e artefatos em couro.

A etapa *wet blue* do processo produtivo avaliado consiste em: o classificador recebe um lote de diferentes tamanhos contendo as matérias-primas e verifica se as características de qualidade satisfazem às especificações, por métodos cognitivos. As matérias-primas que não satisfazem às especificações são classificadas como produtos não-conformes, e a fração de produtos não-conformes às especificações, por lote, é considerada a variável dependente.

Os fatores controláveis definidos como variáveis independentes relevantes para a modelagem da fração de não-conformes foram: a seleção da matéria-prima em cinco diferentes estágios, conforme qualidade e preço (x_1); a procedência da matéria-prima adquirida pela empresa (x_2); o classificador que inspeciona as matérias-primas (x_3) e o estado de rebaixamento da matéria-prima (x_4). Os dados coletados contemplaram uma amostra de 754 lotes.

Tabela 1 – 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	---	---	---

4.1 Análise dos Efeitos

Numa análise preliminar da amostra coletada de 754 lotes, a fração média de produtos não-conformes às especificações foi de 0,185, desvio-padrão de 0,170 e coeficiente de variação de 91%. Nota-se que o histograma da fração de produtos não-conformes ilustra claramente que os dados não seguem a distribuição Normal (ver Figura 1).

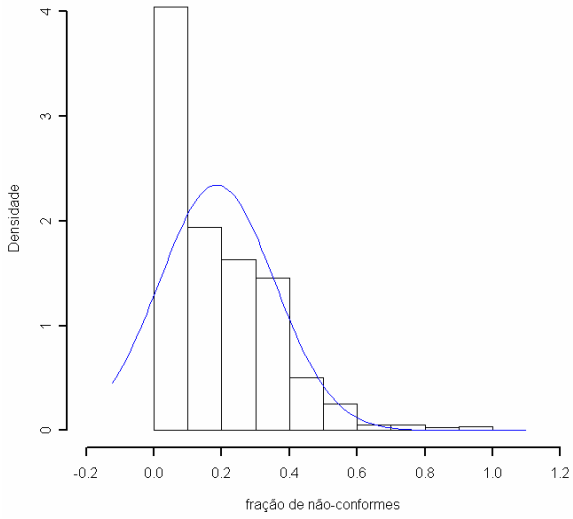
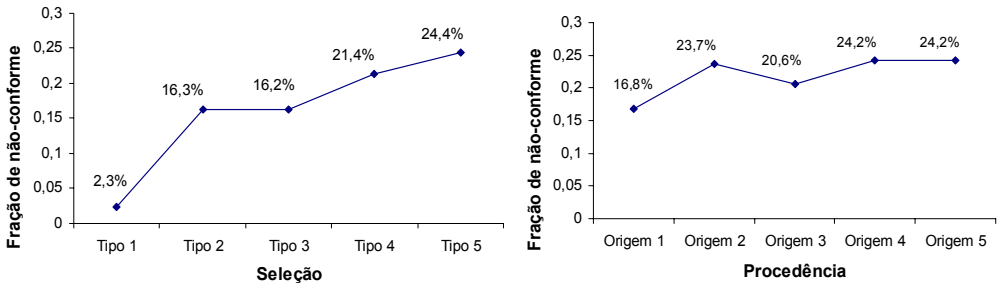


Figura 1 – Distribuição da fração de não-conformes com distribuição Normal sobreposta.

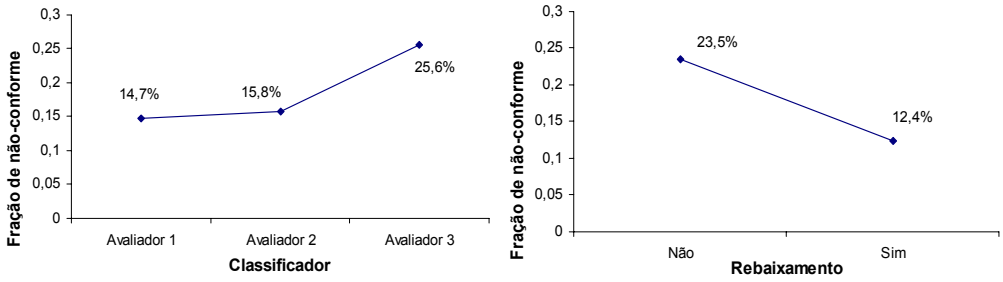
A análise dos efeitos das variáveis independentes sobre a variável dependente "fração de produtos não-conformes", mostra que, a variável independente "seleção" apresenta uma tendência crescente, ou seja, à medida que aumenta o nível da variável "seleção" aumenta a variável dependente fração de produtos não-conformes às especificações (Figura 2(a)). Na Figura 2(b) os níveis da variável "procedência" não apresentam diferenças significativas ($p > 0,05$) para a fração de produtos não-conformes às especificações, segundo o teste estatístico ANOVA.



(a) (b)

Figura 2 – Gráficos da fração de não-conformes em função das variáveis seleção (a) e procedência (b).

Observa-se que o avaliador 3 da variável "classificador" apresentou maior fração de não-conformes que os outros dois avaliadores, sendo esta diferença estatisticamente significante ($p<0,01$) e que o nível rebaixado da variável "rebaixamento" apresenta significativamente ($p<0,01$) menor fração de não-conformes, conforme Figura 3(a) e (b).



(a) (b)

Figura 3 – Gráficos da fração de não-conformes em função das variáveis classificador (a) e rebaixamento (b).

4.2 Estrutura dos Modelos Ajustados

Esta seção apresenta a estrutura dos modelos de regressão Beta e de Quase-verossimilhança utilizada na modelagem da fração de não-conformes às especificações na etapa do processo *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*, para construção dos modelos de regressão. As novas variáveis independentes foram 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}).

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

$$g(\mu_i) = \beta_0 + \sum \beta_j x_{ij} + \varepsilon, \quad \text{com i} = (1, ..., 754) \; ; \; \text{j} = (1, ..., 11) \qquad (23)$$

onde g é a função de ligação, β's os coeficientes do modelo e ε o vetor de erro aleatório.

Observa-se que não é assumido *a priori* que a variável dependente (fração de não-conformes) possui uma distribuição de probabilidade. Segundo Sant'Anna & Caten (2005), para construção dessa classe de modelos outras funções de ligação e de variância podem ser usadas, com objetivo de verificar qual melhor se ajusta aos dados em estudo. Nesta modelagem foram escolhidas a função de ligação "logito" e de variância do tipo $V(\mu) = \mu(1-\mu)$, pois forneceram o melhor ajuste aos dados, ou seja, produziram a menor quase-*deviance*.

Para a segunda modelagem foi ajustado um modelo de regressão Beta, assumiu-se que a variável dependente (y) segue uma distribuição de probabilidade Beta com média (μ) e utilizou-se as variáveis independentes (x_1, x_2, x_3, ..., x_{10} e x_{11}) como estrutura linear dos coeficientes, conforme a equação a seguir,

$$g(\mu_i) = \beta_0 + \sum \beta_j x_{ij} + \varepsilon, \quad \text{com i} = (1, ..., 754) \; ; \; \text{j} = (1, ..., 11) \qquad (24)$$

onde g representa a função de ligação, β's os coeficientes do modelo e ε o vetor de erro aleatório. Para o modelo Beta, foi utilizada a função de ligação "logito" e função de variância do tipo $V(\mu) = \phi^* \{pq / [(p+q)^2 + (p+q+1)]\}$ por gerarem melhor ajuste do modelo aos dados.

4.3 Análise do Ajuste dos Modelos

Ao se modelar com variáveis independentes qualitativas, a estimativa de um dos níveis dessa variável é nula para que os estimadores dos demais níveis possam assumir valores positivos ou negativos, conforme a sua influência na variável dependente. Este nível base, que assume o seu estimador nulo, tornar-se um referencial passível de ser comparado com os outros níveis (Sant'Anna & Caten, 2005).

Na modelagem inicial dos dados, os modelos de regressão propostos apresentaram as variáveis independentes "seleção", "classificador" e "rebaixamento" como estatisticamente significativas para explicar a variável dependente "fração de não-conformes às especificações", ao nível de significância de 5%, baseado no "teste estatístico *Wald-test*". E a variável independente "procedência" não apresentou significância estatística ao nível de 5% nos dois modelos de regressão ajustados, ver Apêndice.

A Tabela 2 apresenta as estimativas dos coeficientes com respectivos erros padrões e critérios de adequação da qualidade do ajuste dos dois modelos de regressão finais.

Tabela 2 – Estimativas dos coeficientes, erros-padrões e critérios de adequação dos modelos de regressão finais.

	Modelo de Quase-verossimilhança		**Modelo Beta**	
Parâmetro	**Estimativa**	**Erro padrão**	**Estimativa**	**Erro padrão**
Intercepto	-3,7692*	0,3844	-2,4225*	0,2225
Seleção de Couro				
Tipo 1	-	-	-	-
Tipo 2	2,0475*	0,4219	1,5822*	0,2761
Tipo 3	2,2857*	0,3886	1,1202*	0,2270
Tipo 4	2,5226*	0,3878	1,2295*	0,2260
Tipo 5	2,7583*	0,3913	1,4469*	0,2308
Classificador				
Avaliador 1	-	-	-	-
Avaliador 2	0,2672*	0,1095	0,2883*	0,0900
Avaliador 3	0,4483*	0,0877	0,3740*	0,0744
Rebaixamento				
Não	-	-	-	-
Sim	-0,8357*	0,0917	-0,6807*	0,0749
Dispersão (ϕ)	0,15512*	0,5015	0,12337*	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 "*Wald-test*" ($p < 0,01$)
gl (graus de liberdade)

A forma de regressão para o modelo de Quase-verossimilhança é descrito como $g(\hat{\mu})$ = -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). Para o modelo Beta a forma de regressão é apresentado como $g(\hat{\mu})$ = -2,4225 + 1,5822(seleção 2) + 1,1202(seleção 3) + 1,2295(seleção 4) + 1,4469(seleção 5) + 0,2883(avaliador 2) + 0,3740(avaliador 3) – 0,6807(rebaixado).

4.4 Análise de Adequabilidade dos Modelos

Avaliando as estimativas encontradas pelos modelos finais (ver Tabela 2), observa-se que para o modelo de Quase-verossimilhança a seleção do tipo 2 tem $exp[\beta_1] = exp[2,0475] = 7,75$ e para o modelo Beta $exp[\beta_1] = exp[1,5822] = 4,87$ o que significa estimar que, para o modelo de Quase-verossimilhança a seleção tipo 2 apresenta 7,75 vezes mais chances de produzir fração de não-conformes às especificações que a seleção tipo 1, já para modelo Beta a seleção tipo 2 apresenta 4,87 vezes mais chances de produzir fração de não-conformes às

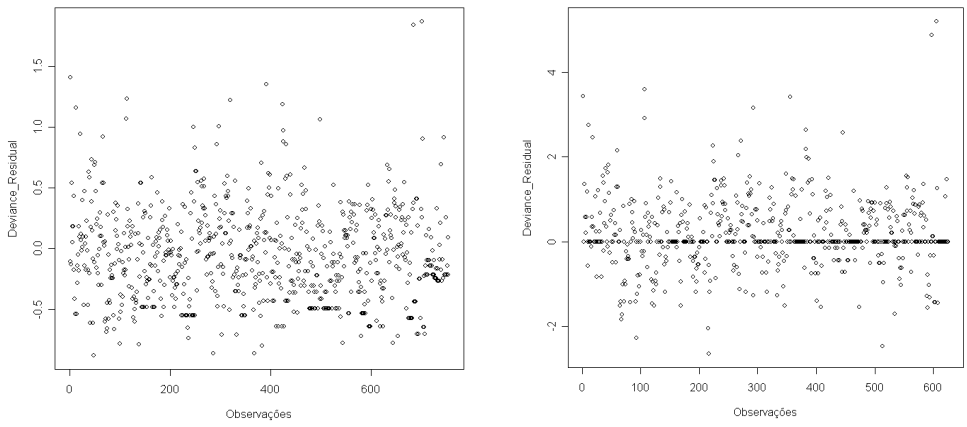
especificações que a seleção tipo 1. Bem como, para o modelo de Quase-verossimilhança, a seleção tipo 5, apresenta 15,77 vezes mais chances ($exp[\beta_4] = exp[2{,}7583] = 15{,}77$) que a seleção tipo 1 e 425 vezes para o modelo Beta ($exp[\beta_4] = exp[1{,}4469] = 4{,}25$).

A variável independente "classificador", no modelo de Quase-verossimilhança, os avaliadores 2 e 3 aumentam as chances de produzir fração de não-conformes em 31% ($\exp[\beta_5] = exp[0{,}2672] = 1{,}31$) e 56% ($exp[\beta_6] = exp[0{,}4483] = 1{,}56$) respectivamente, em relação ao avaliador 1. Enquanto que, para o modelo Beta, as chances de produzir fração de não-conformes dos avaliadores 2 e 3 são de 33% ($exp[\beta_5] = exp[0{,}2883] = 1{,}33$) e 45% ($exp[\beta_6] = exp[0{,}3740] = 1{,}45$) respectivamente, comparando com o avaliador 1.

Para a variável independente "rebaixamento", a estimativa do coeficiente é negativa ($\beta_7 = -0{,}8357$), indicando que o fato do estado de textura da matéria-prima estar rebaixado implica em que as chances de produzir frações de não-conformes às especificações diminuem em 43% ($exp[\beta_7] = exp[-0{,}8357] = 0{,}43$) para o modelo de Quase-verossimilhança. Enquanto o modelo Beta estima que as chances diminuem em 51% ($exp[\beta_7] = exp[-0{,}6807] = 0{,}51$).

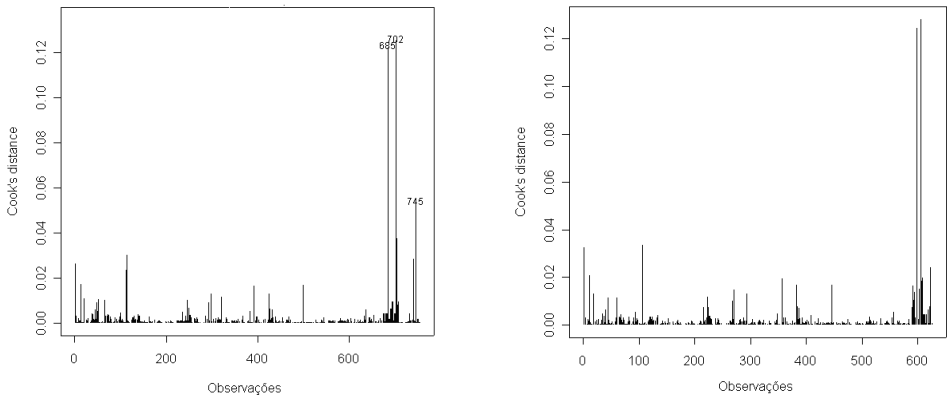
Analisando a qualidade do ajuste do modelo de Quase-verossimilhança, o coeficiente de determinação "pseudo" R_p^2 foi 0,467, a *Deviance* = 115,65 (746 graus de liberdade) que corresponde a um valor de probabilidade (*p-value*) $p < 0{,}01$ e um AIC = -213,3 demonstrando que o modelo se ajustou satisfatoriamente. Para o modelo Beta, o coeficiente de determinação foi de 0,581, a *Deviance* = 121,48 (746 graus de liberdade) e um AIC = -224,96 demonstrando que o modelo foi bem ajustado. Nota-se que as medidas de qualidade de ajuste para os modelos ajustados estão relativamente próximas, porém os critérios para o modelo Beta indicam que este é o melhor modelo (ver Tabela 2).

As Figura 4(a) e (b) apresentam uma aleatoriedade dos resíduos *Deviance*, ou seja, os pontos não apresentam nenhuma tendência indicando que a função de ligação utilizada é adequada. Embora estas revelem que há dois pontos com maior valor residual, correspondendo às observações 685 e 702. Em relação à distância de Cook's, nota-se que as mesmas observações se apresentam altamente influentes ($D_i > 0{,}05$) (Figura 5(a) e (b)).



(a) (b)

Figura 4 – Gráficos de resíduos *Deviance* para os dados com o ajuste dos Modelos de Quase-verossimilhança (a) e Modelo Beta (b).

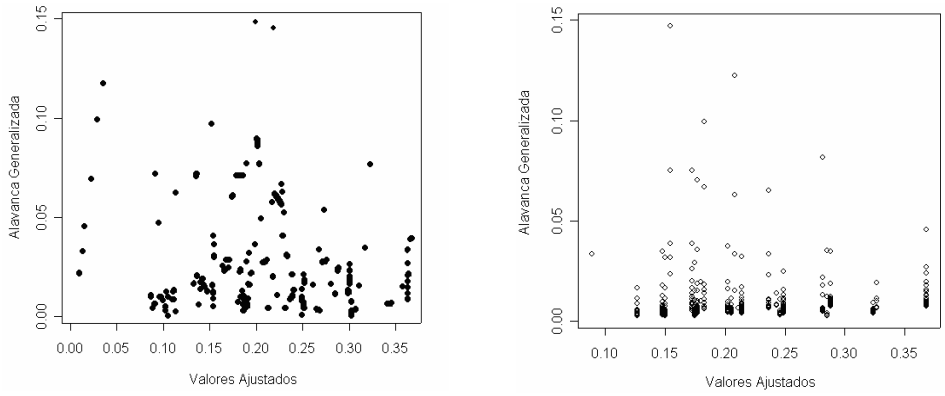


(a) (b)

Figura 5 – Gráficos de distância de Cook para os dados com o ajuste dos Modelos de Quase-verossimilhança (a) e Modelo Beta (b).

Observa-se nas Figura 6(a) e (b), que os Modelos de Quase-verossimilhança e Beta apresentaram similaridade na verificação dos pontos de alavanca no conjunto de dados. Os pontos de alavanca em destaque nas figuras referem-se às observações 685 e 702.

Contudo, após um estudo realizado com as observações 685 e 702 se verificou que, embora estas aparecerem como pontos discrepantes nos gráficos de diagnóstico nos dois modelos ajustados, as mesmas não foram retiradas da modelagem, por não se constituírem observações discrepantes (*outliers*), não modificando as estimativas dos coeficientes dos modelos.

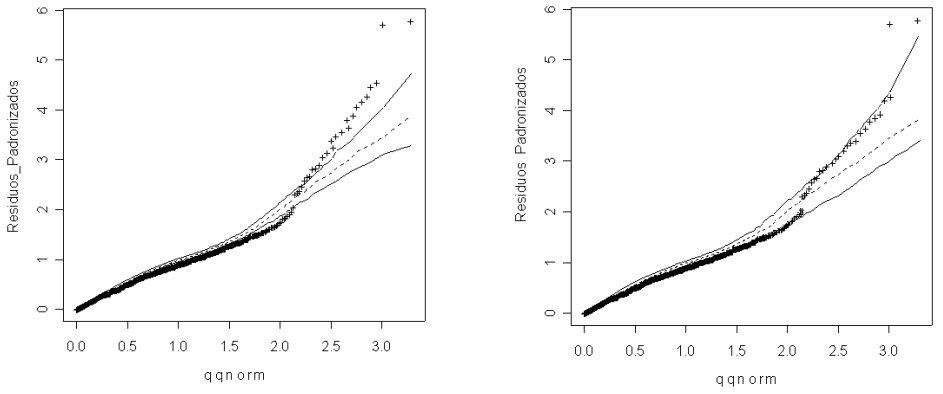


(a) (b)

Figura 6 – Gráficos de alavanca generalizada para os dados com o ajuste dos Modelos de Quase-verossimilhança (a) e Modelo Beta (b).

Conforme Figura 7(a), o gráfico de envelope simulado meio-normal apresenta alguns resíduos fora do envelope, evidenciando a necessidade de um ajuste mais adequado. Uma opção para se buscar um melhor ajuste é investigar a possibilidade de haver outros fatores (variáveis independentes) que estejam influenciando na etapa *wet blue* do processo de produção da empresa, gerando a fração de não-conformes às especificações.

No envelope simulado meio-normal gerado pelo modelo Beta a maioria dos resíduos aderiu aos limites do envelope, embora haja alguns resíduos sobrepostos nos limites do envelope, e dois em destaque. Porém pode-se considerar que o modelo Beta se adequou perfeitamente aos dados (ver Figura 7(b)).



(a) (b)

Figura 7 – Gráficos de envelope simulado para os dados com o ajuste dos Modelos de Quase-verossimilhança (a) e Modelo Beta (b).

5. Considerações Finais

A importância de conhecer e utilizar modelos de regressão vêm da necessidade de investigar o efeito das variáveis independentes, chamadas de fatores controláveis, sobre a variável dependente, chamada de característica de qualidade. O foco deste estudo foi a modelagem da fração ou proporção de produtos não-conformes às especificações de um processo industrial em que a característica de qualidade é mensurada no intervalo [0,1].

Observou-se que a aplicação de modelos de regressão mais corretos para o tipo de dados no processo de modelagem, como modelo de Quase-verossimilhança e modelo Beta, aumenta a precisão das estimativas e fornece uma análise de investigação mais consistente. Ademais, estes modelos consideram a relação não-linear entre as variáveis independentes e dependente e a variância dos erros dependente da média. Estas características são inerentes a dados mensurados em fração ou proporção.

Os modelos de Quase-verossimilhança e Beta apresentam vantagem na modelagem dos dados mensurados em fração ou proporção, por permitir flexibilidade de escolha da função de ligação e de variância que melhor se ajuste ao conjunto de dados. Com base na análise das medidas de diagnóstico os modelos de regressão propostos na modelagem da fração de produtos não-conformes às especificações apresentaram bom desempenho no ajuste e precisão das estimativas dos efeitos dos fatores controláveis.

Espera-se que as comparações realizadas neste artigo forneçam alguma orientação na investigação e análise de dados em trabalhos futuros, pois a natureza dos dados deve ser levada em consideração. Com base nas análises realizadas recomenda-se o uso dos modelos de regressão Beta e de Quase-verossimilhança a processos industriais que mensurem a produção de produtos não-conformes às especificações de manufatura.

Agradecimentos

Os autores agradecem aos revisores anônimos pelos valiosos comentários e sugestões, aprimorando a qualidade deste trabalho.

Referências Bibliográficas

(1) Atkinson, A.C. & Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer-Verlag, New York.

(2) Cook, R.D. & Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall, New York.

(3) Cordeiro, G.M. (1986). *Modelos Lineares Generalizados*. VII SINAPE, Campinas-SP.

(4) Cot, J; Manich, A. & Aramón, C. (1992). Procedimentos e Instalação para o Tratamento Integral de subprodutos da Indústria Curtidora. *Revista do Couro*, ABQTIC, **19**.

(5) Cox, C. (1996). Nonlinear quasi-likehood models: applications to continuous proportions. *Computational Statistical & Data Analysis*, **21**, 449-461.

(6) Dobson, A.J. (2002). *An Introduction to Generalized Linear Models*. 3ª ed., Chapman & Hall, London.

(7) Ferrari, S.L.P & Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. *Journal of Applied Statistics*, **31**, 799-816.

(8) Hamada, M. & Nelder, J.A. (1997). Generalized linear models for quality-improvement experiments. *Journal of Quality Technology*, **29**, 292-304.

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

(10) Lee, Y. & Nelder, J.A. (1998). Generalized linear models for the analysis of quality improvement experiments. *The Canadian Journal of Statistics*, **26**, 95-105.

(11) McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models.* 2ª ed., Chapman & Hall, London.

(12) Montgomery, D.C. (2001). *Introduction Statistical Quality Control*. 4ª ed., John Wiley & Sons, New York.

(13) Montgomery, D.C.; Peck, E.A. & Vining, G.G. (2006). *Introduction to Linear Regression Analysis*. 4ª ed., John Wiley & Sons, New York.

(14) Myers, R.H.; Montgomery, D.C. & Vining, G.H. (2002). *Generalized Linear Models with applications in Engineering and the Sciences*. John Wiley & Sons, New York.

(15) Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society A*, **135**, 370-384.

(16) Prentice, R.L. (1986). 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.

(17) Rao, C.R. & Wu, Y. (2005). Linear model selection by cross-validation. *Journal Statistical Planning and Inference*, **128**, 231-240.

(18) Sant'Anna, A.M.O. & Caten, C.S. (2005). Modelagem da proporção de defeituosos usando Modelo de Quase-verossimilhança. *XXV ENEGEP*, Porto Alegre-RS.

(19) Sant'Anna, A.M.O. (2006). Método de orientação à modelagem de dados mensurados em proporção. Tese de M. Sc. PPGEP – UFRGS, Rio Grande do Sul, RS, Brasil.

(20) Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, **61**, 439-447.

(21) Wei, B-C.; Hu, Y-Q. & Fung, W-K. (1998). Generalized leverage and its applications. *Scandinavian Journal of Statistical*, **25**, 25-37.

Apêndice

A Tabela 3 apresenta um recorte da matriz de dados experimentais utilizados na construção dos modelos de regressão propostos, uma vez que totaliza 754 observações, se tornando inviável publicar. Os autores podem ceder gentilmente a matriz completa dos dados experimentais.

Tabela 3 – Matriz resumo de dados experimentais utilizados na modelagem.

BERTIN LTDA.		**Unidade EV (RS)**					
RESULTADO DA CLASSIFICAÇÃO DE COUROS *WET BLUE*							
Período:		**02 à 06.06.2004**					
Itens	**Seleção**	**Procedência**	**Classificador**	**Rebaixamento**	**Volume**	**Conforme**	**Não Conforme**
1	1	1	1	2	105	0,24	0,76
2	1	1	1	1	200	0,78	0,22
3	1	1	1	1	204	0,79	0,21
4	1	1	1	2	390	0,65	0,35
5	1	1	1	1	400	0,65	0,35
6	1	1	1	1	202	0,71	0,29
7	1	1	1	1	200	0,65	0,35
8	1	1	1	1	209	0,53	0,47
9	1	1	1	2	181	0,92	0,08
10	1	1	1	2	169	0,98	0,02
11	1	1	1	2	123	1	0
12	1	1	1	2	69	1	0
13	1	1	1	2	84	0,36	0,64
14	1	1	1	2	114	1	0
15	1	1	1	2	281	0,91	0,09
16	1	1	1	2	309	0,95	0,05
17	1	1	1	2	201	0,88	0,12
18	1	1	1	1	200	0,81	0,19
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
754	5	4	1	1	197	1	0

A Tabela 4, com as estimativas dos coeficientes, respectivos erros padrões e critérios de adequação da qualidade do ajuste dos modelos de regressão iniciais.

Tabela 4 – Estimativas dos coeficientes, erros-padrões e critérios de adequação dos modelos de regressão propostos.

	Modelo de Quase-verossimilhança		**Modelo Beta**	
Parâmetro	**Estimativa**	**Erro padrão**	**Estimativa**	**Erro padrão**
Intercepto	-3,7702*	0,3862	-3,3144*	0,2217
Seleção de Couro				
Tipo 1	-	-	-	-
Tipo 2	2,0507*	0,4233	1,6806*	0,2761
Tipo 3	2,2832*	0,3899	1,9123*	0,2267
Tipo 4	2,5236*	0,3892	2,1206*	0,2256
Tipo 5	2,7547*	0,3928	2,6724*	0,2304
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	0,1813*	0,0899
Avaliador 3	0,4511*	0,0892	0,3953*	0,0734
Rebaixamento				
Não	-	-	-	-
Sim	-0,8317*	0,0961	-0,8183*	0,0749
Dispersão (ϕ)	0,15271*	0,4989	0,12483*	0,3817
R_p^2	0,361		0,429	
Deviance	113,10 (742 gl)		120,98 (742 gl)	
AIC	– 207,5		– 218,66	

*Nível de significância "teste *Wald-test*" ($p < 0,01$)
gl (graus de liberdade)